Podcast Transcript

The Lion's Den: Demystifying Artificial Intelligence - Episode 3

Rupert Lion, Managing Partner, Boyden United States

Rupert Lion (00:02.222)

Okay, great. So it looks like we're off and running. So I have Sam Jain with me. Sam has had a pretty inspiring career when it comes to data science, machine learning and progressively all things AI. He's been right at the nexus of all of those things for over 20 years now. And we're really excited to have you on Sam and to tell us a little bit about your perspectives on a range of different AI topics. And I think we'll talk about it.

a number of things from your background, including some of the industries you've operated in and perhaps also some of the technologies at play within AI today that you have some interesting concepts and thoughts on. So welcome. And I guess my first question, which is always what I ask most of my guests is, you know, where have you come from? What's your background?

Sam Jain, PhD (00:48.279)

So I graduated from Indian Institute of Technology, Technology, is the top engineering college in India. So even leaders like Google's CEO comes come there. And then I did my PhD at Michigan, and I've been in AI for 20 years.

Rupert Lion (01:06.734)

Great, great. And so I'm guessing this is your passion as well, right? Data science and machine learning and AI and all those things. It sounds like you lived it.

Sam Jain, PhD (01:13.637)

Yeah, of course. I mean, like a working range of industries and really made a difference. I use it for creating $1 billion in sales and enabling 3 ,000 percent annualized profits. So really excited to share more.

Rupert Lion (01:25.454)

Okay, great. So I often start a little bit broad. So excuse me for a moment, but maybe you can just tell me very broadly, how have you kind of used AI to really make a difference in the companies you worked in and the environment you've operated?

Sam Jain, PhD (01:40.869)

Sure, I mean I worked on a range of industries anywhere from energy to manufacturing to capital markets insurance I guess the biggest impact I made was in trading Where I improved annualized win rate from 30 % to 70 % improved annualized profit from 30 % to 3 ,000 % and also in a manufacturing company at Daikin where I enabled 1 billion dollar in sales to continue

Rupert Lion (02:05.902)

Okay, great. And so maybe let's just for a moment talk about capital markets because there's a wealth of data when it comes to anything capital markets. So I'm guessing that that was why there was such a good fit for data science and predominantly machine learning and then latterly AI to support success in those industries. Can you talk a little bit about your capital markets experience and deployment of AI there?

Sam Jain, PhD (02:31.493)

Sure, yes, there is a lot of data there for sure, but you see that alone is not sufficient to create great models. There needs to be a domain expertise as well. And I've seen a lot of implementation of models, but they've not been successful because people didn't know what to use and how to use the data, how to extract features from this information. But I started off in the power trading markets, for example, very simply by understanding market dynamics.

Rupert Lion (02:39.726)

Mm -hmm.

Sam Jain, PhD (03:01.093)

And you think of a two -place marketplace, regardless of capital markets or Uber, it's the same thing. You have buyers and sellers, the more demand, the price goes up. If there are less sellers, the price goes up. Or the reverse, if there are less buyers, the price will fall. So I used fundamentals in the power markets, say temperature or weather or population growth, to determine how the buying increases. And then on the supply side, I tried to understand what...

Where can we get power from, say, wind farms or solar farms or natural gas or other sources? And then use machine learning models to create an equilibrium and say, and this hour we are likely to see a price spike. And then that would be communicated to our traders and then recommended with the probability and then how much probably success we expect and how much capital they should deploy. And that, of course, improved, made it very surgical precision on how.

they knew exactly in which hour to trade and how to trade it. So that really helped them become better in trading.

Rupert Lion (04:09.198)

And that's an interesting one with the weather data, climate data, that sort of stuff. Is there anything that you found out by doing that process that surprised you that you were like, I didn't realize that actually the pricing would change in that way as a result of that particular catalyst?

