UC GIS Week
Thursday, November 18, 2021
11 AM – 12 PM
How To Do Something Spatial

Sessions:

Making Community Connections with Story Maps, Field Surveys, and Dashboards

Speaker: Joseph Kerski, ESRI

Abstract:

Discover how to connect with and solve community problems and issues with by using a combination of 4 key web GIS tools: Field Surveys, story maps, web maps, and dashboards in this engaging workshop.

Transcript: Video Timestamp: 00:01.19

And without further ado, let's get started. Joseph Kursky is our first speaker. I'd like to read his bio for everyone. Joseph is a geographer, GIS professional and educator with a focus on supporting and promoting GIS in all disciplines and in all levels of education. He has served for 30 years in government at USGS, the census and NOAA, academia, nonprofits as president of the National Council for Geographic Education and in private industry as education manager for Esri. He has authored eight books, 1000 lessons and 5000 videos. Welcome Joseph.

Well, greetings all, many thanks.

Let me go ahead and share with you some things that I think are really relevant and show you the power of connecting different apps in this web GIS platform that we all have at our fingertips nowadays. It's an exciting time and I appreciate joining you all in this GIS day week, geography awareness week.

And so let's dig into collecting, mapping, analyzing and communicating field data with ArcGIS. I'm really excited about this topic and I hope that it sparked some conversations amongst our community. And then also I look forward to keeping in touch with you in the future. Just on that note, I do have this contact information but you could probably just find me by just searching for Joseph Kursky.

And again, appreciate the connection that we all have as people passionate about being change agents in the world. And that's exactly what this connecting these components of this platform actually allow people to do is collect, map, analyze, communicate about something they're passionate about. You all know this, but I think it's important to remind students and instructors and others that it's important to be using field tools. And then another thing that I want to weave in here is this whole idea of apps which are spreading

spatial thinking and geotechnologies to non-GIS people, non-GIS instructors, non-GIS students, to data science, business, engineering, health and other fields. It's really an exciting time. And apps are helping us do that. Today, I'm gonna focus on three apps, Story Maps, Dashboards, and Survey123. And you and I have tools, for example, on our phones or on our tablets, laptops, computers

that are disconnected from anything else, right? I have one that is Earthenview created by a developer in Germany. I've used it for 20 years. It's four megabytes in size. It does one thing and it does them very well. It basically edits your photos. It's not connected to anything else. And I use it and you have tools like that that you use as well. But the power here is you're connecting some very easy to use, easy to access tools in powerful ways.

And the workshop here today in the four hours that we have together, just kidding. I know we don't have four hours together, but is this collecting, mapping, analyzing, creating a dashboard, creating a story map that kind of wraps all this up.

Now, there are many ways to map field gathered data and I have lots of notes on that and lots of tutorials and things like that that you can dig into. But today, we're going to work on something like this together and I'm just going to give you highlights of the actual lesson that you can actually go into and if you're an instructor, share with your students.

Survey123, I think one of the innovative things about it is that it's crowdsourcing. So you can crowdsource your surveys and the actual creating of the survey is very much like Google Forms or Survey Monkey or dragging and dropping elements into a form. You can use Excel as well for further power,

for example, additional branching. If I answer this on the survey, then route me to this part of the survey. You also have just like a lot of tools that you use in the modern GIS cloud-based environment nowadays like story maps, you have a zone where you can manage, for example, all of your surveys. The metadata, adjust them, copy them, et cetera. And you can do these things with Survey123, even use higher precision devices with it.

So enabling the citizen science mode is a pretty easy feature under the collaborate tab.

We're going to use the web designer. I'll illustrate the web designer form method of creating a survey. But again, you can use what's called Survey123 Connect and that uses Excel. And again, it gives you additional power. Takes a little more time to learn, but the nice thing about the Excel method is that you don't get just a blank sheet of Excel. You actually get the schema and you populate the schema with elements that you want to collect data on. Invasive weeds, public art, community gardens, water quality, weather, noise. I mean, the sky's the limit with this.

And so I've got an example here of a result that I created with Survey123. But since I've got a very limited amount of time, I'm just going to put this in the chat box later on so you can actually dig into it. But after filling out the Survey123, then you create a web map out of it. And then from that web map, you then do your spatial analysis on it, and then you create a dashboard, and then you create a story map. That is the flow of the actual tutorial that is actually right here that I've got. And I'll put the link in the chat box as well. But it basically guides you through everything that I'm going to cover in the next 10 minutes in a more rigorous way. But when I

say a more rigorous way, we're not talking days, weeks, months. I'll show you an example here as we get done with this. But midway through this, I actually have the step-by-step creation. This is exactly what you're going to do. I have in a zip file that's attached to the lesson all of the pictures even that you need to create. So you don't even have to use your own photos. Well, of course, you're certainly welcome to do that. The bottom line is that we've all seen the JHU COVID map, and for good reason, right? It's very useful for municipalities and individuals to make decisions on. I think the powerful message is, number one, that you have access to these same tools at your own fingertips. So I wrote something that I'm very passionate about as a geographer, Walkable Neighborhoods. So I've visited many of your campuses in the past, hope to do so in the future again. But every time I'm on a work trip or even for fun, I am very tenacious about walking in the community and assessing how walkable the neighborhood or the community is. So I created this essay that has the lesson attached to it that I mentioned earlier. And the results, just to skip to the actual results, is this. This is what you're actually going to create with this lesson. And I've taught with this in numerous places and in numerous levels, from secondary school to community and technical colleges to universities to lifelong learning, libraries, and museums.

