SDS PODCAST EPISODE 894: IN CASE YOU MISSED IT IN MAY 2025

Jon Krohn:	00:00	This is episode number 894, our "In Case You Missed It in May" episode.
	00:21	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.
	00:31	In the first of my four highlights from May, I speak to John Roese, Global Chief Technology Officer and Chief AI Officer at the computing giant Dell, what a guest. John and I had a detailed conversation about multi-agent teams, quantum computing and the future of work in episode 887, and in this clip I ask him to define two terms: "AI agent" and "RAG-based chatbot".
	00:54	Let's talk about the natural next step that has emerged after generative AI, which is agentic systems because as generative AI has become powerful enough, as LLMs have become reliable enough, we've started to be able to rely on them more and more on their own. Do you have, John, your own definition of what an agent is?
John Roese:	01:16	Yeah, I'm going to give you a bigger picture view and then I'll define an agent. So AI attached to the enterprise, applying AI to the enterprise, actually has two different parts to it, of which only one we've done so far. Agents are the second one. And the reason for that is the source of differentiation of an enterprise. A lot of us in the industry have said this over the last couple of years, even though people weren't necessarily paying attention, but there were two parts that make an enterprise an enterprise, the real core source of differentiation. The first is your proprietary data. You know things other people don't know. That's actually very powerful. That's why you don't share your proprietary data with people. My customer list is very valuable. My source code is very valuable. And those are a sustainable source of differentiation. Even if

the people change, the brand changes, the world changes, having proprietary data is very, very important.

02:07 The second source of differentiation is the unique skills in your organization, that you have people that can do things better than other people. At Dell, we have the best thermal and cooling people in the world, the best client developers in the world, the best storage software developers in the world. And the result of that is that translates into better products, interesting innovation, patents. And so if those are the two sources of differentiation, and the journey we're on is to apply AI to an enterprise and those are the two things that matter, it's interesting because for the first couple of years at GenAI, we actually went after the first one. A chatbot, a rag system, all of these things are just tools that allow us to unlock and create value from our proprietary data. What is a rag-based chatbot?

02:52 It is a tool that takes proprietary data and makes it generative. You could take all of your service information and if I gave it all to you in raw format, it would be of no value. If I embed it into a vector database and present it to you through a generative interface, you can ask and answer any question on anything I know, anywhere. That is incredibly powerful, and we have been doing that now for about a year at scale in the industry and it's transforming everything. We're getting huge value out of this. In fact, almost all of our projects that are in production are just that. They're a generative capability to unlock our proprietary data in novel ways that just changes the curve in terms of productivity. That's great.

03:33 Agents are not that. Agents go after the second one. They are about the digitization of a skill. They're about saying, "I'm not just interested in unlocking the data. I'm interested in distributing the work. I actually want an AI that doesn't even require me to do a task, that it can

actually operate autonomously. It can operate without human intervention. In fact, I'm not even going to tell it how to do the job. I'm just going to give it an objective and let it go, and I'm doing this aligned to the skills that I need it to do." So for instance, when we think about agents in the enterprise, now there's two views of this in the current thinking. One thinking out there is that agents will be replacements for multi-dimensional humans that can do everything. That's AGI and ASI. We're a long ways away from that. The reality of agents is that they are actually the digitization of more narrow skills.

04:25 I use the self-driving car example. I do not have a self-driving car today that can drive anywhere in any situation and navigate it successfully. What we do have is self-driving cars. They've been in San Francisco and other places, where if you geo fence it, if you narrow the scope, we see this in the trains and airports, there's no driver on them because it has one job. It moves from terminal to terminal without a human intervention. Well, that's what's going on with agents. The first generation of agents are saying, "Could I take a task, a skill and could I move it into AI not as a tool that a person uses, but as a manifestation of that skill autonomously, that I can just tell it to do something. I can give it an objective and it's smart enough to figure out how to reason through that objective. It has access to a set of data and it can deliver an outcome equivalent or better than what a human would've done for that particular skill."