Sam Jain, PhD (04:25.093)

I would say that in the power trading markets,

it's very hard to predict weather accurately. So we might get a general sense that we might have a cloudy weather in general. But if the solar farm we are particularly focused on, the cloud cover misses it by 100 meters. That's a binary thing. Either you have cloud cover on the solar farm or you don't. So just having a general sense of weather.

is not good enough for us. It needs to be highly precise in location to tell us how much of a power generating capacity we would lose on that plant. So my understanding of granularity of weather changed completely after I got into power trading.

Rupert Lion (05:20.846)

Okay. That's, I think it's one of those things as well where, I guess it depends where you get your weather data from as well, because I'm imagining there's a lot of different sources and so you can feed those different sources into your model, but ultimately you're going to get different results. And as you say, like it might be as binary as the cloud cover misses by 300 feet and therefore you've got a different result. And so I guess with that in mind, were you dealing with lots of different...

feeds and information and having to somehow bring those together within your model. And if so, how did you weight the value of those different inputs?

Sam Jain, PhD (05:57.829)

That's an interesting question and it's a dynamic answer actually. We had to use a general voting scheme in general. You could use that, okay, there's a consensus in weather conditions or we could also use things like recently we are seeing this data feed is being more accurate for this condition. So it was...

both a general rating and also a more, I would say, recency biased rating in terms of which data we would give more weight for certain parameters we're evaluating.

Rupert Lion (06:41.198)

So I guess with that type of work is more the more traditional data science, data science models, I guess, as we think of it today. What about as we progress now into adding things like natural language processing and things like that? How has that influenced some of the sectors you've operated in?

Sam Jain, PhD (07:00.005)

Yeah, that's a very interesting question. And that was actually the reason why as lucrative as trading is, I wanted to incorporate N LP. So a few years ago, I made a shift in my career where I joined an insurance company to not just build an automated underwriting platform, which of course uses classical machine learning, but also we were able to create customer -centric database, which I leveraged to create a Gen AI product for personalized marketing.

So this is in the early, early days when we had GPT -3. I mean, this was even branded as Chat GPT, but still used for personalized marketing where we saw quite a dramatic difference in how customers are responding to it. And we are seeing uptick in email response rate by 2x or even open rate by 2x or response rate 4x, which led to a revenue increase of 4x. So I was very excited to see how it...

like a very nascent technology at that time could still make a big difference.

Rupert Lion (08:01.038)

Okay. And so what do you think that technology was missing that we have today and would have made it ultimately a lot better if you had, if you deployed it today?

Sam Jain, PhD (08:12.005)

So at that time it was still early days. We didn't even know if it would work from a business standpoint. I think there are two parts to your question. I mean, what is the technology missing and how can the business leverage it? Just because you may have a great technology or a substantially better technology doesn't mean that it is usable by business as is. And...

So we're certainly seeing a lot of improvement in the foundational models. We're seeing say GPT 3 .5, 4 .0 and 4 .0 come out now, which is all great. But there are still a lot of of between how that can be productionized. So, Generative A I by its very nature has a known problem hallucination. And we know that Air Canada lost a lawsuit in February because their chatbot hallucinated a bereavement policy.

Google, which is one of the most respected tech companies, Gemini, their chatbot in February, also created toxic images of World War II soldiers. So today people are using things like fine -tuning and RAGs or Retrieval Augmented Generation to circumvent these issues, but there are still production issues in bringing them to life, which I have addressed.

Rupert Lion (09:32.91)

Okay. So let's, let's, back up a little bit to talk about some of these technologies. So I think for some of our listeners, there may not be 100 % up to date and all these things. I, you know, you've talked about chat GPT models, which obviously allow large language, large language models, which are being, which are being used. And then I'm guessing that in the situations you've seen, you've been augmenting them with other data or proprietary data from the business or, or maybe even other models, actually running them side by side.

So maybe you could talk about how those things fit together in the first instance, and then we can talk about probably things like the RAGs and things like that, and you can give a little bit of detail.

