EPISODE 66

[0:00:10.7] SC: Hello and welcome to another episode of TWiML Talk, the podcast where I interview interesting people doing interesting things in machine learning and artificial intelligence. I'm your host, Sam Charrington.

I'd like to start off this show by sending out a huge thank you to everyone listening. We've dropped a ton of great interviews over the past weeks, and through your dedication we continue to see a growing, outpouring of feedback, comments and shares with each release. If you're a regular listener but don't normally send in feedback, we'd really love to hear from you. So please head on over to Apple Podcasts or wherever you listen and leave us a review.

A five-star review is of course appreciated, but what's most important is that your voice is heard. It lets us know what you like or what you feel we can improve on and it also lets those looking for a new machine learning and AI podcast know that they should join the TWiML community.

Speaking of community, the details of our next TWiML online meet up have been posted. On Tuesday, November 14th at 3 p.m. Pacific Time, we'll be joined by Kevin Tee, who'll be presenting his paper; Active Preference Learning For Personalized Portfolio Construction. If you've already registered for the meet up, you should have received an invitation with all the details. If you still need to register, head on over to twimlai.com/meetup to do so. We hope to see you there.

Now, as some of you may know, we spent a few days last week in New York City hosted by our great friends at NYU Future Labs. About six months ago we covered their inaugural AI Summit, an event they hosted to showcase the startups in the first batch of their AI NexusLab program as well as the impressive AI talent in the New York City ecosystem.

We were more than excited when we found out they would be having a second summit so soon. This time we had the pleasure of interviewing the four startups of the second AI NexusLab batch; Mt. Cleverest, bite.ai, Second Mind and Bowtie Labs.

We also interviewed a bunch of the speakers from the event and we'll be sharing those discussions over the upcoming weeks. In this show, I speak with Kul Singh, CEO and founder of Second Mind. Second Mind is building an integration platform for businesses that allows them to bring augmented intelligence to voice conversations. We talked to Kul about the concept behind Second Mind and how the company combines ambient listening with a low latency matching system to help users eliminate and estimated two and half hours of manual searches per day.

Now, on to the show.

[INTERVIEW]

[0:03:01.1] SC: All right everyone, I am here at NYU Future Labs speaking with some of the AI Nexus startups, and I am with Kul Singh right now, founder and CEO of Second Mind. Kul, welcome to this week in machine learning and AI.

[0:03:16.4] KS: Thank you. It's a great time.

[0:03:17.8] SC: Kul, why don't we get started by having you tell me a little bit about your background?

[0:03:20.9] KS: Sure. I think it's worth touching on this, I came to New York actually as a derivatives trader, and so I was doing what was considered now what's considered AI back then. You know what I mean? AI is effectively expected value, and we were doing very intensive Monte Carlos simulations in doing sort of mortgage derivatives and other structured type products. That's where I started, and then I took — Doing that and building the systems — I was on the trading side, but building the systems on that, I got an interesting technology and I started two companies that were in the low latency side of things.

I built two technology companies, and basically one company is actually still operational with the product still being adopted across many enterprises; fortune 100 companies. I have two patents in terms of low latency systems. Then I actually started a lab with a team in Ukraine and in Eastern Europe that was focused on NLP and AI. Sort of built — It's early in terms of now, but they've been doing it for a while, but building early chat bots that are on travel and some other things. I've been doing this for a while, and then I started having interest and idea around this particular company Second Mind.

[0:04:30.8] SC: Okay. What does Second Mind do?

[0:04:34.0] KS: Basically we're looking to solve the time that you're spending on search, and basically we sort of accept the fact that we have to find information and we do have search engines now that's great. The fact is we're now spending about two and a half hours out of our day on average searching for information that's 30%.

If I told you you were stuck in traffic for two and a half hours a day you're like, "Okay. I have to fix my life." But this is what we're doing. This is in 2016, this type of data, so it's not like this is getting worst, the problem.

[0:05:08.7] SC: Just as you say that, like I reject that notion, but it sounds about right.

[0:05:13.5] KS: Yeah. You have to think of it. It's basically as simple as saying, "Are you available for lunch next week?" You're searching for your calendar. Where should we go for lunch? You're searching for restaurants or whatever it may be. It could be for information, like from your CRM or it could be if you're in the trading world, for market data. You're constantly looking for information and the fact is we also don't know what we don't know. That problem takes time.

