SDS PODCAST EPISODE 953: BEYOND "AGENT WASHING": AI SYSTEMS THAT ACTUALLY DELIVER ROI, WITH DELL'S GLOBAL CTO JOHN ROESE

Jon Krohn:	00:00	My mind-blowing guest today helped his company's revenue grow by $10 billion last year, while AI helped cost go down. That decoupling of revenue and cost had not happened in the company's 41-year history, but now thanks to AI, it has, and he tells us exactly how they did it. Welcome to episode number 953 of the SuperDataScience Podcast. I'm your host, Jon Krohn. I'm honored to be joined on the show today by John Roese, global CTO and chief AI officer for Dell Technologies, The Computing Colossus. This episode is densely packed with John's practical AI guidance, including on knowledge layers, MCP, agent to agent protocols, and critically getting an ROI on AI investment. Enjoy.
	00:42	This episode of SuperDataScience is made possible by Anthropic, MongoDB, and Y Carrot.
	00:50	John, welcome back to the SuperDataScience Podcast. It's great to have you back. I couldn't believe that you were on the show just recently, earlier this year, and now you're back.
	01:00	And I know we have planned for this episode and the audience is going to love it. So John, first off, where are you calling in from?
John Roese:	01:08	I am up in the mountains of New Hampshire right now. It's about eight degrees outside.
Jon Krohn:	01:13	See, last time I think you also were calling in from the mountains of New Hampshire, but we had some specialized swords in the background. You must have been in a different spot for recording.
John Roese:	01:22	I was. I've actually graduated to a proper office, so that's a good thing.

Jon Krohn:	01:28	Nice. Great. So we're going to talk about your predictions for 2026. So this is the very final episode of this podcast of 2025. And you are well known for your tech predictions for the upcoming year. So I'm going to have a link in the show notes to your 2025 tech predictions, but they were spot on. And in particular, you talked about 2025 as being the year of agentic AI. So that can be a nebulous term for a lot of people. I think almost every guest I have on the show explaining it explains it in a different way. So give us a definition, John, of Agentic AI and whether 2025 did pan out as the year of Agentic AI.
John Roese:	02:16	Yeah, yeah. Actually, the prediction was that the word of the year would be agentic. And I think that's exactly what we saw. And then that ... Funny enough, in December when we talked about it, I'm not sure anybody even knew what that word was. And now we're a year later and you still may not know what it is, but you know it's a word and it's out there. So Agentic just refers to what we sometimes describe as the second wave of generative AI. The first wave was the era of things that unlock your data, chatbots, rag systems. And they're just tools that use generative systems to take data, whether it's institutional data that's been gathered and trained into a model or rag-based systems and make it useful. Imagine being able to talk to your data, imagine being able to distill from it, create new variants to that data, whether it's imagery or text.
	03:05	And that's fantastic. But there's a second half. And the second half is the digitization of skills. It's not just unlocking data. It's this idea that we're going to use AIs to actually do real work. There isn't going to be a human in the loop. The work is going to be done by the AIs. And that's really what Agentic refers to. It refers to a system that fundamentally is an AI system that's able to autonomously do work on behalf of humans, hopefully. You give it an objective, it reasons, it navigates and

ultimately executes that work. And you as the human are effectively on the loop as opposed to in the loop doing the work with the AI as a tool. And so that has not fully manifested. The term is wildly overused. Many people are calling fancy chatbots agents, and they'll call anything an agent these days, but what is emerging, and they are real in some cases and definitely will be a big deal in 2026, are these fully autonomous systems in which you can delegate a task to an agent, the agent will do the work on your behalf and you will give it intent and validation.

04:02 That's the only involvement you'll have. That is a very powerful disruptive capability. And so we'll see what happens, but it's going to be a big deal. This year we learned the word. We got comfortable with the concept and in 2026, we're probably going to actually experience it firsthand.

Jon Krohn: 04:18 Yeah. I'd say, John, that my biggest gripe of 2025 is people misusing the term agentic and having situations where you're in a conversational agent, like you said, where if every action that this system is taking is based on a prompt from you or some other user, that is not agentic. The key thing of it being agentic system, of it being an agentic system is that it is autonomous.

John Roese: 04:45 Yeah, 100%. I mean, we have a term called agent washing where people want to feel good about what they're doing. Agents are the cool term, but it's not helpful. Honestly, it creates tremendous confusion. I've seen customers claim they have thousands of autonomous agents running. And they'll say this and you just kind of scratch your head saying, "I'm fairly certain you do not have thousands of instances of purely autonomous digital workers working on your behalf doing complete work." You probably have thousands of tools that bring autonomy to lots of people and that's good, but that's different than this idea of a fully autonomous agent. And we need to get prepared for

this because it's going to be a very disruptive technology. It's very powerful. And if you're confused about what it is at the beginning and you misinterpret older technology as being this, you're going to have a tough time navigating it.