So here's the result of what you're actually going to build. Is this place walkable or not? This is when I was in Mitchell, South Dakota for the South Dakota GIS Conference. Some of you might remember that that's where the corn palace is, and that's the dome in the background. But you might put this at a six or a seven, probably, on the walkable one to 10 scale. There's some sidewalks. They're in fairly good condition. There's some ramps. So if I was in a wheelchair or on a skateboard, I could get up and down those ramps. There's some cars, but there's really no traffic to speak of. So more walkable than less walkable. And then the story map, as you'll create, you put a few sentences in there about why communities care about walkability. And remember, by extension, think about other things that you could actually use these same techniques to map. Again, water quality, graffiti, urban greenways, invasive weeds, et cetera. And inside the story map, which is what you're going to create, I have a link to the survey, which you're also going to create in this lesson. It's five simple questions, pedestrian friendly or unfriendly, rate the walkability, tick all the choices, where is it located, which is, of course, germane to what you and I love, the where question. And then

do you have a photograph of it? And then what you'll then build is a map from the survey. The survey creates a editable feature service, which is a bit GIS geeky. But the nice thing about it is that it does it on the side. If you used ArcGIS field maps or a collector in the past, you have to create your data database. So you have to have some knowledge of GIS and how to create a subtype and a domain and so on. Survey123 kind of does that on the side. And so again, it's very approachable, even for your colleagues in non-GIS disciplines. So I've got this web map of the survey points that I and others contributed to with the data from my survey. Pedestrian friendly in this case, good for pedestrians. Ooh, good for kangaroos too. Well, what does that mean? These are the-- this is the actual piece of data that I collected from my own survey. The tails on those kangaroos, they're like those pool noodles that you used when you were a kid. You know, this red and orange and green big old pool noodles. They're thick. Those tails. That was a pretty cool walk. I had some nice company on that one. But the point is, you're creating this web map that houses the data results from your survey. And then finally, you knew this was coming next, right? You create a dashboard, a dashboard just like the JMU dashboard that we looked at a moment ago-- the JHU dashboard that we looked at a moment ago. And roughly 3-quarters of the points in the survey right now-- and I have it crowd sourced. It's open survey. So you can test it out. You can collect some points on your campus into this survey. 607 points have been collected thus far. Roughly 3-quarters are pedestrian friendly, as you see there. And I've got the other pie chart getting data from my survey. So in my case-- and I've got a legend, as you can see there, and some other elements. But the point is, there's two points. One, these are very powerful tools, easy to use and accessible. And number two, to connect these things, the GIS web platform that we have at our fingertips. It's all in ArcGIS. So it's survey123. It's StoryMap. It's a dashboard. And it's an actual web map that's sped from the survey. The JHU COVID dashboard is fed from multiple sources, local governments, WHO, CDC, et cetera. My survey, the one that you'll build and map, the map gets fed from one single survey. It's only one source feeding this. But again, I think it's such a powerful concept. But yet, the other main message that I want to impart to you all today is that this whole thing, the StoryMap, the survey, the dashboard, and the map itself, took me four class periods to create, roughly four hours, an hour on each of those four components. So we're not talking days, weeks, months to create these things. So again, I think

it's very powerful and engaging and ties to what we were talking about at the beginning. Why should we have students go out in the field and collect data? There's lots of things that are fostered by such activities. Richard Louv, last child in the woods, gets students outside frequently and immersed in their entire educational experience for critical thinking, community connections, spatial thinking, problem solving. So it's not just the GIS techie skills that are fostered. It's those broader, deeper goals, having students be empowered to care about issues in their community and become change agents on the landscape. I wish I could get excited about this. But the point is powerful, easy to use, doable, and connected with these broader issues.

OK, comments, questions, concerns. I'm going to put in the chat box now the links, as I promised, to the lesson and the blog, where you can see why I did what I did and some of the results there. But while I'm doing that, if anybody's got any questions for me or any comments, I'd love to hear from you. Thanks.

If we don't have any audience questions, I've-- oh, wait.

Kim, you're raising your hand. Please go ahead in the chat or unmute. Yeah. It was an applause. Oh, applause. Sorry. That's a clap. Sorry.

It's OK.

I have a question.

In your support, if anyone in this session is planning on taking your lesson plan to their students, what are some things that people typically get tripped up on that we should be prepared to think about?

