SUPER DATASCIENCE
MAKING THE COMPLEX SIMPLE

SDS PODCAST EPISODE 745: 2024 DATA SCIENCE TREND PREDICTIONS

Show Notes: http://www.superdatascience.com/745

Jon: 00:00:00 This is episode number 745 on Data Science Trends for 2024 with Sadie St. Lawrence. Today's episode is brought to you by CloudWolf, the Cloud Skills platform.

00:00:15 Welcome to the Super Data Science Podcast, the most listened-to podcast in the data science industry. Each week, we bring you inspiring people and ideas to help you build a successful career in data science. I'm your host, Jon Krohn. Thanks for joining me today. Now, let's make the complex simple.

00:00:47 Happy New Year. Welcome to the year 2024, and welcome back to the Super Data Science Podcast. To kick things off, we've got our annual data science trend prediction special for you again this year, this very day. As usual, we start the episode off by looking back at how our predictions panned out from a year ago, and then we'll dive into our predictions for the year ahead, starting from the hardware layer, and then moving gradually toward the application layer. Specific 2024 trends we discuss include AI hardware accelerators, large language models as operating systems, models like Q* that solve problems along non-linear paths, consolidation of enterprise systems, thanks to function-calling LLMs, and whether generative AI will replace data analysts or potentially other data professionals.

00:01:35 Our very special guest to guide us through these predictions is the magnificent Sadie St. Lawrence, a data science and machine learning instructor whose content has been enjoyed by over half a million students. She's the head of AI at SSL Innovations, and is also founder and CEO of Women in Data, a community of over 60,000 women across 55 countries. On top of all that, Sadie also serves on multiple startup boards, and is host of the Data Bytes podcast. Sadie was previously our guide to both 2022 and the 2023 predictions episodes from two years ago and a year ago respectively. Those are two of the all time most popular episodes of this podcast, and if you listen to them, you already know that you're in for a treat again today. Today's episode will appeal to technical and non-technical folks alike, anyone who'd like to understand the trends that will shape the field of data science in the broader world, not only in 2024, but also in the years beyond. All right, are you ready for this visionary episode? Let's go.

00:02:39 Sadie, welcome to the Super Data Science Podcast. You seem familiar to me. Have you been on a show before? This seems like a well-trodden territory here.

Sadie: 00:02:51 Yeah, I'm feeling a bit of deja vu right now, although we came in with some new assets and some new hardware with these amazing Super Data Science podcast hats, so-

Jon: 00:03:00 All right, that is-

Sadie: 00:03:03 ... appreciate you not just having me back, but getting some swag. Everybody needs some Super Data Science swag.

Jon: 00:03:08 I think this is the first year that we could be doing this. So for people who have not been listening to the show for a while, this is now the third consecutive year where Sadie and I are doing a data science trends episode, so it's the first guest episode of the year, each of the last three years. So, the preceding one was episode number 641. It was the 2023 Data Science Trends episode, and in a second, I'm going to go through our predictions, mostly Sadie's predictions, but that's a really good side point that you make, which is that these having Super Data Science hats, we made these initially as Christmas gifts just for the Super Data Science team last year. So, we shipped them to those eight people that work on the show, and we sent them all over the world to those eight folks.

00:04:00 But people like you, in fact, I think you were the first one is you saw me wearing it, and you're like, "Where can I get this sweet hat?" I had a few extras, so I mailed one to you. Now, we're ordering tons of them, and I bring them with me to events that I go to, and some people really want them. It's pretty wild to see.

Sadie: 00:04:20 I love starting clubs, so I'm always like, "You should start a club. Whoever gets a Super Data Science hat, it's like this token, you get special access into all these things." So, I think you're onto something here.

Jon: 00:04:33 Someday we might. Something that might happen in 2024 is we might create an online store so that people can just go and then somehow... and so that we don't have to... Because obviously the administration, I'd love to just be packing up these hats, and sending them all over the world to people, but it's actually, it's quite... The expense of the hat isn't that much, but then it's the shipping and the time and the customs forms and all that stuff, but so we'll figure out a way to get you these hats, listeners, I think maybe in 2024. Is that my New Year's resolution? I don't do resolutions. It's so weird. If it's that important, just do it whenever you think of this. Don't wait till-

Sadie: 00:05:08 I believe differently. I've much on the contrary of your beliefs, and I respect everyone, and their method, and what works for them, but if you want to make this happen, I know a couple of dropship sites that can make it super easy to make sure that we can get everybody who wants a Super Data Science hat a Super Data Science hat.

Jon: 00:05:26 Sweet. Well, and so obviously this whole conversation is really only relevant to these small percentage of listeners that are watching the video version, and so we should call that out for... We're wearing trucker hats. The Super Data

Science hat is like a trucker hat. It's got a mesh back, and Sadie and I are both sporting them. I think it's the first time on the show. Sometimes I am wearing it, but I think it's the first time that me and the guest are wearing it. Anyway, so episode 641 was the most recent time that you were on the show, Sadie. We were doing a data science trends episode, and let's quickly recap the thoughts that you had. Your first big prediction, there was lots of small predictions, but I've summarized it into four main points.

00:06:07 Your first one is that data as a product will become increasingly popular as API endpoints like DALL·E 2 at that time, and ChatGPT provide a remarkably wide breadth of useful data outputs from large language models. Yes, I don't think we could have predicted how massive the GPT-4 release in March was going to be. I mean, that was for me... I went from being skeptical about AGI in our lifetimes to being, feeling like it could happen very frighteningly soon with that release.

Sadie: 00:06:42 I've never been too skeptical. I am a big fan of Ray Kurzweil and the Singularity, so for me, I'm like, "Oh, we're right on track with all these predictions." I think it's actually happening faster than we think. I was surprised at how big data as a product was, particularly in the startup space. I mean, you are in this space, and talk to people all the time. The amount of AI startups that happened this year was just incredible. I think maybe as many as started will be lost this next year, but I think that just goes to show how much data was a product. More importantly, I wish I would've seen it a little bit more in the enterprises. So when I think about existing companies who have really interesting data, and could have made it a product, I didn't see them do that as much, but what did become number one still was data as a product this year.

Jon: 00:07:35 Yep. Yep. Yep. Certainly, we're seeing amazing data flowing out of these APIs. DALL·E 3 has caught up with Midjourney, and is really great about getting texts in there. It's very cool to see, not as precise as they touted in their marketing materials, but I'm sure it'll get better. Your second prediction was also related to multimodal models actually. So, your second prediction was that multimodal models that leverage large language models, LLMs, under the hood will continue to stun us perhaps by nailing down compelling video. Yes, I mean, we just talked about DALL·E 3 and its ability to have text in there, and to be relatively precise, I think probably the most precise of the text-to-image models out there now, but we did just in the last couple of months, in fact, you recently shared on social media, and I saw that share, I can't remember what the company was, but it was a really stunning demo video of text-to-video.

Sadie: 00:08:35 This is from Pika Labs, and it is-

Jon: 00:08:36 Yes, Pika Labs.

Sadie: 00:08:42 They are doing some really cool work over there, but I've been even blown away with how far multimodal has advanced, particularly with the Gemini demo, I think. I don't have children. I don't know what it's like to have children, but I have felt that this year has been like watching your kid learn to walk and talk, and every week, it does something new. What I'm talking about is AI. I'm looking and reading X and LinkedIn and all these things, and something new comes out, and I'm so proud, extremely proud. I feel like a proud mother of like, "Oh my goodness, it can now see what I'm doing, and can converse back to me."

00:09:23 When the Gemini demo came out this past week, I felt exactly the same. I was just like, "I'm blown away by these capabilities. I'm also super proud. I'm also a little bit scared, because I don't have control anymore. All of the feelings were arisen from that."

Jon: 00:09:38 The Gemini demo was amazing at the time of recording. It came out just this week, and very cool to be able to see it being able to convert video into audio and vice versa. Super, super cool. I think unsurprising to see Gemini with, or sorry, Google, with the amount of resources that they have in terms of brilliant minds as well as compute. It seemed like it was only a matter of time before they had something competitive with GPT-4. Some people who've used Gemini, I have heard or seen people post that they feel like the text-to-text capabilities exceed GPT-4 on the examples that they had, but it's hard for me to evaluate that, and know that for sure, and know that these... I mean, obviously, it's case by case, and I think people, when you have something new like this, you're quite prone to be thinking, "Yeah, everyone's been talking about GPT-4, and this thing that I just got to demo is the new even bigger thing."