Jon Krohn: 05:29 And I think a lot of people out there at conferences, even some guests that I've had on my podcast have complained about too much conversation about agentic AI in 2025 and this feeling that it's been overhyped. But I think it's absolutely true what you just said about how we need to be getting ready for this kind of capability. And something that I've talked about a lot on this show is there's this MTER graph of how the length of a human task that can be replaced by an AI system, and that's doubling every seven months. And now it's constrained to areas where we have really good training data. So things like coding problems, things like machine learning problems. So we're not talking about replacing humans on all kinds of tasks, but for ones that we have great quality training data, we're seeing the length of a human task that can be replaced with 50% accuracy by AI doubling every seven months.

06:33 So now we're sitting at around two, four hour mark and in seven months it'll be more like four to eight. And so that means that way, way, way more tasks in a given enterprise, in a given organization, can be handled by AI systems six months from now, a year from now, 18 months from now, but we need to be setting the stage today.

John Roese: 06:54 Yeah, that's true. But one of the things that autonomous agents do is you start to use them as ensembles where you have collections of agents working together, you can go after tasks that you can't go after today. If I just have traditional AI tools, I can solve a specific task and use that tool to make it better. I can augment a human to do looking something up, writing a piece of code. What I

can't do with that singular tool are tasks that are not constrained to a single person or single organization. And so some of the use cases we're looking at, and probably the first ones that will come live in Dell that are fully autonomous, are actually not simple tasks that are done by a single person using a single tool. They're actually these composites of how do you generate a quote or how do you solve a complex support issue?

07:41 And it's not using a single tool or doing a single task. It might actually not even be within a single organization. In fact, a lot of the real promise of agents is that it dramatically expands the types of work that an AI can go after because they're no longer focused on making a very specific thing better. They're actually able to deal with composites and actually emulate what a team can do. And so that's super important to understand the difference. But if you take that metric of not only is the complexity of an individual task expanding of what can be addressed, but the diversity of the types of tasks and the number of tools used in the task is also expanding. So you got kind of two dimensional expansion of the surface area of AI here, which is pretty exciting and challenging and definitely going to be interesting.

Jon Krohn: 08:27 For sure. Speaking of more access to tools, what are your thoughts on MCP, model context protocol?

John Roese: 08:34 So we published inside of Dell, it's not public yet, but inside for our own purposes, a set of agentic guidelines. So I haven't found many other people have done this yet. We wrote it down and we have rules now about how to use agents, what technology to use, what standards for identity and for ... And one of the areas we focused on was MCP. Now, MCP is a little confusing to people because people sometimes call it an interworking protocol. It is not. Protocols like A to A are interworking protocols. Those are communication protocols between

Agentic systems. MCP, model context protocol, is a very good standard to create an abstraction level for tool use and data access. An agent by itself should be able to perceive the world around it and take action. MCP gives us an interface to do that that's standardized. Now that's great.

09:18 That's the good news part. The bad news part, it's brand new protocol. It was developed in the context of kind of making a public LLM able to consume a lot of data. It isn't really enterprise grade yet. It has a lot of security work to be done. There's good work going on. We're adding OAuth and some other things to it, but to be perfectly honest, as a protocol, it's very early, but it's incredibly powerful and it's incredibly dangerous if you're not careful about how you use it. So in Dell, again, one of our guidelines is we made a decision to centralize all of our MCP servers into a controlled environment. If the protocol itself cannot exert the right kind of controls, I'm going to put it in an environment where I can wrap it with those controls and make sure that it doesn't do anything crazy.

10:00 Because like I said, wonderful protocol, it's lightweight, it's standardized. We think it's a keeper and we expect most agents in the world to use MCP to perceive the world around them and to access tools, but the evolution of it for enterprise grade use is still in front of us. And so people are going to have to be very aware of this powerful protocol can do way more than you think it can do and it might do things you don't expect. And in an enterprise context, you need to put some controls around it. But it's a good problem to have. We'll make it better, but it is in fact, just like A to A, one of, I think, the long-term protocols that agents are going to use.

Jon Krohn: 10:35 Nicely said, as your answers always are, that was really crisp and easy to understand. Lots of potential impact

with MCP, but definitely security concerns. And it sounds like you figured out how to wrap that in an environment to Dell so that you can feel a level of security with MCP. It sounds like the best of both worlds. Really quickly, to kind of just wrap up your 2025 predictions, I think it's interesting that you specifically had said that Agentic would be the word of the year for 2025. And so a publication that I read ... I mean, the publication that I read the most, the news source that I have for 80 or 90% of my information is The Economist. And at the time of us recording today, the economist just announced a couple of days ago that their word of the year is slop. And yeah, that's been another big plague.

11:26 In addition to the misuse of Agentic AI, how are you guys controlling AI slop internally at Dell?

John Roese: 11:32 Yeah, we're a little different than maybe, I guess, the mass hysteria around using AI. As you know, we took a very top down approach. We targeted very specific use cases at scale. We focus on ... The methodology that we use is quite rigorous and it basically just says, look, we're doing this for business outcome. We're identifying the most important processes and the most important parts of our business that are going to have the biggest impact on productivity. And then we're laser focused on getting those particular problems solved at scale. We've only deployed maybe less than probably 30 actual AI use cases in the company, but these use cases have ROIs in the 10 to one, 30 to one, and have transformed our sales services, supply chain and engineering organizations. So we got really big impact from very targeted things. By the way, if you remember that MIT report, what did it say?