Yeah, good question. First of all, yeah, I'd love to chat with you folks more in the future.

I'm all about the student learning and engagement. I don't have any-- and my team, think of us as your industry partner in this. We're all educators on the SRE Education team. We really do care about teaching and learning. And we also work with campus facilities as well on campus about making campus safer and more energy efficient and so on. So think of us-- we're actually real people. You go to some website, since there's a little bot that says, hey, may I help you? We're actually real people. Ooh, I wish I could tighten up my jowls here. But anyway, the

point is we're real folks, and we love getting your challenges. What are you challenged by? What are your success stories? So please take advantage of the fact that the team is here to support you and learn from you. But anyway, good questions. I would say there's two things. Oftentimes, educators from a wide variety of disciplines kind of see story maps first, and they say, ooh, I want to do that. And that's good that they serve that purpose to get people excited about, hey, there's a story map on climate. There's a story map on typhoons, et cetera.

But there's the things that go underneath it. OK, great that you want to have your students create story maps, for example. But what about those web maps that go underneath them that they point to? OK, so we need to have that included in such a workshop and a workflow. It's not just making the story map. It's the-- there are applications, as you all know, that point to the maps and the analyses that happens underneath. So I think that's one comment. The second comment is, how do I say this? You all know what I'm talking about next. But the goal is not to map-- to put your points from the field on the map. That's an enabling, empowering kind of thing. But there's a greater goal, not to put just the data on the map, but to understand something better and maybe to inspire action about urban greenways or whatever you're studying. And then leads to additional questions. So I just want to nudge people and keep nudging your students and fellow faculty members in. The goal is not to learn more GIS. It's fascinating. It's very empowering and so on. But the goal is to understand our world better and to make it a better place, more resilient, more safer, et cetera. And then the last thing I wanted to share about this is that there is, I think, sometimes a bit of a short shrift still on the analysis part. OK, once you map things-- there are some-- even in ArcGIS online, there's like 40 analytical tools. Don't ignore those.

And so don't--

interesting. Don't rush from the mapping part to creating the dashboard and the story map. Make sure you do some analysis. I kind of gave it, admittedly, a short nod here today just because of time constraints. But make sure you do a little bit of analysis. So if you're mapping, let's say, I don't know, water quality and quantity, even just running the trace downstream tool. And where does that water go on its route to the ocean? So those are the things that I wanted to share about in response to your question. There's lots more we could say, I'm sure, about that.

Thank you.

Should we pause for more questions? Or did you have more to present?

I'm going to put a few more things in the chat box. But I want to be respectful of everybody else's time. And I want to hear what you all have to say, the other presenters too. So anyway, thanks again for having me. And let's keep in touch. Thank you. Thank you very much.

All right, next speakers.

Protected Areas Databases in California

Speaker: Kimberly Becerril & Saba Gebreamlak, GreenInfo Network

Abstract:

The California Protected Areas Database (CPAD) and the California Conservation Easement Database (CCED) are authoritative GIS databases of all protected open space lands in California and are the California element of their national counterparts. These databases are publicly accessible and they are updated twice per year. GreenInfo Network, the non-profit maintaining the databases, will conduct an overview of the data, use cases, recent updates, and key resources.

Transcript: Video Timestamp: 00:16.51

Kimberly Besseril is a geospatial analyst and cartographer at Green Info Network and is part of the team that maintains the California Protected Areas databases in California. She received her BA from UC Berkeley.

Saba Gabriemlak is a geospatial analyst at Green Info Network and is also part of the team that maintains the Protected Areas databases in California. Take it away, Kim and Saba.

All right.

Thank you for that introduction, Mayana. And thank you all for making time to join us for this presentation on the California Protected Areas database and the California Conservation Easement database. I'm Kimberly Besseril.

Hi, and I'm Saba Gabriemlak. And we're part of the team that edits and publishes CPAAT and CCD.

So before getting into the databases, here's a short introduction of the organization that created and maintains them. There's around a dozen of us staff, and we work for a wide range of organizations all over the US, but mostly in California.

In more than two decades, Green Info has supported over 1,000 public interest organizations and agencies with a wide range of information technology to enable these groups to better meet their environmental conservation, public health, social justice, and other goals.

Those who want to learn more can go to greeninfo.org, our website, which features a lot of the projects we've been working on. In particular, we've been working on the California Protected Areas database, CPAAT, and the California Conservation Easement database, CCD, for quite some time. They form the basis of a lot of what we do here.

I'll also note that there's a couple of different ways to check in on the things that we're doing. We have a blog started by our executive director, Dan Roddenmacher, and Instagram feed for those interested in cartography and design, which features content from our side of the map making process, and you can find us on Twitter.

Now, let's take a look at CPAAT. The California Protected Areas database is now the most extensive state data set on protected lands in the United States, with over 49 million acres of protected lands represented.