05:60 And yeah, there might actually be humans doing those specific jobs that might not do them anymore because agents can absorb them. But what you don't have is a fully well-rounded entity that is the equivalent of a full human being that can do lots of different things. Think about in your life, how many different things can you do? Well today the manifestation of agents can probably pick

off a few of those, but what they can't do is pick off all of them and create a completely equivalent of your whole well-rounded human being, including your ethics, your morality. That's a really hard problem. That's AGI and ASI, a different journey. And so bottom line is you take these two technologies, first gen GenAI, which is what we call reactive AI, that a human is in the loop and the human asks the AI to do something and it gives it an immediate response, but ultimately the human is the doer of the work and these are tools around the human.

06:07 And then you move over to this kind of second generation of agentic AI, which are complementary, and now you have a situation where the human is on the loop, they're the supervisor and all they're doing is creating objectives and delegating work. And now the AI independently is able to take that task, figure it out, run with it, and even run with it in perpetuity that it may never go back to the human being because it's been delegated below the machine line. The reason it's so important to distinguish these is that one, they aren't even the same technology. Well, this one, the center of the universe is a large language model with some data around it. It's a very static data set. An agentic environment has large language models, but they're used for part of the equation. They act as somewhat of its brain, but it has a body, it has a knowledge graph where it creates its own representation of data that it represents what it's learned and its memories and its evolution of skills.

06:57 It has interfaces around it that allow it to reach out into the real world, something called tool use and function serving, where it can actually go and activate a tool and interact with the world and perceive things. Very different technical architecture and quite frankly, appropriately so because it's solving a different problem. Now fast-forward into the future of an enterprise. Well, yep, still got proprietary data and still got unique skills, except now I

		have a path to digitize both of them. And that's the thing that's going to profoundly change most enterprises.
Jon Krohn:	07:26	John touches on an important point about the future of AI being in the interconnectivity between tools such as AI agents and RAG-based chatbots, and I recommend listening to the entire episode to hear more about how John's team applies such integrations at Dell. Based on the social-media response, listeners absolutely loved the episode.
	07:47	My next clip is from 885 with Jeroen Janssens and Thijs Nieuwdorp. In maybe one of the liveliest conversations I've had on the show, we talk about the coauthors' latest O'Reilly book on Polars, and how their work with the major Dutch utility Alliander helped them to write Python Polars: The Definitive Guide.
Jon Krohn:	08:05	So earlier in the episode, you mentioned about a real-world implementation of Polars and maybe as you said, maybe the first ever production instance of Polars. And so am I right in understanding that's Alliander? I'm probably butchering the pronunciation of that.
Thijs Nieuwdorp:	08:19	Yeah. Alliander, it's a power grid provider in the Netherlands. Also, they provide the infrastructure for both electricity and gas in a third to half of the Netherlands, I believe.
Jon Krohn:	09:31	Yeah. So the largest utility company in the Netherlands therefore. I can't even say Netherlands. That's how bad I am at Dutch pronunciation. Netherlands, that's actually easier, isn't it that way, isn't it?
Jeroen Janssens:	08:43	For us, it is.
Thijs Nieuwdorp:	08:46	Oh, that's what you're talking about. I was wondering what is this country?

Jon Krohn:	08:49	Where are these Netherlands?
Jeroen Janssens:	08:51	That ain't no country I haven't heard of.
Jon Krohn:	08:53	Yeah. So tell us about that project and what it was like. And actually, it'd be interesting to know was there overlap in working on the book and working on that project and did working on a Polars book help with a real-world implementation? Anyway, that's an interesting side question.
Thijs Nieuwdorp:	09:09	Yeah.
Jeroen Janssens:	09:10	Yeah. So the origin story here is that Thijs and I, we were both very excited about Polars. We were writing a book about it. And then all of a sudden, it became clear that at Alliander we needed to speed up the pipeline, we need to lower cost, we needed to process much more data. And in the current state, that just wasn't possible.
	09:34	It was a combination of not only Python and Pandas, but also R Code. So it was very inefficient. To give you an idea, we were running this on a single AWS instance that had over 700 gigs of RAM, 700 gigs of RAM. And so yeah, we can provide you a link with more backstory to this with some actual numbers, but we were very excited, and we were like, "Hey, let's try this out. Let's do this."
	10:02	At first, the team was very hesitant where there are two people or three actually, we had another colleague, three people promoting Polars that is being developed at Xomnia. So they were very skeptic, understandably.
	10:18	So what we did in order to convince them is to just take on a very small piece of code, some low-hanging and benchmark it and re-implement the Pandas code into Polars and then just show the numbers. And by then, they were immediately convinced, "All right, this is indeed

way faster, uses way less memory. Let's try this out. Let's take on this huge code base piece by piece, by translating not one-to-one, because you can't do that. You really have to reason about the inputs and the outputs and then do it in an idiomatic way."

