SDS PODCAST EPISODE 948: IN CASE YOU MISSED IT IN NOVEMBER 2025

Jon Krohn:	00:00	This is episode number 948 our, in ICYMI in November episode. Welcome back to the SuperDataScience Podcast. I'm your host, Jon Krohn. This is an in case you missed it episode that highlights the best parts of conversations we had on the show over the past month. First off, I invited Dell's ish Gupta and Tyler Cox back to the show for episode number 939. In this clip, I ask Tyler, a distinguished engineer, what state space models are. To give you a little context in this clip, we're specifically discussing IBM's open source granite 4.0 language models, which include mamba state space layers that theoretically allow infinitely large context windows. Here's my conversation with Tyler. Now I want to dig into state space models and hybrid architectures, which you touched on Tyler. So the Granite 4.0 release introduced a hybrid architecture that combines states-based models with transformers, and that sounds important too, but maybe I've gone too far. Maybe we should start with digging a bit more into what states-based models are first.
Tyler Cox:	01:10	Yeah, so states-based models are a really important family of models, even pre AI usage, right? So if you go back to the 1960s, these were being used in space flight control. They've been used in population studies and economic modeling for a long time. The core construct is that you have an input signal that you map onto a hidden or latent state space with one set of equations, and then you have a second set of equations that translate that state into an output that's observable. So there's lots of different ways that you can construct a state space model to represent different problems.
	02:03	What's happened over the last 10 years or so is that state space models were investigated is part of the deep learning revolution is how do you construct your matrices for state space models to better model different tasks without as much classical feature engineering approach to it, right? So that was one important thing, but then

there's a couple of researchers at Carnegie Mellon and Princeton who really kind of drove this home over the last five years or so. So there's a great set of papers. I invite everybody listening who wants to know, learn more to look up the work here on mamba and structured state-based models and things like that. There's a series of papers from 2021 all the way up to 2025 still working on it, that introduced some really great optimizations inside of state space models to make them appropriate for sequence transformation and language modeling tasks. So you get this evolution of structured state space models, your S four paper, and then you go into structured state space models with selection and computation by scanning your S six, which turns into, hey, that's a lot of s's, sounds like a snake. Now we've got mamba and the mamba architecture really optimizes states-based models for the kind of compute profile that's needed to be relevant for the language model tasks that we're applying here. And so some great optimization work that happened there, some great mathematical insights into the matrix properties of those.

04:06 I won't be able to do those full justice here, but really, really interesting work that kind of accumulated into the Mamba and mamba two language model blocks that IBM pulled in to the granite four H series models, right? So composition here, you've got a nine to one ratio of mamba layers to attention layers in the granite four H family. So quite a bit of state space model in that hybrid. Those are, like I said earlier, those are linear context scaling. They are the granite four H models are no positional embedding. So they have in the dataset out to 512 K context represented in the samples. They're validated out to 1 28 K. The IBM team says, theoretically you should be able to push it past that, right? So some really great long context performance. I think one of the key measurement points in the release notes are if you take eight sessions at 128 K context on a micro, so a 3 billion parameter

model, you get about a 15 gigabyte of memory usage versus about 80 on a pure transformer architecture, right?

05:34 So some really great context reduction, which means that on more constrained devices on the edge, you can make use of more useful context in rag workflows and multi turn workflows and things of that nature. Just to here, re had mentioned if F eval, that's a really great benchmark for instruction following structured output tasks. The other thing that granite supports really well on is the Berkeley for function calling leaderboard, specifically B-F-C-L-V three that it shows up in the top five as we sit here recording among a bunch of other frontier models and hundreds of billion parameter models even at for the small 32 billion parameter footprint. So really, really punching above weight class there.