For information on CPAAT and CCD, your go-to resource is calands.org. In the opening page, you'll find hyperlinks to learn more about these databases, and if you scroll down, you'll see case studies of CPAAT in use.

In the contribute data tab, you'll find a map collaborator hyperlink. If you have any issues with anything in the databases, go to the map collaborator application and tell us about it, or you can contact us directly by emailing cpad at calands.org.

In the about our work tab, you can subscribe to our newsletter,

and you can also download a poster of the latest CPAAT and CCD release.

And lastly, in the get data button, you can download the data and subscribe to our newsletter.

So, yeah, please sign up for the CPAAT newsletter because it's our best way of seeing in touch. And we don't send newsletters out very often, just updates on the new releases.

All right, let's get into the data.

In the data set itself, we have a very expansive view of what lands are protected. It includes the big wilderness areas and the deserts, but it also includes all the way down to the small urban parks, even as small as the neighborhood mini parks.

We want a broad inventory that covers everything because there's a wide range of users and uses of the data.

So, what's in CPAAT? It's all of these different sorts of lands, and we've really made an effort to be broad and inclusive.

I'll just give you a second to observe this list.

All right, now, I'll go through the CPAAT structure, data statistics, and explain a little bit about how it works and what's under the hood.

But before getting into specifics about CPAAT's structure, we'll note that with each release, we provide an updated user manual on calands.org. It is also included in the general database download zip files.

The manual is a very complete resource,

so it is great for all of your CPAAT questions, including its structure.

And we encourage users to take the time to read through it. We aim to be clear about what's in the data set, how lands are added, and document unique scenarios.

The manual is a terrific resource. We use it ourselves here in the office on a daily basis while editing CPAAT.

All right, let's get into the structure.

Sending back for a second and going a little into the technical details here.

CPAAT is actually three different types of geometry.

There are holdings, units, and super units.

Each of those is a separate shape file in the release, and they behave this way. Holdings are the most detailed, which represent individual ownership usually at the parcel level.

Those get aggregated into units, dissolved by the list of attributes listed here under core attributes.

Finally, we aggregate units into super units, the most simplified version of the database.

These are dissolved on the core attributes with an asterisk next to them. The way you think of super units is the way we ask the public to interact with a park.

We know the name. We interact with the park managing agency, with their website maybe, or through paying a fee to a park, et cetera.

And finally, we know whether we have access to it. The parcel owner and its county is not identified here.

Units are a good level for those who are interested in understanding how ownership and county boundaries change across the landscape.

This holdings, units, and super units relationship is really important for how CPAD works, and it gives us a great deal of analytical strength, as well as makes it much easier to track over time.

An example is Las Trampas Regional Wilderness in the Bay Area.

In the holdings level view, individual parts are divided up at the parcel level. In the units level view, the data is more generalized, but you can see the split at the county line.

And in the super units level view, all of the details are dissolved to only differentiate between name, manager, and access. This is the recreation of focus, where the most important details are public access and who manages the land.

CPAD is a GIS database. It's managed in QGIS and some related programs like ARCPro. It includes about 49 million acres, which is roughly half the state. We

published CPAD twice a year and just recently began publishing CCD twice a year as well.

We changed the naming some years ago so that people could know clearly what year the data was from and used the A and B format to distinguish between the two annual releases. A is typically released mid-year. B is released near the end of the calendar year.

A way to look at the inventory is by the different agencies that we have holdings for.

There are 1,000 agencies and nonprofit landowners represented in CPAD.

It's a very large array of institutions that own all of this land.

Another perspective is comparing the acres. CPAD is 49 million acres, and the easement database is now up to 2 million acres.

While easements are incredibly important, they're eclipsed by the feline data, one considering A-bridge.

If we look at just CPAD and ask how much of that is open for the public to use, almost all of it is. The amount that's restricted in the dark yellow here, instead of permit is required. It has very seasonal hours or is very limited in some way. And the amount that's actually close to the public is also quite small. Most of the lands is either an ecological reserve or areas that are not yet open as parks, but are owned by public agencies and may someday be open to the public.

Another way of looking at CPAD is in terms of who owns these lands. So if we look at the top 50 agencies by A-bridge, we end up with a huge percentage acreage owned by the top 50 and very little owned by the remaining 950 agencies.

Three different lands, if we look at the number of parks, the top 50 agencies by acres owned own less than half of the total number of parks. And the share of CPAD owned by the other roughly 950 agencies jumps up.

The other 950 agencies are typically serving local recreation needs. In those instances, the number of small to medium parks helps distribute access to users throughout communities. The size is less important in this case. Depending on

how you look at CPAD, you get a different perspective on what's important and for what reasons. All right, I'll pass a presentation of Susava.

OK, thank you, Kim.

OK, in terms of data quality, CPAD's pretty good.

But anybody who works on a large data set has to maintain a pretty humble approach. CPAD has been updated and improved over decades. And along the way, we've added new attributes that make the data set increasingly robust,

as well as retired those that were both cumbersome to maintain and less useful to users.

