SDS PODCAST EPISODE 785: MATH, QUANTUM ML AND LANGUAGE EMBEDDINGS, WITH DR. LUIS SERRANO

Jon Krohn: 00:00:00 This is episode number 785 with Dr. Luis Serrano of the Serrano Academy. Today's episode is brought to you by AWS Cloud Computing Services.

00:00:12 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. And now let's make the complex simple.

00:00:43 Welcome back to the Super Data Science podcast, prepared to have your mind expanded by today's guest, Dr. Luis Serrano, a master at making complex math and machine learning topics friendly and approachable. Luis is the beloved creator behind the Serrano Academy, an educational YouTube channel with over 145,000 subscribers. Until this month, he worked as head of developer relations at Cohere, one of the world's few AI labs that is actually at the frontier of large language models. Prior to that, he was a quantum AI research scientist at Zapata Computing, lead AI educator at Apple, head of content for AI at Udacity, and ML engineer at Google. He holds a PhD in math from the University of Michigan.

00:01:23 Luis is also the author of the popular book, Grokking Machine Learning. I will personally ship five physical copies of Grokking Machine Learning to people who comment or reshare the LinkedIn post that I publish about Luis's episode from my personal LinkedIn account today. Simply mention in your comment or reshare that you'd like the book. I'll hold a draw to select the five book winners next week. So you have until Sunday, May 26th to get involved with this book contest.

00:01:48 Today's episode should be appealing to just about anyone. In it Luis details how supposedly complex topics

like math and AI can be made easy to understand, how Cohere's focus on enterprise use cases for LLMs has led it to specialize in embeddings, the most important component of LLMs, the promising application areas for quantum machine learning, and what the next big things in AI will be. All right, you ready for this fascinating episode? Let's go.

00:02:20 Luis, welcome to the Super Data Science podcast. I'm ecstatic to have you here. Where are you calling in from today?

Luis Serrano: 00:02:26 Thank you, Jon. I'm happy to be here. I am in Toronto, Canada.

Jon Krohn: 00:02:30 Toronto, my hometown. Nice. I do miss Toronto so much and our regular listeners will know that I'm based in New York, but something that they wouldn't know is that lately I have been homesick for going back home and I'm thinking of starting to spend a lot more time back home in the Toronto area. So Luis, maybe we'll be hanging out soon.

Luis Serrano: 00:02:50 Definitely. Let me know when you're in the area.

Jon Krohn: 00:02:51 Nice.

Luis Serrano: 00:02:52 I'm normally here when it's warm and in the very winter I go to Colombia, which is where I grew up, and so I've been avoiding winter for some time, which is nice.

Jon Krohn: 00:03:02 Nice. Yeah, that makes sense. That is also, I'm planning on this Northern Hemisphere summer, is when I'm planning on having my first big foray in Toronto and maybe I'll come to Columbia with you for the winter.

Luis Serrano: 00:03:14 Oh, even better. Yeah.

Jon Krohn: 00:03:16 Nice. So I actually don't know how I met you or first heard about you. It feels like I've kind of always been aware of you, Luis. You are an extraordinary content creator, particularly for data science content and also for math content, which overlaps a lot with... I don't know if you've seen, but that's mostly what I post on YouTube is math for machine learning content, which some people out there listening to this are probably really annoyed with me because I haven't published a new video from that series on YouTube or the corresponding Udemy course for two years now. But I do have a plan for fixing that and it's actually already in motion. And so listeners who are annoyed with me, I think you'll be less annoyed hopefully soon.

Luis Serrano: 00:04:04 Perfect.

Jon Krohn: 00:04:05 So you are a very popular math and AI instructor. Much more popular than my channel. Your Serrano Academy has 145,000 subscribers at the time of recording. Probably by the time this episode is published, you'll be over 150. And who knows, by the time people are hearing this, it could be hundreds of thousands or a million subscribers. And yes, super well-produced channel and it does such a great job of surveying lots of math and machine learning topics. I suspect in some ways, maybe getting started with this YouTube channel and you can fill us in, is related to your Grokking Machine Learning book, which was published by Manning and very popular book as well. This idea of grokking is kind of like... Maybe you should explain grokking. I'm sure you can do it better than me.

Luis Serrano: 00:04:56 Yeah, it's interesting. I didn't know what the word grokking meant when they suggested that I write the book, but there's a Grokking series. I think it means to understand something really well, like to really break it down, which makes sense because that's what I enjoy

doing. I enjoy taking these concepts, really, really breaking them down to the basics.

Jon Krohn: 00:05:17 I think it's kind this idea of... I think it's kind of hacking it together from the parts, so you kind of-

Luis Serrano: 00:05:24 Exactly.

Jon Krohn: 00:05:25 ... by practicing it, you understand it. And so I think the inverse kind of approach to learning would be doing everything theoretically.

Luis Serrano: 00:05:36 Exactly.

Jon Krohn: 00:05:37 Learning all the theory with grokking, your kind of an application forward way of learning, which is certainly one that I agree with as well. And in a space like ours, it makes so much sense because you can have so much fun, whether it's linear algebra, calculus, probability theory, computer science, machine learning itself, statistics topics, any of these things, if you throw together a Jupyter Notebook and have people being able to execute the code and be able to change things and understand how things work, I find it's so much easier for people to understand than if you're just working with equation set.

Luis Serrano: 00:06:10 I agree completely. Equations always kind confuse me. I feel like sometimes people teach and they just teach the abstract concept and they're just steps away from a really concrete example that can make people really understand and appreciate it because we appreciate things when we know them, when they relate to something we already understand well like the real world.

Jon Krohn: 00:06:33 And so you've actually claimed that AI and math are easy, but they are made unnecessarily hard by many layers of abstraction and your goal is to pull back the curtain. And so in the words of our researcher, Serg Masis-

Luis Serrano: 00:06:51 Those are my words I remember.

Jon Krohn: 00:06:53 Yeah. And then Serg Masis who does our research said, "Not to sound conspiratorial, but why is there a curtain and who put it there?"

Luis Serrano: 00:07:03 Yeah, that puzzles me and it happens in every field that I've studied in the past. There seems to be a barrier and those who teach it, whether it's consciously or subconsciously, protect it from the rest by speaking in very, very high level. Because I think if you're... I don't know. Maybe if you're a researcher or a practitioner, at some point you have to do things at high level, but you never learn them that way. You have to always come back to the basics. And so I have the problem that I cannot think high level. When I think high level, I have to bring it down to the basics. So when I teach, I always, always come back to the basics.