12:21 95% of projects fail. The ones that don't are the ones that you focus on a few projects and get them done at scale. Because of that, we don't have a lot of slop because what we didn't do, and some people chose to do it, we chose

not to. We did not just let these technologies run loose inside of our organization. Of course, you're going to get slop. Everything has a distribution curve. If you give, I don't know, a half a million people access to a random AI tool and hope for the best, you're probably going to have some pretty cool things happen and you're going to have a lot of really bad things happen. And unfortunately, that's not a strategy in our mind. So we didn't take that approach. And so at least I feel comfortable that the projects that we implemented and the processes we went after largely achieved their objective because the objective was very clear from the beginning and we were able to control the AI system around it, make sure that it had accuracy and relevance, et cetera.

13:12 So yeah, slop's the word of the year for them, I agree. There's a lot of slop out there, but the way you mitigate slop is to not be sloppy, be focused, be targeted, be disciplined. And that is something that unfortunately with all the hysteria around AI, a lot of people have kind of just said, "I got to move. I got to do something. It's better to do something than nothing." I would argue it's better to do nothing than to do chaos. And so that's been the dialogue. But yeah, again, not everybody has benefited from that, but what we at least have slot countermeasures with discipline and focus.

Jon Krohn: 13:47 I agree with you 100% on all those points. For people who are unfamiliar with that MIT study, the 95% of gen AI projects fail. I've got episode number 924 for you listeners dedicated to that, but even more important than that reference to a back episode is John Roese's previous episode on the show, which is 887. And in 887, he provides tons of information on what he just gave an overview of where you can be basically guaranteed to have successful projects within your enterprise, how to select the projects that are going to have the best ROI. And you just gave some citations there, which I don't

think you did in the previous episode of kind of 10 to one, 30 to one ROI ratios, which is amazing. And as soon as you have a few of these projects, then you start this AI ROI flywheel spinning, which is a term that I learned from you and I have used in countless business meetings since we did that episode 887 together.

John Roese: 14:46 It seemed to work. By the way, it's been a while, but look, the big impact that we had at the end of that process once we got the four big areas running is in our fiscal results last fiscal year, which we reported in March, that was the first year that we'd kind of got all this stuff up and running. Something happened at Dell that hadn't happened in 41 years as a company. Our revenue grew at $10 billion and our absolute cost declined by 4%. We have never decoupled revenue growth from cost. Every other time the revenue grew, the cost grew with it. It didn't happen last year. And it was a combination of people, process and technology, but it was really a positioning and redesigning the company to actually get a better productivity as we did work. And if you do that, boy, it's a pretty powerful tool when your revenue grows and your costs don't grow with it.

15:31 That results in margin expansion, that results in just a better performing company. And so we now have a lot of empirical data. Those ROI metrics can measure. We measure everything. And so I feel good that, hey, not only was it good theory and not only was it discipline, but it actually seems to work in terms of producing really strong ROI, which, hey, that flywheel works better if the ratios are higher.

Jon Krohn: 15:52 For sure, 100%. You get a lot more attention when you have a 30 to one ROI ratio for sure. And it's great to hear those kinds of stats around revenue increasing without costs going up or costs even decreasing because I'll keep this completely ... This isn't a comment on anything

related to Google necessarily, but a friend of mine, a good friend of mine, just the other day at the time of us recording, said to me, "Do you really think you're getting better productivity from this AI stuff?" And yeah, it's fantastic for you to be able to cite these hard numbers on how if you do it right, you absolutely can be getting a great return on investment in your organization.

John Roese: 16:31 Yeah. I remember one of the first principles was have a hard ROI metric. It's hard to measure goodwill and happiness. It's a lot easier to measure things like, I don't know, cost or margin or revenue or things that you actually can measure and understand. And so by focusing your AI efforts on things that are empirically measurable, you can actually get some level of confidence that one, you're either achieving success or two, you're not. If the ROI is nothing, and there's actually a term that I won't use the full language, but I think it was a McKinsey report and it described what they call BS work. And they basically said generative AI is a BS work detector because it turns out that if you use a whole bunch of AI tools on a piece of work that you currently do without them, and after you use them, even if you feel good about it, nothing changed.

17:23 Your revenue didn't go up, your cost structure didn't go down, nothing really changed. What it's telling you is maybe that work isn't worth doing. Maybe it's BS work. And it turns out that there's probably a lot of that out there where just doing AI to something doesn't necessarily result in an impact. And I'm pretty, let's call it focused on, you want an impact from this stuff. You're investing heavily in the technology. It's costly, it's complex. And if you're not getting an impact that disrupts your business in some way in a positive sense, then why are you doing it? And trust me, there's a lot of targets to go after, so spend the time there.

Jon Krohn: 18:01 Right. So what you're saying is in a lot of cases, the reason why AI projects are unfruitful or don't appear to move the needle internally is because it's automating something that was BS in the first place when humans were doing it.