Sam Jain, PhD (10:10.917)

Yeah, sure. So while we have a very strong, say, foundation model, like Chat GPT anthropic, or other models out there, they don't have access to any proprietary information that's in a company's database or in a company's literature. So for a customer of a company to get a useful answer, they don't need a general answer. They need an answer from this company. So things like,

They would not, for example, say in an asset management firm, if you're a customer and you ask a question, hey, what are my 401k options? You don't expect a general answer like you can invest in equities and bonds and money market funds, right? You expect a very specific answer in terms of the products this company offers for your 401k. And that information is not available to a foundational model.

So these models have to be, we need a second step, it'll be an additional step after the foundation model to answer your specific question, right? And there are two ways of doing it. One is fine -tuning and one is RAGs. So we can talk about that in detail. But the idea basically is two -step, that you have a foundational model and then you combine it with proprietary data from a company which is not shared with the foundational model company.

It remains within the confines of the company itself. But still, combining these two enables us to address issues like hallucinations.

Rupert Lion (11:46.574)

Yeah. So we'll talk about hallucinations in a second, but I think you've also touched on that important point around the kind of propriety nature of data and the fact that you don't, you know, everyone is scared of sharing their data in the wrong way, particularly if it's, you know, overlaid. I mean, if you look in Europe, if there's GDPR associated with it, obviously not so much in the U S but this is super important stuff, the safety of that data. So it's good for everyone to understand that there is a, there's, there's models out there that actually process, analyze and generate.

but then those are augmented by proprietary data, which is not shared. And I know I'm just making the same point again, but I just think it's a really important one. You obviously mentioned those two things about fine tuning and basically training and things like that as well, I guess. Can you maybe talk us through how that actually works in practice? So you're an insurance company, let's say, and you're using ChatGPT 4 .0.

And you're then applying a whole load of underwriting data about each of your individuals who have got a policy with you. And you're then suggesting to them what they should do in terms of their next policy or their next, you know, I don't know, the way in which they value their home or something. How does that actually work in practice when you bring those components together and fine tune that model?

Sam Jain, PhD (13:08.933)

Yeah, so there are two ways of doing it. One is fine -tuning, which you asked. The other is also RAGs or retrieval augmented generation. So the way fine -tuning works is we take a proprietary model and then we further train it with proprietary data within the confines of a company. Now, this post -training model is not shared outside of the company's bounds.

And that sort of enables it to remain safe. The issue is that every time you update your basic data, you have a lag between when your data was updated versus when your model was updated. And training this over and over on GPUs can get costly. So there's a bit of a trade -off here in terms of you can get really accurate answers, but if...

as long as your data doesn't change. And if you're okay with a bit of a lag between your data being updated and your model answers, then that's fine. And you're okay with the cost part of it. But in terms of adopting the corporate language, the way a company speaks or communicates to its customers, this is probably the best customized way of doing it. But there are sort of cautions that, hey, you have to be...

a ware of the lag part of it and you have to be aware of the cost part of it of training. As long as they're acceptable, then that's okay.

Rupert Lion (14:43.534)

Yeah, the cost part comes out quite a lot in conversation. I think there's to the layperson, there's this panacea of, you know, Gen .ai will do everything for me. But the reality is that even with the advent of GPUs with these kind of, you know, ability to do multiple parallel processes, the reality is that the cost of delivering on that is colossal for the simplest or

what seems to be the simplest of tasks. I remember someone talking about from a customer service perspective that in theory, you could have a extremely intelligent

GenAI assistant that does all of your sales calls for you is like an SDR for example. And it could be really convincing. But the problem is that to do it at that level, the cost of doing that versus the benefit of a slightly increased conversion rate is completely the wrong way around. And I think that's, and I'm sure that's the same in other industries. Have you come across that? Any other specific examples in the industries you've operated in?

Sam Jain, PhD (15:39.109)

Yeah, I mean, that actually is a key consideration, right? I mean, unless you have sort of free or unlimited training power, you just, in my opinion, an ordinary company, it's not feasible. So the approach I suggest for an average company is that we want to fine tune the model once to learn the corporate way of speaking or corporate language.

but not for information retrieval. And if your data is constantly changing, which it is, then you don't want to use this part for it. But for this part, you want to use something, a different part, which is known as RAGs, or retrieval augmented generation, which is a fraction of the cost of fine -tuning if implemented correctly.