Jon Krohn: 00:07:45 Yeah, it makes a lot of sense. I suspect that part of what makes language so powerful is that it allows you to group together more ideas or more abstract ideas into terms that could potentially be short to say. And so as you end up being more and more in some area, you end up more and more having this language that makes no sense to people from outside because it allows you to speak more efficiently while you're inside of it.

Luis Serrano: 00:08:19 But I try to bring it down... So I always keep music as my example of things done well. We could go to a concert, classical music concert, and Mozart could be there and a baby could be there and you play the same music to both of them. And Mozart would understand it in higher level and a beginner music student will understand on a level, and a person who knows nothing would appreciate it at a different level. But at the end of the day, there's some music that encompasses everything. And I think that with math, and by math I mean also with STEM, with machine

learning and physics and everything. You can always play the music. You can always tell the story in a fun conceptual way that a beginner would appreciate and an advance like an expert would also appreciate in a different level.

Jon Krohn: 00:09:15 That is a really cool analogy. I am going to be saying that. I'm going to continue to say that.

Luis Serrano: 00:09:21 Anytime.

Jon Krohn: 00:09:22 You actually yourself, you had a journey from disliking math due to the way that it was taught conventionally and you overcame that. Somehow you were able to become kind of like a Mozart. I realized that's probably a very... I mean that's very generous, but I think one, the bronze medal at the International Maths Olympiad, I mean that's amazing. I mean that's like you really obviously were able to get into math and understand it well. So what was the moment or experience that transformed your perception of math and made you from a disliker to an international Olympiad winner?

Luis Serrano: 00:10:04 Yeah, yeah, that was a very interesting story and a pivotal moment in my life because I was very, very bad at math. I remember when I was in grade A, I was the only one failing it in the class. I could not understand it.

Jon Krohn: 00:10:18 Really?

Luis Serrano: 00:10:18 I was very distracted. The formulas spoke nothing to me, meant nothing. But I was always kind of a nerd, just a nerd with bad grades. I liked puzzles, I liked to watch, I don't know, watch shows about space. I liked games that were sort of brainy games, but math somehow or most subjects never spoke to me and I was a terrible student. But one day I took a math contest because the incentive was if you take the math contest, you skip two hours of

class because you would sign up... And I was like, of course I'll do it and I'll just scribble for two hours.

00:10:59 And then when I went into this national math contest, I realized that it was not the math I knew; it was a bunch of puzzles. So I skipped everything that had formula, but I did the puzzles and I really enjoyed it. I stayed until the very end. People left and I stayed until the very end because I enjoyed it. And then a few days later, the teacher comes to me and goes, "Hey, congratulations, you did very well." And I was like, "I never do well in any..." "You are first." And I was like, "Oh, I've never been first in the school." And he said, "No, you're first in the country." And I was like, "Ooh, that's not supposed to happen." And it was very surprising, but it turns out that I did like math, I liked the puzzles and I didn't know that puzzles were math. The thinking, the logic was math. I just didn't know the language. It's like they're speaking to me something really cool, but in a language I don't speak. So I didn't catch it.

00:11:56 And so after that, I started participating in the Olympiads more seriously. And then when you get to the higher grades, you get to travel. If you make a certain number, six people get to travel to the International Math Olympiad. And it's the best experience I've had in my life and that made me want to be a mathematician. So then I went to university for math and grad school and everything. Later in life things switched to machine learning, but I always stayed in that sort of mind frame. And yeah, it was just the best experience in my life.

Jon Krohn: 00:12:24 That's such a cool story. I didn't actually know that. That's wild. What a cool experience. It's kind of like the premise of so many math movies out there, like Good Will Hunting.

Luis Serrano: 00:12:37 It's like a movie a little bit. Yeah. It definitely feels like an ugly duckling of math or something. And I try to do that now. I try to say, okay, maybe some people are not as lucky as I had this sort of chain of events that worked out, but maybe that doesn't happen for everybody. And many people just spend all their lives saying the mantra that everybody says, "Oh, I'm really bad at math." I want to tell them, no, you are really bad at abstraction and so am I to this day. You may be really, really good at math because math is not what you think it is. So I try to bring that moment that I had of wonder, of realizing that math was a completely different thing. And kind of hearing the music of math, I want to try to give it to others. And if I can give it to one person, then I'm fulfilled.

Jon Krohn: 00:13:29 Well, you have hundreds of thousands of people who are being fulfilled by that pedagogical approach on your YouTube channel, again, the Serrano Academy where people can check that out. What are some of the biggest challenges that you've faced in distilling complex topics like machine learning or maybe quantum AI, which is something... We're going to talk about quantum AI in more detail later, which isn't something we often have on the episode. So yeah, machine learning, quantum AI, some of these complex topics. What are challenges in making those beginner-friendly? I mean, a lot of your videos on YouTube it's about the friendly introduction to whatever topic.

Luis Serrano: 00:14:08 Some of these videos take me many years of thinking because the thing is, I actually do it for me. If there was no way to do YouTube channels, I would still do my own thing like that because I don't understand a topic if I don't put it the way I put it in my channels. So there's a lot of stuff that I still don't understand well and that I may have to work with it. So some things take me 10 years. I've been thinking about certain topics for a long, long time and I still don't have a good explanation for

them. And yeah, I feel like some things may be just harder. I have my own ways of... If it's a math, I try to bring it down to a couple numbers that I add and multiply. If it's physics, I try to bring it down to maybe like balls bouncing. If it's machine learning, I try to bring it down to something like a small image. I always try to bring it down...

00:15:06 Some things are hard to bring down to basics. Some things in deep learning, you kind of have to take by faith, like a big neural network, still not clear to people why it works so well, what happens in the layers. And so I look at so many examples. I scribble numbers, everything, and some things are just not yet fully understood. So I scratch my head a lot with things. So it's pretty much a challenge 99% of the time until something clicks and then I make the content.