John Roese: 18:14 Yeah, exactly. And like I said, I think it was McKinsey, one of the consulting firms put this article out. You go Google, BS work and AI. And it makes a lot of sense. I mean, look, hey, you can apply AI to anything you want. Great. That might make it better, might not, but you should probably pick the things that if you make them better, your business will improve. Something material will happen. And that level of discipline is something that people have to spend a little time on upfront. And if you do, yeah, seems to result in things that matter.

Jon Krohn: 18:43 That makes a lot of sense. And I feel like that's going to be a really big arrow in my quiver now going forward, just like the AI ROI flywheel has been in the past six months since I last had you on the show. Thank you for that, John. All right. So let's move ahead now to your predictions for 2026. My understanding is that your first one is related to governance.

John Roese: 19:05 Yeah, I'll say it's a little bit of a soapbox more than a prediction because look, the general concept of governance, everything we just talked about about picking your projects, having discipline around it, all of those are under this rubric of governance, but so is the regulatory regime, government involvement. And what I will tell you as a person implementing this stuff at scale, the governance structures of the world external to an enterprise are a disaster. I have over a thousand jurisdictions around the world telling me what to do without talking to each other. And so you can't possibly achieve success with a thousand different jurisdictions with conflicting governance structures and rules. So part

A is, look, governance is going to become a bigger and bigger either impediment or enabler, but governance is going to be very central to every AI discussion in the enterprise. On the external side, it is incumbent on us to work with governments to try to make sure that we either reduce the number of distinct regulations or get them coordinated or harmonized.

20:06 Now, the NIST frameworks and some of the EU frameworks can be normalization layers, but it's messy. And I don't have a good answer about what's going to happen to fix it, but the prediction is we're going to see as people really start to go into production, governance, managing, navigating external governance, becoming a bigger and bigger piece of their activity. There's a tactic you can use. We've used it. Stay away from highly regulated use cases. There's plenty of low hanging fruit. Go after that first. Don't go after the most regulated environments because candidly, the governance burden is very, very high right now. On the internal governance side, we do think that a lot of people have done governance free AI, give everybody a tool and hope for the best. That doesn't work. And so as we go into this year, having that internal governance framework, things we talked about on the previous episode is incredibly important because it allows you to prioritize where you actually want to focus.

20:59 And that's not a picking favorites. It's just a governance structure. It's a methodology to make sure that the work you do gets focused on the most important outcomes. And so this term governance, whether it's external governance becoming more and more of a factor in slowing us down or navigating what we prioritize or internal governance, which is the act of actually having an orderly structure about what you pursue and how you pursue it, those will become much, much more dominant as we go into 2026 than they were in 25.

Jon Krohn: 21:26 It sounds like that internal governance aligns exactly with the AI slop conversation we were just having. 1000%.

John Roese: 21:32 Yep. You want to avoid AI slop, you better have a governance structure to pick the things that are not sloppy. And by the way, even then, some of the slop, the implication of slop isn't just that it's potentially messy, it's potentially dangerous. If the slop touches something that's kind of regulatory third rail and you don't realize that, that is an enormous problem. There were a few articles about some consulting companies working with governments that inadvertently, some of their team members used AI tools to generate content and they got sued. That's slop. And it's not slop because it was in a vacuum. There was a governance breakdown there that should never have happened. And so governance is a counteraction to slop and it definitely keeps you at least more deliberate about what you're doing. No one will choose to be sloppy. They will choose to do the right thing, but if they don't even make a choice because there's no governance, then you get what you get.

Jon Krohn: 22:27 Yeah. Yeah. Very well said again there. So I should have mentioned probably before I talked about your first prediction for 2026 is that you have five for 2026 plus a bonus. And I think that the bonus is something we talked about your preceding episode in your preceding episode a fair bit as well, but just so that people have some sense of kind of where we are. So first one was governance, your governance soapbox point. Number two is data management, which you call the true backbone of AI innovation. Fill us in on that, John.

John Roese: 23:01 Yeah. Yeah. So we've been obsessed about large language models and AI compute. Those are great, but we're doing pretty well in those spaces. And we have a lot of business around building out the compute infrastructure of the world. But as you put these into production, what you

realize is that the compute can move around. You can do compute on a device, you can do it at edge, you can do it in a cloud, you can do it in a public service, you can do it in a data center. You have lots of choices to compute, and that's actually in pretty good shape. What's hard, especially as you move to things like Agenttech is you have to create entirely new data structures to make these things work. An agent by itself does not actually work without the addition of things like knowledge graphs and additional information.

23:39 And those are not created randomly. Those are created with a new data layer. Sometimes we call that layer the knowledge layer. You have your primary data, your systems of record, and you have your compute layer, which is the other end of the system. But right in the middle of it, there's a new knowledge layer forming. And that's where things like graph databases, vector databases, knowledge graphs, all of the necessary data, sometimes mathematical representation of data, sometimes graphical representation. But the bottom line is that new data layer is really the thing that allows systems to have context. If you just take an LLM and try to run your business with it, it's not going to know anything about your business. You have to attach a vector database via rag to make it intelligent. If you're an agent, it's different. You might use a rag, but you might create knowledge graphs that are very specific to particular roles and also instantiate things like memory, long and short-term memory, the ability to accumulate information and skills.

