SDS PODCAST

EPISODE 871:

NOSQL IS IDEAL FOR AI APPLICATIONS, WITH MONGODB'S RICHMOND ALAKE

Jon Krohn: 00:00:07 This is episode number 871 with Richmond Alake, staff developer advocate at MongoDB. Today's episode is brought to you by the Dell AI Factory with NVIDIA and by ODSC, the Open Data Science Conference.

00:00:18 Welcome to the SuperDataScience Podcast, the most listened to podcast in the data science industry. Each week we bring you fun and inspiring people and ideas, exploring the cutting edge of machine learning, AI, and related technologies that are transforming our world for the better. I'm your host, Jon Krohn, thanks for joining me today. And now let's make the complex simple.

00:00:52 Welcome back to the SuperDataScience Podcast. Today we've got the tremendously gifted writer, speaker and machine learning developer, Richmond Alake on the show. Richmond is a staff developer advocate for AI and machine learning at MongoDB, a huge publicly listed database company with over 5,000 employees and over $1,000,000,000 dollars in annual revenue. With Andrew Ng, he co-developed the DeepLearning.AI course Prompt compression and query optimization that has been undertaken by over 13,000 people since its release last year. He's delivered courses also on platforms such as Coursera, Datacamp, and O'Reilly. He's authored over 200 technical articles with over a million total views, including as a writer for NVIDIA. He previously held roles as an ML architect, computer vision engineer and web developer at a range of London-based companies. He holds a master's in computer vision, machine learning and robotics from the University of Surrey, in the UK.

00:01:46 Today's episode will appeal most to hands-on practitioners like data scientists, ML engineers, and software developers. But Richmond does a stellar job of introducing technical concepts, so any interested listener should enjoy the episode. In today's episode, Richmond details how NoSQL databases like MongoDB differ from

relational SQL style databases. Why NoSQL databases like MongoDB are particularly well suited for developing modern AI applications, including agentic AI applications. How Mongo incorporates a native vector database making it particularly well suited to RAG, retrieval augmented generation. Why 2025 marks the beginning of the multi-era that will transform how we build AI systems, and Richmond provides his powerful framework for building winning AI strategies in today's hypercompetitive landscape. All right, you ready for this fun and informative episode? Let's go.

00:02:41 Richmond, welcome to the SuperDataScience Podcast, we're filming live, in person, in London, in your beautiful MongoDB office. Thank you for having me.

Richmond Alake: 00:02:50 Thank you for coming here. It's a pleasure to have you in London. It's nice to do this in person.

Jon Krohn: 00:02:54 It's been nice. I haven't been in London long. I did a flight just yesterday and so obviously with the jet lag from New York, I couldn't fall asleep until 4:00 AM. And then the craziest thing was where I'm staying, I already told Richmond this, but then they were doing fire alarm testing in the morning, that started at 10:00 AM. which I get is fine for most people, but when you've just come from New York and you're going to bed at 4:00 AM, I was like, "Oh my god, this is crazy." And it was in my room, the fire alarm in my room, it starts going off at 10:00 AM. It was like, "Yeah, such a..." So anyway, but it's been very nice... Since I've gotten to the MongoDB office, it's been a very nice day.

Richmond Alake: 00:03:35 Nice.

Jon Krohn: 00:03:36 A beautiful space. Here you can see a little bit of the background, maybe even a little bit of the Shard, which is a iconic London building, in the background. So

Richmond, you previously were on the show in episode number 685, that came out in June, two years ago, June 2023, and in that we focused on building real-time machine learning applications. We are going to talk a little bit in the episode about building machine learning applications, but all of that now is going to be colored by the perspective that you've developed over the past year and a bit, as staff developer advocate for AIML at MongoDB. So tell us about how that came about. What is a staff developer advocate role and how'd you get into it?

Richmond Alake: 00:04:23 Yeah, that's a very good question and again, massive pleasure to be on the show and it feels like a lot has happened. It's only been two years right since the last time I was here, but a lot has happened in that time. So obviously the major thing is I'm at MongoDB as a staff developer advocate, and when I came on your show the last time, I think I spoke about some of the external writing that I do, outside of my day job. So back then I was a machine learning architect at a consultancy company and I did a lot of... After my 9:00 to 5:00, I did a lot of writing and I did a lot of teaching as well.

00:05:03 So it turns out it's a full-time job, that's a full-time job. So it's called developer advocacy, and people like me reside within the developer relations of certain organizations. So what I used to do for fun is actually now my full-time job. So I get to teach our customers both externally and internally on the latest and greatest on AI. And I do that in various forms. So it could be writing content, it could be going on podcasts, it could be speaking directly to our customers as well, over at MongoDB. And we have a lot of customers that are building very interesting AI application. Now the staff level of it is just I'm of experience, and the expectations that's expected of you as well.

Jon Krohn:	00:05:54	That was my next question. So being a staff developer advocate, it's not like that's an internal facing role, it's like when somebody gets a staff machine learning engineer role, it's a level of seniority?
Richmond Alake:	00:06:05	Yes, exactly.
Jon Krohn:	00:06:06	Cool. So let's talk about MongoDB, specifically. So I have built startups that leverage MongoDB in the background for a decade now. And so I can speak about it really vaguely, which is this... The key difference for people who are already familiar with SQL, SQL is a structured table, so it looks like when you see an Excel spreadsheet or a Google sheet, where you have specific columns with names and a certain number of rows, that is similar to how SQL tables are structured. A NoSQL database, like Mongo, doesn't have that structure. So fill us in.
Richmond Alake:	00:06:53	It has a structure, but it's-
Jon Krohn:	00:06:54	It has a structure.
Richmond Alake:	00:06:55	... not the tabular structure that you get in relational database. And you point on one thing, which is, MongoDB has been within the space for maybe over a decade now, and we've been supporting startups and large enterprise organization with just application development.
Jon Krohn:	00:07:13	I'd also say really quickly is that Mongo is the NoSQL company. When I think in my own vector representation in my brain, a meaning-
Richmond Alake:	00:07:24	Love that.
Jon Krohn:	00:07:25	... representation, NoSQL and Mongo, they occupy the same location, in my internal vector space.

Richmond Alake: 00:07:30 That's very good. And I'm hoping after this conversation you're going to see MongoDB is the database for your AI application and you can make that... And that takes up another, I guess, space within the latent space of your brain.

Jon Krohn: 00:07:45 Nice. Yeah. And so sorry, I interrupted you as you were about to explain how NoSQL database differs from a relational database like SQL.

Richmond Alake: 00:07:53 Yeah, so relational database, will use data stored in that tabular format, so you have rows and columns. And relational database has been... They've been within the whole space since the '80s, so they work well for the use cases and some of the applications that are built. Well, MongoDB is a database that is based on the document data model, which essentially you're storing data in key value pairs. And one thing that resonated with a lot of startup developers back in the early days of MongoDB, was how easy it was to get started in MongoDB. And how when you use the MongoDB database, it reflects how the developers think, which is in that JSON-like structure.

