FEDERAL TECH PODCAST

Ep. 40 The Rise of the Machines . . . in Federal Technology

This is John Gilroy from the Federal tech podcast.

00:21
And this is Dr. Elsa shaper with linquist.

00:24
Today we're going to talk about machine learning, hit the music cloud. Welcome to the federal tech podcast. My name is John Gilroy and I will be your moderator. Today we're gonna talk about some mysterious question and, and some concepts and terms that you've had heard about and maybe have a fuzzy understanding about it. And this whole concept we're gonna talk about today is called machine learning. And I thought it'd be great to bring in someone who is a certified Brainiac to explain machine learning to us. We have Dr. Elsa Schaefer. She's a corporate data scientist for linquist. Elsa, how are you?

01:01
I'm doing great, John, thanks for having me here today.

01:04
Do you have that certification on your desk certified Brainiac little piece of paper there?

01:08
Oh, absolutely. I don't want to forget.

01:11
That's good. Well, when we deal with machine learning, we deal with large, large datasets. And of course, you have large datasets, we have certain mathematical problems that have to be resolved in his handling of them, whether you're going to take large datasets and apply it to John's donut shop, or to the DOD. And so need a mathematician. And so Dr. Schaefer has a PhD in mathematics, it makes sense for her to come in and give us her observations on, you know, here's the good and bad, the ugly about machine learning, here's what you can do, here's what you can't do. And if you're the federal agency, here are the questions you should ask. Because believe it or not, Elsa, there are some people who try to dazzle my listeners with terminology and statistics and numbers and NGOs, who, like they're not a space or something. But I don't think we do that. We're gonna have some fun, tell some jokes, and maybe give some practical tips, we have a lot of practical advice during this little conversation. So before we begin, once a little bit about your background, and how you want to work in for

02:08

linquist. So I was an academic for several decades, and I did a lot of research while I taught and my main interest in recent years was computational epidemiology. And when that earthquake hit Haiti, the second one ago, while back, I realized in order to model while you have to understand data, and that started me moving towards understanding data better. And I ended up leaving academia for data science startup, learned all about visualization and data science and started moving into machine learning. And since that time, I've switched to link quest. And I have the privilege of working with the CTO and trying to build synergy and community among our data scientists and operations research, folks. And it's just tremendous fun.

03:04
Let me play up on that word, synergy. Whenever I hear the word machine learning, I hear of here, artificial intelligence, kind of like that's, that's the planets kind of circling each other. Right? And so that kind of one is a subset of the other. What is the sequence? What's, what's the order here?

03:20
I was doing artificial intelligence back as an academic, it's sort of the big circle in a Venn diagram, if you're picturing it. And it's much more general than machine learning where in machine learning, you're really combining your input data in specific ways to try to make prediction. So machine learning, and then deep learning is a subset of that is a specialized kind of artificial intelligence. But AI dates way back. Yeah.

03:51
And so let's talk about the people in our community here. The DoD, I think we all know the DoD is getting attacked 5000 times a second. I don't know how many times getting attacked, maybe more than that. And they have a tremendous amount of information they have on on who's attacking him and why, and 1000s and 1000s. I mean, they're talking about petabytes of data here. This is that it's how many books can you read in the library? I mean, how much of this data can you actually get a hold of how much can you wrangle got his data wrangling? You got to put a cowboy cowgirl hat on Yeah, and get a rope going and wrangle that data? So that's the term that is that all these different days? Which ones do you bring in and start in the funnel? Is that right?

04:29
Yeah, absolutely. And that and that's the fun. And I think another part of that equation, John is, how much money do you have? Because the bigger the data set you're wrangling and playing with, the more compute time and resources you're going to need. And so figuring out what is the size of the data set that you need to work with in order to get good results and validating the veracity of that dataset and the use of that data set? Those are challenges to heart noticing this big data.

05:02
I remember years ago, I was dealing with a company and they had a set of data at one location. They want it mirrored at the other location. And I said, Well, this was going to cost. What? So I'd say it's, what's it worth to you? And okay, how much do you want? Is I want everything. No, you talk to a general, what should I choose? You just bring everything in? Well, you know, you can't afford everything. But you have to be careful about selecting it. And that's where the rub, what would Shakespeare say i That's the rub. Yeah,

05:29
absolutely. So

05:32
we all know, there's no faultless data set. So what guidelines can we give our listeners maybe looking at big datasets? And which ones to bring in? And which ones? Which ones to hold? Which ones to fold?