Sadie: 00:10:43 This isn't part of my predictions, but I'm going to throw a prediction in here, which is I think in this case, Google has been the tortoise, and OpenAI has been the hare. If we know the story, the tortoise wins in the long run, right? So, I look at this as when you look at the architecture Gemini, they thought about it from the ground-up instead of having an LLM model that they add additional things to. They started with multimodal. So, to me, a lot of people have been sleeping on Google this last year, and like. "What happened to Google? Where'd they go?" This was their first real example of like, "Hey, we're still in this 100%. We've just been slow playing it, and thinking it out, and well, maybe OpenAI is fighting about lots of things. Google's just slow and steady going to win this race." So, not a prediction, just a observation, I'll say.

Jon: 00:11:39 Nice. I mean, we could talk about this, I guess, probably the whole episode. I mean, we probably are going to be talking about this. I guess we're going to be talking about 2024, so let's get past the 2023 recap, and just get into the 2024 stuff soon. Your third of four big predictions for 2023 was that while senior data administrators might not have access to a true data mesh, or the inventor of the data mesh is Zhamak Dehghani. That kind of data mesh the sophistication of it, we talked about it last year. I'm not going to dig all back into it, and we did actually have Zhamak on the podcast as well in episode number 609. So, people want to learn all about the data mesh, but it's basically it's a sophisticated way of an organization being able to access data, lots of privacy settings.

00:12:30 While the technology for that doesn't exist yet, there will nevertheless need to be, in 2023 was your prediction, data literacy and data accessibility across the enterprise in order to realize the commercial potential of data and data models. It seems to me like the big breakthroughs in AI have made it a very easy sell. It's now suddenly so obvious that people need to have data literacy skills, and these generative AI models make it so much easier, because you can be asking questions about things in the flow of work, or you can be creating charts with your ChatGPT code interpreter now called advanced data analysis or whatever, but just generating charts with natural language, generating code with natural language and vice versa, being able to understand things. I feel like these kinds of tools make data literacy and data models so easy.

Sadie: 00:13:21 I think the other side of things is just because AI really took all the oxygen out of the room in 2023, because of that, the positive side, so many individuals who maybe weren't interested in data or machine learning, and how all of this works became interested in it. So, not only did the interest peaked and the awareness peaked, but now we have these tools that are amazing on wraps for individuals. I know you and I both posted about GPT data analysis and how amazing it is, but I think it's incredible for an in-business user. I'm so excited for what they're going to be able to do with it.

Jon: 00:14:00 I think it's... It was originally called the ChatGPT code interpreter, and my guess is that they changed the name to advanced data analysis, because when somebody sees... If you're a lay person, and you don't write code, and you see the code interpreter, it probably seems intimidating, but it's a shame because advanced data analytics only describes such a tiny part of what that code interpreter can do. So, you posted about it.

Sadie: 00:14:28 But I think you're hinting on why the whole GPT store got released was you have this tool that can do so many things in one. It's almost overwhelming for people, so we need to segment it into a data analysis or code interpreter. There's probably somebody who's going to write code creator or data science GPT. It's really all just the GPT on the back end, but segmenting it, so individuals know that it has that capability, and gives them a quick on-ramp to that.

Jon: 00:15:00 Data science and machine learning jobs increasingly demand cloud skills. With over 30% of job postings, listing cloud skills as a requirement today, and that percentage set to continue growing. Thankfully, Kirill and Hadelin, who have taught machine learning to millions of students, have now launched CloudWolf to efficiently provide you with the essential cloud computing skills. With CloudWolf, commit just 30 minutes a day for 30 days, and you can obtain your official AWS certification badge. Secure your career's future, join now at CloudWolf.com/SDS for a whopping 30% membership discount. Again, that's CloudWolf.com/SDS to start your cloud journey today.

00:15:42 Great points, really great points. Things have exploded so much now that you have these GPTs that you can be creating yourself instantly without writing any code in just a few minutes. I haven't done a whole episode on that, but it is a wild thing that that store opens up in terms of capabilities. If people are interested in my five hacks for data scientists, specifically using the ChatGPT Code Interpreter or now called ChatGPT Advanced Data Analytics, that was episode number 708 that I released in August. It has a 20-minute hands-on video so that you can see these hacks in action. As I was creating it, my first time using it, it blew my mind its ability to...

00:16:21 When you Upload a CSV file, it automatically can make guesses about what the column names are. It'll suggest to you, "Oh, thanks, I can see that there's some data here. Do you want me to do some exploratory data analysis, or maybe create a machine learning model?" You're like, "Okay, make a machine learning model." Then it makes thoughtful choices all the way along, and if it runs into bugs, it fixes them itself. It doesn't even prompt you like, "I think there's a bug. Do you want me to try to fix it?" It's just like, "Oh, I got a runtime error, but I know how to fix it," and it fixes it. So, it runs all right there in your browser. Very cool capability. Anyway, we could talk about that all day.

00:16:57 Your fourth and final prediction from 2023 was that many countries will adopt frameworks like the AI Bill of Rights, and enshrine data privacy and AI algorithm ethics into law. There has been a huge amount of activity all across the world in terms of AI regulation, this bringing together of academics, the commercial leaders, politicians in a way that is wild, and countries seem to be competing for. You see in the U.K., in Downing Street, their White House, as well as in the actual White House and various parliaments and governments all around the world trying to bring in Sam Altman and these other big players, Demis Hassabis to save us from across the AI world, and have them talk about AI legislation. One thing that I don't like about what's going on is it seems like these entrenched players are making a case for a fair bit of regulation that would obviously keep out innovation.

00:18:02 I get it. I get it that there's risks around open source, and you don't want to be providing bad state actors just with access to these super powerful tools. I realize that there's problems there, but it seems so obvious that these same big companies that now have entrenched interests are the same ones that are saying, "Yeah, you know what? I think we should have lots of regulation that only we'll have the budget for."

Sadie: 00:18:28 Andrew Ng talked about... I know he was pretty vocal when the AI executive order came out from the U.S. government about this, just in terms of the testing that is required for some of these models, and how it'll make it harder for startups or for open source communities to comply with these guidelines. I also agree, and I also think that it has a potential issue too just for the amount of diversity of who can enter these fields. So, I think that there's going to need to be an addition to this executive order. I don't know if you read it. It's a little bit over 150 pages long.

Jon: 00:19:08 I have not.

Sadie: 00:19:10 It's not super clear. It's not the most exciting read I will admit, but I think there's going to be a lot more to come in this space. But it's a start saying, "Hey, this is not something where we're just going to let tech this time go and build whatever they want. I do think that government is going to be involved in more ways than we've seen in the past."

Jon: 00:19:35 I guess they learned the lesson from the polarization that we've seen in newsfeeds, for example, with the proliferation of the internet, and so there's definitely things to watch out for. I mean, I wouldn't want even more polarized politics because of chatbots that, yes, foreign actors are creating that compel you to convincingly feel like you should trust them, and that you should vote in the way that's convenient for the foreign power. That's so obviously something that you could do today already. So, we should be trying to minimize those kinds of things, for sure. Also, something that happened in the past year is we got context windows large enough that that 150-page document you could put into a single call of Claude 2 from Anthropic, and most recently GPT-4.5 would be able to handle that as well, so from OpenAI.

00:20:26 That's a cool thing in the past year. I wouldn't have to read that 150-page document, just get a summary. Heaven, I don't know. I'm not even that interested in the summary. I think, probably regular listeners on the show know that I love listening to Last Week in AI. It's a weekly new show. Jeremie Harris, he's one of the co-hosts whom I know you know. He's been on your podcast and vice versa.

Sadie: 00:20:49 Yes.