Jon Krohn: 06:36 Benchmarks have reappeared with considerable frequency in my interviews this year, and I expect this to continue into 2026 given the number of LLMs coming onto the market with ever more impressive capabilities. Permissions have been another big topic on the show this past year. This is especially important now that we're often working on projects that involve teams of AI agents working together where we need to be careful about information flows between these agents and the tools they use. Dr. Vijoy Pandey is developing Cisco's open source platform for the internet of agents tackling exactly these kinds of problems. In episode 941, I ask him about privacy and security when using AI agents, when we're working with something like tac, it sounds like it might be tricky. I can't wrap my head around exactly how permissions are granted. Just in time when you need to grant permission to an agent to be able to do a particular task and then you need to revoke that afterward. That sounds like it could be complicated. How do you handle it?

Vijoy Pandey: 07:39 We have to bring in a whole bunch of infrastructure around this notion of T back to enable the end goal, which is the zero trust for agents, which is I give you permissions to do something specific for that just in time and then for that duration of that task two or transaction, and then I revoke that token or revoke those permissions the moment you're done. So what else do you need? So in my head, the equation runs like this, which is zero trust agency is I don't trust any agent and I just trust it once it's proven that it's supposed to do X, Y, or Z for the duration of that X, Y, or Z, and then I revoke those permissions. So that is zero Trust for agents is a combination of the availability of trust, task tool, transaction-based access control. So all identity providers, authorization servers need to support tac.

08:44 That's step one. We need to have a passing entity for all communication that's taking place between agents and agents and humans that can pass that communication, that pass that discourse and figure out the tasks or the tool access or the transactions that are taking place between agents. So that's the second piece because that'll help us define what those tasks tool access and transactions are. So there's a semantic passing element, and then your adjusted in time comment basically implies that you get a token, you do that task in a very contained sandbox jail environment, and then you're taken out the moment you're done. So the analogy I draw here is I want to access a safe, which has a lot of money, but I want to withdraw $10 from the safe. Now you can give me, since I vjo, you can give me access as VJO to go and open that safe.

09:55 But then you can say, you know what? There are other people's money in that safe, so I'm going to give V Oy just enough to withdraw cash for the next 10 seconds and then move out. So that's like a task-based access control. But then I need to pass the communication that's

happening between myself and somebody else where we are talking about withdrawing 10 bucks and say, okay, vitro is allowed only to withdraw 10 bucks and that is the task he's doing. So let me just give you access for that $10 withdrawal and not sit around to withdraw all of the money from the safe. So that's the task based passing that needs to happen. And finally, I'll let you in into the sandbox environment, give you that authorization to withdraw $10 and then I'm going to shut the door. I don't trust you beyond that point, I will not let you linger around. That's safe. So that is a sandbox runtime environment that needs to happen. So is the hooks in identity providers to provide task-based access control is a semantic passing of the discourse of the communication to figure out what that task is and then a runtime sandbox environment to just do that task with that authority and then get out. So those are three things that need to come together for zero trust for agents to happen

Jon Krohn: 11:20 Semantic parsing, ephemeral runtimes and human in the loop approvals. And overall, you gave me a really clear picture now of what this all involves. One thing that I guess that I still don't quite get logistically, you said that you won't trust the agent even for the particular task tool or single transaction that you're going to approve it for until the agent has proven itself. How do agents prove themselves trustworthy in the first place?

Vijoy Pandey: 11:51 This is where the entire pipeline comes to the picture. So we are looking at identity, which is the first stumbling block that everybody's running into. And so one of the things that we're seeing is in the agency framework, the identity piece is the problem to solve first, even before you can start deploying agents at scale within the enterprise. But then as you pointed out, there are other aspects of trust, there are other aspects of semantic parsing. So there are these other aspects of their entire pipeline that we need to solve for. So coming back to trust, the simplest

way you can start with is saying, is there a directory somewhere that allows me to discover agents that are trusted? So I want to find a financial agent not from a particular vendor maybe, but also if there are 10 vendors, I want the best of breed financial agent from a vendor that is highly reputable and highly trusted.