00:08:38 And for me, it was a very, very massive moment in my career, and I don't think I touched on it the last time I was on this podcast, but I was a full stack web developer about maybe five, maybe seven years ago, but I used to hate programming. But I fell in love with programming because MongoDB made it easier for me to think about the stack. So I was a front-end developer, but when I understood... When I took MongoDB on the data layer and I saw that I can represent the data in the application layer, in the same way within the data layer, which is that JSON structure I was talking about, that document data model, everything clicked, it changed my career, and I became a MEAN stack developer, a full stack developer, using the MEAN stack to build web application.

Jon Krohn: 00:09:34 And so MEAN, M-E-A-N, an acronym where each of those letters represents a different layer in the stack. And the M is-

Richmond Alake: 00:09:41 Exactly.

Jon Krohn: 00:09:42 ... MongoDB, right?

Richmond Alake: 00:09:42 Exactly.

Jon Krohn: 00:09:43 It's the base of the stack.

Richmond Alake: 00:09:44 And the E was Express, and A was Angular, and N was Node.js, and it was actually Angular JS, then you have the Angular. So before Angular that we have now, there was another library called Angular JS, which was a whole... Much harder to work with.

Jon Krohn: 00:10:01 One of the things that I remember, as to why we went with Mongo databases, way back, yeah, this would be about 2014, we were making database decisions, so 11 years ago, and one of the reasons why we went with Mongo and this, NoSQL JSON-like structure key value pairs that you're describing, is that it allows you to have way more flexibility. So a given part of your application of your website, you might not know for sure, long term, what information you're going to have on that page or in that part of the application. And so MongoDB makes it really easy, instead of needing to plan in advance and be really rigid about how some data are structured or some part of your application is structured, you can just say, "Okay, we're going to have... Today, we're going to have this key, it's going to be called this, it's going to be populated with this kind of data." And if in the future you change that, that's okay.

Richmond Alake: 00:10:56 Yeah, and where... You hit the nail on the head, firstly. As in, you hit the nail... You said it very well. And we're

seeing the same thing within AI today. So the AI space is moving incredibly fast. I think that's an understatement. It felt like yesterday ChatGPT came out, and it was just two years ago when ChatGPT came out in November 2020, '22. But one thing that we do at MongoDB is we have that document data model, which allows you to have a flexible schema, which means that as your understanding of the requirement of your application evolves, your data layer can evolve with it. But at the same time, once you are stable, once you have a stable understanding of the data representation, that you want you're... The way data is represented within the application layer, you can actually lock your schema in MongoDB. So it gives you that fixed structure once you go into production, but you get the flexibility once you're experimenting, which is what a lot of people are doing with AI today. So they need that flexible data structure, they need what MongoDB brings.

Jon Krohn: 00:12:08 Makes perfect sense. Something that I'm kind of thinking off the top of my head, if you were building some kind of application that leveraged LLMs today, you might be experimenting with calling different LLM models. Sometimes you're calling an open AI API, sometimes you're calling Cohere. Other times you have your own Llama based model running on your own infrastructure. And so you might have... Depending on which LLM you're calling, even if your user experience is remaining relatively consistent under the covers, you might have different fields. Maybe when you're calling your Llama LLM, it has different outputs, it maybe it has some additional field that you don't have when you're calling the OpenAI API.

Richmond Alake: 00:12:48 Yeah.

Jon Krohn: 00:12:49 And so Mongo allows you to very flexibly, under the hood just be like, "Okay, we're going to store that extra

information." Whereas if you were trying to do everything in the tabular way, you would've had to have already had that column in place-

Richmond Alake: 00:13:01 Yes.

Jon Krohn: 00:13:01 ... for that LLM that you ended up using later on.

Richmond Alake: 00:13:03 Yes, we have... And one thing is, again, we have a lot of evolving and experimentation happening within AI workload today, and having a rigid schema having to do schema migration just because some things change in the application space that you didn't realize at the beginning, it causes a lot of pain for developers. And in the age of AI that we're in today, speed is a competitive advantage, enabling your developer, developer productivity is a competitive advantage. So if you have a flexible database and you have MongoDB, it allows your developers to move a lot quicker.

00:13:42 And we also realized that we also wanted these LLMs to give us structured output. So we have the structured output in OpenAI or JSON mode, is what most of the LLMs have. And that resonates with what we were talking about earlier, which is what resonated for developers, like yourself, 10 years ago was, "Hey, JSON makes a lot of sense in the application layer, now it makes sense in a data layer." MongoDB right now, if you look at it, we have LLMs enabled in JSON mode, now your LLMs... It's just a way to think of the whole AI stack in a unified manner. And you think about JSON in your application layer, you think about JSON in your data layer, you think about JSON in your LLM, it just makes everything easier.

Jon Krohn: 00:14:30 Nice. That's a great sound bite. Maybe we'll have to convert that into a YouTube short or something. I love it, it's perfect. So yeah, so we've now covered what NoSQL databases are, how Mongo is the archetypal NoSQL

database, and how it varies from SQL relational databases. And there's ways that people can get access to MongoDB for free right away, so there's a MongoDB community version, which I'll provide a link to in the show notes, but that is free. It also may not have the latest features that MongoDB proper offers.

Richmond Alake: 00:15:08 Yeah.

Jon Krohn: 00:15:09 So it sounds like you're going to kindly provide me with a link-

Richmond Alake: 00:15:11 Yeah.

Jon Krohn: 00:15:12 ... to MongoDB Atlas, which is a MongoDB cloud platform, it's free to start, and that allows you to get going with MongoDB.

Richmond Alake: 00:15:21 Yes, exactly. So in the show notes, there's going to be a link for you to create a free MongoDB account, and you can just track MongoDB for your AI application. So storing your operational data, your metadata, and also, which we're going to talk about, storing your vector data as well.

Jon Krohn: 00:15:37 Yes, that is super exciting. In fact, I think we should talk about that next. But I want to also quickly say, it sounds like this has almost been a big long infomercial so far because I'm such a big Mongo fanboy. Obviously you're a Mongo developer-

Richmond Alake: 00:15:50 I'm too.

Jon Krohn: 00:15:50 ... advocate, so you are as well, but I am not being paid to say any of this. Mongo has in no way sponsored this episode or the podcast-

Richmond Alake: 00:15:59 Yeah.

Jon Krohn:	00:15:59	... this is just genuine passion about this tool.
Richmond Alake:	00:16:03	Exactly. And I think that's what Mongo does, because it resonates deeply with developers building these applications. And we understand how developers think, and now in this AI era, we also understand how LLMs are providing outputs, it just makes building application fun.
Jon Krohn:	00:16:20	Yeah.
Richmond Alake:	00:16:21	That's what Mongo... No one wants to be a database administrator. No developer wants to be a database administrator. So MongoDB lets the database... It just makes it fun for the developer.
Jon Krohn:	00:16:32	Nice.
Richmond Alake:	00:16:32	And yeah, one thing... Sorry, sorry to interrupt, but one thing about fanboying I wanted to point out, is a couple of years ago I used to write a lot about MongoDB and there's some stuff seven years ago on Quora, me telling people to go use MongoDB. So it's actually... It's a full circle moment for me, to come back and have an impact.
Jon Krohn:	00:16:53	This episode of SuperDataScience is brought to you by the Dell AI Factory with NVIDIA, helping you fast-track your AI adoption - from the desktop to the data center. The Dell AI Factory with NVIDIA provides a simple development launch pad that allows you to: perform local prototyping in a safe and secure environment. Next, develop and prepare to scale by rapidly building AI and data workflows with container-based microservices. Then, deploy and optimize in the enterprise with a scalable infrastructure framework. Visit www.Dell.com/superdatascience to learn more. That's Dell.com/superdatascience.