Jon: 00:20:51 That's how I get my summaries of important news. Usually that's... Unless I'm like, "Oh, this is something I'm going to need for work, or something I should make a whole podcast episode about." When it's the regulatory stuff, I don't know. I feel like that's enough, because Jeremy in particular, because he is doing work with governments all over the world, but particularly in the U.S. and Canada, he is so on top of that. I feel like I get such an informed opinion. I don't need more of my own opinion involved in there. Anyway. All right, so that's

2023. Congratulations on being four for four on your big prediction, Sadie.

00:21:27 Let's see if you can do it again this year in 2024. Nice. All right. So for 2024, Sadie, I understand you have some structure, some overarching structure. Last year, you had a really cool structure as well. What's the structure this year for your predictions?

Sadie: 00:21:42 This year, I think of it like a tech stack where you have your hardware, your software, your end users. So, I really structured this in terms of a tech stack all the way from the beginning fundamental base stages of hardware. That really leads me into my first prediction, which is we're going to see an increased demand for data and hardware. I call this just data and hardware hungry. The reason for this is we all know that everyone is trying to get their hands on more GPUs, and there's a shortage, but the main reason for this is we have a really well-behaved predictable function of how to improve LLMs today. So, it's a function of N, the number of parameters in the network, and D, the amount of texts we train on.

00:22:36 Now, this is for just LLMs. It's not all models out there. But that being said, we're not looking at improvements from the different algorithms that we're using. It's pretty tried and true. Give it more compute, and give it more data, and it will perform better. So, I just see the need for hardware and data to continue to increase, and two byproducts of that are, one, I think we're going to start to see some innovation in the hardware space. Nvidia today has been the leader, but they're reaching this demand point. In addition to that, there's this opportunity where now additional startups are saying, "Hey, there's so much business in this space. We want to innovate, and we want to come into this market."

00:23:26 So, I think that there's going to be some innovation in the hardware space. Then on that data hunger space, we've seen that with these companies such as OpenAI and Google and Claude, they've scraped the internet. We have the base models that have pretty much taken everything from the internet, and been used to train on, but now we really need to delve into that area of fine-tuning and fine-tuning some of the niche areas for these models. So, I still think that there's going to be this incredible hunger for data, but not as much on that mass scrape the internet train the base model, but more of that fine-tuning of the top layer of the model. So, that's my first prediction, just an increased still appetite for hardware and data.

Jon: 00:24:18 For sure. Agreed. That feels like a really safe prediction. I feel like we're going to be taking that box off in a year's time. No problem. I mean, there's no question that is the way that things have been going so far, so as you're saying there. So, with the modern LLMs since GPT-3.5 and also the InstructGPT approach that OpenAI published on before GPT-3.5, since then, we've been using reinforcement learning from human feedback, RLHF, to add this third step onto the LLM training process. So, step one is this pre-training. It's unsupervised, so we don't need label data for that. We can just take a scrape of all of the internet as you're saying, and we need to do some cleaning up, because you don't want just duplicates of the same information.

00:25:11 Ideally, you want to avoid having misinformation, so you do some cleaning up on all of the internet, but there's no labels. You just use that to train up a general understanding of vocabulary in the world across all languages and programming languages and everything. It's wild with so many parameters, hundreds of millions, hundreds of billions, sorry, of parameters. We can encode all these different meanings. It's just in that unsupervised way. Then the next step is fine-tuning. This had been going on for a few years for LLMs, so GPT-2, GPT-3, of course, GPT-1, and other models around that time. They were fine-tuning in a supervised learning way.

00:25:57 So, you were using a traditional machine learning, supervised machine learning cost function, usually cross-entropy cost to be able to get closer to the kinds of outputs that a model should have given some input. So, this requires lots of relatively expensive data to obtain as opposed to that unlabeled data that's used for the pre-training, but it was this third step. So, first step, pre-training. Second step, fine-tuning, that had been going on for several years. It was a year ago in 2022 with the InstructGPT paper and then the GPT-3.5 model that was released under the hood in ChatGPT that it's this third step of reinforcement learning from human feedback, RLHF, that was this big step change in how aligned outputs are with the kinds of things that humans would like to see outputs be.

00:26:58 So, it's a more nuanced. It's a more dynamic. This reinforcement learning from human feedback is more nuanced, more dynamic, more aligned with the kinds of things that humans want than just the supervised learning that had dominated that second step before. So, I don't know, just a bit of history there, I guess, for people who didn't know it. But in all of these cases, more data is absolutely better at all three of those stages. As you're saying, we're running out of that first step, because there just isn't really much more data to have. But for steps two and three, there's still a huge amount of potential. So, in addition to the data sets going up, it's also the number of parameters which so we're scaling up by orders of magnitude. So, things like the Chinchilla scaling laws give us a sense of how many tokens we'll want given a certain model size, but that's just for doing it in the most cost-effective way.

00:27:54 You can... There's no... You always want more data. You always want more parameters. You're going to get better results, and so because of that, the demand for GPUs and other kinds of accelerators that are used for training AI models, that is definitely going to go up in the coming year. But all of that said, and I've been talking way too long. I'm so sorry, but all of that is to say that I think also we're going to see in 2024... For the first time, I think we're going to see big alternative approaches that require much less data. So, you might've already had this on your roadmap of things to talk about. If so, we can just put a pin in it, and come back to it later. Were you going to talk about Q*?

Sadie: 00:28:39 No, I was not going to talk about Q*, but I know you've been doing some research on this, so I'm excited to hear your prediction on this side of things.

Jon: 00:28:46 Well, I mean, so I've got a whole episode on it, which actually at the time of us recording this, I hadn't recorded yet, so I'm sure it's going to be a whole bunch more that I know by the time I publish episode number 740, which now from a listener's perspective has occurred in the past couple of weeks ago. But Q*, so Q* is this... The idea of Q* has been around since at least the early '90s. So, as a mathematical concept, it's the optimal state that we're predicting with something called a deep... Well, from the '90s, it was just a Q-learning algorithm. Then in the recent years, once we started using neural networks to approximate this optimal state Q*, then we started calling that deep reinforcement learning.

00:29:37 But the rumor is that using some new approach involving Q* at OpenAI supposedly might have played a role in Sam Altman being briefly fired, because apparently, with very small, with remarkably small amounts of training data, augmenting a large language model with this deep reinforcement learning approach, this Q* approach, has elicited apparently mind-boggling capabilities and accuracy on high school level mathematics. So, it was apparently... If something like a week before Sam Altman was fired, he publicly talked about being in a room. He's like, "Several times in my life in recent years, I've been in the room when some new capabilities seen," and apparently, he's talking about this Q* thing. Apparently, I guess, because of the board structure being designed to try to prevent potentially harmful releases like AGI releases happening too fast, apparently, that's why that might've played a role.

00:30:49 It's a rumor that it might've played a role in Sam Altman being let go from OpenAI. I mean, that's really interesting. So if we can be in the way that RLHF already played a huge role in the jump from GPT-3 to GPT-3.5, and alignment, and then it was scaling that up by an order of magnitude roughly that led to GPT-4, and those capabilities, which are certainly a big step in the direction of AGI as I already said at the beginning of this episode, we're ending up in this situation where potentially, this kind of Q* approach could be very powerful if all of a sudden it's way more efficient with the data that we have.

Sadie: 00:31:27 I think there is an opportunity to be more efficient with the data that we have, particularly when you start to look at niche industries. So, think of manufacturing as one. If you're making a product, and saying making a Super Data Science hat, and you have robots doing this portion of the task, you're in a very niche market. This is not something where the data is just widely available on the web. The same thing with healthcare is a great example of this, where we have HIPAA protection laws in the U.S. where you can't get access to this data on the web, thankfully for patient privacy, but there's so much that can be done with this information, but the availability of it is not what we see from social platforms from the internet.

00:32:15 So, I think that there is... If Q* does prove to be true, and the rumors are true of what it can do, we will see an enormous breakthrough just in the capabilities of what these systems can do. Because as you mentioned, it'll be able to go into all these niche areas, where the amount of data is less, and which will be helpful from two sides of things, one, more use cases to be tackled, more understanding for the system, but secondly, also able to train it at a cheaper and lower rate as well.