Jon Krohn: 00:15:41 Are you stuck between optimizing latency and lowering your inference costs as you build your generative AI applications? Find out why more ML developers are moving toward AWS Trainium and Inferentia to build and serve their large language models. You can save up to 50% on training costs with AWS Trainium chips, and up to 40% on inference costs with AWS Inferentia chips. Trainium and Inferentia will help you achieve higher performance, lower costs, and be more sustainable. Check out the links in the show notes to learn more. All right, now back to our show.

00:16:19 Nice. Makes a lot of sense. We kind of already alluded to this, but your book Grokking Machine Learning, because it's application-forward approach, it also includes lots of Python-based exercises and mini projects. Do you think that that kind of hands-on learning is also critical to mastering machine learning concepts?

Luis Serrano: 00:16:38 Absolutely, yeah. When I was at Udacity, we studied a lot of our learners, a lot of data, and we noticed there was three main ways people understand things. One is sort of the theoretical, like the formulaic one. So whoever's a good student at school has that, I don't. The other one is like me, which is the visual. I like to see a little example, a cartoonish example with little animals walking around and then I get it. And then there's the builder. Many engineers are that it's just like I need to build something. So I understand the image recognition when I build a Python and a Jupyter Notebook with analyzing a big data set. So there's sort of the builder, the visual and the formulaic. Of course there's more, but those three big families, I try to cater to those three. And I think when you have a course that has those three aspects, it's successful.

Jon Krohn: 00:17:44 It's interesting that you say that. I'd never thought about it in those terms as having those three specific categories of learners. And yes, as you're saying, I'm sure there's more. Maybe there's some people out there that are auditory math learners or something. But just say the equation to me. I can't look at it. I don't know. I really need to smell an equation, you know?

Luis Serrano: 00:18:05 Well, that'd be cool.

Jon Krohn: 00:18:08 Actually, that's really interesting. We had on the show last year, we had a color number [inaudible 00:18:17]. Do you know about these people?

Luis Serrano: 00:18:18 I've heard they can, I don't know, smell numbers or see colors. Yeah.

Jon Krohn: 00:18:25 Yeah. So in episode number 719, we had Margot Gerritsen. She is Professor Emeritus at a university in California, and I'm forgetting which one off the top of my head. I think it might be Berkeley. And she always felt

very good... She was kind of the inverse of your experience with math growing up. She was always very, very good at math because she had a superpower, which was that she sees numbers, individual digits as colors, but then she can do math very, very quickly because they can merge together and form the new color. So she doesn't even have to do the math in her head. It's just like they're kind of happening automatically.

Luis Serrano: 00:19:07 Oh my God, that's fascinating. Wish I had that.

Jon Krohn: 00:19:13 Yeah, so they're kind of [inaudible 00:19:14].

Luis Serrano: 00:19:14 If I had that, I wouldn't tell anyone. They'll just [inaudible 00:19:17].

Jon Krohn: 00:19:20 Yeah, it's really interesting. She described that she can even have it for... While the effect is strongest for individual digits, any number could theoretically have a color. And so numbers that she sees a lot kind of get their own hue. Many digit numbers can have that. So yeah, very interesting thing there. But I get off on a tangent. The main thing is that I think I agree with you that these three main types of learner are formulaic, visual and engineer. And I think you and I probably share a very similar pedagogical approach because whether it is my YouTube videos or the courses that I create or the book that I wrote, Deep Learning Illustrated, it blends all three on purpose. And for me, I guess I'm a kind of learner that flourishes when I get all three. When I can have all three together, it helps me cement my kind of the engineering part of it, like being able to build it. It kind of helps me build an intuition around how it works.

00:20:26 The equations, the formulas make it easier I think to remember how things work because you get that abstraction, you get the conciseness in the end, there's less to remember. And the visuals, I don't know. For me,

that's always been... My book is called Deep Learning Illustrated, because I found that for me, like you're saying, a lot of the content I create, it's because I want it for me. And for me, visual was a very easy way to be able to express something as I was trying to learn it. And then it seemed to be also that way for students that I was teaching. And so this visual thing seems to be key for a lot of people.

Luis Serrano: 00:21:02 Yeah, I definitely need it. I feel like learning has always been difficult for me, and I feel like that gives me a strength in teaching because I am so slow at understanding certain things and I need them in such a clear picture, otherwise I don't get them that it makes me do the extra effort of drawing the picture. And then I've learned that I used to teach with no formulas, just the pictures. And people were like, "Please, please give me the formula," until I started putting them later. And the same thing with the building part, I started adding that. So I think it's good to have feedback from students a lot because each one of us has one aspect of learning and it's really when you put everybody together that you get all the perspective.

Jon Krohn: 00:21:45 For sure, for sure. So you've expressed previously in an interview that every human deserves a world-class education. And we're going to switch into this topic next. We're going to talk about LLMs a lot. But before we get into LLMs in detail, it's interesting to me, just before we started recording, we were talking about utopia and how some technologies could enable that, assuming we avoid dystopia. And education seems like one of those places where it's not even science fiction now, to be able to say with generative AI tools that we have available today, you can be allowing anyone in the world to be getting as good an education as anybody else and tailored specifically to them. So this is very interesting. It's just about scaling up and providing access.

00:22:35 But following on from the visual thing that we've just been talking about, I wonder how long it will be before we have generative AI systems that not only can they explain things to you like we can get with Google Gemini or GPT-4 or Claude 3 Opus. We get these amazing explanations in natural language. We can get code printed out, we can get equations printed out. But I wonder how far off we are from a system that in the same way that you and I like to be able to teach things, can visually create graphics on the fly, custom for the user based on the situation that they're in. I hadn't had that thought before. I don't know. Does this sound interesting to you?

Luis Serrano: 00:23:20 Yeah, absolutely. I hope soon. I mean, I think we're not that far. And I think the key is personalization. I mean, I think technology, even without the AI, the fact that we can bring education with videos and with labs and things like that, it already brings a barrier because it brings it to many places in the world and gives it to people that would not have the opportunity otherwise. But there's one more level with AI, which is the personalization level. Because learning, I mean, when I look at education, it's a field that has been steady for hundreds if not thousands of years. Everything changes, like medicine changes. Everything changes, but education stays the same. We still have the same methods. We sit down in the most passive possible way, look at someone doing stuff, which is kind of going to the gym and looking at someone exercising and thinking, we're going to get bigger like that. That makes no sense, but why does it make sense in education? And so we have many...