Rupert Lion (16:32.142)

So tell us about, because you've mentioned it quite a number of times, retrieval augmented generation or RAGs. For us laypeople who aren't in depth in this stuff, what is that? How does it work?

Sam Jain, PhD (16:42.853)

So sort of like fine -tuning, it's also a two -step process, right? The first part is that you have a foundational model and then you ask a question, like in our case, for example, like what option do you have for your 401k, right? Or what options do I have for my insurance policy? That question is sort of parsed semantically and that information is retrieved in real time.

from the company's proprietary database or proprietary data sources. But then the foundational model comes in and then threads it together to answer your question. The challenge is how is the data that you have in the company, the proprietary data that you have, how is that being chunked? How is that being broken up so that it can be picked up by the semantic search?

engine, right? And that's when I start seeing some production level issues. So devil is in the details here. Every model that you talk about, like machine learning, AI, or otherwise, the question is, how do you measure success of a model? All right? In your classical machine learning, like in trading, for example, we had issues like accuracy or F1 scores or true positive, false positive. But in Gen. AI, well, how do you

The answers that are going to be created in the future are different than what we have today. So the classical metrics don't work. When I have conversations with leaders of the top companies even in America, the answers I've received are things like, well, we let our users tell us. Thumbs up, thumbs down. And in my opinion, that's not a convincing approach.

because your users don't have to tell you an answer and they don't have to give you a correct answer. And even if they give you a correct answer, things that are going to create in the future are brand new. So you have no metrics to measure the success. So that's a foundational part of it. Like how do you measure success of your models?

Rupert Lion (18:59.022)

And how have you tried to do that? Because you must have tried.

Sam Jain, PhD (19:03.717)

Yes, so in my opinion this approach is not a way to do it The asking the user for feedback is not the way to do it So I had to create quantitative metrics for measuring the success of every answer in real time So that's sort of my secret sauce of my success. I created those

metrics to Share are the user answers being is a user is a query being answered completely is a query being answered thoroughly and

Are we using all the data? So that's part one. And the other part of it is in production, I'm also seeing very superficial level of information being answered. So things like, well, where does Rupert live? I mean, that question can be answered by a search on your LinkedIn or your Facebook or Instagram and say, hey, city search and find Rupert city and you can answer it. To your point earlier, does it have value? Yes, but it's like using a sledgehammer to, you know,

hit a nail. The real power comes in asking deeper questions like, say, what's the best selling dish in the best restaurant within five miles of where Rupert lives? That information is not on your Facebook, Instagram, LinkedIn, anywhere. It's not in any one particular place either. You really have to go through multiple sources of data to get it. And similarly in a company,

Companies care about deeper information, telling them things like, how many customers do we have? Sure, something. But what they really care about is, which customer are we at risk of losing in the next three months, and how we should intervene. And that information requires a completely different architecture than what is being used in the marketplace today.

Rupert Lion (20:55.918)

That's an interesting one because I think there's a generative element to that, right? Because that's it's a future prediction about what it is that's going to make that customer happy or sad in the future, or rather stay or go in the future.

Sam Jain, PhD (21:09.157)

Yeah, I think it's that part, but I think it's also the part on how we get to that information. How do you answer which customer is it? And the searches that are being done today are, for the most part, very superficial, like text only. They don't really even dig into a company's

database. They don't really even dig into a document which has text and tables. So my point is,

that the architects that have been constructed today, for the most part, don't even get to that answer. And that, in my opinion, is a major gap in how things are being constructed today from a basic architecture standpoint, which I had to do. That's the first part of it. The second part of it is, it's the first part of it, like the quality of the answer, right? The second part of it is how fast you get there.

Think of typing a query in Google and you get a response in 10 seconds. I mean, do you think it would be an acceptable speed? Probably not. Everyone's used to getting answers in microseconds, a second at best. So that's the latency part of it. And that comes again to the cost part of it. How are you architecting your thing so that the user experience is good, is acceptable? And the way things are being architected today, in my opinion, will not

you have low latency.