Jon: 00:32:49 Eager to learn about Large Language Models and Generative A.I. but don't know where to start? Check out my comprehensive two-hour training, which is available in its entirety on YouTube; yep, that means not only is it totally free, but it's ad-free as well. It's a pure educational resource. In the training, we introduce deep learning transformer architectures and how these enable the extraordinary capabilities of state-of-the-art LLMs. And it isn't just theory. My hands-on code demos, which feature the Hugging Face and PyTorch Lightning Python libraries, guide you through the entire lifecycle of LLM development—from training to real-world deployment. Check out my "Generative AI with Large Language Models: Hands-On Training" today on YouTube. We've got a link for you in the show notes.

00:33:32 Exactly. Which would be, obviously, breakthrough. If we can be having AI capabilities more cheaply, more widespread in more use cases, the fire of AI will spread even faster, so really exciting. It is interesting that this Q* breakthrough is supposedly a big step in the direction of AGI again, which something to keep an eye on. Maybe 2025, we'll just be like, "Okay, it happened."

Sadie: 00:34:02 I don't know if we'll know when we get to AGI. I think it's interesting. I really love how thoughtful Alan Turing was in the fact that he came up with his own test. I know there's criticism for the Turing test, and that yes, just talking to a computer, and thinking of it as a human is not AGI, but I think it is very insightful that he came up with this test, what, in the '70s, so quite some time ago, and realized that once we came to this point, we wouldn't even realize how magical it was, how much progress we had made. So, I think we're going to somewhat see a similar progress when we hit AGI too is at first, I don't know if we're going to be 100% aware that we've made that. I could go into a whole subset of neuroscience and psychology philosophy on that, because then you get into what actually makes a human, and is it conscious? Let's stick with predictions, because that would lead us down a rabbit hole.

Jon: 00:35:07 It probably doesn't really matter all that much, but the Turing test was devised in 1950 by Alan Turing. So, we're talking about big transistors, and I mean relative to today, very unsophisticated computers. Some of the first computers were all that were around back then. He called it the Imitation Game back then, which is also the name of the feature film starring Alan Turing. He died in 1954 tragically also. I don't want to ruin that film for you, but still a lot of question marks around that death. But no matter what, he was very wrongfully treated, at least in the social norms that we have today by the British government. It is shocking from this vantage point. All right, so we've got hardware was your big one. Demand for GPU is going up, of course. Then I threw one in there around... I mean, I don't know what my prediction is with Q*. That's not really a prediction. I guess, I just talked about it for a while, so let's just go to your number two. What's number two?

Sadie: 00:36:15 Number two is LLM operating system. One of the things that, I think, is becoming clear is we are in the midst of a platform shift. So, I think of this from the desktop era to the mobile era to what I would call like an LLM, or really more broadly, because they're becoming multimodal and

AI assistant. I can share the tweet, so there's a graphic for everyone, but Andrew Karpathy, or sorry, Andrej Karpathy released an image of what it would look like if an LLM did replace an operating system. So, I think this is really the base layer of where we're headed as a whole. I don't think we're ready for an LLM to replace Linux or our traditional operating systems today, but I think the thing we can all agree on is we are headed for a platform shift. Just as the same way that mobile came out and apps came out, the way we are working today is greatly going to be disrupted by LLMs and what I more likely broadly call AI assistants, and we are entering a new operating system era.

Jon: 00:37:34 I mean, how does that look and feel? If I think about a Linux operating system or a Mac operating system, what is an LLM operating system? So, it's this idea of conversation encompassing the system. So, instead of, "Ah, ah, I see where you're going. Okay. Okay." Up until the '90s, we had just command line interfaces. Then since the '90s, we had click and point interfaces. So, what you're saying with the LLMOS is this idea of get rid of the click and point, get rid of the screen, just talk and listen.

Sadie: 00:38:10 Yes. It'll be a little bit more, I think, than talk, and this is why it's really more than an LLM operating system. It's AI assistant-

Jon: 00:38:20 Right. It's showing assistant.

Sadie: 00:38:21 ... because it'll be able to see and hear. I mean, it already is today. I think we saw this exactly with the Gemini demo video. The demoer would draw on a piece of paper, and it would converse with it, and it would show images. They would talk to each other back and forth. So, just as we went from desktop where everything had to be on a keyboard, and you had your monitor to... Everything was on the go. Now, it's really more almost like a Jarvis, where you just talk to it, and it's this assistant that can hear and see you, and I don't know quite feel you yet. I'm not sure if we're into that phase yet, but definitely hear and see-

Jon: 00:39:04 What? Feel you?

Sadie: 00:39:04 To feel, right? It can feel how you are feeling to understand, but it's-

Jon: 00:39:11 That doesn't seem that hard to me reading facial expressions. I feel like that's something... I feel like there's a real opportunity there ranging from lie detectors to just having an emotional buddy for geriatric people or lonely people or whatever, autistic people. There's probably so many different kinds of use cases of machines being able to recognize facial expressions, and being able to feel you. So, J.A.R.V.I.S, that's Iron Man, Tony Stark thing, right?

Sadie: 00:39:49 Yes, exactly. I think that's a good-

Jon: 00:39:53 I'm not very good at the Marvel universe or any of those fantasy things.

Sadie: 00:39:57 I'm pretty good at the Marvel Universe. I'm not going to lie. Dr. Strange is definitely my favorite. We aren't ready to travel through the multiverse yet, so that will be phase two, but I definitely think that we're just at a base layer in a platform shift. So, this has a lot of implications, particularly for companies. I think back when mobile came out, and people were like, "Oh, we need an app," and everybody was just rushing to create an app, but the ones that did really well are the ones who had a purpose behind why they had an app, and why it was needed on the go, and why it was mobile, and was really still purposefully driven versus just trend driven.

00:40:38 I'm starting to already see that happen with AI and particularly large language models where everybody's jumping on board, and wanting to integrate it into their systems in some shape or form, which is important. It's where we're all headed. It's how we as individuals are going to be used to working with technology here in the next few years in terms of prompting and asking instead of moving around and clicking, but I still think we need to move forward very thoughtfully in this space. It's not a matter of like, "Let's just add an LLM to our product, because that's a hot trendy thing, and that's what we need to do, and this is the new operating system that we're moving towards," but more importantly, still having those design principles, the human experience principles. That is how you're going to be successful in operating in this new world.

Jon: 00:41:31 This makes perfect sense. I think the AI pin that was released by Humane in recent weeks at the time of recording also speaks to exactly where you're going. Hardware companies now are seriously trying to come up with a device that will supplant the phone and, as you say, LLMs or foundation models more broadly that can see, and hear, and speak allow you to potentially get rid of that screen, even if it's something like the AI pin, which is projecting a screen onto the palm of your hand, which seems a bit funny. I don't know if it's going to be AI pin, but I definitely agree that the phone is not going to be around forever just as the desktop was not around forever. Cool. Love it.

Sadie: 00:42:15 I'm curious. I've noticed that I have stopped using Google as much as I used to. I definitely use ChatGPT a lot more. I enjoy the process of asking a question, and getting an answer back right away, and so that's just changed my behavior. I'm curious, have you done the same? Do you notice that you're using Google, and how we used to search for things versus ask now is what we're doing?

Jon: 00:42:42 Certainly. Absolutely. I mean, if you're listening to this show, and you're not paying for a ChatGPT plus subscription $20 a month to be using GPT-4, I don't know. I hope that you're then using an API call, and doing it out of a command line interface or something that you've rigged up, because you've got to be augmenting with GPT-4 today in terms of answering questions, understanding the world, getting a sounding board, coming up with ideas. It makes so many things so much easier. Yes, absolutely, I definitely prefer, in most circumstances, having a conversation with GPT-4 over doing a web search with Google, because it synthesizes responses, and it anticipates.

00:43:30 I can crudely craft the RLHF factor in this. It allows me to crudely get at... I'm like, "Ah, it's probably not going to get at what my question is here where I... If I did that with a Google search, I would get back a page of results, and I was like, "Okay, I know exactly why I got back these results, but I'm going to have to change my query, and try again." Then maybe on the first page, some of the articles are going to be the kind of thing that I want, and I can click on a few of those, and synthesize with Google." On GPT-4, it's taking a leap by removing a whole bunch of those steps, because it's that crude thing that I type in that I'm like, "Ah, I probably haven't given enough context. I can't imagine how it's going to guess what I want here." It does, remarkably.