00:24:20 So education, if it's not interactive, it doesn't work if you're not actively doing something. And I've learned that with language models, it's enhanced my education a lot. Now, I talk a lot to it. I talk a lot to the model. Sometimes I want to understand a paper, I just put the paper there and ask it questions and it actually picks up your way of

thinking. I sort of say, "Okay, I want to work out this example. Help me out. Ask me questions about this." It's very interactive. I think we're very close... I think we already have that. We can learn using a language model pretty well already and it's getting better. So that's an aspect that I'm very optimistic about.

Jon Krohn: 00:25:07 It's unreal for me, the experience of being able to speak to these LLMs since the release of GPT-4.

Luis Serrano: 00:25:18 Yeah.

Jon Krohn: 00:25:19 That release last year, it made me exactly the kind of thing you're describing. Its ability to anticipate through, I guess great RLHF training. Its ability to anticipate what I am looking for with me seemingly providing not nearly enough context for it to be guessing that. And it's right on the money so often with really subtle things about what I'm looking for. And it was that moment of interacting with GPT-4 for about a year ago now, that it made me flip from being a skeptic about AGI to being optimistic about it happening very soon.

Luis Serrano: 00:25:58 Yeah, it blew my mind. When this language model started speaking so well, it really blew my mind. I knew we're going to get there one day, but when I started asking questions and it started answering, I thought, wow, I didn't expect it to be like that so quickly.

Jon Krohn: 00:26:16 And so speaking of which you had been until very recently, you were the head of developer relations at Cohere, which is one of the world's leading... They're right at the frontier. They're a frontier lab for large language models. But I haven't talked about them as much on air, and I suspect that to people in general, they're less aware of Cohere relative to OpenAI, of course, probably Anthropic. And so tell us about Cohere and its unique

niche and perhaps why despite them being a frontier lab, they're not as well known as some of these other groups.

Luis Serrano: 00:26:59 Cohere is a great place and they do great work. I think the main shift in Cohere is that they look at these models as tools. And I agree with that. I think there's always the search for AGI, but I think these models should be taken as tools before we see what they can do. A big decision should not be taken by a model. But if you look at them as tools that help me do my own things, then I think they're being well-used. So Cohere focuses on tools for enterprise, and so they have the... Because when we think of large-scale models, we think of the chatbot. That's where our mind goes because that's the cool one. But there's a lot of other stuff, for example, embeddings. Embeddings are really, really fundamental. A good embedding makes things so much easier for anybody. And the embeddings are Cohere top. Then you have things that become easy with embeddings, like classification or clustering. So they have all those products' summarization, et cetera.

00:28:07 And then if you're doing tools for enterprise, then you need them to be very exact. So then you focus on things like RAG, Retrieval Augmented Generation, which is basically a model hallucinates a lot. They're meant to hallucinate, they're meant to talk. They're not storing facts. ChatGPT doesn't store any facts. It's not a database of facts, it just talks. I like to say that actually, it's not that it always says is correct, and sometimes it hallucinates. It always hallucinates, and sometimes it accidentally hallucinates the truth because it doesn't store stuff. And so you can use RAG for that, which is search. You make the model, search the answer, whether it's on Google or on Wikipedia or in your own database, and then bring it back and answer from there. So it's much more exact. So search is optimized with things like

Rerank. There's a lot of tools. So as a tool set, I was very impressed with the toolkit I acquired.

Jon Krohn: 00:29:13 Ready to master some of the most powerful machine learning tools used in business and in industry? Kirill and Hadelin who have taught millions of students worldwide, bring you their newest course, Machine Learning Level 2. Packed with over six hours of content and hands-on exercises, this course will transform you into an expert in the ultra-popular gradient boosting models, XGBoost, LightGBM, and CatBoost. Tackle real-world challenges and gain expertise in ensemble methods, decision trees, and advanced techniques for solving complex regression and classification problems. Available exclusively at superdatascience.com, this course is your key to advancing your machine learning career. Enroll now at superdatascience.com/level2.

00:29:55 Yeah, very well said. And you've actually partnered with Andrew Ng's deeplearning.ai to deliver a course on semantic search and large language models that is based on these embeddings and how important they are. I end up talking about them a fair bit on air because that's something that at my company and Nebula that I co-founded, embeddings are the key to our intellectual property as well. So that course on semantic search and LLMs that you did with your colleague, Jay Alammar-

Luis Serrano: 00:30:27 [inaudible 00:30:28].

Jon Krohn: 00:30:30 Yeah, as well. So let's talk a bit about semantic search more.

Luis Serrano: 00:30:34 Sure.

Jon Krohn: 00:30:35 Let's dive into embeddings more. So how would you describe embeddings and say how semantic search differs from a traditional keyword search?

Luis Serrano: 00:30:46 Yeah, yeah, yeah, great question. So to me, embeddings are the most important object in LLMs. And I would go as far as saying in many fields in machine learning is the most important object because it's really where the rubber meets the road. It's where we translate our language to computer language. Computer language is numbers and only numbers. And so if we're working with embeddings, we need an image embedding that turns into numbers. If we're working with text, we need a text embedding that turns text into numbers or sounds into numbers. Or anything we want, we need an embedding. And if that embedding is not super strong, you're not doing anything. So the better embeddings get, the better LLMs get. If somebody comes up with a better embedding tomorrow, believe me, all the models are better because it's just a better way to turn text into numbers. So some problems that were hard 10 years ago, like classification, now are much easier because embeddings are better.

00:31:36 What's an embedding? An embedding is just you're literally associating your text through your words to a bunch of numbers. So a vector, like a list of numbers. I like to see them graphically. I always bring everything down to a simplest example. I like to see sending words to pairs of numbers. What's a pair of numbers? It's a coordinate in a plane. And so I imagine all the words flying around and words that are similar get put close. If I have an apple and a pear, they get sent to similar pairs of numbers, which is a similar location.

00:32:13 And I also like to think as each coordinate as a description of the word. So in a very toyish example, I could think of the first one as the size and the second one as the flavor or something like that. But if you have thousands of numbers, then you're pretty much describing your sentence or your word or your piece of text in a very, very, very detailed way using a lot of numbers. And maybe these numbers mean something to

us, maybe they don't, maybe they mean something to the computer, but it's really that.