Rupert Lion (22:41.199)

But I assume with the architecture point, if you get the architecture right, you can actually deal with both those issues. You can, number one, provide better answers and number two, do it in a way that's much quicker and more effective, right? So you can deal with the cost of processing power and you can also deal with the, I guess, the negative cost of not giving an answer in good time that's good enough quality.

Sam Jain, PhD (23:04.165)

Yes, yes, you are right. I mean, if the right architecture is deployed, but people aren't doing it.

Rupert Lion (23:08.654)

So why aren't people doing it then? I mean that's that's always the question, right? So it's a bold statement is probably true I'm sure it is but but but why aren't they and is there something that's prohibiting them? Is it because there's a legacy, you know, a legacy stack of code and models that they have to use and they're just building on and building on and building on and no one wants to start from scratch or is there something else?

Sam Jain, PhD (23:30.661)

I obviously don't have an insight into what everyone is doing. This is of course my public understanding of material that I come across and the conversations I've had. So I can't conclusively answer your question. My suspicion is the things you mentioned. The architecture that they have is probably legacy or the vendors they're using are legacy. And they are not thinking in terms of

relationships across multiple domains. Even today, people think of a solution for marketing, people for sales, or solution for underwriting. But I think in terms of enterprise level, how are we threading information across the entire company? And rarely do I come across leaders who are discussing at enterprise level how we're creating models that have that

are processing data across enterprise. So my guess to your answer is, the way I would guess your answer to your question is, A, I think it's a legacy architecture that they're using or they have used and they continue to use it. And B, I think they're still thinking in silo terms.

Rupert Lion (24:54.382)

I think it's also the nature of relatively new technologies in that they start off relatively fragmented. That's because very few businesses or companies are large enough and have enough resources, enough money to do a very large scale addressing of the new technology. What happens is you have someone who's left Google, who's done some AI stuff there, will start up an AI startup that's highly focused on a specific slither of the market.

I think over the time what happens is, and we've already seen this, let's face it, those businesses, if they come to fruition and actually have a good product, get bought up and get subsumed into the larger organizations dealing with AI. Then you get a consolidation in the market. I think we're still at that phase of just a highly, highly fragmented industry. It's not to say it's a bad thing because you need all of that innovation in pockets everywhere, but I would guess that's why we don't have a systematized way of

of rebuilding and restructuring that architecture. And maybe it just needs all of the raw materials to be in place first, which bubbles up through that process of fragmentation to consolidation. Does that kind of resonate as well?

Sam Jain, PhD (26:04.709)

Sure, I think that's a fair point. I think you're right, and we'll see how the market matures. There's still very early days in the GEN .AI space. I was also thinking in terms of how is a company storing its data itself? So it's not necessarily a solution architecture sort of perspective, but internally, how is a company storing or accessing its data?

How is it deriving relationships between different aspects? So I've seen or heard stories where a CEO asks a question about, hey, tell me about this customer, and he or she gets six answers because the marketing team has one answer, sales another answer, finance another answer. So, and you really want one answer or one representation of a customer. And I think it's that piece as well.

Rupert Lion (26:59.982)

Okay, okay, now that makes sense. So let's change tack for a moment. You mentioned before hallucinations and I do love hallucinations because it's quite entertaining. So I guess I have two questions on this one. The first one is, you know, have you got some good stories apart from some of the obvious ones that we know about from Gemini and things like that. And the second question is, how can we reduce hallucinations or I wonder whether those hallucinations become...

fulfilling the same function as dreaming does for humans, which is actually a beneficial thing to help to iterate the brain processes and memory and all those things.

Sam Jain, PhD (27:35.813)

I think if you're using it for internal corporate humor, then yes. But I think in general when we are answering customer questions, we really need to be very, very careful. And it's better to be safe than sorry.

our sense of humor might not be appreciated by everybody else. So I think Air Canada was certainly for a cautionary tale and then you really don't want to be telling your customers that, hey, you can have a free ticket.