00:44:13 Most of the time, I'm just like, "Whoa, that is exactly what I was looking for." I don't know. People talk about hallucinations, but I don't think they've been using GPT-4, because it doesn't. I don't see problems. I don't see hallucinations. I don't see errors like I was with GPT-3.5, and I think that's quickly getting better all the time.

Sadie: 00:44:36 The other thing I'll add to that too is I 100% agree with you. If you're not paying for a subscription right now, why are you making life harder than it needs to be? It's my analysis to anybody. I'm like, "Your life could be so much easier, which is $20 a month." I don't work for the company. I don't get anything from it, but it's just the benefit that I have. The other side of things that I would say is not only how poorly my prompts are, but you don't need to watch any videos, or listen to anybody on Twitter or X or whatever about the best way to prompt. Just talk to it. I think that's the thing that is so surprising to me is however you talk much better than what we saw with Amazon's Alexa or all of those, where you would be yelling at it. Whatever your voice is, it does something with it, and it understands you, and it comes back to provide value. So, get ready for the new operating system. We're seeing lots of benefit.

Jon: 00:45:39 I mean, we talk about machines as being the driver of all this ingenuity, but ultimately, it isn't just the human engineers who are scaling up this transformer architecture to an order of magnitude more parameters, and an order of magnitude more data. The other really important thing here is that for that... We talked about three steps in training an LLM, step one being unsupervised, step two being supervised, step three being the reinforcement learning from human feedback. For step two and particularly for step three, that requires... There are people, there are humans who are all over the world and usually in places that have good internet connectivity, good language skills, but aren't particularly well paid.

00:46:26 Those jurisdictions are having these tasks being farmed up for them, creating these RLHF datasets that we are then training our models on. So, the way that the model is able to make that leap from your garbage input into exactly what you wanted, that is not just a machine innovation, and not just an AI engineer innovation, but that is also just pure hard sweat work that people have done for you all over the world. So, I hope people appreciate that about how some of this stuff works. Also, I just wanted to quickly give a shout-out here for Google Bard, which when I do need to do a real-time web search, I do not use Bing in GPT-4, because I mean, I don't know. I'm sure there's things about Microsoft that aren't bad.

00:47:18 I don't really actually use any Microsoft products at all ever, and have been like that for a very long time, but Bing seriously underperforms relative to a Google search. If you want to do a real-time web search, you're probably going to find better results by going to Google Bard, which is free with your Google login, and you can do a real-time search that way. Anyway, there's my practical advice for you, not you, Sadie, but [inaudible 00:47:47].

Sadie: 00:47:46 I love it. I do use Bard every once in a while.

Jon: 00:47:51 My custom instructions for GPT-4, the only custom instruction I have is never use real-time Bing search. That's not something I want you ever to do. That's my custom instruction. Okay, sweet. So, hardware, more demand for it than ever before. More players getting involved in it than before, not just NVIDIA. LLMOS is number two. What's number three, Sadie?

Sadie: 00:48:19 Number three is, I think, we're finally going to see what I'm calling a slow thinking model. Most people are familiar with the book Thinking Fast and Thinking Slow. We have built large language models that are really good at the thinking fast, right? They're good at the creative. They're good at the unconscious if you think about the hallucinations, right? They're just throwing words together, and a lot of times, they're very helpful, and it makes sense. We have gotten amazing use out of it. But if you look at things like how well these models do at performing math, they're still not at a passing grade level, so probably the best report-

Jon: 00:49:08 So, you do have Q* in here. You may not have done it, but this is it. This is it. This is the Q* thing.

Sadie: 00:49:13 Once you explain it more, I was like, "Oh, that fits in number three." I thought you were talking about Q*s being needed for having less data. SO, I was like, "No, I haven't heard [inaudible 00:49:25]."

Jon: 00:49:25 It ends up, I think, having that side effect, because it is thinking more... It less about regurgitating. It's less about auto completing the next sentence. It's more about... I think it is this. I think this was using Daniel Kahneman in his book, Thinking Fast and Slow, as the analogy here was a brilliant way to do it, but this is it. I think, this is exactly what you're talking about with Q*.

Sadie: 00:49:47 So, whether it's Q* or whether some other company comes out with it, we're all talking about the same thing, which is really a slower thinking model, so one that can be rational, one that has logic and more reasoning, and more importantly, performs better on mathematical tests. So, if you look at the latest stats from Google Gemini where they're comparing GPT-4 to Gemini Ultra, I mean, pretty much everything, we're in the at least minimum 70% to 90% on different tasks. But when we get to mathematics, we're not even at a passing grade level at 50%. So, we have a lot of progress left to make in this area. I think this is where the focus is going to be. Now, I have a side prediction on this, which is if we are able to create a slow thinking model, I don't know if we will see it released to the public in 2024.

00:50:49 The reason that I say that is because if a model can actually perform mathematics at a higher functioning way, there is concerns about encryptions being broken, and lots more security concerns, and we've seen with the traditional LLMs today. So, that's a side token, which is I think we're going to have the breakthrough of it. If let's say, for instance, OpenAI does have it, and that it is working, I think one of the reasons we may not see it yet is just some of the safety concerns of what happens when you have a ultra powerful model that can run mathematical equations and predictions released to the public.

Jon: 00:51:34 Right, or just more CEOs might get fired from Big Corp. All the big tech companies, their CEO is going to get fired in 2024, because all of them are going to come up with a Q*-like solution, and everyone's going to revolt and say, "It's the end of the world." I mean, it's absolutely, there are risks, but also enormous positive potential new scientific discoveries, physics understanding if... I mean, this really is... I mean, people don't... What the word AGI means is quite nebulous, and we've talked about that in a number of recent episodes on the show, most particularly with episode number 729 with Professor Blake Richards, if you want some nuanced discussion of the concept of AGI. But in terms of a general idea, it really does seem like if this is happening, it could be 2024 that you're right. Maybe not in the public, but it could at least be in the private that we have this key piece of AGI figured out.

00:52:35 Maybe there will be some scaling up or some fine-tuning or more data required after that, but potentially super game changing, and hopefully we'll make a big, big positive difference more than Armageddon-like difference. Cool. All right. You got something else? I see you nailed.

Sadie: 00:52:54 I'm just excited... You can hear the deep breath in. I'm very excited about this, because I think this is where we can possibly reach some additional scientific breakthroughs. So, the impacts of this on health and physics and biology, I think, are really tremendous. So, that excites me the most about some of these slow thinking models. But more importantly, I think that this has a really good opportunity to... If you think about what we're building, essentially, we're building the brain of what will soon be a robot. So, this goes on to a little bit of a side prediction, but I love watching Tesla's AI conference day or AI day.

00:53:47 While their bot doesn't work that great today, if we can see where it's headed, and think about these models that we're building with the large language capabilities and some of the reasoning capabilities, what happens when we take that? That truly is the brain that goes into some of these robots. To me, that's when we get into a very ultra futuristic state, because not only is it something that's locked in my computer or my phone, but now has the capability to walk and talk and move.

Jon: 00:54:19 For sure. Some people also hypothesize that it's this ability to explore that really allows something to be an AGI, because what are you if you just memorized a whole bunch of facts, and it's an experience of the world that... Anyway, that's part of the nuance that gets into these arguments, but you're absolutely right. Awesome. All right, so hardware, more of it, LLMOS, slow thinking models. What's next, Sadie, for 2024 after that?

Sadie: 00:54:50 As we move up the tech stack, what I see is a tool consolidation, and this is happening for a couple of reasons. I think that a lot of organizations are still going to have budget constraints, but if you take a simple tool like ChatGPT, I can already think of how many tools I've had to get rid of because it can do it all in one, right? We've also started to see this with things. I know you're not a Microsoft fan, but Microsoft came out with what they call their Fabric. This is a tool that essentially connects to any data cloud. It does your data governance. It does your ETL. It does your visualization. It's really your whole data tech stack all in one. So, in the past, instead of having individual vendors that may provide governance or ETL capabilities, it's so much cheaper to be able to have this in one.