00:32:47 And then one of the things that embeddings are really useful for is semantic search. When we search, for example, if I am going to search for... I always use this example. Let's say you search for a visa to travel from Brazil to USA, and I want to find the article that has as many words as possible in common with those. And I find one that says visa to travel from USA to Brazil. It's the complete opposite. It doesn't help me at all because it's different, it's flying in the wrong direction, but it matched all the words. So that's keyword search. If I find that the documents that match all my words or as many as my words, it's decent, but it doesn't get you there because I could reorganize the words and change the meaning of sentences. That's when we use semantic search.

00:33:43 So in these embeddings, you're locating sentences in a plane or in space. And if I have all my documents and they're flying around and I send the question to my embedding, then it's very likely that the answer is going to be close. So embeddings are just a way of mapping, of putting all the text flying in space and then it's easy too to search in that space instead and find an answer. So we've seen a lot of improvement with semantic search versus keyword search. And then when you throw in things like Rerank for example, it's very useful because see, the closest sentence to something is not the answer. If you ask me a question, you ask me what color's an apple? And I answer to you, what color is an orange? Then you say, well, what the hell? And I say, well, I just gave you the closest sentence, but it's not the answer. So Rerank helps you actually locate the answer with an extra training. But yeah, semantics is in a great place. I've seen some great results. So yeah, I've been happy to see that.

Jon Krohn: 00:34:52 Very nice. So I'll try to repeat back to you some of what you said. So embeddings, they allow us to have a language of computing where we speak in say, natural language or math or code. So we represent things in that kind of form. When we want a machine to be able to do something helpful for us with that language, we need to convert it into some kind of numeric format. And I love that idea of explaining the kind of simplest embedding would probably have just two numbers, two float values that you represent some word or some sentence or some document by. And then you can very easily imagine that in a two-dimensional space.

00:35:41 If you all of a sudden have three dimensions, three numbers that you're using to describe the location of that word or document or piece of code or whatever it is, however you are embedding that, if you go from two to three, then it's imagining it in the visual world that we encompass in a 3D space. And then it becomes hard for us to visualize in four or more dimensions, but for a machine to be able to represent this is trivial. The math is all the same, the linear algebra is the same. And so it's very common with organizations like Cohere to have... Your embeddings have thousands of dimensions. And as you're saying, that then allows a huge amount of granularity for semantic search.

00:36:24 So we can embed any natural language code, math, whatever it is that we want to represent. Its meaning can be understood well in this high dimensional numeric space that the machine has and that allows it to fulfill lots of different kinds of things. So you mentioned earlier clustering, but I think one of the most interesting applications that we see today with generative AI is things like being able to answer your questions. And so by using this semantic search, you can get into the right region based on the person's query, like you said, what color is an apple or something like that, that can bring you to

some region of that high dimensional... There's many thousands of dimensions of space where everything's about the color of fruits. But then once we get to that space, Rerank allows us to find a great answer to the question as opposed to just some phrase related to the question.

Luis Serrano: 00:37:24 Exactly. Yeah, very well put.

Jon Krohn: 00:37:26 Nice. Awesome. So that helps us understand embeddings and Cohere's business model, how they're helping out the enterprise by ensuring that they have great embeddings, facilitating all kinds of these enterprise capabilities. So what kind of experience... If I'm a user of an application that is powered by Cohere embeddings versus maybe some old or poorly trained embeddings, maybe some embeddings that I tried to make myself that weren't well-made. As a user, how can my user experience... Are you able to give some examples of how my user experience will be improved by better embeddings? I guess one we already talked about is question answering.

Luis Serrano: 00:38:09 Yeah. I mean, everything becomes easier. The analogy I used to... Imagine if you had a great book, but a really awful translation. So you can only enjoy it so much. You can only understand so much. And so a good embedding just translates your data really well into numbers. For example, 10 years ago, let's say we had embeddings that were so-so, or even no embeddings, you can just use every coordinate as a different word and you would have a huge space. But anyway, the fact is, let's say we have poor embeddings 10 years ago, and you want to do a classification model that tells you, I don't know, if emails are spam or not, but the region between the spam emails and the non-spam emails is very, very complex. It's very curvy. In order to separate them well, you have to use a really big neural network to come up with a really

complicated boundary. There may be problems or fitting, et cetera.

00:39:09 Let's say you have an amazing embedding that picks up things so well, that puts all the spam emails on one side and all the non-spam emails on the other side, and it's really a line that cuts on our plane. I'm exaggerating, but a good embedding makes these problems easier. I've done classification problems with three or four examples, so you can do much better now as opposed to before that I needed thousands of examples, a huge model. So that's just classification, but you know, clustering is the same thing. If I have a really good embedding, the clustering becomes much easier because things are located in the right place. So the embedding is the most fundamental thing, and many companies just want the embedding. They're like, "Okay, we have a bunch of machine learning stuff. I just want your embedding and I can do wonders with it."

Jon Krohn: 00:40:01 Nice. Makes perfect sense. That was a really great analogy. Another analogy that you've used to explain some Cohere technology is the idea that it's like a power plant for large language models. So it makes LLMs accessible to end users, including enterprise users without having to worry about the backend complexities. Do you want to talk about that a bit more?

Luis Serrano: 00:40:21 Yeah, yeah. I think that's a breakthrough because a few years ago we would sort of train our own models. Now I tell people, you want to train your own language model that's kind trying to make your own electricity with a lemon or something to power something at home. Just plug it to the wall because somebody's taking care of that electricity, makes sure that it works. And for things like responsible AI it's very important to look at those layers. We like to have this analogy where the power plant is Cohere, Anthropic, OpenAI, they create the electricity.

Then you have the people who create the apps. So in the analogy with electricity, I would say that these are the people who build appliances. Let's say they build an oven. And then we have the users, which are the ones who use the oven. So I use the oven at home, or a chef uses an oven or something, and those are the people who use the apps.

00:41:12 So you have the power plant is the LLM, the one who builds the LLMs, makes sure it's good and it's responsible and it does what it should do. Then there's the level of the people who build the apps using the LLM or build appliances using electricity. And then on the other level, there's the users and all of them have to use things responsibly too in the electricity side to not make a fire and in the LLM side to not have some big problem. We know there's many.