24:33 All of that layer is ... In 2025 wasn't really talked about, but as you go into 2026, it's insufficient to just have your primary data and a bunch of AI compute. You've got to build out an effective data management layer in the middle called a knowledge layer. One of the interesting subtleties about that knowledge layer is usually in the

past when we reflect or represent primary data to other systems, we do it close to the primary data, meaning you got your data in a database, then you put a front end on it to express it a different way. The knowledge layer is different. It's not an API. It's actually persisting data. It's a layer that lives. It's always warm or hot. And what we're seeing is you almost have to make a choice. Do you want the knowledge layer close to the static primary data or do you want it close to the dynamic compute environment, the edge, the device?

25:24 And it turns out it makes more sense to put the knowledge layer close to where the action is because that's where all the transactions are happening. The feeding of the knowledge layer from like a database is like a one-time event. It's non-real time. The consuming of a knowledge layer by an AI system is hot and real time and perpetual. And so all of those dimensions of what does the data architecture look like for a robust AI system, pre-agentic or agentic? And it's different than anything we've ever built. It's an entirely new layer and it has different characteristics and we shouldn't walk in there with an assumption that just because it's called a vector database, it belongs where all the other databases are. It might actually belong close to where the compute is happening, and that's definitely going to be true for agents. So if you thought data was interesting before, now we have a whole bunch of new data to deploy, but we have to do it in an intelligent, architecturally consistent way.

26:17 And we have to think about it from an AI first perspective, not a legacy data perspective. And that's going to be interesting this year, but I think it's going to occupy a lot of oxygen in the dialogue about how to do AI properly, not just compute, not just data, but what is that knowledge layer in the middle? How do you build it? How do you make it optimized? How do you make it work best? Where

should it be? All of those questions have to be answered for you to be successful, especially with Logantic.

Jon Krohn: 26:41 I love this idea of terming this, the knowledge layer. I've never come across a term so nicely, because I think about vector databases, reg database as a knowledge graph. I've always had those as kind of different types of knowledge that are being derived from raw data and that are going to be needed downstream by my generative or my Agentic application. But by terming those all as the same kind of thing, as a knowledge layer that sits between your raw data, your system of record and your compute, that's a really nice mental framework for thinking about these kinds of things. And in particular, the point that you made there around having that knowledge layer be close, be on the edge, be on Dell makes a lot of AI PCs. It's been a big thing that you guys have launched in 2025 as new AIPC devices that allow you to have quite a bit of compute on the edge and quite a bit of data storage as well.

27:37 So more plenty of data storage to be having your knowledge layer maybe updated on ... I'm guessing here you might even have some real world use cases, John, but in my mind it would kind of make sense to me to have a 24 hour process. Every 24 hours, a cron job goes off that goes over the systems of records and updates the knowledge layer on some edge device at midnight or something. And then that updated knowledge layer information is available for the generative system or the agentic system to be able to use it regularly with low latency.

John Roese: 28:14 It gets even more dynamic when we start talking about fully autonomous agents because the agent uses that knowledge layer, specifically the things like knowledge graphs as a place to keep track of what it's learned. We built autonomous agents where we've given them an ensemble, a task to develop a skill, to run a CNC

machine, to do a research report, to whatever it is. And it turns out that it starts with that batch representation of knowledge. And then as it starts to work through the problem, it actually creates pathways through the knowledge graph that they look like vectors that essentially allow it to understand how it got from point A to point B, how it turned on the CNC machine, how it found an answer. And all of that is dynamically being added to the knowledge graph as the agent is actually working. And so that's why I keep saying it's not a system that ever goes cold.

29:01 It's hot or warm at the worst, and that's just a different way to think about it. I think you talk a lot about data in this environment, this dialogue, and the takeaway is there's a new layer forming and it's not just, you shouldn't treat these as just another database and use the same approach you do for any type of database on this AI stuff because it's not doing the same thing. It doesn't exist in the same place and it doesn't have the same behaviors. And so that's going to be fun. There hasn't been a lot of new data stuff for a while. This is all new data stuff and it's actually pretty exciting to figure out, okay, what does it do? Where do I apply it? Then how do I optimize it? How do I make it work well? And that's where things like deploying knowledge graphs close to the point of presence and making them incredibly dynamic becomes incredibly powerful.

Jon Krohn: 29:45 Yeah. And I can't believe I was thinking of it so simple mindedly to think of data only flowing from system of record, data layer to knowledge layer. But of course, in this agentic AI world that we're in now, the agent's going to be updating that knowledge layer as well the other way.

John Roese: 29:59 You bet Yep. Bidirectional.

Jon Krohn: 30:02 Yeah. And then I guess then you might have a cron job taking things out of the knowledge layer and putting that into your long-term system of record. Exactly. Remotely. Cool. All right. Onto number three. This one is about Agentic AI. And it looks like there's some topics here that we've already kind of brought up and maybe you want to dig into a bit more. You mentioned knowledge graphs, LLMs, MCP, and A2A, which you touched on briefly earlier in this episode, but we didn't go into too much detail.