00:55:48 With budget constraints, I think that we're going to start to see more organizations consolidate their tech stack, and are also going to start starting to look at their enterprise tech stack as a whole anyways. As we're changing our operating system, people are reevaluating what they need to have in their stack, and I think that's going to lead to a lot of consolidation.

Jon: 00:56:14 Nice. I think that's a solid prediction for sure, and these kinds of tools like the GPT-4 function, well, actually it was originally GPT-3.5 function calling API, more recently, GPT-4 function calling API, allowing these LLMs to be the glue between these systems. It makes it so much easier, because it can be taking the API definitions from one system, and then converting them into an API definition for another system without a human needing to be involved in all these decisions anymore. It's also allowing, as you say, in the LLMOS conversation, allowing a human to be able to use natural language to make queries about these data, or to get these systems to do something. There could be a robotic endpoint, and so yeah, absolutely agree with you 100% consolidation of systems.

00:57:06 A really cool idea here is that you could have one voice command or one typed command in natural language that then goes out and calls multiple different kinds of systems in the enterprise, integrates those results together from different systems. Whether that's something like a SQL table join, or whether it's conceptual, or whether it's creating a data visualization from multiple different sources blended together, or whether it's using information from one API call to realize that actually you're going to need to make some other API calls, all of these kinds of things become possible with LLMs. It's very exciting.

Sadie: 00:57:44 It is an exciting time. I mean, I think it's just going to even speed up our work process much faster, because instead of switching from one application to another, it's all in one, and it's making all the calls on the back end for us that we don't need to worry about. So, I think it's also just going to speed up our work time, and make us each much more productive as well.

Jon: 00:58:06 Nice. All right. What's next up the tech stack, Sadie?

Sadie: 00:58:09 All right, so now, we get into the end users, right? We've consolidated the tools. This one, I'm calling workforce upheaval. The reason for this is I was speaking with one of our mutual friends, Hapreet, who we were just even questioning what it means to be a data scientist even more.

Jon: 00:58:32 Hapreet Sahota.

Sadie: 00:58:33 Yes.

Jon: 00:58:36 He was on a... Sorry, I'm interrupting you completely, but I'm just going to blurt it out, and then I'll shut back up. He was most recently on our show in episode number 693, and if people want to know more about Hapreet. Anyway, continue. You guys were talking.

Sadie: 00:58:50 So, we were just talking about where does a data scientist fit today? When I got into data science over 10 years ago, it was really expected. The field was new. There was this idea like a full stack data scientist, right? You were supposed to be able to talk to business stakeholders. You needed to be able to clean and manipulate your data. You needed to understand it and build visualizations, and story tell with it, and then build a predictive model, and then deploy that model. You needed to do what is now split up into multiple job functions from data engineering to data product management to data analysts. So, as we are talking about this to machine learning engineers, to MLOps engineers, the job has expanded quite drastically so much so that I would call it as a data science job family, but I don't particularly know. Maybe you can help me with this is where does a data scientist fit today?

01:00:01 Is it somebody who is just doing AB testing, because more of the ML engineers are doing the model development, and the MLOps are doing the deployment? Are they really just a glorified analyst? But I think this whole concept creates a broader conversation too, which is what I'm calling work upheaval, because we're all now working alongside these AI assistants, and just as we started this conversation talking about ChatGPT's data analysis tool, what happens when business users can just upload their own data set, and start to talk to an AI assistant, and get the answers that they need? Are they now data analysts? Does this mean there aren't a need for data analysts?

01:00:51 I think there is a need and a place for everyone. More importantly, from most of the reports that I read in terms of job loss and job created, there's going to be more jobs created. I just don't think we have a clear vision of what that looks like. I don't think roles anymore are going to be as siloed as they have been in the past. When we have AI assistants that are multimodal, we as individuals are going to be expected to also have multimodal roles. I'll call it multimodal roles. I don't think that's a real thing yet, but what I'm trying to get at here is it's not going to be a siloed approach anymore. I think this next year, we're going to fill a lot of contention just societally, but also in the workplace of what different roles mean, what should they be doing? We will get it all figured out. I just think that this next year, it's going to feel a little awkward, and that's because we are in the midst of a transformation, particularly in the way we work, what it means to work, and how we work.

Jon: 01:02:03 Totally agree. For sure. I think things like retraining programs, talking to your local elected politician, and encouraging them to be investing in retraining programs, I think, is critical. Organizations like yours, Women in Data, Sadie, are also great for allowing people to retrain in their job in an active or in a more passive way. Hopefully podcasts like this and like yours as well allow people to continue to evolve with these tools, because absolutely, as we were talking about with the 20 bucks a month, it's like if you're not doing that, you are being left behind by people who are. In roles that require more creativity and domain expertise like data scientists and associated jobs, you might be buffered a little bit, and it's a little bit more abstract.

01:02:57 It's a little bit less obvious how these automation tools are impacting you. There's lots of other kinds of... The more that your role is something that could be encoded in model weights already today, the more that you need to be thinking about how you can be adding value as the human to the chain, and a GPT-4 conversation can figure out exactly for your conversation, how you could be doing that, exactly for your position, how you could be doing that. So, it's interesting. Something that I noted here is that what you're describing actually in your previous prediction for 2024, you were talking about tool consolidation facilitated by LLMs acting as a glue between different systems, and as well as between humans in the systems. This is actually the same thing, but it's the impact on the worker.

01:03:48 So, that same kind of consolidation, LLM acting as a glue, yes, it is allowing a business person who doesn't have any coding experience now to be doing sophisticated queries from doing data polls, doing their own data analysis, getting the suggested insights. The downside of all this, I think, Sadie, is that it's going to be these... We've all worked with them, these arrogant people who do a Google search that proves their point that you know is wrong, and they dig their heels in, and they're such a pain to work with. This is going to potentially make that worse, because you'll have situations where they feel like they have all this extra data, and quantification, and expertise that they can be pushing down people's throats when in fact-

Sadie: 01:04:46 Oh yeah, we've all worked with that person, or the person who's also the loudest in the room, "We're really good at selling things, right?" You're like, I know that you don't do any work, but you also talk the most, and so everybody listens to you kind of thing. But I think if they're using it, let's say you're an individual who does the deep analysis, and truly questions things, and makes sure that the data is accurate, that doesn't mean you shouldn't be using these tools. That's the sad part that I see for a lot of individuals. If they are that individual who puts in the work, they put off using an AI assistant, because they attribute it to that person who talks about it all the time, doesn't do any work. All I'm saying is don't throw the baby out with the bath water here, which is just because some people may use it poorly doesn't mean that you using it are going to be put in the same campus. You can still remain true to who you are and your deep analysis.

Jon: 01:05:47 Absolutely. Absolutely.

Sadie: 01:05:48 So, just use it, and don't worry about the other person, because somebody needs to take them on.

Jon: 01:05:55 Exactly. It was a corollary. Nice. All right, so I can see how we've been working up the tech stack here. We started with hardware. We're talking about the underlying operating system. We're talking about the model. We're talking about system consolidation. We're talking about the workers being impacted on the user side of the tech stack. How much higher can we go? Are we reaching the end, the top?

Sadie: 01:06:18 We're reaching the pinnacle here. I'm curious if there's anything you feel like we've missed adding to this tech stack.

Jon: 01:06:25 I mean, I guess I've got a few separate points. I don't know how neatly I can weave them into your tech stack. Do you want me to go? Do you want me to try to [inaudible 01:06:39]?

Sadie: 01:06:42 Yeah, or we can build a whole new tech stack over here. We got to always update our enterprise architecture. So, let's see what we got.

Jon: 01:06:46 So, people who regularly listen to the show may already know that Serg Masis, a brilliant data scientist and author and speaker is our researcher on the Super Data Science podcast. So often, most of the time for most guests, like 90% or 95% of them or more, he's doing research. So, he compiles together for me information from papers they've written, patents they've published, from video lectures they've given, books they've written. He will... When we have guests on the show where they recently wrote a book, Serg will seemingly read the entire book, and create questions from all over the book. It seems like our guests are often very impressed. Hopefully you as a listener are also impressed, and a lot of that is I'm like Serg's marionette.