12:58 So the simplest version of the question, the answer to your question is, is there a directory which tells me which agents are trusted agents? But then the next question is, so that's where the directory comes in. So we have a directory where you can discover agents through capabilities and through things like reputation and trust. But then the next question would be how is that trust enforced or attributed? Is it crowdsourced? Crowdsourced could be one thing. So Ted people have used it, it's like the Apple app store which says Five stars trust or Trustpilot score, and it's like, yes, it's awesome, but if you want to be a little bit more mathematical and provide some rigor, then you go to the last pillar of that four pillar thing that I talked about earlier, discovery, identity, communication and evaluation. You look towards evaluations and you look towards evaluating agents and multi-agent workflows and saying, over time, I've built trust by evaluating this agent. And that trust can then feed back into the directory's reputation score and say, yep, all good to go till something put ups in the system. Because these things are constantly evolving

Jon Krohn: 14:18 From multiple agents. We move to multiple metrics is more always better. In episode 937, I speak to fabi.ai co-founder Marc Dupuis, about how to manage and evaluate several variables in a way that our teams don't lose sight of their goals and crucially that the AI they're using stays on the same track as well. You have a decade of experience prior to co-founding Fabi as a product manager for lots of great companies, triada, clar assembled, I'm probably mispronouncing all of those

company names. That's all good. And so at those companies, you build platforms, APIs, self-service analytics, drawing from that experience when you are managing metrics platforms like Fabi would allow product managers, if you think about yourself in those previous roles, if you had a tool like fabi, you would be able to very quickly spin up complex engagement metrics for your platform or your APIs on the fly. Would you be worried then about, or how would you prevent metrics sprawl where all of a sudden you just have tons of different measures flying around, you're not really maybe people on the team that aren't going to be clear about what we're really building towards. Does that question make sense? Do you see that as a potential problem or Yeah,

Marc Dupuis: 15:40 It makes a lot of sense. Yeah, the question makes perfect sense and I do think that there's always this risk and also fear, and I think justified fear that if you give everyone ai, then the AI is going to go and reinvent the metric every single time you ask a different question. And so we could talk about semantic layers. That's probably,

Jon Krohn: 16:00 Oh yeah, I hadn't even thought that's a huge problem too. Yeah, exactly. Where you could end up, you could have each person in the organization say, just ask their AI enhanced BI tool, how are we doing on this metric? And every time it's coming up with a new way and pulling different data. Yeah. Oh my goodness. Yeah, that's an even bigger potential problem here.

Marc Dupuis: 16:19 Yeah, so that's maybe another episode for us. Talk about semantic layers, and that sort of ties back to the guardrails as well that we talked about Fabian, make sure the team's actually supervising and collaborating and working with the business stakeholder. But to go back to your maybe original question, if that's not a hundred percent what you had in mind, which is how do

you make sure that, and actually let me make sure I ask the question back, is the question about how we actually, lemme ask you for clarification of the question to make sure I understand because that's what I was thinking about when you asked the question.

Jon Krohn: 16:54 Yeah, I mean we should definitely answer the question that you brought up. But what I was thinking about you would end up with the problem that you just brought up in what I described, where if project managers, the people across the organization, executives, product leads on individual products, if everybody can be using automated BI tools to be creating metrics on usage, you could end up having a lot of, it could start to muddy the water around what the organization as a whole is building towards. But even if there was agreement, so what I think is more interesting about the question that you heard is that even if you have agreement across everyone and humans are being consistent with their definitions and the humans all kind of have a clear idea of what they're building towards, what the metrics that they're trying to optimize are with the product that they're building, the AI could be surprising them by recalculating things a different way. So does that help kind of clarify what I was originally asking?