00:17:33 For sure. That's great, I'm glad you're getting to enjoy that. So MongoDB Atlas, tell us about why somebody... What advantages you get working with a cloud system like MongoDB Atlas versus trying to set up maybe your own MongoDB, running on your own infrastructure?

Richmond Alake: 00:17:52 So with MongoDB Atlas, you get a managed offering of MongoDB. And one of the key thing is you can use MongoDB on any of the large cloud providers today, so AWS, Azure and Google Cloud, across several different regions. So we have a multi-region, multi-cloud infrastructure, and the database layer is also distributed within MongoDB. So you get all of this... Again, what I was saying is no developer wants to be a database administrator, you get all of this within MongoDB Atlas, alongside other products such as streaming, and we have dedicated search nodes, which we can talk about when we're talking about the vector data capabilities of the MongoDB database. So that's why essentially you would.

Jon Krohn: 00:18:40 Nice. Yeah, let's talk about those vector databases right now. So I already made a joke earlier about how... So if you are not already an AI developer, then you may not be aware of vector spaces, but essentially vectors are a high-dimensional space, so it's very easy to imagine a three-dimensional space because that's the visual world that we live in. And so if you imagine a three-dimensional space, you can take any kind of object, and so it was a big thing around 2011, 2012 to be taking words and putting them into a vector space. So there was this very famous algorithm, Word2vec by Thomas Mikolov and that was an algorithm for taking words in natural language and using some really interesting properties related to words, which is that the meaning of a word tends to be the average of the words around it. Which is a trippy thing to think about, but ends up being true because you can train these computer science algorithms to say, "Okay, just based on the words around," Without any labels in your

data, you can just take a large amount of natural language and use something like Word2vec to create a word cloud, in however many dimensions you want, where the closer you are in that space, the more similar the meaning of the words are.

00:20:03 So the British versus the American spelling of a word would typically end up in exactly the same location because whether you say honor with or without a U, it has the same meaning. And then synonyms would be close by. You could end up with a region of database languages, so you'd have, NoSQL and MongoDB would be very close together in a well-trained space, but close by you'd find SQL and MySQL, and relational databases. And then in another part, not too far away, you might have Angular and React, programming languages. In a completely different area of the 3D space, you might have dates and times, and so just you end up having meaning represented in this three-dimensional space, but with the vector space, you could have 100 dimensions or 1,000 dimensions. And so that allows you to have a lot more granularity in that space where moving along one of those dimensions represents some kind of meaning, to the way the data are stored.

00:21:06 And so anyway, vector databases have been hugely important over the past decade, starting mostly with words, but now whole documents. So for RAG, for example, really important for retrieval augmented generation, you typically take, let's say you have 1,000,000 documents, you could convert each of those documents into a location, into some high-dimensional vector space, and then you could in real time, based on some query that somebody asks, you could retrieve the right relevant documents, based on the semantic meaning of the query. So yeah, vector databases are huge these days for applications like RAG, and I did not know until we were preparing to do this interview, that Mongo offers

a Vector database as well. Because I always think of Mongo as just that JSON structure, but yeah, you have vector databases now as well, tell us about that.

Richmond Alake: 00:21:57 Yeah, so with MongoDB, we work with a variety of customers and our roots started obviously with storing data within a database. So we saw most of our customers over a decade ago using search engines on top of MongoDB, and to make it easier because that's what we're focused on, which is our customer needs, we brought search into the database. So we had Atlas Search. Then we also noticed that a lot of our customers were dabbling within use cases that were related to Semantic search, with their vector data. And most of these folks were bolting on other extensions or maybe another database, to implement vector search, but we brought that into MongoDB. So now you have one database where you can do your normal lexical search, but also you can do vector search as well, and you can actually combine both and you have hybrid search. So in a situation where you would have, and I've seen this in some cases, two, three databases, with MongoDB, you have one database that can power your entire AI application. So that is the beauty of it.

Jon Krohn: 00:23:13 It makes a lot of sense. You could imagine if you were starting from scratch and you knew that, you would probably think, "That seems easier than getting some separate vector database." And having to be confident that the linkages between them are always going to work. Because you can end up having version issues, where that vector database has some kind of update where it no longer brings in the field like you'd expect from Mongo or Back. Whereas when Mongo's doing all of it, there's probably people testing to make sure that everything's working properly.

Richmond Alake: 00:23:44 Exactly. Spot on.

Jon Krohn: 00:23:47 Nice. All right, so let's now talk about something that's an extension of this idea of vector databases, AI and MongoDB. You recently wrote a blog post on AI stacks, and it's actually right now at the time of recording, if you Google the term AI stack, your blog post comes up as the number one hit. So I'll have a link to that in the show notes. We talked a little bit earlier in the episode about things like the MEAN stack, which was this idea of back-end all the way through to front-end technologies, for the developer. Is an AI stack somehow related to that kind of thing?

Richmond Alake: 00:24:23 Yes. So we said the MEAN stack was... Well, we didn't say, but we know that the MEAN stack is a composition of a bunch of tools and libraries to build application. So the AI stack is a composition of tools and library to build an AI application. One thing I would say is, the AI stack is different, in terms of how you visualize it, depending on the persona you're talking to. So when I'm talking to developers, and when you see the... When you look at the article, when I'm talking to developers, there are more layers in the AI stack than when I explained the AI stack to a C-suite or VP-like person. And that's because I feel developers need to really go... We really need to dive deep into what is making the AI applications today and understand the composition. But for some CEOs and VP-like folks, they don't need to know the intrinsic detail, they need to know the high-level information.

00:25:23 So just to make the point is, most of VPs or high-level execs within companies would describe the AI stack as, you having the application layer, you have your data layer and you have your compute layer. Very easy, so application would be... Sorry, you have your application layer, you have your tooling layer, then you have your compute layer. So application would be any of the products you see today, so Cursor, a very popular IDE that is powered by AI, would lie in the application layer.

Then in your tooling layer you have folks like MongoDB or any of the tools that enable the application layer, then in your compute you have NVIDIA. But when I'm talking to developers, I double click into that and we talk about the other layers of the stack. I'm not going to remember everything now, but programming language is very important.

00:26:14 When you're developing this AI stack AI application, the languages you select is very crucial because not all libraries that you're going to be using further in the stack, are written in all the languages you have available. Some are just Python or maybe some have... Or just JavaScript or there's some that are evolving to have both now. But your programming language is crucial, you have your model provider, you have your database, which would be MongoDB, then you have your model provider. So you have... There have several layers to that stack, and when I'm talking to developers, I tend to dive deep into that.