John Roese: 30:29 Yeah. This is one that I've had a lot of agentic discussions over the last year and a half. And you can define it at a high level, but we thought we'd go a little deeper. And it's important to recognize that when you start thinking about an agent, and here's why it's in the predictions. There are two different philosophical approaches to what an agent is. There are a cast of characters who think that an agent is a feature of an LLM. The central piece is the large language model and the large language model has agentic behavior. That might be true. I don't think that's very useful, to be perfectly honest. I think that isn't really a fully exploiting the capability. We on the other hand, and we are not alone, there are a lot of people who absolutely believe exactly what I'm about to say, but not everybody, that an agent is actually a software system that has multiple dimensions to it, of which an LLM is one of them.

31:24 And in both cases, both groups are saying this should be an autonomous system. It should be able to do work independently. But the difference is, is it just a feature of an LLM or is it a software system that uses LLMs and other stuff to accomplish that task? We're in the second camp. And so we drew a little picture that I think has worked really well for people to say, okay, at a technical level, an agent that is that software system actually has four components. One of the components is in fact a large

language model, and it needs that because that's where it gets basic communication skills, world knowledge. And in many cases, its reasoning capability is a function of the LLM. There are reasoning models that work really well. So you can't really build agents without LLMs. They might be large language or small language models, but you do need them and they're one of the four fundamental components of an agent.

32:11 The second part of an agent is what we just talked about in the previous prediction, which is that in addition to world knowledge, agents benefit by having specialized knowledge. And that specialized knowledge is expressed in some cases as a gentic rag, but in many cases as things like knowledge graphs where you populate it with very specific information, but you don't just give it an information set. You give it a graph that it can actually manipulate and it can use not just to see specialized information, but to actually develop skills around that specialized information, pathways through the knowledge graph that help it repeat tasks. And sometimes we call that short and long-term memory, that it can actually keep state over a very long period of time by instancing what it's learned in the knowledge graph, not in the LLM.

32:57 The third component though is that, and we talked about this earlier also, is that agents, this kind of system need to perceive the world around them, meaning they have to access external data and they need to be able to do things, which means they have to be able to use tools. MCP becomes the protocol of choice to do that. So in addition to having large language models and knowledge graphs, MCP or a like interface that allows them to access external data sources and interact with tools is a critical component of this system. And then the fourth component, which is probably the most important in my view, is agents can talk to each other. Unlike traditional AI tools that operate in isolation, agents actually do better

when you put them into an ensemble and you actually allow them to work together, that they bring different skills and they collaborate.

33:39 And the reason you do that isn't just because you want to distribute the work, it's because it allows you to start to introduce behaviors like consensus into their decision making. And we actually did some studies where we've seen that when you take a collection of agents, an ensemble of agents and they share a base truth with a common knowledge graph. The knowledge graph is truth. The LLM is opinion and the agents all share that base truth. If you give them a governmental system, because ensembles you tell them, is this a democracy, an autocracy? How do you want these systems to behave? If you tell them that consensus matters, if one of the agents and their LLMs start to hallucinate and do crazy things that are inconsistent with ground truth in the knowledge graph, the other agents ignore it. They don't let it impact the outcome.

34:24 It's like imagine you take five people in a team and one of them is absolutely nuts and start saying crazy things. "You have countermeasures. You agree collectively, that's a dumb idea. We're going to do the right thing. "And so we've seen that behavior in agent ensembles, but that brings us to that fourth component, which is they have interworking protocols and that's where agent to agent or a to A primarily lives. And it's an ability for you to actually express communication between agents so that they can work as teams, they can delegate tasks. And that's incredibly powerful to go after these bigger sets of activities that we want to pursue beyond just single process, single. So that picture, and I'm sure we can throw it up on the screen or something and you can see it, has become incredibly helpful to help people understand these are not just an LLM.

35:10 These are these four capabilities, general world knowledge and communication skills, specialized knowledge and state memory, the ability to perceive and act in the real world and the ability to communicate together in ensembles. When you take those four capabilities, put them together, you have an autonomous agent. And I think that's going to be kind of a fairly enduring definition that most of the world agrees with. And just caution, there's a few people who think all that stuff can just be done as magic inside of the black box of magic called an LLM. We disagree.

Jon Krohn: 35:40 Yeah. We do have a sizable chunk of listeners who are listening to the show in an audio only format, so they won't be able to benefit from this diagram, but we will get the diagram up on YouTube for those of you watching it on YouTube. And just to kind of summarize, basically there's four components to this agentic software system that John went through. So the LLM is a component of the software system as opposed to the center of the system. It has a knowledge graph or some other kind of graph for being able to store information and refer back to that. MCP for accessing databases and tools and then A2A for allowing agents to talk to each other. And a really cool thing that I like about the collaboration point that you made there, where if one of the agents takes magic mushrooms and is completely starts hallucinating all the time, the other ones start to ignore it. And that reminds me how, when I was talking about that mTurgraph of how you have tasks now kind of on the two to four hour range human tasks being handled at a 50% accuracy rate, if you're willing to spend a bit more on inference costs and have multiple agents working together on the problem, you can very quickly get much, much, much higher than 50% accuracy rate.