Marc Dupuis: 18:04 Yeah, I think it does. And when you asked that question, I actually have, there's two things that come to mind for me. The first is that the role of the data team, I don't think changes in the sense that data team still needs to be there to help make sure there are consistent metrics and that we're all working towards the same North Star as a whole as an organization, as a department or whatever. And that's not changing. What you don't want is you don't want AI that this is why we talk about dashboards. I don't think dashboards are disappearing in that sense. I think that dashboards are actually going to be much more powerful and useful because there's going

to be fewer of them, but the ones that are actually built are going to be curated and managed by the data team because they are tracking the actual core metrics that matter for the business is going to be tracking your churn, your A RR, your retention or whatever it is, and the data team is going to be spending a lot more time making sure that those are correct.

18:57 Now, when I think about AI for the business, what I think about is all the other questions that haven't yet made your way into your North Star metrics or your OKRs. So as a product manager, maybe going back to your original question as a product manager, and I still do today as a founder slash product manager, I'm constantly exploring new ways to ask to think about the data. So you're thinking about your user activation for example, maybe at some point if you're a mature organization that's growing, your activation is very well defined, you have to friend seven people on Facebook and then that's your activation metric or that's our North star that's set, we're good. A lot of organizations don't have that or it's evolving or you're interested in your product. And so you don't want to go and model what's effectively a hypothesis into your data and pull in the data, create these new tables and then feed that and go through the entire process and feed that through your BI solution before you've actually taken the time to figure out that's actually what you want.

19:54 That's where you can, thanks to AI actually let someone, ideally a pair of a data scientist and a product manager or a data scientist and a CSM or whoever work together to explore that messy phase and see like, okay, does this metric actually make sense? Is this how we want to think about it? And a lot of times I think what you'll find when you do that is, again, I'll draw my notes experience to the product manager, you'll look at data, I'll be like, actually, that was the wrong question I was asking. Let me rethink about what I'm really asking and I'll get back to you. And

as a product manager, I can start asking you my own questions off the data that data s security for me or off the raw data, I'm going to get to my own answer much faster. And then we can also just sort of experiment and sit with that metric for a minute, for a month or a quarter.

20:40 And then if it's like, okay guys, guys and gals we're looking at this metric and it's been the same metric for the last three months, now maybe it's the time for us to go and build this into, add this to a dashboard. That's the point where you can say, okay, well how do we pipe this data into our data warehouse and how do we add this to our data modeling and feed it all the way through? So I think you just need to carefully think about, okay, is this a metric that we know that we've established and that is set and if so, and let's go through the proper channels that have the right guardrails. If not, let's actually take the room to explore that before we go and overly invest in the measurement.

Jon Krohn: 21:22 It's important that we return to thinking about the human motivations behind every AI project. And in my final clip from the month that was Santa Clara University, professor Maya Ackerman and I talk about the importance of keeping human interests and welfare at the center of the conversation. So you've previously argued that the sci-fi dream of AI as, yeah, I mean we talked about earlier as an infallible answer, machine is disempowering and it's misguided. And in chapter 10 of your book, you note that disempowerment doesn't scale, but human flourishing does. And so how can investors and companies reframe their perception here? So reframe engagement metrics away from stickiness and towards sustained creative empowerment kind of like you and David have with wave ai. Yeah,

Maya Ackerman: 22:17 It's interesting that this addiction model persists even when the biggest problem with gen AI products is this kind of one hit wonder phenomenon. When a person feels really like they're really unnecessary, let's say there's a little app where I upload my photos and then it does something fun with them, I might enjoy it once or twice, but if I feel like I'm not really doing anything, I don't have any control over what comes out, even if it's kind of cool, I'm very unlikely to come back. So we had a whole bunch of one hit Wonder Gen AI products. People want to feel that they're doing something. People want to express themselves, they want to realize their own ideas. And industry and investors in particular have been very, very, very, very slow to realize that. I believe that the reason Chad GBT is successful is because it for those who want to give of themselves, for those who want to really collaborate, it makes that possible and that's why it's so successful. So it's sort like we need to sort of snap out of these old outdated exploitative practices that are, a lot of us sort believe in them as gospel and really explore what's meaningful and appropriate with this technology, which I believe from my experience in business otherwise, that the real potential is to really elevate humans in our capacity.