One common request is for the ability to show the change in parkland over time. We are hard at work on this, but it's a long project. For example, in this image, we can see Clayton Ranch, Regional Preserve, and the year each parcel was protected. This information is very helpful, and we are working continuously to include this information when it's available.

This is our map collaborator tool found at mapcollaborator.org slash CPAD or through greeninfo.org. Map collaborator allows users to easily view the latest data and to contribute to the data by adding notes or drawing polygons on a web map or by uploading their own data for review.

CPAD editors here at Green Info look at these notes and use it to inform our edits to the database before each release.

In terms of what the users really are, as you can see, there are a very wide range of users, and we encourage that.

We want to hear from people who have used it. So if you use this data, please feel free to share with us what you've done with that.

Here are some maps to illustrate different ways that CPAD is used for distance from parks analysis, trail planning, brochure for recreational access, and habitat conservation.

OK, now switching over to easements for a minute. Those lands are tracked in CCD database. Many of us may already be familiar with how easements work, but just

in case, here's a quick depiction. On any particular landscape, some public land or open space can be owned upright. There will be a CPAD shown here with the bright green outline.

Others would retain private ownership, but would have an easement. A conservation or an open space easement placed upon them. Those are shown here in orange. The private owner retains ownership of the land, but voluntarily agrees to limit what they can do on their land. Most notably, development rights are limited or restrictive or removed.

Is the longer term conservation of open space side of things that we're interested in? We focus on those agencies and the development of the land. We focus on those agencies and lands maintained for open space or conservation purposes.

We're not looking at short term easements. We're looking at the permanent ones.

OK, at the moment in the database, there are approximately 2 million acres of lands and 250 easement holding agencies represented.

However, the work on CCDs tell in progress. We know we are missing easements and have others that are not quite right. Addressing overlaps, for example, is complex when multiple agencies hold an easement over one section of land. In addition, some lands overlap with CPAD because a property can be held in fee title and have an easement.

There is also a manual, just in case you're interested. There is a manual that's available for CPAD. It's for CCD data set similar to that of CPAD, which you can download and check out.

Be sure to read the guidance about representing

representing the data as a sensitive data because easements often involve private land. It is important to respect them as private lands. Many of which have no public access. Having said that, it is important to track these lands. They are critical to our landscape in California, and they are typically funded at least partially with public dollars. But we are also we also want our users to follow policies on how the data is shown on NUTS and in the public regarding their access. There is a balance between public and private information.

OK, now a quick note about CPAD and CCD is national data set equivalence. Nationally, CPAD is a California element for the protected areas that are based off the United States, Padua for short, publishes it's published by USGS.

We work very closely with them on a variety of things. Ideally, most states would have something on the order of CPAD. But at this point, not all states do just a handful.

And ambition is to get there someday.

The California easements data set is contributed to the National easements data set and set for short, which is published by a group of nonprofits Conservation easements dot US.

It's very similar to the relationship CPAD has to padua.

In 2020, CCD was updated to align to the data set schema up and set. And we continue to contribute to the national data set each year after the releases.

So what can you do if you have used it already? Go to the site, download it, pop it into an ArcGIS or any other GIS application. If you have any issues with any of that, go to the map collaborator tool and let us know. Or you can contact us directly.

And just a reminder, please sign up for CPAD newsletter.

OK, with that, we come to the end of our presentation and thank you for coming. I'll open it up for anyone who has questions and or any comments or anything else to me and Kim.

Wonderful. Thank you, Saba and Kim. There is one question in the chat.

CPAD seems amazing. What is the funding and is the funding sustainable?

Currently, it's being funded by the state and CDPR state parks.

But we're also open to receiving funding whenever we can find it. But there's a lot of things we could improve in CPAD. Similarly, in CCD as well. So, yeah, we're very open to receive more funding on it.

Yeah. And then just to add on to that, so CNR already also funds CPAD, which is again a state agency. And typically, like if there are particular regions in California that,

for example, are working on park equity,

then the county or a city tends to turn to us to like really

update that area within California, and then they provide additional funding.

Yeah, but I was open to more funding.

Let's see any other questions.

Well, I have a question if none are coming yet.

If someone here downloads CPAD or CCD today, is there anything you recommend them testing out with the data? An element like an attribute, maybe, or a region that people tend to miss when they're first using it that you think is interesting to share?

You want to go, Kim? Sure. So we had a slide on it, but it's like the year protected. I think a time series of seeing a protected area expand over time is always really fun to play with and like a pretty

like simple

map series that one can create. So that's what I would recommend.

Yeah, on that note, I would like to

people to know that we were trying to do that because people tend to are confused the day it was entered into the CPAD database as the day it was protected, which is not the case, because we find out about some parks like a decade after they were protected, depending on where the information was made. We were made aware of the information.