Jon Krohn: 00:26:50 Nice. So that if people want to have a framework for describing the different tiers, like you said, from the bottom all the way to the top, from your compute layer all the way through to your application layer, this AI stack framework that you provided, and again, we'll have a link in the show notes, to make that a piece of cake for people to follow along with. Another thing that you published recently is something called Memo Riz, and So it's like memories, but it also has... Is that like riz-

Richmond Alake: 00:27:18 Yeah, that's like riz.

Jon Krohn: 00:27:18 Yeah. So that's-

Richmond Alake: 00:27:21 That is me coding late in the night and just coming up with a funny name for a library. So-

Jon Krohn: 00:27:30 That's like a Gen Z term.

Richmond Alake: 00:27:32 Yeah.

Jon Krohn: 00:27:32 It just means good?

Richmond Alake: 00:27:34 I don't know. That was my attempt-

Jon Krohn: 00:27:37 Please-

Richmond Alake: 00:27:38 ... at trying to be young.

Jon Krohn: 00:27:38 ... some Gen Z listener provide us with information about that.

Richmond Alake: 00:27:45 Yeah, riz is... Yeah, I thought it was funny. I don't know if people think it's funny, but really that's an interesting project that I dabbled in, and it was mainly trying to solve a bunch of issues I was seeing around some of our customers building AI application. And it had everything to do with memory. So as we're going into, or we're already in this agentic era, the way... Memory becomes important, data becomes important, and the way you structure, store and retrieve memory, actually affects the reliability and the performance of your AI systems, of your agentic systems. So memory is becoming an important aspect of our AI application today. So we had RAG in 2003, in 2004, vector databases made RAG a popular.

Jon Krohn: 00:28:34 In 2023-

Richmond Alake: 00:28:34 [inaudible]

Jon Krohn: 00:28:36 ... 2024.

Richmond Alake: 00:28:37 In 2023, what did I say?

Jon Krohn: 00:28:37 2003 I think.

Richmond Alake:	00:28:40	Well, 2023 and 2024, vector databases made RAG a very popular mechanisms within AI application. But as we move into late 2024 and 2025, we have more the form factor of agentic systems more prevalent within the space, and agentic RAG became a thing that people were implementing. But now what we're seeing is, there are a bunch of companies and open source libraries that are focused on how data is structured, and retrieved within this agentic system, and the different scoring mechanism that goes beyond the relevance. There's a bunch of research papers on this, and one company that comes to mind will be MemGPT, the company behind, there's Letta, there's another paper called HippoRag. And the main problem they're trying to solve is how do you store memory efficiently within a storage layer, but retrieve it efficiently at the right time, for your agentic system? So Memories was my attempt to solve some of the problems I was seen around memory management, in a agentic system.
Jon Krohn:	00:29:54	Yeah. What kinds of issues do people run into when... When you want to have an agentic system, is it common to run out of memory when you're trying to develop some RAG application? Or what are the common issues-
Richmond Alake:	00:30:07	Well-
Jon Krohn:	00:30:07	... that people run into with memory?
Richmond Alake:	00:30:10	... in respect to the, is it common to run out of memory? It's common to run out of context window. Obviously we've got large context window now, 1,000,000 context window. How deep does your pocket go? Is a joke I make to folks that are trying to stuff 1,000,000 worth of tokens at every inference group.
Jon Krohn:	00:30:30	Yeah, yeah, yeah. Oh my goodness.

Richmond Alake: 00:30:33 But one of the issues or the problems I saw, relating to that context window, is, within the agentic space, the LLMs are now capable of tool use, which is... Tool use is when you provide an LLM, a collection of JSON schema tools, that it can use to complete an objective. So this would be JSON schema of maybe Python functions or microservices within your system, that the LLM is aware of, and now can actually select the right tool to use, select the right parameters, and then you invoke it on your system. That is tool use essentially, or function calling, whatever you want to call it. One thing is, if you're using some open AI models on the documentation, the guidance is you only put between 10 to 20 JSON schema of these tools within the context window at a time, because obviously you have limited context window. And the more tools you put in, the harder it gets for the LLM to pick.

00:31:38 So in the library memories, what I implemented was a design pattern that we call MongoDB as a toolbox. Essentially what you're doing is you're using MongoDB, and this is where the MongoDB affinity for JSON comes very useful, again, that document data model, is we store all the tools you can have within your system. It could be a thousands tools, you can store it within MongoDB and have a vector representation in MongoDB as well, alongside your metadata of the tool. The vector representation could be the function and some of the Google Docs string of the tool that describes when the tool is used, represented in numerical format. And what you can do is just before you make a call to the LLM, you can actually make a call to your database to get the right tools, and that way you can maybe send maybe two or three tools.

Jon Krohn: 00:32:33 In today's ever-changing AI landscape, your data demand more than the narrow applications and single-model solutions most companies offer. Domo's AI and Data Products Platform is a more robust, all-in-one solution for

your data. With Domo, you and your team can channel AI and data into innovative uses that deliver measurable impact. While many companies focus on narrow applications or single-model solutions, Domo's all-in-one platform brings you: trustworthy AI results, secure AI agents that connect, prepare and automate your workflows, helping you and your team gain insights, receive alerts and act with ease through guided apps tailored to your role and the flexibility to choose which AI models to use. Domo goes beyond productivity… it transforms your processes, helps you make smarter, faster decisions and drive real growth. Data can be hard, Domo is easy. Learn more today at AI dot Domo dot com. That's AI dot Domo dot com.

00:33:34 Nice. That makes a lot of sense.

Richmond Alake: 00:33:36 Yeah, it's an interesting design pattern and it allows your agent system to scale. So now you can have as much tools as you want and you're bypassing the limit.

Jon Krohn: 00:33:48 And it's a great tie together, right from the very beginning, of how JSON structures that MongoDB inherently recapitulates with the key value pairs, how that's always as a developer, since you discovered it's made developing a breeze-

Richmond Alake: 00:34:04 Easier.

Jon Krohn: 00:34:04 … turned you into a full stack developer, and now it's allowing you to come up with clever ways of using LLMs to be calling tools as well. Nice. Speaking of agents, many folks have been saying, including probably myself at times, that 2025 is the year of the agent, of agentic AI. But before we started recording, you said to me that it's the year of multi, tell us about that.

Richmond Alake: 00:34:30 So 2025 is the year of agents, in different perspective, and I think in the mainstream commercial perspective. I was, and a lot of people within the space have been screaming about agents before this year, last year, even some people in 2023, now it's mainstream, it's commercial, now everyone is aware of it. But I think it's our job as leaders within the space and being at the forefront and helping our customers, to be several months ahead of our customers, if we're able to provide good services to them. So agents, they're good, they're nice, and we should all be building them and learning about them, but I think that 2012 is the multi-era of AI.

00:35:12 So what I mean by that is multi-agent architecture is going to be very popular, they're going to be very popular. We're going to have multimodal embeddings. Now you have embeddings that it's not just text, they're images, videos, audio, all captured using one embeddings, multilingual embeddings that can understand different languages, and you're going to have multiple retrievals mechanisms within your agentic system. So vector search is not all you need, vector databases were popular, but you can't just rely on vector representation to retrieve the right information, from your database to give to your LLM, you need more various different search mechanism. And we are seeing a lot of people moving to this memory management space I was talking about. So multi-agent, the multimodal, multilingual, multiple retrieval mechanisms, there's multi-step reasoning as well.