Jon Krohn: 00:41:47 Nice. So you can trust Cohere or these other big frontier labs to have a great LLM for you. And you can typically, I think, choose between multiple different options. So you might have a very, very large LLM in situations where accuracy is very important and you're less concerned about cost, but then you might choose a smaller LLM where accuracy is a bit less important, but being able to provide real-time answers in a cost-effective manner is more important. And so yeah, Cohere has these off-the-shelf available for you, and you can use them to create embeddings or to generate text, and you don't have to worry about making sure that there's enough GPUs out there for you. You don't need to worry about spinning up some cluster in Kubernetes and have it running. It's just all there happening.

Luis Serrano: 00:42:37 Yep.

Jon Krohn: 00:42:37 Very cool. So you mentioned in a recent Cohere presentation actually that every 10 or 15 years there's a

huge advance in technology and you highlighted generative chat as that current big thing. So I already know from our pre-conversation that you do like thinking about the future. So can you speculate on what the next big technological breakthrough might be after generative chat?

Luis Serrano: 00:43:06 Technological breakthrough.

Jon Krohn: 00:43:09 That's a tricky one. It's a tricky one.

Luis Serrano: 00:43:09 That's a tricky one.

Jon Krohn: 00:43:10 You need a crystal ball.

Luis Serrano: 00:43:13 I need that. Maybe that could be the breakthrough. A crystal ball to get me through. Well, I see some things in the horizon. For example, multimodality is the next thing I believe, because right now it's text in, text out. And humans we're not just text in, text out. We see images, we talk. I could know something that I don't know if I saw it in an image or I heard it. So that all the types of media coming in and also coming out, but then there's also doing things.

Jon Krohn: 00:43:48 Like the example that we were talking about earlier with being able to teach with visuals.

Luis Serrano: 00:43:54 Exactly. So multimodality is definitely next. And then also doing things because a lot of it is, I put the question, I get the answer and then I go do things. But you could just kind of plug it in into apps or into things and say, "Do this for me." So that's agents, agent use and tool use. So that's the next, and that's being developed right now. And I think I would say things with reinforcement learning with machines or something that I can... When you plug that in, we have some of that with autonomous cars and

all that. But I think a next level of plugging the AI to machines will-

Jon Krohn: 00:44:42 Bringing things into the real world. So this same kind of LLM technology that is now scaled up well in software that being more and more embedded in hardware, allowing for real world changes to happen, robotics, autonomous vehicles and so on. Yeah.

Luis Serrano: 00:44:57 Yeah. I wouldn't say those are the next one in 10 or 15 years, big breakthrough, but they're definitely the subsequent ones, the ones that follow this big breakthrough.

Jon Krohn: 00:45:12 Yeah, it makes perfect sense. I agree with you a hundred percent. And I also think that... I'd be interested to hear your thoughts on this as well, but I think that these major transformative technologies are emerging at an accelerating pace.

Luis Serrano: 00:45:26 Very much. Now things change so often that I don't even know what we'll be working on in five years because it probably doesn't exist. And I think that gets something good out of us, which is it makes us be ready for change. It makes us be ready for learning out of things. I think previous generations studied one thing, made one decision at age 20 and stayed with that, worked with that forever. Two to three years in advance is the most we can get ready for.

Jon Krohn: 00:45:58 Yeah, hopefully that means that folks like us who do a lot of educating, that will still be useful, but we'll see.

Luis Serrano: 00:46:05 It will keep us busy.

Jon Krohn: 00:46:08 It'll keep us busy, at least in the short term.

Luis Serrano: 00:46:09 I feel like I have to make videos faster because the technology changes. I'm like, "Okay, I need to get this out because it'll become irrelevant."

Jon Krohn: 00:46:17 That's part of how I got into doing so much content on math for machine learning, because I was thinking to myself, I mean, one, it had to be something that obviously interests me a lot and something where there's still a lot more for me to learn, and there's still so much more in math. But the other thing was I was like, if I teach linear algebra, partial derivative calculus-

Luis Serrano: 00:46:38 That doesn't change.

Jon Krohn: 00:46:38 ... that's probably not going to change. That would be crazy. If in 10 years we're not using linear algebra or calculus in machine learning, that would be mind-blowing.

Luis Serrano: 00:46:48 That would blow my mind. Yeah. That would definitely. No, that's why I did the Coursera course on math for machine learning. It took a while to build, but it's stuff that will stay forever. And my book is also in sort of the most basic machine learning stuff that sticks around. Linear regression will always be linear regression, base improbability will always be the same. So I think, yeah. And then for anything that's shorter, and I've seen that with Deep Learning AI, for anything that's cutting edge, it's a short course. You cannot build a specialization on something as big as LLMs or something, because by the time you finish it, it's going to change completely. So you have to make bit-sized content.

Jon Krohn: 00:47:30 Yeah, it makes sense. Although it's interesting, there are some aspects that end up still being still having a lifespan of say, at least decades. So there's interestingly like yes, while specific ways of implementing LLMs, this is going to get old very quickly, not only is the linear algebra and the

partial derivative calculus associated with the operations and the training of those models the same, but so too, the deep learning, the neural network theory that all these LLMs are based on, transformer architectures are based on, striped hyena. Whatever thing is going to replace the transformer, it's still probably going to use neural networks and deep learning, which have been around since the 1950s. And so it is interesting how there are some aspects of it of what we're doing that is relatively consistent. Linear regression was another great example there, Bayesian statistics. There's some things that you can be relatively confident about, lasting on a decade scale, even if they won't be around forever.

Luis Serrano: 00:48:38 That is true. Definitely the concept of a neural network, it'll probably be around for a while. And when you think of transformers, they're just a bigger neural network with some extra stuff like attention or things like that. People figure out how to make them work better, but if you understand... A beginner ML course should always have a neural network built from scratch, and that will take a while to change. It always has to be using some linear algebra, some calculus. I'll be surprised when those stop being used. So the basics are consistent.

Jon Krohn: 00:49:16 Now, thinking of some technology that could potentially have a huge impact on how everything in machine learning happens, something that is around today though very much in its infancy is quantum machine learning. And you have a fair bit of exposure to this because prior to Cohere, you were at Zapata Computing where you were a quantum AI research scientist.

Luis Serrano: 00:49:42 Yes.