So that is something we're constantly working on, acquiring all those information as much as we can get. But protected land in California go as far as the 19th century, late 19th century. So there is a lot of information that we would like to get our hands on. And yeah, and another thing we're trying to like, hopefully, ambitiously trying to improve and see if that is the accessibility

tag or information to parks. So anyone who's interested in providing those information are it is privy or has those information. They're pretty happy to receive those.

Wonderful, thank you. And people can go to www.calands.org. OK, yes, exactly. Thank you so much. Both of you.

Towards Scalable Spatial Data Science

Speaker: Amr Magdy, UC Riverside

Abstract:

NA

Transcript: Video Timestamp: 00:38.00

All right, our next speaker is Amr Magdi. Amr is an assistant professor of computer science and engineering and a co-founding faculty member of the Center for Geospatial Sciences at UC Riverside. His research interests include database systems, spatial data management, big data management, large scale data analytics, indexing and main memory management. Welcome, Amr.

Oh, thank you very much for the introduction and thank you all for your time. So good morning. It's almost noon, but let's say good morning, everyone. So today I'm going to share with you some thoughts about our work along with my grad students and UC Riverside towards building scalable spatial data science.

So let me start with a story, a somehow old story. When we start to doe maps, dates that we mark now as BC. So we doe maps for a long time, for navigation, for military applications, for a lot of different things. And in the 19th century, we have this use case, which is the Calleah epidemic

that started in Asia and spread in different continents, at the British London in 1954.

At this time, we had the impression that Calleah is spreading over air or breathing like COVID.

Dr. John Snow started to doe the Calleah cases in London, and he noticed that it's around Broad Street where we have a water bomb.

They started to put hypotheses and study that Calleah actually spread through water, and this started a total new understanding for this epidemic. At this date, we

started to understand that maps is not only for navigation and military, it can be used in a lot and a lot of use cases, and we started what's so-called the human geography. And now you can imagine what are the applications of maps in health and politics and urban planning, and it's endless. You can say it. A lot and a lot of applications are using maps.

So as the maps and mapping literature evolve over time, we have a parallel world that have a disrupting technology evolving as well, which is computing. So in the 19th century, in the 20th century, the computing technology started to evolve from mechanical computers to electronic computers to the chip we have now.

And it happened that we have a lot for geographic data, and in the 60s, we started to have a player in the middle. And the player here is GIS software. When we started to combine geographic data to be handled by computing technology, and this is a startup GIS, it's certainly 60s. In late 60s, ESO is founded. ESO is the neighbor for us here in New Riverside. They are in the Netherlands.

And the three worlds started to evolve.

We have a world of computing, computing technology, continue to evolve over time in various aspects and now computers and everything.

Also that geography world evolved over time and the GIS becomes more advanced and you have different products and different software packages.

And if you look at the users of GIS, basically, you will find different types of users, so users are not like it's very wide user page.

And there are users who use GIS and this is the majority of users use GIS as is from the user interface, geographical user interface, and do a lot of analysis with this. This is definitely a reach package of software and other users who do more analysis in data. So GIS, as you know, is a general package. So it's not designed for earth scientists only or geographers only, or it's designed for everyone who can do mapping and using analyzing data. So there are type of users who need more functionality and they can use a computing technology in a different way. They can use something like Python, like our language. So those people start to integrate with what's called the open source initiative. So developing software that is free and available for everyone to contribute and use. And over time, we had very each

set of tools using these tools, Python and R and open source software to analyze the spatial data, what I call here spatial data science tools, something like Geodas, something like PiSound, JOPandas, JOPi and so on. So these tools providing still spatial data analysis similar to GIS, but they are more tailored. You can contribute to these tools. You can write your own functions. You can use them side by side to GIS. It's not a placement and it's not totally orthogonal. It's somehow having utility.

Over time, the history has not ended and will not end soon. So we have more and more of the geographical side of the data. We have smartphones and smartphones in 2008 started to have Internet on the smartphones. So you can imagine that everyone had a smartphone, has an Internet and has sensors and equipment to generate geographic data. So now even geographic data is not only from scientists or governmental agencies, but also from everyone, from everyone. So that includes social media data, that includes mapping data, that includes reviews, that includes a lot and a lot of content that include geographical content. And this kind of data is of interest for social scientists. You can imagine you as a social scientist, you have this amount of data comes from users and you study your work to study users. So this is a very rich source of data. So now people who are analyzing data, spatial data, do spatial data science want to ask questions? Very important question. Can we actually include these rich sources of data under the umbrella of a spatial data science? That's actually a question. We want to answer this question in this talk.

So let me first introduce spatial data science in both terms. So it's not settled. It's a not settled term yet. So I cannot imagine this or I cannot like claim that there is one definition for these terms. But alternatively, let me introduce what do I mean with the spatial data science. From a user perspective, data science is understanding a phenomena or a certain situation through analyzing data. OK, so it's another term for data analysis.

From a technical perspective, data science is usually an intersection between three things, between domain science or domain expertise that need to analyze a certain problem and statistical analysis. So you use statistics to analyze data and computing technology. So we are using the computing to automate the analysis and scale it up.