10:56 You just translate Pandas to Polars. And I think it took us, well, what, six months, a year? I don't even remember. But eventually, I left that client at that time. But there was a moment like, "Okay. We can now get rid of R and Pandas as a dependency of this project." And it's been running smooth ever since.

Thijs Nieuwdorp: 31:40 Yeah, definitely. Yeah. I think ultimately, the size of jobs at the beginning was about 500 gigabytes for just that task of doing one calculation, and we shrunk it down both being a consequence of implementing Polars, but also, as we were going rehashing some of the code structure that we were using in the project, we hashed it all the way down from 500 to 40 gigabytes, which makes it a lot more doable to-

Jon Krohn: 11:29 Wow, 10X.

Thijs Nieuwdorp: 11:48 ... my calculations.

Jeroen Janssens: 11:55 And so the second part of your question was, okay, how did this influence each other, the book writing and putting it into production? And yeah, it was a perfect match because when you actually need to put it into production, when you have a real problem to solve, that's also when you start to notice the limits or maybe inconsistencies or missing functionality.

12:23 For example, there was this random sampling with weights. That's something that you can do in Pandas. You just give it another column that indicates the weights for the sampling. That's something maybe even up until this

point, something that Polars doesn't have. Luckily, that was for an ad hoc analysis that we had to do. But at that point, it becomes clear what Polars can and cannot do.

12:53 Also, when you write, you start to look at things from a little bit of a higher level. So sometimes, we noticed inconsistencies in naming or missing methods like, "Hey, why is there no inline operator for the XOR operation?" That's something that nobody ever thinks about. But when you need to put in a table in your book and you need to fill in all the pieces, that's when you start noticing these kind of things. So we were able to also submit some issues, maybe even a few pull requests to Polars itself along the way.

Jon Krohn: 11:29 From writing nonfiction I turn to CEEK, C -E -E -K a new platform for education with VR capabilities. In episode 889 I talk to CEEK's founder, the space engineer Mary Spio, about the potential for CEEK to revitalise the way we learn and make even specialist education accessible worldwide.

13:53 Another thing that's really cool about CEEK is how it could potentially, I mean platforms like CEEK or VR in general, how it could help with education.

14:00 So in the US for example, there's a shortage of 400,000 kindergarten to grade 12 teachers. So primary and secondary school teachers and post-secondary institutions, there will also be shortages coming because those workforces are unusually old. And so college and university faculty are going to be retiring. And so education is failing to attract and retain faculty and has equity failings a cascade into long-term disadvantages for students. And you've described previously how one professor can teach a hundred thousand students via CEEK. So where do you see VR CEEK potentially having a long-term impact on improving educational outcomes?

Mary Spio: 14:45 Right. Where we see it is in multiple ways, right? Because right now, like you said, the shortage, I saw a stat somewhere that was like, Europe needs this much US North America needs this much. It was like 4.6 million total. And then it says Africa needs a miracle because it's like when you look at the rest of the world, the shortage is so dire. And so a platform like CEEK, we allow its single individual to be able to teach at scale so they can present their course virtually, and then people are also able to experience it. The reason why we're getting interest from the likes of the eVTOLs of the world is because this is a brand new industry. So for example, you have the mass displacement of a lot of the current jobs as we know it, and they also have to train people for all these new autonomous vehicles, new, all these new industries that are coming out as a result of automation.

15:54 And you need, for example, a hundred thousand pilots, eVTOL pilots within the next few years, which means you have to train a million, you cannot put a million people in these very expensive aircraft. It's also very dangerous. It's a danger to the person, it's a risk to the aircraft. But on CEEK, you can have a million people training at the same time with the VR headsets, and so that person is able to really scale themselves and have all these people train around the world. And that's basically what we're building today. And this isn't just a sample scenario. We are working with the leading electric aircraft company and they're exploding. They have a massive backlog because right now, when you look at fuel-based airplanes, helicopters that are being used for logistics and delivery and stuff like that, that costs about 4,000 an hour in fuel.