Jon Krohn: 00:36:04 Yeah, yeah.

Richmond Alake: 00:36:04 So 2025 is the multi-era for AI.

Jon Krohn: 00:36:10 Yeah, yeah. We're taking up some of the foundations that were honed over 2023 and 2024, and scaling them up in a multi-way.

Richmond Alake: 00:36:18 Exactly.

Jon Krohn: 00:36:19 Yeah. You're absolutely spot on. Looking beyond 2025, what are your predictions for agents beyond this year?

Richmond Alake: 00:36:27 Predictions in AI are... They're very funny to make, because everyone's always wrong. No one could have predicted it will be here, maybe three, four years ago. If I was to make a sound prediction, it's not going to be wild, I think voice agents are going to be... They're going to rise in popularity. I have a very strong opinion that voice agent might not take off as much people think it would, they're not going to be the dominant way that we engage with agents. But voice agents... Everyone's going to want to implement voice within their AI application because just experiment. Like I said, the experimental nature of AI in general. I see memory management becoming an important topic. And I met Andrew Ng, who's one of... He's an AI pioneer, and the first thing he said is, "Sir Richmond, how do you model memory within agenetic system?" This was last year, April, and he's way ahead in the space than I am. And I see a lot of people starting to focus on memory management and modeling memory within agenetic system. Because that's how humans work, we don't have just one form of retrieval mechanism within our brain, we have several different forms of recalling memory. So agenetic system will need the same as well.

Jon Krohn: 00:37:53 It's perfect that you mentioned Andrew Ng because my very next question for you is related to that. So Andrew Ng, it's an understatement to say that he's one of the most important people in data science and AI, one of the most known. And he was on this show in episode number 841 a few months ago, and you had a course with him. So Andrew founded a platform called DeepLearning.AI, which is a learning platform that probably a lot of our listeners are aware of and have maybe taken courses

from there. And you did a generative AI course in DeepLearning.AI with Andrew Ng, how did that happen?

Richmond Alake: 00:38:33 Well, it happened from my manager actually just contacted, I think DeepLearning.AI directly, and just proposed we do a course. Because MongoDB is, we believe that we are one of the best... We are the best database out there, for AI application, and we wanted to spread the word. And so yeah, they were interested in what we had, they were fans of MongoDB, so Coursera in the early days was built on MongoDB. So they were interested in what we had to say. And funny enough, what we had to say in that course back then, it was released in June last year, is even more relevant today. Because we don't just talk about RAG, we talk about RAG within agentic system, but we talk about a pain-point, which is having large context to give within LLM, so we talk about the topic, which is prompt compression. So the course is, Prompt Compression and RAG Within AI Systems, or something along the lines, the link would be in the show notes. But yeah, so Andrew was excited about the course. We were excited to deliver the course, and it was a pleasure to actually meet him in person.

00:39:50 Funny story, Andrew knew about me, well, the digital version about me, the digital version of me before actually meeting me, because back when I used to write for fun, I wrote a lot on Medium, and Andrew has one of the best videos on reading research papers on YouTube. He was teaching a Stanford class, and he describes to the students how to read research papers. So the extent of my education is to the master's level, but I still had to read a lot of research paper and I still read a lot of research papers today. So I found that video very useful around, it would be five years ago, and I wrote a Medium article on it, breaking down and creating a framework or notion, on how people can use it. And he saw that article.

Jon Krohn: 00:40:40 No kidding.

Richmond Alake: 00:40:41 Yeah.

Jon Krohn: 00:40:41 That's cool.

Richmond Alake: 00:40:43 And the article went viral as well on Medium and all the good stuff. And yeah, he just mentioned, and we spoke about the article as well, he was like, "Wow, you wrote article?" I was like, "Yeah, I did." And so he met the digital version of Richmond before-

Jon Krohn: 00:40:55 Nice.

Richmond Alake: 00:40:55 ... he met the real version.

Jon Krohn: 00:40:56 Yeah, me too. I was talking to pixels of you the entire preceding episode, they were just Richmond pixels, a few seconds behind the real Richmond moving somewhere else in the world. Nice. So what kind of listener of my show should go and take your Gen.AI course, in DeepLearning.Ai? Who's it targeted at?

Richmond Alake: 00:41:18 I think most software developers should dabble in what people are building today. So if you are a software developer, go take the course. It's very easy to... I break things down from first principles of, "Hey, this is what relational looks like, but this is what MongoDB is. And this is MongoDB within AI and this is RAG, and this is a problem that you are probably going to face at different level." So I break it down. So I really think most software developers should take that course.

Jon Krohn: 00:41:54 Nice. All right. So now let's zoom out a bit. We've been focused mostly in this episode so far on specific AI tools, functionality, features that people could be using. Let's talk now more broadly about AI strategy in general. So you are a leader in this space, you are leading teams and

you are mentoring broadly within MongoDB, outside of MongoDB, on how people can build successful AI teams and products. So you're a great person to speak to about this, how can one build an effective AI strategy, in such a competitive environment where there's so many people that are trying to build something similar?

Richmond Alake: 00:42:38 Yeah. That's a good question and it's one that I had to think a lot about, in a short amount of time, by the way, as in AI moves very quickly. And I'm sure there's a bunch of AI leaders and enterprise AI leaders that are having sleepless nights just trying to keep up with... Coming up with an AI strategy and showing return on investment on AI initiatives. The years also the year of, show me the money within AI, but AI strategy with... AI strategy within the space is very difficult. But one thing that I've seen that's worked is laying the groundwork of your strategy in pillars and objective truths. So one thing that I... When I'm coming up with a strategy or if I'm thinking of how am I going to put a strategy to win a certain... For any mission or task I'm given, I always think about first principle and I look for pillars that I can base my tactics on, and I look for objective truth within those pillars.

00:43:41 So for example, if a pillar would be, let's say I was trying to... We're talking about developers and engineers, let's say I was trying to think of an AI strategy that is around developers. What I would do is look at what pillars are around that developer persona. So one thing we know is developers live on GitHub. So GitHub would be a pillar. Then I'll think to myself, "What are the objective truths of GitHub? So what do developers actually do on GitHub? What is something within GitHub that would never change? And if it does change, then that means there's a massive platform shift." So those are the way I think about things. I look at a persona, I look at other attributes, I try to form the pillars, try to form objective truth within a pillar, try to form some convictions and

assumptions that I'm making, and I try to form some tactics from all of this information.

Jon Krohn: 00:44:37 Excited to announce, my friends, that the 10th annual ODSC East (Open Data Science Conference East), the one conference you don't want to miss in 2025, is returning to Boston from May 13-15! And I'll be there leading a hands-on workshop on Agentic AI! ODSC East is three days packed with hands-on sessions, and deep dives into cutting-edge AI topics, all taught by world-class AI experts. Plus, there will be many great networking opportunities.No matter your skill level, ODSC East will help you gain the AI expertise to take your career to the next level. Don’t miss out — the Early bird discount ends soon! Learn more at odsc.com/boston.