05:45
That's a hard question. I think that's where you really need an engineer, a data engineer and case by case basis, helping you decide. One of the things that we need for a lot of our machine learning solutions is labeled datasets, we're trying to predict something and we need accurate labels in order to make those predictions in a training data set. So I think one thing that you'd really want to look at is which data sets can you track the provenance of and you understand how they were created and what the labels are? And do they apply to your future problems?

06:21
Well, I think we set the stage here, we got the first part of the acronym down the M L part. And now we go back towards synergy and talk about the other part one focus in Dante is, is the way people develop code. And you know what I know back in the day, you'd bring someone in, they'd actually write lines of code, and then you'd have been testing it. And then all of a sudden, they move up in the stack, and they get more and more sophisticated. Now, what's happened is, is that it's all becoming into the cloud, it has been cloudify. There's a word for the cloudification. And so then we have development people working with operations people to cloudify solutions, and then we have the, you know, the portmanteau is called DevOps. Okay, and so people taking some playing around with DevOps and, and security's dev SEC ops, and some, like Elsa comes in and goes, Hey, let's put ml in front of ops. So we'll have mL ops. So this is machine learning operations. So what is ml ops mean for you and for our audience?

07:17
Yeah, so one of the most exciting things for me leaving academia was working with a team of people who understood all different parts of a production process and being able to do just the math, just the modeling and getting this great product out in the end. And ml Ops is taking that to a new level, where do I need to recode a particular ml algorithm? Again, no, it's repeated. It's a repeatable process. Let's code it once really well, each kind of ml architecture. And let's have a stable of really strong code available to junior engineers to use. And the same with data cleaning data wrangling a lot of the same processes used repeatedly. And when we're rushing, a lot of times, we might not notice little problems in our data, little blips or little, we might just train one kind of model where maybe another model would have worked better. And the idea of ml Ops is to walk an engineer from importing your data through deploying your final model, and helping that engineer detect any little blips that they might have been in too big of a rush to notice both at the beginning stage of looking at the data and at the end stage of choosing the best model. And part of that includes understanding, is your model explainable? Is your model going to be believable to the person using it? Can you justify that it makes intuitive sense, the way this model is doing predictions and ml ops can help you pick all of those little pieces together.

08:59
Your company, of course, is Lin ques LINQU, e. S t.com. And we had a previous conversation. And you said, well, linquist is actually making a foray into helping others understand this with some kind of a platform. And so tell us about this platform you have.

09:15
Yeah, so I have been in love with this platform. Since I first came to linguist it's how the harnessed for adaptive learning, and a bunch of our very clever operations research data science engineers wanted to help our customers in the intelligence community and Department of Defense, analyze their simulation models. And this platform uses adaptive sampling, to very rapidly create a stable of machine learning proxies for sometimes quite complex simulation models, and through the stable of machine learning models are able to understand the explainability Have these simulation models and do some basic validation and verification help helping our customers to improve their models quickly. So it's often an iterative process where we look at a model and we can say, oh, it's not performing very well here. And here, it's here are the kinds of intelligence you can get from the model that you're running currently. So I looked at that, and I thought, wow, this is so cool. And you basically have half of an ML ops solution already there. Because you've already created this stable of models you've already automated tuning of your hyper parameters for those models. Let's, let's do some internal research, let's build this into a platform that could do a full ml ops solution. And one of my big focuses with that, and this goes back to my days in a startup, but also the incredible user experience, people that we have here is, let's build an ML ops solution that has a low point of entry, so that people who are not trained as data scientists or machine learning engineers, can use this platform successfully and safely and create trustworthy solutions.

11:21
No Else, if they're listening to podcasts, and there's snow on the ground or something you want to learn more, there is a fact sheet about how you can download and it's at Lynn quest.com/products, and solutions slash Howell kind of hard to say in a podcast, I'll have a link in the show notes page to that. But it's, it's a nice way to to filter out some information that may not be accurate with information that it has is for discrete is that discrete, filtered information that may not be accurate with information that is accurate, because this is the real deal. I mean, if somebody else is putting her shoulder behind this, he is going to have some authenticity to it. And and in in the in the field, what they say is that 80% of the time is spent with just preparing the data. And so if you get a leg up on that, all of a sudden, you can be more efficient with this. This firehose of data everyone's experienced here.

12:12
Yeah, absolutely. And there are a lot of several very large commercial ml ops solutions out there. And so I think one of the things that companies and agencies will struggle with is Do we want a big generic solution that is already dealing with a whole lot of the pitfalls we're encountering? Or are we special? Do we need special solutions? Do we already have great engineers who can tackle problems that are really relevant to us and our organization. And in that case, you might want to work with a company like lean quest and develop specialized solutions that really target your specific needs.