16:55 It's costing about 300 an hour. So even beyond being good for the environment, it's also good for business, which is why they have these massive backlogs and they have this need to train people at scale. And these are things that you just can't do physically, which is why our

platform is now in demand. And then the other aspect is the fact that in VR, the brain hasn't developed the ability to differentiate between what you do in VR for the first time, we're creating memories. So it's almost as if you're actually flying the aircraft. It's almost as if you are actually moving the equipment and doing all these things. So you're building memory, which means you're building experience. So you can now show up day one, now able to train in that helicopter because you've gotten that a thousand hours or however many hours that you need before you can step inside the real deal.

18:00 And the same thing applies whether it's elementary education, primary education. For the CPR, we built CPR for the children's hospital for adult, infant and child CPR. The interesting thing is this was for new mothers because actually before I did the CPR program, I didn't even know there was a difference between infant child and adult CPR. And a lot of new moms don't either. And so by them putting on the headsets, they were able to learn and train, and they felt more confident than watching a video because they were actually holding the baby and they were doing all the different actions. And the clinicians and the EMT that we worked with also felt better equipped. The reason Baptist Health was looking at the nursing residency is because there's such a huge gap between nurses. The average age of a nurse today is 50 years old. That's how big the gap is because a lot of people are staying a year or two and then they're leaving.

19:08 And the reason there's such a high turnover is not because of competence, but rather confidence. A lot of nurses by nature are very caring, so a lot of them are afraid that they don't want to hurt someone. So by now being able to learn and make the mistakes, they don't want to make a mistake on a real person. Now they can make the mistake, they can practice, they can do all of these things in VR and feel more confident to be able to

do it in real life. Then there are other areas that you just need to do in VR versus in person like intubation where they're learning how to insert a tube into somebody's throat, and a lot of times they'll perforate the throat. Today, what some hospitals do to train is they hire low income people and pay them, and then they can test intubation on them.

Jon Krohn: 20:05 Oh my goodness.

Mary Spio: 20:09 Who wants to perforate their organs for $50? I mean, not me. I don't want to do that. And unfortunately you have the homeless, the elderly. Some people also do the testing on the elderly with Alzheimer's and to train the nurses how to do the intubation. Yeah.

Jon Krohn: 20:34 Oh my goodness. That's shocking. Yeah, VR definitely seems like a more humane way to be laying that. Oh my goodness.

Mary Spio: 20:42 And you could do that without the risk of perforating anybody's organs. Yeah.

Jon Krohn: 20:46 Being able to make mistakes and learn from them is such a core part of education. Learning from past mistakes is also an unavoidable part of running a business. My last clip is from episode 891, where I speak to Martin Brunthaler about the lessons he learned as the founder of Adverity.

Jon Krohn: 21:02 We have listeners at home. A lot of our listeners are either hands-on data science practitioners like machine learning engineers, AI engineers, data scientists themselves, or people who are interested in building products or companies that leverage generative AI. What are the kinds of lessons that you've learned in implementing a product like data conversations at Adverity? What do you need to do? What are all the things you need to line up in

advance of bringing in a large language model and having conversations work effectively with data? You talked a moment ago about the issues that you typically see without this conversation in place where people have a dashboard and it's not exactly the information you needed.

21:51 It's too fixed in its outputs, and so then people end up going and digging under the covers into the raw source data to try to really find answers, which adds strain onto the data analyst team. So, I get all of the advantages of being able to have a conversation with your data, but what are the things that you at Adverity, that are listeners if they want to be making this similar transition, what do they need to get right in order for that conversational aspect to work out?

Martin B.: 22:19 So I think one really critical piece is the quality of the data underneath. So, there's many aspects of data quality if you will. Also from an academic perspective, you can list those out, but from a more practical perspective, you need a complete data set that is also very well aligned with all the various sources that you have. So, harmonization plays a role in this as well. We built up actually a data quality component in our platform that helps you monitor all those issues that you can have in your data. There's specific monitors for data quality in marketing.