Rupert Lion (28:20.238)

Well, yeah, that's true. Right. And I think it's funny you bring that up, actually, because obviously one of the things that Gen AI is able to do, but it's still limited, is to really truly understand emotions and intent. And, you know, I think I saw something the other day where you can take one sentence with 10 words. And if you put the emphasis on a different word, the first word, second word, third word, across all those 10 words, every sentence means a different thing.

And I don't do you think we'll ever be able to deal with that? Do you think Gen AI will ever be able to overcome that level of complexity?

Sam Jain, PhD (28:55.365)

I think we're already seeing it. I mean, I think we're already seeing it in GPT -4 .0, the Omni - model model which just got released. So I have not played with it myself, but I've read reports that it is very capable of not just reading or analyzing emotions, but also responding in a humorous fashion. So we're already seeing that.

come out. Of course, it's good to be tested and seen in a corporate environment how it works, but I think the technology is certainly progressing to that point. So there's a bit of a difference here, like in the base technology advancing to that point where it can do all the things you mentioned. And from a business standpoint,

how do we leverage it in a safe, legal way? So I think we'll see a bit of a lag here in terms of can the basic technology do it, but how do we use it?

Rupert Lion (30:10.542)

Yeah, and it's interesting, right, that you see the two sides of the coin in Europe and the US, for example. So in Europe, you've got this bill that says you can and can't use AI like this, and half of it is predicated on something. You don't even know how AI is actually going to evolve. On the flip side, you've got the US, where, OK, admittedly, there are some things in the motion at the moment, but ultimately, there isn't something today. And in that situation, it's great for...

innovation within AI, but it opens up to all sorts of ethical and data -led issues. I guess we'll end up with some kind of happy medium. Who knows? But let's see. I just want to get back to that humor point for a moment. You talked about chatbots earlier and you obviously developed and launched a chatbot a few years ago. Technology has completely changed in Gen AI now.

If you had to launch a chat bot today, what do you, what do you think would be the core things that you would change or add or iterate about those earlier chatbots?

Sam Jain, PhD (31:10.885)

What I would change are things like having a much deeper, richer architecture which can dig into much deeper level of insights for customer, which are spread across multiple data sources. So that's both an architectural design issue and also the vendor selection issue.

is being very cost -conscious. Just because we have GPT 4 .0 or the latest foundation model doesn't necessarily mean it is best for our purposes. That comes to the cost point of it. And what model is good enough? What cost are we paying for the computation? And what is ROI? So I have a very clear focus there. And that's sort of a...

I would say it's a continuous process because every day you see more improvements, but the cost is coming down as well in some cases, or cost is changing, I would say. So we have to continuously keep abreast and check on that front. But that would be sort of my, and the third piece of it is being very, very aware of my metrics. Like I don't want to diminish that part. That is a very, very important part of it, is how are my metrics changing?

in terms of completeness, thoroughness of the answers to ensure that we answer the customer's question, we satisfy the customer, but also don't get to any either ethical or legal issue. Just because it's not a law doesn't mean we still ought to be very cautious on the ethical side of it.

Rupert Lion (33:03.374)

Yeah, I think we do. I think there's a lot of stuff outside of what the technology can do. And it's more about how it can be used and how it can be, you know, carefully watched and considered within the community as a whole and the global community and ethics and all these sorts of things. So, look, I think it's, it's, look, there's obviously a bright future ahead. You're obviously heavily embedded in it. It's been really, it's been fantastic chatting with you, Sam. I think, you know, there's some, certainly some things to think about deeply. I'm.

I'm very curious now, I'm going to be thinking a lot about this architecture point, because I'd love to see whether the larger organizations start talking about their models in that way, that they've actually got to a point where they need to re -engineer that architecture or think about how they work proprietary data. So super interesting. So look, thank you so much for joining us again. It's been fabulous speaking to you and we look forward to hearing more from you in the future. But for now, thank you very much and appreciate your time Sam.

Sam Jain, PhD (33:56.613)

Good much.

Sam Jain, PhD (34:01.604)

Thank you very much. Bye.

Rupert Lion (34:02.734)

Bye bye.