01:07:43 For this episode, initially, I said, "Oh, we don't have a guest next week. It's just Sadie and me. We're doing our Data Science transprediction episode again, and so you don't need to worry about next week's episode." But then I wrote back about an hour later, and I was like, "Oh, but if you have some insightful ideas on 2024, feel free to let us know." Of course, I got many pages back, which is the standard here. I'm not going to go through all of Serg's predictions, but maybe the ones that I like the most and agree with the most. So, one of is here, it aligns with what you were talking about a little bit back with the hardware stuff, which is that it seems like there's a bubble that's going to burst here, an AI bubble.

01:08:27 So, we are in, as pointed out by Serg in his research here, what the Gardner hype cycle calls the peak of inflated expectations. So, there are some headwinds ahead of us, like the physical limits of chip design in terms of how small you could make transistors without electrons hopping over in between circuits. He likens the moment we're in right now potentially to being the dotcom bubble around 2000, where everybody is jumping in. Everybody that was doing crypto before now has an AI startup, and there's probably way more AI startups than we need, but he also makes the argument here that it'll be good. So, it's like after the dotcom bubble, the ship got tightened. You're focusing more on better practices, standards, safeguards, businesses that are actually generating revenue and profit margins as opposed to just being hopeful.

01:09:27 Then capital will pour into needed infrastructure to scale it, and support use cases that were only pipe dreams. So, these kinds of things could happen in 2024 to address model size, processing issues, energy consumption, latency limitations. There's all kinds of things that could end up becoming good. I don't know if he's predicting that we'll have a crash in 2024, but I do agree that there will be some... Things can't continue in 2024 in terms of investment in AI, I don't think, like they did in 2023.

Sadie: 01:09:59 No. Eventually, people want to see some returns on their investments, so the collectors always call, I'll say that, whether it's for VC or not.

Jon: 01:10:10 Then we touched on this a little bit with a pin released by Humane, and so he makes a big point here that with Apple's Vision Pro coming out in... I guess we'll get our hands on them as consumers for the first time in 2024. Yes, they have a big price tag. I think it's 3,500 US dollars, but I think people just like and trust Apple so much more than Meta. I absolutely... I was an early Oculus user, and when all of a sudden everything I had to be... They switched me over to doing a Facebook login to use my Oculus. I'm not allowed to swear on this show, but it makes me want to... Just, they're just going to bleep it out. It's just when something like that happens, it makes me say, "F*** you, Meta." There we go. We can bleep. I guess that was bleeped now. It's just I don't... Although I love Llama too, I love what they're doing for the open source LOM ecosystem,

Sadie: 01:11:22 Do you have to use your Facebook login to access it? That's my question.

Jon: 01:11:26 That's a really funny question. You don't, but I'm surprised that they didn't make you do that, because they do actually have an approval process. You do have to fill in a form to get access to those models, because you have to prove that you don't have 700 million users in your platform, which maybe that's going to be a problem I have in 2024. I'd love to have that problem for my Nebula platform in 2024, to have 700 million users. Then I can no longer use Llama 2. That'd be great.

Sadie: 01:11:54 I feel like it's a good kind of benchmark, because if you do get to that point, you can afford and pay to train your own base model. I think they set up the structure of it, like somewhat open source, not fully if you grow big enough, but if you grow big enough, hey, you can pay for it yourself then. So, I hope you can pay for your own model then someday.

Jon: 01:12:17 I hope so too. Another one that I like to hear from here, my final one that I'll go into is breakthroughs in Edge AI. He had a lot more detail here than I'll go into. Maybe Serg will post these on his own personal LinkedIn or something, or maybe he'll let me post them on on my LinkedIn, and just say, "This is Serg's work," because he had 10 of them, and all of them were incredibly detailed, but this one, breakthroughs in Edge AI, I thought was great. So, this is the idea of being able to run LLMs on a phone in common use cases. So, this is obviously... We are today limited by being able to send the query along an API to some central server, which does all the heavy lifting, all the heavy compute, and then sends the response back to that edge device.

01:13:04 But if smartphones, Raspberry Pis, these kinds of small devices could be running their own LLMs, this would open up... This would revolutionize, in his words, industries like agriculture, mining, healthcare in particular, elder care, and retail, all industries that have insufficient boots on the ground. That could be augmented by the kinds of LLMOPs, AI automation across the board that you're describing, Sadie, happening without information needing to be sent back to a central server would open up a lot more possibilities. So, that's what we've got between us. Then we also... I did announce that you would be on the show, Sadie, a few days ago, and quite a popular post, 150 reactions, a couple dozen comments, 14,000 impressions, and some people did actually... Some of those comments are related to my question in the post, which is let us know your predictions for 2024. It's actually only a small minority of the posts were actually on that.

Sadie: 01:14:13 It's always encouraging when you get comments on what you asked to be commented on versus just...

Jon: 01:14:19 Also, we did have a couple, so Rupinder Kaur, who is founder at The Sherry Code, she says, "Love listening to you two. I love the 2023 episode. That's how I started my year. Looking forward to your 2024 predictions episode." As for her predictions, she says that there will be additional focus on scaling systems, AI scaling systems, scaling AI chips, and teaching small language models to reason well. So, great points. Scaling, obviously, we talked about the demand for AI chips going up. That was your very first big prediction, Sadie, but something that we didn't talk about that is absolutely right. I mean, I guess it is actually Serg's point here, teaching small language models to reason well, so that's related to actually the edge device kind of thing that we were talking about with Serg. So, Rupinder, we are on the same page as you. Thank you very much.

01:15:11 Then we have Sarah S. I can't actually see her last name, because of I guess the way her privacy settings are set up, but it says that she's a data science slash engineering student and a musician. So, Sarah S. predicts that under the ethics category, there will be a push for image generators having permissions or having to pay for licensing to use artist's work for source content. She says it would be a win-win situation for artists to be paid for AI to create work from their body of work. So, there you go. It ends up being that her musician background potentially is playing a role in that perspective. I mean, Sarah, that would be great.

01:15:53 I guess, it's similar to what we're seeing in some countries now. In Canada, the federal government imposed a law that said that the big news aggregators like Google and Facebook have to pay the news creators for sharing that content. Initially, there was pushback where things like

Facebook was just like, "Okay, there will be no sharing of public posts in the Facebook feed, or news items in the Facebook feed." But now, I think it's coming around, and Google at least I think has decided to... They're making now a big investment in Canadian news. So, hopefully that allows us to have a bit less bias, a bit more perspective, a bit more local in our news reporting, and so yeah, the same kind of thing with artist's content. That would be a cool thing to see.

Sadie: 01:16:45 I think it'll be interesting just because we don't have a lot of case law in this area yet. So, typically, how it works is whatever is a law that has previously been decided upon then gets cited, and interpretations get taken after that. So, it'll be interesting to see what the rulings are on some of these first instances brought to trial. I've seen a couple of them. I don't know if you're familiar with Sarah Silverman had a lawsuit, and I believe she lost, and then there was another recent one in terms of artwork that also lost. There wasn't enough direct replication of the artist's work. So if those continue to become law, it's not looking good for artists, but hopefully they can get a couple of wins here or there, and reverse some of the case law hearings.

Jon: 01:17:43 Awesome. I'm a big fan of your predictions again this year, Sadie. Tell us though, before I let you go, what you are going to be up to personally in 2024. I know, for example, that you've been working on The Observer.

Sadie: 01:18:00 Yes. Most of the times, when I tell people this, they look at me like I'm a little bit crazy, but I promise that... I shouldn't promise that I'm not crazy. I won't make any prediction to say that, but I have data on what I've been doing for the past 10 years in half an hour increments. I did not mean to start to do this. Essentially, I started using a planner almost 10 years ago. What I found was that I would make plans, and they would never fall through. So, I would plan out my day, and say, "From this time, this one-

Jon: 01:18:41 They would never come through.

Sadie: 01:18:41 Yes, they'd never come through. They'd always fall through. They never came through. Thank you. So, it'd be frustrating, right? You set out your agenda for the day, and interruptions happen, and everything just goes chaotic, and you have little motivation to add to your planner again, because you don't have control, right? I love the quote like, "We make plans, and God laughs." That is how I felt every day. So, what I started doing instead was today, I would write what I did the previous day, and so my planner was broken down in half an hour increments. Slowly, over time as I wrote down what I did that previous day, I started to change my choices, and I noticed was going through that process that I started to observe myself, and hence the name The Observer.