00:45:22 That makes a huge amount of sense. I really like that idea. I don't think I was doing it in such a structured way before, I was doing it more in a NoSQL way in my head, where was... But I do... There are those kinds of things. Even though we live in this very fast moving space, there are absolutely these pillars and these things that don't change, GitHub is a really good example. And for me, even in my career, going back to when I was doing my PhD between 2007 and 2012, even back then I was like, "Wow, the world is changing fast. Where can I find solid ground in my career?" And when other people in my... I did a neuroscience PhD, and so some people are growing cell tissue cultures, other people are doing recordings from the brains of ferrets, so you have very specific skills. And I was like, "Well, the amount of data that are being stored on the planet, that's going to keep increasing, compute to be analyzing those data is going to get cheaper and cheaper. And so I was like, "I should probably learn statistical computing and machine learning, that's going to be a useful thing." So there are these kinds of mega trends or pillars, that even in this very fast moving space you can feel secure around.

Richmond Alake: 00:46:42 Yeah. It just keeps you grounded. I did the same thing when I was a web developer, and wanted to... I was looking for the next step in my career, looking for the next challenge, and I thought to myself, "What's an interesting problem that I can get involved in today, that will be relevant for a very, very long time?" And there is no more interesting problem than the replication of human intelligence. So I said, "Okay, I'm going to go to university and get a master's in AI." And that's going to be relevant... It might take different shapes or form-

Jon Krohn: 00:47:14 For sure.

Richmond Alake: 00:47:14 ... but it's going to be relevant until we maybe get to AGI or some form of super intelligence.

Jon Krohn: 00:47:20 Yeah, it's interesting, something that I've been thinking recently... I don't think I've said this out loud yet, so it'll be interesting to hear your thoughts on this given your strong AI background, in terms of education as well as commercial applications. Something I've been thinking recently is even when AGI comes about, I think that the way that the system thinks, it'll still be quite a bit different from the way that we do. And so I think more and more like we're seeing today where some coding problems or math problems, you can get a multistep reasoning system to be able to be at or near PhD level math students or chemistry students, on these kinds of problems, computer science students. But there's all kinds of problems that are more difficult to package into that way, it'll be more difficult to train AI systems in that way.

00:48:18 And so it seems to me like there's going to be places where human intelligence still has strengths, where we're complemented by these AI systems. In the same way that you would rarely do long division today, you would pick up your calculator and have it do it for you. There's going

to be more and more things that we can use a computer to solve for us. But even when, I don't know, I mean, I guess the idea in many ways of an AGI system, and it depends on how you define it, is that it can theoretically do any of the kind of thinking that humans do, but it still seems like there's going to be relative strengths. And so yeah, it's just something I've been thinking about lately, that even... There was a time a few years ago where I was like, "Wow, an AGI system or an artificial superintelligence system, does this pose an existential threat to humans?" And there's still... There's absolutely some percentage chance that that's true, but more and more, for some reason, in a way that I'm maybe struggling to articulate really well, and you might be able to articulate better than me, it doesn't scare me. It seems to me like we're going to be this natural complement to each other.

Richmond Alake: 00:49:31 One thing I do understand is, when two intelligent system meet, it could either be chaos or there could be collaboration, one of the two Cs, whichever one is. Right now we're creating artificial intelligence, and right now it's collaborative, that's great and I hope it keeps going that way. But even if you look at humans, we have the same form of intelligence within humankind, but we still have collaboration and chaos. I think there's going to be a point where, hopefully not, but there is a probability where it switches to that chaos state, who knows? But another C that I think that AI won't be able to replicate, is curiosity. I really think that humans are one of the forms of intelligence that are just so curious about the environment, and we still have space exploration, we haven't even left this planet yet. So we're still at the beginning of this journey of... So I think the curiosity of space is... And the universe is vast and we're not going to send AI to go explore the universe for us ,because one thing that we want to do is experience as well. We want to add the human experience into it as well. So I know we

send a bunch of satellites to probe space, but they send the information back to us so we can experience, and learn from it.

00:51:05 So curiosity is something that AI will probably never take from us. They might take what we define as work today, and I look forward to that, I don't want to work, but like I said, the things I used to do for fun, now I do it as a job. So I think human beings in a few years will have a choice. Most human beings will have a choice or the opportunity to do most things that they consider fun and get paid for it. It might be a... Well, I'm not a doomarism, but I do think that we have to consider a possible probability of there being a chaos state between our form of intelligence and the artificial form of intelligence.

Jon Krohn: 00:51:44 And it would be nice... And I realize this is just extremely naive, but it would be nice if we could have all of the key players in AI development, like the US, China, UK, if everyone could be working together towards the same goals, instead of being in opposition.

Richmond Alake: 00:52:05 Yeah. And we are, we are working together towards the same goals.

Jon Krohn: 00:52:09 It is interesting how the research community continues to operate collaboratively, despite sanctions.

Richmond Alake: 00:52:17 Exactly. Innovation will find a way.

Jon Krohn: 00:52:21 Yeah. Nice. That's another good sound bite. Nice. All right, so cool. That was an interesting tangent that we went on there. Another thing that I'd like to talk about with respect to AI strategy, so we just were talking about how to build an AI strategy in a competitive environment with your pillars. Another framework that you've developed alongside your pillars is something that you call the ARENA framework. Tell us about that.

Richmond Alake: 00:52:50 So the ARENA framework is for the strategy aspect, and it's something that I think about when I'm trying to formulate strategy, and it came off thinking about, what stood the test of time? And well, not entirety of time, but some form of time. And the Colosseums where gladiators used to fight, those have stood the test of time, you can see some of them today. And the architectural buildings around them are very solid. So you look the forms of pillars, it's a circle and they have a bunch of pillars and the pillars have segments. So you can think about that as your AI strategy, where your tactics get to battle it out in the arena. The arena is the market, whatever you want to call it a free market. So formulating your strategy, one with pillars and having segments of objective truth, understanding your resources and formulating your tactics, creates very good strategy and tactics that you can put to work in the arena.

Jon Krohn: 00:53:46 Nice.

Richmond Alake: 00:53:47 So then in terms of the team aspect of it, I think that was going to be your next question, right?

Jon Krohn: 00:53:53 Yeah.

Richmond Alake: 00:53:53 So you have your strategy, you have your tactics, now you have a team, and I think the AI space moves relatively quickly. And I think for those of you that are leading teams, I think you need to be aggressive, in terms of your pursuit of innovation, you need to be aggressive in terms of the release and going into the market. So being aggressive, centralized, coordinated, resourceful, efficient and data-driven, ACRED, is the short form of it. I think those principles, and if you embody those principles, you'll be successful in a fast-moving space, because if you aggressive, you'll approach most problem with your full potential. If you're centralized, you'll be focused, if you're coordinated, you'll be in harmony with your environment

and with the resources in your environment. And if you're resourceful, you're going to use everything within your environment. If you're efficient, you're going to use it to the best of your capabilities without expending cost, such as time or money, and your data driven, which allows you to prove your results or prove your assumption and refine. That was a long aspect of the way I'm thinking about things, but I think about this maybe more often than not.