12:53
Everyone in the federal government now has to as a mandate to say zero trust every 32 seconds, or you get out of me, you can find. So I have to say zero trust. But one part of that word is trust. And I think that is a concern that many people have with these large datasets, because how good is Elsa? Really? Did she really pick the right ones? And and so this is a I think it's serious question here about how do you establish trust in the algorithm itself? And how do you know you're not biased? I mean, the classic study with with I think it was done with Amazon, maybe with filtering for resumes or something, there's a big bias in that. And so how can we make sure that it's, it's not but maybe a little art to the science? Ah,

13:39
yeah, I mean, I love this question, because it's something that we're really worried about, certainly within the DOD and intelligence communities, and there's a lot of guidelines that can help us avoid bias, or be aware of what our biases are, and build trustworthy models. The DoD is currently really focused on building out their authoritative sources of truth and, and starting to have data governance rules. And the beauty of an ML ops solution is that it can really mandate compliance with those rules, and be part of that metadata collection. That adds to the transparency of a model. So a purely technical ml ops solution would simply show you hey, here are some graphics about what your dataset was like, what the correlations were, the models that you train, what their accuracies were, where they were explainable where they had bad accuracy. But if you add in this ml ops, now this SRE data governance, then now you're really expanding that ML ops tool to add transparency to who created that dataset. Why did they create it? How are we applying it? How does that relate to what we're doing? Are their privacy concerns or their ownership concerns? Now we could really have all of that in one place. Better want to be

15:01
too serious here, but in the Pentagon, they make decisions that result in lives, life and death of people. And they have to be very serious about making sure this is solid and actionable data. And and I am certain that this new Chief Data Officers organization the government has, I mean, when they close the door, and it's just them, they go, Well, who do we trust? What about this? How do we how do we make sure that we're responsibly handling? I think in the civilians, their response for Florida, like health care information, the responsible and the DOD, those people can die. But it's, it's people can die in the healthcare system, too. So this is probably more serious than you would think it's, it's, you know, we can joke about data wrangling and this and that, but there's some serious decisions are going to be made based on these assumptions.

15:50
Yeah, I agree. And fortunately, the people I've worked with when the within the DOD take this very seriously, I think one of the things that ML ops can offer that's beyond the the coolness in the moment, like it's a pretty great helper in the moment to rapidly create a good model. But let's say six months from now, one of our engineers figures out a way to validate our data from a new lens. And to find problems, like one problem that can occur in data is that you accidentally trained it with the answer key embedded in it, it's called data leakage. Let's say that in an organization, they find a way to detect that so that they can be sure that they're not deploying models in life and death situations that were trained with the answer key that isn't available in the

real world. Well, the nice thing about ml Ops is it provides this repeatable pipeline for your model. And so you can go back and test your old models. And if you've got remediation abilities, you can apply those very rapidly. And you can keep your old models from being stale. And you can also be testing to see if you're deploying them in situations where you were not training them. And so having this kind of tool can really help keep people safe. Because it is so easy. Once you've moved on to the next model, as an engineer to completely forget about the one you did before, we really need a system that is ensuring that we're staying fresh. And we're using all of our state of the art tools to apply to all of the models that we are using.

17:29
Every human being is subjected to being tired, and being tapped at 530 in the afternoon. And a little bias may creep into their selection. I mean, it has to happen with everyone, you know, I mean, I, I was once talking to a PhD in computer science, and he focuses on cybersecurity. And he was fished, he got caught on a phishing email, he said, Look, I was up for like, I don't know, the two days, and I was tapped. And I was walking. And as in my phone bag, I get killed. And so I think if you have a net, a net between you and your phone, you're not gonna get fit or a net between you and making those decisions. Or, Oh, we know that problem that you didn't, did you realize that you're just taking data from North America and not you know, the rest of the world? Because most of the day you see, that's the bias and, and in podcasting, it's the bias to most podcast listeners are in the United States. And you get that but this is plenty in Brazil and overseas, but most of them are here. But you just get lulled into that bias. And I think it's it's a way to at least admit that you're in a not an academic style, maybe a mental silo of data collection there. It's a it's an acquisition silo. Hmm.

18:42
Well, I'm beyond the bias, John. You know, although you gave me a brainiac award, I am a mathematician. And I've gotten a lot better at coding, because I've been working with my young colleagues to learn better practices, but I'm a mathematician. So code that a software engineer writes with my understanding of machine learning in partnership is going to be way better than code I'm going to write by myself. So, you know, not just bias, but also a test based carefully created code driving or model creation. I think it'd be critical. Yeah.