And this goes in the computing part from all the stack, from managing data, cleaning data, doing machine learning, visualization, security, privacy, a lot and a lot of tasks.

And this actually comes with along with the sides with academic programs of data science are designed. So if you look at academic programs of data science at all these universities, you will find them the degrees, joint degrees between computer science and statistics. This here in UC Riverside, it's also in UCI. It's also in other campuses other than UC like Purdue, Michigan, Minnesota, Washington and so on. If we want to put spatial data science, you add the spatial work into every of these circles. So you have a spatial science or a spatial expertise combined with spatial statistics, combined with spatial computing. And in the intersection, you will find the spatial data science. So you basically analyze the spatial data to understand the certain phenomena situation. And also from a technical side, this comes with the whole task of spatial data cleaning, spatial data management, spatial data mining, machine learning, geo visualization and so on. So in UC Riverside, we have a center for geospatial sciences where we work on an interdisciplinary field of research. I work from the computing side and I also have collaborators from the social sciences.

And let me focus on social sciences here as a prime use case for spatial data science. So it's easy to list books, not only papers or research work, books in analyzing spatial data and social sciences. So this is a very popular topic used a lot in studying societies and phenomenons. And the common operations here that's needed that you use a spatial statistical analysis to analyze the regions, to analyze the states and counties and so on. Let me give you an example. I assume we have our social scientist here and she is using now something like GIS software like Python libraries that analyze spatial data, PySal, JAPANDA. And she wants to use these new sources of data, the large data, the big data.

So like social media data, like satellite data, like IOT data.

So an example of this, if you have this amount of social media data and you want to count simple statistics, compute simple statistics in this, for example, having one billion tweets, one billion tweets. So can I compute the location quotient? Can I

compute the local indicator for some special association? This is special statistics measures.

So if you see the definition for these measures, you will find the equations here. I will not go into the details of the equation, but the equations here talk about the count, the count of a certain type of data in this county, in this state, in this city, the percentage, which is also count over the total time. So it ends up that this query is given one billion tweets, count the tweets in each state and each county and each city. Okay.

So one billion tweet, the question, how can our social scientists handle this one billion tweets? So she has the count in each of the 50 states, for example.

That's a scalability issue. That's obviously a scalability issue. If you try to use the existing tools to handle this query, you are going to run in a very bad performance. And this limits our social scientists from using spatial data for analysis.

But the good news that GIS and Python libraries are not the only tools available. We have scalable big data systems, something like Apache Sedona, something like Amazon, Amazon Cloud Computing, JS Perk, which is another name for Apache Sedona, NVIDIA GPUs they are using. I don't want to go into technical terms, but the bottom line that there are tools to handle big data. Big data is all over the news, right? Why not using these two data sets on these big data developments and our social scientists can analyze its big data sets?

The problem that these tools are not designed to support spatially statistical operations.

This is the problem. They are spatial tools, yes, but not for spatially statistical operations. And from a computer science perspective, we focus on three fundamental queries, which is spatial K-NIN engine joint.

But aside from the technical part, these three queries are not spatially statistical. So the existing capabilities of big data systems is limited to support a wide variety of spatial data science operations that are actually currently used by social scientists and other domain experts. So the idea here is we connect these two walls together. We have the wallet of big data systems and we have the wallet of

spatial data science tools. And we want to our social scientists to use the tools he used to use usually. OK, we use Python interfaces, he used ArcGIS or QGIS.

And at the back end, we handle the large data sets on a scalable computing platform. So the work of bridging these two walls together is supporting the scalable statistical queries in big data systems. And this is not just a mere integration between two packages of software. This is a major effort to develop modules that enable these two softwares to interact altogether. These softwares now are totally independent from each other, developed independent, used independently for each other. So to have components, things from a technical side, called query processing components, indexing components, integration components between these two walls of software, that's a major effort that we are trying to achieve now in our group.

And I am going to quote a research report from the database community, saying that the database community must do a better job integrating database research with data science ecosystem. We must collaborate with other disciplines so that our technological contributions are recognized and adopted widely, meaning here by people who analyze data, data analysts.

And NSF report on data science, they recognize social sciences as a field that's emerging and leverage the capacity to collect and analyze data at scale and that may reveal patterns of individual and goal behaviors. So obviously, working in the intersection between data science and social sciences here is a very rewarding field and we can make a lot of predictions. What we propose here is called Pineapple. And Pineapple here is a system that enables our social scientists to analyze large data sets, large spatial data sets seamlessly without getting any of the complications, technical complications, underlying complications of big data management.

And under the hood, Pineapple will be responsible to connect the two walls, something like PySal, a Python library for analyzing spatial data, and something like Apache Sedona, which is distributed and scalable system to analyze big data. Pineapple will be integrating these two walls together and providing the users with a seamless interface and friendly interface to analyze large data.