Jon Krohn: 00:55:08 Nice. You have some sound principles for AI strategy there, obviously a lot to take in. Maybe it'll be a future course or book.

Richmond Alake: 00:55:15 Maybe a future course.

Jon Krohn: 00:55:17 We do have... I wanted to quickly mention before the episode ends, that you have actually, in addition to all of this other work that you've been doing since we last spoke, there's a couple other things that you published that our listeners might be interested in. So in the O'Reilly platform, you published a computer vision course, which actually is orthogonal more or less to all the MongoDB stuff we've been talking about today.

Richmond Alake: 00:55:39 Yeah, so computer vision is what I studied or was my strong point in my master's. So I've been doing a bunch of teaching, like I said, while I was a machine learning architect or computer... My first role in the AI was a computer vision engineer actually, then I became a machine learning architect, now I'm a developer advocate for AI MML. So I released a computer vision course in collaboration with O'Reilly last year, October 2024. And it just touches on the foundations of computer vision, so object detection, pose, estimation, we're looking at different techniques and I'm explaining the models. It's actually a practical course, so there's about four hours, or maybe two hours of... It got edited down, but there's

about two hours of me coding using PyTorch and showing you how to build neural networks from scratch. Then we train the neural networks and we do some transfer learning. So it's a bunch of the good... Weird to say, but the good old stuff. It's not generative AI, it's computer vision.

Jon Krohn: 00:56:41 Yes.

Richmond Alake: 00:56:41 But it's really good.

Jon Krohn: 00:56:42 Yeah, I mean it's interesting with how much we can end up in so many episodes, including today's episode, we ended up talking so much about generative AI and AI agents, which is very much what people are talking about these days, it's new, it's cool. But absolutely things like computer vision are not going to go away, it's not a generative thing, it's not necessarily an AI agent thing, although it can obviously be tied into a system with either of those kinds of things-

Richmond Alake: 00:57:06 Exactly.

Jon Krohn: 00:57:07 ... compliment it, provide an AI system with the ability to see and navigate the world, or make decisions, come to conclusions ,based on information in its surroundings. Cool. And then the very last thing, so I always end the episode asking the guest for a book recommendation, and you are still going to have to do that, but before you do that, I'm also going to mention that with a number of other co-authors, it's about 10 of you, I didn't count the exact number, you co-authored a book published by Packt in 2024, called Building AI Intensive Python Applications, and it's a MongoDB book effectively.

Richmond Alake: 00:57:48 Yes. So MongoDB, when I joined this company, there were so many amazing people residing within MongoDB that are very creative, they're very intelligent and eager to

learn, and also eager to contribute just to the general intelligence of this space. So that's why you see that there is a lot of authors in that book, because everyone is so eager to contribute in a space that moves very fast. So there's a lot of learners here, we're in the trenches, we're in the arena, we're learning from our customers, we're helping our customers, and we brought that learning into a book that we feel resonates with a lot of developers today. And that's what we're doing, so we build learning materials that hopefully stand the test of time, maybe two years, three years.

Jon Krohn: 00:58:39 Nice. Yeah. Yeah, so I guess if people want to be learning more about the kinds of things that we were talking about in today's episode, they've obviously got the DeepLearning.AI course-

Richmond Alake: 00:58:47 Yes.

Jon Krohn: 00:58:47 ... on Gen.AI, we'll have a link to that, but if you would prefer to be doing learning from a book, I'll also have a link in the show notes to your Building AI intensive Python Applications book.

Richmond Alake: 00:58:58 Excellent.

Jon Krohn: 00:58:58 All right, so then that brings us Richmond to my penultimate question, which is do you have a book recommendation for us?

Richmond Alake: 00:59:06 Yes. Can I do two?

Jon Krohn: 00:59:08 You can.

Richmond Alake: 00:59:08 Okay, excellent. So one, there's a book by Seth Godin called This is Strategy. PS, it is not Strategy, the book is actually not strategy. The book is actually a bunch of riffs from Seth Godin, who's a very popular marketing person.

So it's a bunch of riffs really. You can see there's a bunch of long-form tweets that is placed as a book and some people might not like that form or structure, but he talks a lot about formulating and building strategy, within marketing and sales environment. And it's very relevant. I'm within a marketing organization within MongoDB, but I think developers should also read those book as well, and it just contains a bunch of just gems. It's not really structured, so people hate that, but I love it. Yeah, so that's my first book recommendation, This is Strategy, by Seth Godin.

01:00:15 The second one is Crucial Conversation, and I think a lot of... That's a popular book. It got recommended to me a few months ago because one thing I found myself, within my career, was having a lot of conversation with different types of people, MongoDB is a massive company, we have over 5,000 intelligent people working here, and the AI space is growing. It was just a few years ago, it felt so small, like a close-knit community, now everyone's in this space, but people like yourself and other influencers and experts within the space, we still have that small group. But the AI space is growing. I communicate with different forms of people for different purposes. Crucial Conversation allows me to speak in a way that gets me what I want. And that sounds very simple, but it is a very hard thing to do where you can have a conversation with someone and actually communicate your objective and not get lost in emotions or in a moment, and make sure you come out with your desired outcome, while not make enemies as well. And doing this in the most effective times, in the most crucial times. So that book really helped me with that, and I've only read about half of it.

Jon Krohn: 01:01:38 What was the book again?

Richmond Alake: 01:01:39 Crucial Conversations.

Jon Krohn:	01:01:40	Crucial Conversations. Nice. Very cool.
Richmond Alake:	01:01:42	I think it's a very popular book, and This is Strategy, is I think also a fairly popular book as well.
Jon Krohn:	01:01:48	Yeah, yeah, yeah. Nice. All right, Richmond, so-
Richmond Alake:	01:01:51	Oh, I was going to ask, do you have any book recommendation for me?
Jon Krohn:	01:01:55	That's so interesting. I actually am currently... On the flight yesterday I was reading a great book from, do you know Cole Nussbaumer Knaflic, have you ever heard of her?
Richmond Alake:	01:02:04	No.
Jon Krohn:	01:02:06	She wrote a book years ago, maybe a decade ago, called Storytelling with Data.
Richmond Alake:	01:02:13	Yeah, I've seen-
Jon Krohn:	01:02:14	It's hugely popular. It's crazy. I don't know how many copies she sold, but based on the number of Amazon reviews, it's like thousands and thousands, or tens of thousands, I can't remember, of Amazon reviews. And so it seems safe to say that there's been hundreds of thousands of copies of that book sold. Now, I actually haven't read that book, I'm sure it's a great book, lots of people have read it. But at the time of recording, I am in the midst of preparing for a big opening keynote, you and I were talking about this, I don't know if it will have passed now at the time of publishing this episode, but on March 19th, I'm doing this opening keynote at the RVA Tech Data and AI Summit in Richmond, Virginia. And I've done keynotes before, but I had Cole Nussbaumer Knaflic on the show. I don't know the episode number off the top of my head, but maybe a year ago or two years ago, as

she was announcing the release of her latest book called Storytelling With You, and so it's a guide to effective presentation. And it reminds me of this Crucial conversations, part of that is... A big part of this book that she's... What she's talking about in this book is how you can use a presentation to switch people's way of thinking, to your way of thinking.