	36:53	You can probably get 99% or more if you're willing to spend on the inference and have agents working together on the problem.
John Roese:	36:59	Yeah. We don't know if this solves hallucination, but it sure mitigates a lot of it. And we're not inventing a different way to do it. In humanity, how do you mitigate insane behavior and ridiculous statements? You collectively look at it and you dismiss it because it's obviously not consistent with ground truth. So the same principles apply to agents if you build them the way we just described it. So we think it's promising. Nothing's going to be 100%. Nothing's going to be absolutely accurate. That's just not possible in the world in a probabilistic system. But boy, having a few agents work together gives you a completely different outcome than hoping a single LLM in a vacuum is perfect. That's not going to happen. Nice.
Jon Krohn:	37:44	All right. And so I'm conscious that we might only have a few minutes left of your time for this recording. And it seems to me like number four is a critical one to be talking about because this is one that you and I have been talking about all year round. You and I recorded an ad for television about the Dell AI factory with Nvidia and your point number four is that AI factories will redefine resilience and disaster recovery. Tell us about AI factories.
John Roese:	38:07	Yeah, this one's simple. We have about 3,000 customers building out AI factories today with Dell and in different states, but we haven't figured out how to make them resilient yet. And the reality is we could just back the whole thing up, but it turns out that AI factories are different than traditional environment. Kind of going back to that data discussion. If the primary data sources are not actually used in most of the AI interaction, let's take an example of, I take all my service information, I

vectorize it, and I create an agent that can use all that information to solve service problems. At no point does that agent go back to the primary data while it's doing its job. And so it opens up some really interesting thoughts about where do we provide resiliency, which things need to be resilient. If I lose access to my primary data, but I'm not even using it to do the AI stuff, then maybe I don't need 79's reliability for the primary data anymore.

39:01 I need it for the AI side. In addition, where that backup occurs, what does a backup to an AI factory look like? It needs to have access to a lot of GPUs, but you don't want to have them sit in their idle. So there's interesting opportunities about using kind of elastic infrastructure in some of the CSPs and other environments. But this discussion of now that we've built our AI factories, we need to make it resilient and we shouldn't go into that discussion thinking an AI factory is a traditional IT infrastructure. So we have to look at different models, find different efficiencies, use some of the new infrastructure that's being built where GPUs exist at scale to maybe create different kinds of resiliency models. In addition to that, the whole cyber resiliency world of ransomware protection and data vaults and cyber vaults all has to be applied to protect knowledge graph, vector databases, agents.

39:49 Agents are a software system. You need to store them somewhere. And so we think this is going to be a big topic. We think that the guidance is not to go into it blindly thinking we will bring cyber resiliency to our AI factories exactly how we did it 20 years ago. That is not going to be the case. And there's a huge opportunity to be highly disruptive and innovative and to really build resiliency for AI factories in a modern way, which we're pretty excited about.

Jon Krohn: 40:13 And so you mentioned that 3000 different Dell clients are using AI factories. It sounds to me like an AI factory, it's bringing together a lot of the services that Dell and then Nvidia GPUs as well can offer. So it's data, it's services, it's open ecosystem infrastructure, all combined together to allow whoever's taking advantage of an AI factory to get an AI ROI flywheel going on their own.

John Roese: 40:41 Yeah. And that's where NVIDIA and Dell got together going almost two years ago with this AI factory concept to say, "Hey, we had to absorb the complexity to us." Customers shouldn't have to figure this stuff out. They should be able to take reference architectures and templates and make these things fast moving. No one should have to design their own server and figure out their own software stack. These things should be as consumable as possible so you can spend your time on the use cases and the data and the application. So NVIDIA and Dell have a long history here. And now our next challenge is, okay, now they exist. Let's make them super resilient and survivable in a modern way, which should be a very exciting journey.

Jon Krohn: 41:14 Cool. Yeah. So number four about AI factories redefining resilience. Awesome. And five, your fifth of five predictions for 2026, it's around sovereign AI. Tell us about that.

John Roese: 41:26 Yeah. Last year we talked about sovereign a lot saying there's different ways to do it. You could build government for government where you build data centers for your government, you could build government or for industry where you build data centers to enable your indigenous industries to be able to move fast on infrastructure they couldn't afford to build themselves. And then there was government with industry where the government plays a convening role. All good, been happening all year long. What we're predicting is clearly there is a dramatic acceleration. There will be no country

in the world that does not have a sovereign AI strategy as we go into 2026 and into 2027. It's mainstream. What's exciting about it though is we've now ... And that's all maybe predictable and understandable. What's exciting is now that you build out these sovereign infrastructures, what do you do with them?

42:15 And it turns out you do a lot more than just government stuff. We're looking at in the future, what do we need them for? I don't know, resiliency. If I want to basically create a backup for a critical infrastructure AI capability, maybe doing it in a sovereign infrastructure in the country that the critical infrastructure's in would be a good idea. That seems like a good use case. If I want to create agents that live somewhere long term, but those agents are in some way tied to a government. They are authorized or certified by that government. They're a Swiss lawyer or they're given validation by the fact that they meet a expectation and a certification a government provides, maybe their knowledge graphs and all of their backend systems and data should be certified and live physically in an environment that has some connection to that sovereignty and on and on.