Definitely, we have a lot of modules in the system, evolving modules. We're still developing that.

Roughly, I'm like dividing these modules into four parts, scalable distributed queries, scalable generalization, learning-based modules, and social data models. I'm not going to go into all the details here. I just want to give you the idea. Maybe I give a highlight of a couple of modules if you have time. I may check with our moderator here. Do I have a couple of minutes to give some highlights about one of the modules or we're on out of time? We've got four minutes to cover wrapping up and questions, but no questions yet. So I say yes. Go ahead.

Okay, cool. So let me talk about regionalization. This is something that we actively work in now and we're trying to support scalable

regionalization. And the regionalization here is basically dividing a space into a set of regions. Basically, if you have a US, I can divide it into four or five regions depending on what criteria. And definitely different criteria are defined based on the application domain. So if you are defining school districts, this is different from economic studies or environmental studies like air quality and water assessment and so on. A major problem now that you as a user need to define the spatial scale is my spatial phenomena is a state level or a county level or a city level. So you need to define how many regions you need to divide. So this is the problem. What is the appropriate value of P? For example, if you have P equal four, that's an nation-level phenomena. It's like snowing. If P equals 50, that's a state-level phenomena like wind speed, for example. It's different from each state to another. If P equals 3000, that's a county level. So defining the P here means you have brilliant information. Some special studies does not know that. I do not know what is the phenomena, what is the spread rate of coverage. That's a pretty new thing I need to discover what the scale here. So that's why what's so-called max-P regionalization comes into. So it's dividing the space into unknown number of regions, but you have a certain characteristic, you have a user defined constraint in this. For example, divide the whole U.S. and the region so that each region has at least 50,000 COVID infections. Okay. Something like this. The problem that there's a major scalability issue here. So if you both one of this way, it takes from two minutes to several hours for only 3000 areas. So social scientists actually have more than 70,000 of areas to analyze and this never ends. So we have a scalability

issue. We have the module that can help in solving this scalability issue. And our module gets much better performance. It highlights about the performance like 20 to 30 times faster on time. So our social scientists can analyze much larger data. And this is part of our vision of scalable analysis of spatial data.

And we still maintain high quality solutions. So our performance here does not compromise the quality of the solution. We still have a good quality. We also have other variations of this regionalization, like having generalized constraint and in nature of clustering and using machine learning in adapting streaming data. I'm going to stop here having any of your questions. So please, if you have questions, let me know.

I need to skip this set of slides.

So in summary today, I motivated for having scalable tools to analyze spatial data. And we briefly introduced highlights about the performance. I would like to thank all my collaborators, recent collaborators and the past collaborators, and also thank the National Science Foundation for their support for our research. And this is my web page and email. If you have any questions or any kind of potential collaboration anytime, please shoot me an email and I would like to have your questions now if you have any.

Thank you. Thank you so much, Dr. Magdi. Any questions?

Can send them in the chat and I'll read them out or you are free to unmute and ask.

I can, I can start us off with a question. Since I'm guessing our audience here is a little on the geography and GIS side, but let's say they're interested in a career in spatial data science.

What tools do you recommend them branching out maybe into statistics or computer science departments to enhance their skills? Yeah, I think the Python library that analyze spatial data is very like, like J-openders and PyCell and J-opi and this ecosystem that's very each with a lot of functionality that's not yet incorporated in computer science tools. And that's why I work in this area to incorporate this kind of functionality. So having knowledge about this, besides the GIS, GIS is a fundamental thing, right? So it has a lot of functionality, but side by

side with GIS, if they get familiar with the Python libraries, I think that that would be a great engagement for their skills. Thank you. And I see we have a raised hand, please go ahead and unmute.

Hi there. Thank you so much for your presentation.

I am actually asking a question that kind of contradicts what you had just asked. I'm thinking about I'm a student who's coming more from theory and I'm looking at how to incorporate GIS and data analysis into my work. I'm wondering if you have any recommendations for reading literature or courses for sort of a beginner in the GIS world, moving over to more data science. Thank you.

I think our guests from Eze here would be excellent to answer this question.

For Mathias, I definitely special statistics in your area to read about, but from tools, I think Eze has a great introductory tutorial to start about GIS and data science. Maybe Joseph want to comment here. Yes, thanks, Amr. I totally agree with what you're saying. I'll put a couple of links in the chat box. I would say if you only had time to do one thing in terms of additional learning, I would say one of those SRE MOOCs. We only have about six or seven of them because we want to do them really well, you know, like a university course that you go through as a cohort. It's fun. It's engaging. It's very rigorous. They're free. I'll put those in the chat. I would say that would be the most valuable. We do have one on spatial analysis that might be suitable for you. Okay, will do. Thanks.

Thank you. Any other questions? Okay. Well, we're only a couple of minutes over. I think we made great time and we can continue conversation in the Slack channel that we've been sending out links to. Thank you. One more time to all of our speakers.

Thank you. Thank you.

You all.