19:21
You know, I like to tell my kids that they all should take statistics in high school, my wife took graduate level statistics and I think if you're going to be a car mechanic, you should know statistics. I think if you're gonna drive a bus, I mean, I think it's like, you know, a one to one or one today, maybe maybe better than algebra to have that because you can look and say, well, this probably would have been probably what this is my happen I'd happen at least at least you won't be bamboozled by a lot of sophisticates with with some, you know, petabytes of information tossed about here and there. Speaking of petabytes of information, I think you have a meet up where if you're listening to this and you're at that exalted level, and you want to discuss papers on machine learning, you You can go to meetup.com and search for a machine learning paper club. Is that right? Yeah. So if

20:05
you want to have a dialogue and learn with me and my colleague, Julius, we just do this in order to keep ourselves fresh and make sure we know what is the state of the art.

20:19
I'm thinking, maybe, maybe you should bring in the developers, the developer should just learn and see what's going on there. Because that's really the the best way to get, you know, I've said this on my podcast many times the, the best Chief Technical Officer ever saw was a guy with a degree in ecology. And they said he could learn to code, but they said, Look, I see the whole system, you guys are just, you know, you argue about hot sauce over there. You don't realize there's this whole building that and everyone gets a little too focused on these things. So there's a term I want to bring up that kind of a catchy term is called deal brittleness on a few talked about or not so so. So what is data data brittleness?

21:00
So data brittleness is the idea that, if you perturb your data just a little bit, you're gonna get a vastly different prediction from your original piece of data. So I think we've all seen examples, like people put little pixels on a stop sign, and the self driving car no longer recognizes as a stop sign. And there's a lot of research currently about how to avoid that or how to detect that. And one way to avoid it might be to do an ensemble model so that you're using two different machine learning architectures that are using different information to make a prediction. So that a little change in one corner doesn't impact the final prediction is globally. I think that that's an area where many organizations might have proprietary special ways to identify data brittleness, and to address that, based on what their applications and their specific data needs are. And I think we're that can feed into ml Ops is that the ML ops tool then can have those those advanced solutions available to all engineers who may be trying to create solutions, or an organization.

22:16
And maybe there's a bias there where, well, the system has always worked. It's bulletproof. It's always worked. And then I would assume that the data in this system will just flow right into the system be sure why not? So you don't you don't know if you don't know what you're talking about. So this has worked fine for five years, and maybe it has, but guess what doesn't doesn't mean it'll work in the next system does it.

22:37
And I think this is one of our danger zones, when we're writing ml, for defense purposes, is that we're all using open source software, and we should, but maybe we need to tweak it a little bit, because our adversaries can guess pretty successfully, exactly what we're using for image recognition or for, you know, any other solution. And so we need to be aware that it would be possible to reverse engineer the predictions that we're likely to make.

23:11
Well speaking predictions, a lot of machine learning is used for predictive analytics. And so I'm gonna take and put the predictive analytics hat on your head, and say, Okay, now I want you to predict what's going to happen for five years with this whole concept of ml ops, you think it's gonna be widely adapted, you think there's going to be an incident people are going to back off or so when you see this whole concept going the next five years. So

23:36
I think that it's absolutely going to grow. I think that it's anybody who has sat at their computer and tried a million different things to get a model to work well, and then can't remember at the end of the day, what they did, once ml ops to track what experiments they did and how they performed. And I think the real question is going to be, are we going to see the few very large one size fits all solutions? Being the ones that are prevalent in five years? I think they'll certainly be prevalent today and in the next year? Or are organizations going to decide that their proprietary needs need proprietary solutions? And particularly within the DOD? How are they going to create architectures that are less predictable, and more resilient to the very specialized needs that they have?

24:31
I keep thinking of this dark hole, a black hole called Google or Amazon or are they going we are all we'd all well, do you worry a little head about this? Well do all the machine learning for and I can see them talk about a good salary they can convince but not Oh, don't you worry about this. What's and there's I think it's more of being like being a chef than being a McDonald's. I mean, there's there's some human elements Do you have to consider it and maybe a big monster company like Google can do some things but not not everything?

25:06
Yeah, I agree. I mean, what they offer is a candy shop of little pieces that I'd love to combine. But I suspect that final solution maybe needs to be a little more nuanced. Wow.

25:17
Well, thank you very much for your time this morning. You've been listening to fiddle tech podcast with John Gilroy. I'd like to thank my guest, Dr. Elsa shaver, corporate data scientist at linquist.

25:26
Thanks, John. It's been a lot of fun talking to you today.