Jon Krohn: 23:50 Yeah, it's a great idea and I agree with it wholeheartedly. It does seem like it's going to be hard to convince product managers. Do you have any sense, I realize this is a really tricky question. I don't have a good answer to this, but you're a lot more creative than I am. How do we convince big enterprises product managers to move away from just stickiness to empowering people?

Maya Ackerman: 24:16 I think we need to explain to them why something like why the products that are successful, why they are successful. Chacha PTI has not been beat yet. It's the number one, it's one of the most successful products in history. And that's the case, not because you press a

button and it replaces you, but because of power users, you always have to look at power users because those are the ones who are the customers that love you, show you where the product needs to go, and those are the users that go deep that become better as a result of using che pt. I've heard people say that their vocabulary expanded, the writing style got more diverse as a result of using CHE PT when you use it with a certain kind of intention. And also to show examples of how this sort of replace of paradigm has often failed in a lot of smaller products, but also how it ultimately fails collectively. This whole, everyone optimizes for themselves really fails when we want to apply AI to replace human workers because if we don't have any more human workers, the entire economy collapses. And so it's really time to wake up from these incredibly greedy principles that are guiding our economy and to think more holistically in this moment, not assume that the old principles are just going to keep working and somehow magically everything is going to work out.

Jon Krohn: 25:41 Perhaps gen AI will kind of force this shift that you're hoping for in product managers and in enterprises. You mentioned in your most recent response this idea of diverse responses. And so in interviews you have previously contrasted convergent systems that gravitate towards safe average outputs with divergent co-creative tools, presumably like your lyric studio and melody studio that widen the search space. And like you just said, broaden people's vocabularies, increase the diversity of their writing styles. So as AI scales to creating music, to creating video, how do we keep models from nudging creators toward averaged risk averse aesthetics where you hear another pop song that sounds the same, how do we instead get diversity?

Maya Ackerman: 26:43 Yeah, it's really impressive what industry has done to gen ai. Gen AI has always been about expanding possibilities.

That's how it was born. Going to new spaces, we think of creativity as this massive search of possibilities. The next line of lyrics, there are billions of options. How do we search, right? You are about to play the piano. John, you were talking about playing other people's pieces, like we're all told, where do we even start? What could be my first note? What could be my second note? It's a search space problem, and machines are phenomenal at just exploring different unlikely possibilities and even in trying to take you to places that are pretty good, but not very common. The problem is that science fiction has convinced investors and some entrepreneurs that it's better to create an all-knowing Oracle that gives you the most expected, most safe, most.

27:41 You've seen it a million times already answer, which is maybe good for a fact lookup table, but it really limits what gen AI can do. And it's never actually going to be very good at being an all-knowing Oracle regardless of what we do. And so when you look at chat chip pt and you're like, oh, AI is not creative, that's because chat is optimized for the opposite. Inherently, from a technical standpoint, you have to take creativity into account from the beginning in the way that you construct this machine brain. But in the grand scheme of things, in the grand scheme of things, it's not that hard. If we figured it out as a team of three, granted we have David, but still a team of three people originally at Wave ai, then I'm sure somebody like OpenAI, Microsoft and Google can figure it out. It's just not their goal. Their goal is not to open our minds. Their goal is to replace search, and I think that goal needs to be modified.

Jon Krohn: 28:41 Alright, that's it for today's, ICYMI an episode, to be sure not to miss any of our exciting upcoming episodes. Subscribe to this podcast if you haven't already. But most importantly, I hope you'll just keep on listening. Until next time, keep on rocking it out there. And I'm looking

forward to enjoying another round of the SuperDataScience Podcast with you very soon.