Jon Krohn: 00:49:42 So tell us about quantum AI. I know that you actually already listened to Amira Abbas's episode. So Amira Abbas, incredible communicator on quantum machine

learning. We had her in episode number 721. I was pursuing her for literally years to get her on the show. And so that whole episode is an introduction to quantum machine learning, but let's do a mini one here, especially because you might have different perspectives from her. So Amira, it was interesting, her overall sentiment was that at this time, with the number of qubits that quantum computers today can handle and the number of qubits that they're going to be able to handle over the next several years, she didn't think that there were practical, real-world machine learning applications that we could handle with quantum computing. But I'd love to hear your thoughts.

Luis Serrano: 00:50:39 Yeah, no, it's great you had Amira. Amira is really a top scientist in the field, a great communicator as well, which is rare to find. So I've asked her a lot of questions in the past few years, and she's very, very knowledgeable. Yeah, I think I agree with that. Definitely right now you can't do very much. So our research was based on two things basically. Question one was, is there anything we can do that is meaningful with what we have right now, whether it's a simulation or a small quantum computer? And two, what can we do when the computers come? So a lot of the research is what can we do when we have computers this big, the Shor's algorithm, like the Bayes' cryptography, things like that. But our work wasn't what can you do in machine learning? And so Amira, for example, has done a lot of neural networks, like supervised machine learning. They built quantum neural networks for supervised learning, and they've got some great results. We were focusing on unsupervised. And the reason is I kind of look at...

00:51:58 In my head, a quantum computer is a generator whether you want it or not, and a classical computer is more of a supervised learning machine. Because let's think of the most basic generative machine learning problem. Let's

think of the most basic generator in existence. It would be flipping a coin. Generating a random bit is the most basic generative machine learning problem. And a classical computer cannot do that. It cannot generate a random bit, it cannot flip a coin. It can pretend to flip a coin and give you pseudo-random numbers, but they're not random at all. They're predetermined because everything a classical computer does is predetermined. It's deterministic, which is good for supervised learning because you need a lot of operations to work.

00:52:52 But in generative machine learning, if you look at classical generative algorithms, which are very good, like GANs or RBMs or even transformers, LLMs, they always have a random part. There's always a part where you pick a random number and you plug it to the network, or you do something like that, you pick out of a distribution, and those things aren't perfect for classical computers. So we started noticing that. We started experimenting with quantum circuit and just measuring it because the flipping coin problem, which cannot be done with a classical computer, is a one-on-one for quantum computers. The first thing you learn in quantum computing is for a circuit, how to measure one qubit. And measuring one qubit is exactly flipping a coin, and there's true randomness. So we were just making that more complicated, having bigger quantum circuits, trying to train them in the exact same way that you train a neural network by flipping the angles a little bit, by changing the parameters a little bit in order to generate stuff that comes from a data set.

00:54:08 And we were getting some results that showed that they were a little better. We were getting some results that showed that they kind of looked farther. You can use things like quantum entanglement to make them look in places where a classical computer just kind of stays within its lane. It kind of looks things that are close, but

the quantum computer started looking far away and finding patterns that the classical one didn't. Of course, these are all empiric results, and we were comparing a very small quantum, a very small classical algorithm, but we had some promising results, and I found it very exciting.

Jon Krohn: 00:54:39 That is very cool. So there's potentially opportunities in quantum machine learning where that true randomness can be leveraged to create some kind of efficiencies.

Luis Serrano: 00:54:52 That's what we thought, yeah.

Jon Krohn: 00:54:53 Nice. Nice. Nice. Given that you're from Columbia, you're a Spanish speaker.

Luis Serrano: 00:54:58 Yep.

Jon Krohn: 00:54:58 Zapata, that sounds a lot like shoe.

Luis Serrano: 00:55:01 It is, yeah. There's a funny story about the company. Zapata is shoe. Shoe is zapato. So there's a male Zapata and a female... It's a last name. So it's used as a last name. And then big Mexican revolutionary was Emiliano Zapata. And the story how it's called Zapata is because Alán Aspuru, he's a professor at University of Toronto. He was at Harvard when they founded the company. He's Mexican, and he wanted a name that matched the quantum revolution with the Mexican Revolution. So he called it Zapata Computing because of Emiliano Zapata.

Jon Krohn: 00:55:43 Nice. Very cool. That's great to know that etymology. I love etymologies. That also helps me understand things because once you learn... I don't really speak Spanish, but somehow I know that zapato means shoe, and then it allows you to make more connections in my neural network because I also don't have any facts. I'm just a bunch of weights.

Luis Serrano: 00:56:05 I'm like that. I love etymology. I look at words and I try to see where they come from, especially names where they come from. And you can learn a lot from there.

Jon Krohn: 00:56:13 Nice. Yeah, for sure. Serrano, that's like Serrano ham. What does that mean?

Luis Serrano: 00:56:17 Yeah, well, there's also Serrano pepper. So Serrano comes from a mountain. It literally means guy from the mountains. So sierra is a saw, like a saw to saw wood, but then they have the little spikes. And so when you look at a mountain range, it looks like a sierra, like a saw. So when you use the word sierra for mountain range, like Sierra Nevada or Sierra Leona or Sierra... And so the person who lives in the mountains are the serranos. So I'm like guy from the mountain.

Jon Krohn: 00:56:47 That's very cool. And there's an easy way to remember that because a serrated knife is exactly what you just described.

Luis Serrano: 00:56:54 There you go. I never made a-

Jon Krohn: 00:56:56 So I'm just saying the same etymology there.

Luis Serrano: 00:56:58 Which is in English too.

Jon Krohn: 00:56:59 Very cool. Well, so we do have an audience question for you. You and I put together having you on the show very rapidly, which I really appreciate you doing, by the way, being so available for us to film quickly. But that meant that my post about you being a guest on the show, we only had one day of time. Typically, I try to have at least a week of people to be able to notice the post and ask questions. Despite it having been less than 24 hours from when I made the post, you have 150 reactions, 10,000 impressions. So yeah, lots of people. Very excited about your episode coming out.

Luis Serrano: 00:57:37 Yeah.

Jon Krohn: 00:57:38 Most of the comments are just about how people can't wait to listen and wishing you luck. Hopefully this isn't a very hardball interview. I don't know how much luck you need.

Luis Serrano: 00:57:48 [inaudible 00:57:48].