43:01 And so what we're excited about is now that we've built a few of these things, sure, it's making government better, it's accelerating training infrastructure, but there's a lot more we can do with them. And that's pretty exciting when we think about the fact that government's role may in fact increase as we bring more and more AI systems online, especially in the enterprise, and especially where they intersect with things that have a national interest, critical infrastructure, defense, intelligence, healthcare, these kinds of services, it seems like it's almost impossible that we're going to be able to fully separate government from private sector in AI. And that's probably a good thing because collectively we have a lot more resources to apply to it.

Jon Krohn: 43:37 Nice. Thank you for that perspective. I don't spend enough time thinking about sovereign AI and we don't talk about it that much on the show. So it's nice to have your input there as your fifth and final prediction for 2026. Now, I do see that you also had a bonus one on keeping an eye on quantum, but that is actually something we talked about a fair bit at the end of your previous episode on the show, 887. And so it looks like, I think you might have a hard stop coming up in terms of recording. So we might jump right to your book recommendation for us.

John Roese: 44:05 Okay. So my, by the way, quantum, it's coming, it's just accelerating. Don't take your eye off of it. But for a book recommendation, this is a book that I got recently and it's called The Book, The Ultimate Guide to Rebuilding Civilization. And what's inside of it is if the world ends and civilization collapses, it has the knowledge of mills, drills, and lathes, spices, how fungus works, different medical systems, how to build an x-ray machine. And so I found that amusing because it's kind of like plan B. If the AI thing fails, maybe civilization will collapse. And if it's successful, maybe civilization will be so different that we have to rebuild it. But having a book to help me do that seemed like a good idea to have my coffee table.

Jon Krohn: 44:53 I like it a lot. It seems like the kind of knowledge that I would love to have for myself. And I don't have kids yet, but someday would like to. And it seems like kind of a fun thing to go over and help them understand how the world works. For people who didn't see it in the video, it is gigantic. It is maybe the biggest book I've ever seen. Absolutely. Fantastic. Okay. The book, The Ultimate Guide to Rebuilding Civilization. We'll have a link to that in the show notes as well as anything else we talked about in this episode. And just the final thing from you, John, is obviously it is crystal clear to anybody who's been listening to this episode that you are a fountain of

		crisp knowledge on AI. Where else should they be following you after this episode to get more of your thoughts?
John Roese:	45:34	Well, obviously anything Dell related, but we do a YouTube AI insights with John Roese. It's really just a narrative. We've been running it for over a year and it's kind of the weekly updates of what we're learning as we're going and very specific stuff. But one of the things we decided to do is to share what we were doing because it seems like we're figuring out some things that might be helpful to people. So AI insights with John Roese is probably the best way to see what I'm up to and podcasts like this.
Jon Krohn:	46:02	Nice. I love that. Thanks, John. Thank you so much for your time again today. It's an honor to get some time of yours and your no doubt extremely busy schedule. And yeah, hopefully we'll be able to check in with you again soon.
John Roese:	46:15	Thanks, John.
Jon Krohn:	46:19	What a brilliant mind. What a brilliant leader in today's episode, Dell's global CTO and chief AI officer, John Roese, detailed how we should be wary of agent washing, where companies rebrand existing tools as Agentic, despite a lack of autonomy. How Dell achieved something unprecedented in their 41-year history, revenue grew by $10 billion last year, while absolute costs declined by 4%. Thanks to disciplined AI deployment with 10 to one and even 30 to one ROI ratios on some projects. John talked about a new knowledge layer and how it's forming between primary data and AI compute, including things like vector databases and knowledge graphs, and how this knowledge layer should live close to where the AI action happens. He also talked about how true autonomous agents have four components, an LLM for

reasoning and communication, knowledge graphs for specialized knowledge and memory, MCP for perceiving and acting in the world, and A2A protocols for agent to agent communication.

47:16 I hope you enjoyed all of John's predictions for 2026. As always, you can get all the show notes, including the transcript for this episode, the video recording, any materials mentioned on the show, the URLs for John Roese's social media profiles, as well as my own at superdatascience.com/953. Thanks to everyone on the Super Data Science podcast team, podcast manager Sonja Brajovic, media editor, Mario Pombo, our partnerships team, which is Nathan Daly and Natalie Ziajski, our researcher, Serg Masís writer, Dr. Zara Karschay, and our founder Kirill Eremenko. Thanks to all of them for producing another super episode today for enabling that super team to create this free Super Data Science podcast for you. We are deeply grateful to our sponsors. You can support this show by checking out our sponsor's links in the show notes. And if you ever want to sponsor the show yourself, you can see how to do that at johncrone.com/podcast.

48:03 Otherwise, help us out by sharing this episode with someone who would enjoy listening to it, review it on your favorite podcasting app or on YouTube. Subscribe, obviously, if you're not a subscriber, but most importantly, just keep on listening. I'm so grateful to have you listening, and I hope I can continue to make episodes you'd love for years and years to come. Till next time, keep on rocking it out there, and I'm looking forward to enjoying another round of the Super Data Science Podcast with you very soon.