01:03:34 And so I've been really engrossed by this book so far. She's an amazing communicator, and so of course it's a great book. She's this huge multi-best-selling author, and so unsurprisingly, it's a really tight book, really well-written, and she makes great use of story within this Storytelling With You book as well. So I'm enjoying it a lot so far. I still haven't finished it, but it's been great. It's been perfect. If you are listening to the show or you specifically Richmond asked for this recommendation, so now you're getting it. If you are preparing for a talk, I'd sketched down a couple of pages of ideas, of things that I wanted to do, and I'm a month out from doing the talk at the time of recording, and so starting in this book is perfect because it has tons of exercises. And so I'm writing lots of things down as I go with, "Okay, that's helpful for defining my audience, and for developing the materials." And I'm really excited, I think it's going to allow me to have my best talk ever.

Richmond Alake: 01:04:40 Nice. One... I'm not going to be in Richmond, Virginia, but I'm hoping the talk is recorded and I get to see it. One thing I wanted to mention was soft skills are very crucial within AI or within any tech role. Most developers should look beyond the technical capabilities and look at soft skills as well. And I actually wrote an article a few years ago, maybe four years ago or three years ago on NVIDIA, called Data Scientists Should Learn Storytelling. Yeah, I think that's what I phrased the title, but it talks about basically storytelling with data. It's a short article, maybe 2000, 3000 words with a bunch of imagery, and it talks

about the same thing, using data to convey ideas to the right audience. And I think I provided a framework as well, as to how people should approach it. But that is all to say it was a relevant topic then, or she wrote the book 10 years ago, so I think that is something, again, that is a pillar, that you can base your career around, which is effective communication.

Jon Krohn: 01:05:53 For sure.

Richmond Alake: 01:05:54 That's-

Jon Krohn: 01:05:54 I used to... When I first started hosting this show five years ago now, four years ago, a very common question that I would ask guests was, what do you look for most in people that you're interviewing? Because I figured a lot of our audience would be interested in hearing what are the key skills that people are looking for, and communication was the first thing people were looking for, probably most of the time.

Richmond Alake: 01:06:17 Exactly, and even talking to LLMs, communication is a big thing. Prompt engineering, it's all about communications, man.

Jon Krohn: 01:06:24 Yeah, we had a funny episode recently with Varun Godbole. It was published... I don't know, I don't have in front of me. One of the interesting things about doing an episode live like this, is I don't want to just be sitting behind my computer, but usually when I'm recording remotely, I have my computer right there-

Richmond Alake: 01:06:40 You can just check it.

Jon Krohn: 01:06:40 ... so I can look up, "This is the episode number." So same thing, [inaudible] or Netflix episode, I don't have the episode number memorized. Varun Godbole, same thing, and it was a really funny thing, Varun Godbole spent 10

years as a software developer at Google, and in the final years he was a core developer on Gemini, so he's an LLM engineer. And he had this really funny moment in the episode, I think it's one of the funniest maybe things that's ever happened on the podcast, was he was talking about how he became a better prompt engineer, as a result of going to relationship therapy with his partner. And I gave him a hard time because I was like, "That is such a Google developer thing, that the main outcome from your relationship therapy was you became a better developer."

Richmond Alake: 01:07:30 That is actually-

Jon Krohn: 01:07:32 Better at talking to machines.

Richmond Alake: 01:07:33 ... very funny.

Jon Krohn: 01:07:33 Nice work.

Richmond Alake: 01:07:33 Oh wow, that's actually good. I feel that's... That was very good. I love that.

Jon Krohn: 01:07:39 Nice. All right, so final question for you, Richmond. How can people follow you after the show?

Richmond Alake: 01:07:43 Yes, LinkedIn is the platform of choice for me, you can follow my LinkedIn. Also, I publish a lot of articles within MongoDB, we do a lot of articles here as well, so there'll be a link to my profile within MongoDB. Yeah, LinkedIn is the most popular way to follow me. And also I do really encourage some developers and developers listening to try out MongoDB for the AI application, and there'll be a link below where you can actually get a free MongoDB Atlas cluster and try it out. Bring your vector data in, bring your metadata in, and let's build a future together.

Jon Krohn:	01:08:27	Nice. Great message there at the end of Richmond. Well said. You're such a great speaker. It was awesome having you on the show and to meet you here in person.
Richmond Alake:	01:08:34	Yes.
Jon Krohn:	01:08:35	Thank you so much. And yeah, maybe in a couple of years we can check in again, see how you're doing.
Richmond Alake:	01:08:40	Thanks for having me.
Jon Krohn:	01:08:47	Nice. What an excellent episode with Richmond Alake. In it, he covered MongoDB's document data model, which uses a JSON-like structure, making it ideal for AI applications, because it offers flexible schema during experimentation, while allowing the schema to be locked in during production. He also talked about vector search capabilities and how they're now built directly into MongoDB, enabling hybrid search, combining both lexical and vector search in a single database. He talked about how 2025 is the beginning of the multi-era that encompasses multi-agent architectures becoming mainstream, multistep problem solving models, multimodal embeddings handling, text, images, video and audio, multilingual capabilities and multiple retrieval mechanisms, beyond just vector search. He also provided the ARENA framework for AI strategy, which involves building on solid pillars with objective truths, following the ACRED principles, aggressive, centralized, resourceful, efficient, and data-driven, and testing tactics in the arena of the market. He also talked about how memory management is becoming increasingly critical for AI systems requiring sophisticated approaches to storing and retrieving information efficiently.
	01:09:59	As always, you can get all the show notes including the transcript for this episode of the video recording, any materials mentioned on the show, the URLs for

Richmond's social media profiles, as well as my own at superdatascience.com/871. And I can't believe it's finally here, if you'd like to connect in real life as opposed to online, tomorrow on March 19th, I'll be giving the opening keynote at the RVA Tech Data and AI Summit, in Richmond, Virginia. There's a ton of great speakers, so check it out. You don't have much time to get involved, so if you live in the Richmond area, that will be tomorrow, the RVA Tech Data and AI Summit, it'd be awesome to meet you there.

01:10:39 Thanks of course. To everyone on the SuperDataScience Podcast team, our podcast manager, Sonja Brajovic, media editor, Mario Pombo, partnerships manager, Natalie Ziajski, researcher, Serg Masis, our writer, Dr. Zara Karschay, and our founder, Kirill Eremenko. Thanks to all of them for producing another fun and informative 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 sponsor's links, which you can find in the show notes. And if you'd like to sponsor the show, you can get the details on how to do that at jonkrohn.com/podcast.

01:11:13 Otherwise, share the episode with people who'd like it and review the episode on your favorite podcasting platform. Subscribe. Feel free to edit our videos into shorts, to your heart's content, but most importantly, just keep on tuning in. I'm so grateful to have you listening and hope I can continue to make episodes you'll 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 SuperDataScience Podcast with you very soon.