Jon Krohn: 00:57:51 But yeah, so cool to see all that. We only actually in that ended up having one question asker, who is a frequent question asker on the show and a regular listener, also a winner of many books in our recent book contests. So Jonathan Bowne is our question asker, and Jonathan, he asks, how long does it take you to develop a course and what advice would you give to people that want to develop their own?

Luis Serrano: 00:58:18 Oh, that's a great question. It depends. I mean , the Coursera specialization took us a while, like two years, more than two years, and with a team, so that can take a while. When we were at Udacity, we were getting them out much quicker because we had a team where everybody had their own section. So I would say a few months would be to build a machine learning course or something. So it really depends on how long the course. The course we taught with Jay Alammar and Hadelin, that was quick. That was maybe two, three weeks. So it depends on what team you have and what resources you have. The stuff that I do myself, it takes me a while. I take a long time thinking about the stuff and I'm normally having a few wheels spinning in my head about different videos. Once I have the idea, I would say maybe a month of building the slides and recording and all that stuff. So yeah, it basically varies. It's a lot of variance.

00:59:26 Advice for what you like. If you like teaching, just start putting your content out there. I mean, that's how I

started kind of randomly. I put one video and people watched it, and so I kept going. And that's how people start. Just start putting your videos out there or your blog posts. Find what you like to do the most. What you do best is what you enjoy the most. Some people enjoy writing, some people enjoy making videos. Do it that way in whatever format you enjoy better. Some people do it just talking like this. Some people do animations. Just do what you like. And start putting them out there and gathering an audience because the most important thing is to gather an audience. Once you have an audience, then you can build a course with a company or something. But I would say just put your content out there and bless the world with your content because everybody has their way of seeing things and so everybody has a new perspective. So just don't deny, give that to the world. That's the first step.

Jon Krohn: 01:00:31 There you go. Very nice. Thank you, man from the mountains, for that great question answer.

Luis Serrano: 01:00:36 [inaudible 01:00:38].

Jon Krohn: 01:00:41 Nice. Well, thank you so much for taking so much time with us today. Really great, interesting interview. And before I let my guests go, they must provide a book recommendation. Obviously we already know about Grokking Machine Learning, so maybe if you have something beyond your own book.

Luis Serrano: 01:00:58 Okay, great. Yeah. Oh, a lot of great books. I'm going to give you two that are in my head right now that actually are affecting a lot of how I think and how I see the world. One is Sapiens. I think Sapiens is really good.

Jon Krohn: 01:01:14 Oh, I love that book so much.

Luis Serrano: 01:01:16 It's giving me ideas of what's next. Talks about the Agricultural Revolution, how that changed us, how all these things... And I see the technological revolution as one more step there. So this is a book that I sometimes read again because it gives some good perspective of humanity.

01:01:33 The other one that actually influences my teaching a lot is Pedagogy of the Oppressed by Paulo Freire. I don't know if you know it. That one really, the philosophy of teaching that it has really, really influences mine. I'll give you a small example. For example, one thing I do when I teach is I don't quiz people afterwards. I quiz people beforehand because you don't want to turn people into instruction followers, you want to turn them into creative thinkers. If I give you instructions and then ask you a quiz, I turn you into an instruction follower. Whereas if I first ask and then you have all the creativity and then I tell you my way, but yours could be different, better, or the same. I always did that just unconsciously. And then when I read this book, he talks about that. Education should not turn people into instruction followers, it should turn them into creative thinkers. So I found that I just found myself in that book. So that's Pedagogy of the Oppressed by Paulo Freire.

Jon Krohn: 01:02:33 Oh, I love that. Really, I would much prefer to be that kind of teacher, but I'm absolutely the person who asks questions at the end. So there you go. Changed my life today and I guess my students' lives. Fantastic. All right, Luis, so how should people follow you after this episode? Obviously your YouTube channel is a great place to follow your content, and then it sounds like LinkedIn is probably your primary social media.

Luis Serrano: 01:03:02 My primary is LinkedIn, yeah. My name, Luis Serrano. Twitter, Serrano Academy is a little smaller, but you can follow me on Twitter as well. My page has everything. So

it's actually serrano.academy is the page, and so if you go there, you see videos, blogs, et cetera. So that's it. But my main interest right now is going into doing a lot of work in the YouTube channel, so that's where you see basically whatever I'm up to.

Jon Krohn: 01:03:28 Nice. We'll make sure to include links to all of those things in our show notes. Luis, thank you so much for taking the time. I look forward to hanging out with you in Toronto soon.

Luis Serrano: 01:03:37 Thank you, Jon. I had a great time. Been family focused for a while, so it's an honor to be here and thank you for having me.

Jon Krohn: 01:03:50 Wow, love that convo today. In today's episode, Luis filled us in on the three major categories of learners, formulaic, visual, and builder, and how catering to all three of them thoughtfully can make supposedly complex topics like math and AI friendly and approachable. He also talked about how Cohere focuses on embeddings, making their products powerful, enterprise-grade options for not just generative AI, but also semantic search, retrieval augmented generation and clustering. He talked about the most exciting emerging application areas for AI, namely multimodality, agents, and real world interaction through, for example, robots and autonomous vehicles. And he talked about how quantum computing's actual randomness relative to classical computing's mere simulations of randomness suggest promising quantum ML applications may develop as the number of qubits quantum computers can handle increases exponentially in the coming years.

01:04:45 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 Luis's

social media profiles, as well as my own at superdatascience.com/785.

01:04:57 Thanks to my colleagues at Nebula for supporting me while I create content like this Super Data Science episode for you. And thanks of course to Ivana, Mario, Natalie, Serg, Sylvia, Zara, and Kirill on the Super Data Science team for producing yet another fascinating episode for all of us today. For enabling that super-duper team to create this free podcast for you, we're so grateful to our sponsors. You can support the show by checking out our sponsors' links, which you can find in the show notes. And if you yourself are ever interested in sponsoring this podcast, you can get the details on how you can do that by heading to jonkrohn.com/podcast.

01:05:32 Otherwise, please share this episode with folks who might like to hear it, review the episode on your favorite podcasting app or on YouTube or wherever you listen to this podcast. Subscribe if you're not already a subscriber. And most importantly of all, just keep on listening. So grateful to have you listening and I hope I can continue to make episodes you love for years and years. 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.