23:01 There's a concept called naming conventions, for example, for campaign names that we can monitor and act on in an intelligent manner. But there's also simple things like if you onboard a generic source from a database or from a REST API, all the data types need to be aligned, data formats need to be aligned. You want all your data to be harmonized in UTC, for example. You need to clean up some stuff. This is also why there's some transformations going on usually either by splitting up, combining various

sources, and all those things, but I think it's very critical to get the quality right. You need to be alerted. If something's going wrong, you want to prevent, not saying dirty, but problematic data sets to hit your production environment. I think we can help in this discipline quite a bit.

Jon Krohn: 23:50 You could help in the discipline by having these data quality reportings built into the platform.

Martin B.: 23:56 Yeah, but also the multi-layer approach to this. So, we keep always a raw data set that can then be used as a starting point to reiterate on transformations for example. So, you can always go back to the previous state and improve your transformations. There's also obviously today an AI system helping you to compose those transformations. This is specifically always very useful for those type of generic sources, but it's a simplified data of wrangling exercise, if you will. Then once you're satisfied with that, there's a component that helps you monitor the quality as it flows through the system. There's an anomaly detection and all the things that you want to monitor.

Jon Krohn: 24:41 Right, right. Yeah. So, built-in anomaly detection would be key to this working out. There's a huge amount of breadth of capabilities that you could potentially get from a conversational interface. When you are designing a conversational product, how do you figure out, okay, this is the range of things that we're going to support or not support, and then how do you select the right large language model for that breadth of features that you decide to support? Yeah, let's start there. I have more follow on questions from that, but I feel like that's a good starting point.

Martin B.: 25:23 Yeah, I think it's useful and maybe one thing to add to the previous question in terms of quality, like I already

said, the data dictionary descriptions, understanding of lineage is very critical as well. This goes also into the design of our conversations interface and how people can interact with that. We iterate very quickly. So, we are going through, I'd say, a pretty fast-paced development cycle with adding features every week. We have a dedicated team taking care of benchmarking and analyzing the quality of responses. So, using frameworks to monitor that and the data science team is having a continuous test on...

00:24:44 We have a predefined set of responses that we expect from our questions and we can monitor on those and improve and test models as we go. To be fair, at the moment, we're committed to one model, but there is also the plan to use different models for different aspects of our capability. So, for example, we could use a different model to compile our SQL query, a different model to do the preflight qualification of a question, a different model to do the actual conversation. So, yeah, that's also possible.

Jon Krohn: 26:45 Nice. Obviously, the questions that I asked you were tricky because I'm trying to get at what are the things that people need to be doing in order to build these kinds of conversational interfaces like you did, but obviously, there's proprietary things involved.

Martin B.: 27:02 Yeah, I think there's no trade secreting in building, if you will. A lot of LLMs and the type of APIs they offer are similar in regards to their capabilities and you see all models reaching the same capability and basically the leaderboards change. Just every other month you'd have another leader, but everyone's catching up to the same state of quality, if you will. I think where it then boils down to is how you put the components together to create a compelling and exciting use case on top of that. I think in terms of how this works from a technical perspective, it's pretty straightforward.

	27:51	You can qualify a user input into a type of question, select the model that you want to run with, basically feed it with a system prompt and additional information about the model, which is very critical to get the answer right, use this to create a SQL query, verify it's actually a valid query that can be executed, fire the query, use the data to run some basic analysis and create a decent nice answer for the user. For us, the use case then circles a lot around the table that we generate from that response.
	28:26	Because what our approach to this is, first of all, in terms of democratization, we are targeting two sides of the business. One of which is IT, and the other one is in the business user. Both have a requirement to access data. So, rather than going through a full chain of various teams, so it used to be that you had to create a ticket to get access to a set. The data set would then be prepared within two weeks and put onto a Snowflake table or whatever today, a Snowflake table. It used to be something entirely different. With this, you can actually run the query, create a table in near real time available for your further analysis and that's exciting for us.
Jon Krohn:	29:10	All right, that's it for today's ICYMI episode. To be sure not to miss any of our exciting upcoming episodes, subscribe to this podcast but most importantly, I hope you'll just keep on listening! Until next time, keep on rockin' it out there and I'm looking forward to enjoying another round of the SuperDataScience podcast with you very soon.