01:19:31 What The Observer is it's a method that now I've developed having done this for 10 years, where you break it up into sections. You dream what you want to do in your lifetime. It has this process of being able to focus, but the biggest portion of it is to reflect. For me, it's a very data-driven method. As individuals, we talk about making decisions with data, and being data-driven, but very little do we do it personally in our own life. Maybe you have some fitness metrics. There may be a subset of your life that you have data on, but I doubt you have data on yourself in half an hour increments for 10 years. If you do, please call me, because I really want to talk to somebody on this, and compare notes.

01:20:12 So, the whole idea is that you observe yourself. You keep track of what you did, and by doing that, you start to make different choices by recognizing where your time is going to see if it aligns with your goals and priorities. I shared about this on my Instagram and on my story, and people started asking, "Hey, can I get access to this?" So, I ended up creating a drop shipping site. I really wasn't planning on launching a whole product this year, but it's been super fun. I've had a ton of fun being able to create the website, do some different marketing material, and now I'm just super excited to chat with people about the method and what it does in their life.

Jon: 01:20:53 I think you have it right with you, don't you?

Sadie: 01:20:54 I do.

Jon: 01:20:55 You hold it up for the YouTube. There we go. Nice. The Observer looks beautiful. I don't exactly do what you do with the half hour increments, but I do log how many half hour increments in a day are spent on deep work. So, I log that, but it's just generally... I was in the habit until recently of knowing exactly what deep work each of those half hour increments was, but I don't know. I recently decided that that was... I was adding process. I don't know.

Sadie: 01:21:29 Well, I will say you were the one who inspired me from last year's conversation to make more time for deep work. So, this year, I totally changed all my meeting schedule, where I only take meetings on Tuesday afternoon and Thursday mornings to allow for maximum deep work time.

Jon: 01:21:49 That is a lot of deep work. I'm so jealous. Hopefully I can get to 700 million users soon so that I can have that kind of deep work in my life. That would be amazing. Sweet. You also have a new LinkedIn class coming out?

Sadie: 01:22:03 I do have a new LinkedIn learning class coming out. It is AI for data professionals. This is my first LinkedIn learning class. I had a ton of fun creating it. It has been five years since I taught a class, and I am so happy to be teaching again. I didn't realize how much I missed it, and I can't wait to just connect with students. This class is designed to provide individuals who are working in the data space, just an overview to the whole world of AI. There is so much more than ChatGPT and reinforcement learning, and so expanding where everything fits in one. The idea is to be a launchpad for individuals, so really excited to connect with students, and hear their thoughts, and see where they launch after taking the class.

Jon: 01:22:55 Nice, and expanding interest even beyond just data science and machine learning and so on. You and I are also both very interested in CrossFit. So, Sadie, who are your predictions for the games winners in CrossFit in 2024?

Sadie: 01:23:12 I think that one is actually harder than the data science trends. I do think it is harder, because you just never know who's going to really show up. I don't know if I have a hard or strong prediction. Do you have a strong prediction on this one?

Jon: 01:23:31 I think Jeff Adler will probably win the men's again, but I would love for Pat Vellner to win. I would just love... At some point, he's been around and so close so many times. I'd love to see Pat Vellner win. Roman Khrennikov could also win, but I don't know. Jeff Adler is probably my favorite for winning the men's side. On the women's side, Tia Toomey back from her first child after... She missed a year, but she was very strong at the Rogue Invitational this autumn in the Northern Hemisphere, and getting a second place just narrowly. Tia Toomey is probably the favorite. I work out in Canada with Emma Lawson who came third in this year's game, so I'd love to see her chip away.

01:24:14 I think that it's possible that someday she will stand on top of the podium, not just on the podium. I think at... I guess she'll be 19 for next year's games, or 18 still. She'll be around that age. I can't remember exactly. The strength events are tough because building muscle mass just takes time. It's just an investment of years and years and years. So, I don't know if she'll be able to win the games in 2024, but I wouldn't be surprised if that happened in the not too distant future.

Sadie: 01:24:49 I'm definitely cheering for her. I knew that you worked out with her, and I think you have the same coach too, or follow the same program.

Jon: 01:24:57 Same cross. Well, yeah. My family is in Waterloo, Ontario, Canada, famous for the University of Waterloo. I spend long stretches at home in the summer or Christmas as much as I can really. When I'm doing that, I'm working out at a CrossFit gym called Polsky's Strength and Conditioning, and Nick Anapolsky who runs it. I mean, he's the coach of the gym, although Emma now does... There was a time... This may still happen from time to time that she's doing just the programmed workouts for the gym. I do them all year round. I do Polsky's programming all year round, which is just the strength and conditioning work that they have from their box. There was a time where she would sometimes just do those with us, but I think she's gotten to a point where...

01:25:54 She works out in the facility. You can work out with her, or alongside her slowly, but she's typically following her own bespoke programming that's created for her, something called the Mammoth method, which is... He's a coach that operates out of the same facility, out of the same gym. I can't remember the guy's name right now, but his approach is called the Mammoth Method, so people can look that up.

Sadie: 01:26:20 Well, I will say that was a highlight of my last year was getting to do a workout with you when I was in New York, so I'll have to make that a yearly tradition. I don't know.

Jon: 01:26:34 For sure. Even more so, or maybe I should be coming to California. I should be spending part of the winter in California working out there. It seems like that would be a good move.

Sadie: 01:26:43 Yes, it is a nice time of year to be in California, I'll say that.

Jon: 01:26:48 I bet. All right, Sadie, well, thank you so much for being on the show yet again this year. I'm looking forward to seeing how your predictions pan out over the course of the year. Super insightful as always. So, if people have listened to you before, they'll already know this, but if they haven't, what's the best place to follow you on social media? What are the ways we should follow you?

Sadie: 01:27:11 LinkedIn, I'm fairly active. Also, Instagram is a good place to follow me. I am not as active as I should be probably on X, but if that's your platform of choice, I will meet you there.

Jon: 01:27:26 There is a name change we didn't see coming a year ago. All right, Sadie, thank you so much. Super generous with your time, your attention. I'm so delighted that you agreed to do this again this year. Maybe next year, we'll be checking back in with you again. Thanks, Sadie.

Sadie: 01:27:44 Thanks, Jon. Bye, everybody.

Jon: 01:27:52 What a great start to the new year. In today's episode, Sadie predicted that in 2024, demand for GPUs and other AI hardware accelerators will continue to increase, perhaps creating an opportunity for players other than Nvidia to find a foothold. She also predicted that LLMs will begin to act as a sort of new operating system, allowing us to interact with a broader range of applications using our faces and voices instead of typing on a screen or keyboard. She talked about approaches more clever than just scaling up LLMs, and that these will facilitate multiple paths to be explored, emulating human slow thinking and markedly improve performance on logic, reasoning and math tasks perhaps with fewer training data required.

01:28:33 She predicted that the function calling APIs of LLMs will act as a glue between different systems and applications, facilitating enterprise tool consolidation, and that business users will be able to use LLM interpreters like ChatGPT's Advanced. She predicted that business users will be able to use code interpreters like ChatGPT's advanced data analysis to do data analytics themselves, potentially displacing analysts, or freeing them up to do more sophisticated work. 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 Sadie's social media profiles, as well as my own at SuperDataScience.com/745. Thanks to my colleagues at Nebula for supporting me while I create content like this Super Data Science episode for you.

01:29:18 Thanks of course to Ivana, Mario, Natalie, Serg, Sylvia, Zara, and Kirill on the Super Data Science team for producing another visionary episode for us today. For enabling that super team to create this free podcast for you, we are deeply grateful to our sponsors. You can support this show by checking out our sponsors' links, which are in the show notes. If you yourself are interested in sponsoring an episode, you can get the details on how by making your way to johnkrohn.com/podcast. Otherwise, please share, review, subscribe and all that good stuff. But most importantly, just keep on tuning in. I'm so grateful to have you listening, and I hope I can continue to make episodes you love for years and years to come. Until 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.