That whole situation creates this situation where we're constantly looking and searching for data. People are overwhelmed in that thing. We've heard of this information overload, but this actually a different thing where, yes, we have so much data in so many different places and sort of manage it and so forth. We have search engines, but that's still not working. What we're looking to do is say, "Hey, can we solve that problem specific to conversation?"

Because the fact is the input for many of our searches are driven by a conversation, just as I gave examples for. Clients are saying, "Can you have that information?" The fact is when you're no a call or an in-person conversation like we are now and someone asked for that, you're either having to multitask, which is obviously stressful, or you're saying, "I have to get back to you later on that." What's worse when you have clients here to say I have to get back to you, someone else is going to close that business.

From a business standpoint, this is a very costly problem, because of spending, you know, not only the money to employees. It works out, it's about from an $80,000 average to about $14,000 annually that's just gone and wasted. On top of that —

[0:06:46.1] SC: That's two hours a day of searching?

[0:06:47.4] KS: Exactly, that type. Then you're also losing the opportunity cost in terms of the revenue of not closing the clients when you have that information. That's the problem we're looking to solve. What we basically have developed is you can think of this matching engine. I'm sort of using a Wall Street prerogative here, but basically think of it as a [inaudible 0:07:04.8] exchange with you for stocks, matching buyers and sellers. What we're doing is we're identifying events in a conversation and mapping it to any data, whether it's internal or external and pushing it to people in real time.

That can be information from you CRM, from your market data, from, say, Bloomberg or Reuters. It can be from your email or files. If someone says, "Hey, did you get that file I sent you?" It show up right away. You don't have to sit there and fidget. The typing part takes time. That's sort of the thing. The way you can think of the product, it's sort of like you can think of like a Slack for conversation or the glue to provide all those applications to people and developers and companies can tie in, and so to feed as you're speaking.

[0:07:50.8] SC: You're using conversation and speaking as the way you describe these interactions, but then you reference Slack. Is it primarily for textual communications or are we literally talking about speaking?

[0:08:04.2] KS: Our focus is around voice conversation. That's sort of where our core IP is. We can do regular conversation, I mean text and chat type conversations as well. Where we see the real opportunity and the unique differentiation is around voice conversations.

[0:08:18.3] SC: When I think of the way we've envisioned some of these interactions like from a sci-fi perspective and kind of this notion of the augmented human. It's like you want something that's kind of passively listening to all these interactions that you are having and then just making you smarter by popping that email up or, "Oh, you just mentioned lunch. Here are three places that fit the profile, all the kind of place you like to go to." Without having to pull out the phone and the keyboard and all that kind of stuff.

[0:08:49.2] KS: Exactly. That's sort of the big vision we have. It's around the AR side. Sort of what people can be effectively J.A.R.V.I.S and Iron Man where you can show that. I mean that's not that far-fetched from our standpoint, because it's just a screen from our standpoint. We can display that information on any screen and basically it's just what we're running as a low latency matching system to provide that data in any screen, it can be in your phone, your mobile phone, it could be on a desktop. If we can connect with an enterprise void, it can be on video calls or it can be in AR.

[0:09:20.1] SC: Okay. Do you end up taking a position on kind of the pervasive listening aspect of that?

[0:09:28.0] KS: That's a great question. One of the unique IP aspects of our product is that we don't need to record conversations. Optionally, we can. A lot of the companies that sort of are in the voice type space are recording data. What we've really — Again, it goes back to sort of my background with sort of like in-line streaming of this data throw away processes and throw it away. In fact, our core data that we're capturing is the metadata of the mapping engine, of the maps. That's where we're really focused on.

We basically are looking to process inline. Everything is in memory. Where a lot of the companies are saying, "Hey, we're building out this unique part of the components and so forth." Our unique things that we're doing is focused on the real-time low latency on the frontend, not on the training side so much, but on the frontend where very few comes.

Basically, you can think of them like — Again, the best way to think of it is like a high frequency trading system or what you'd see on Wall Street where you're processing trades very, very fast and where it's looking at caching that data very fast and processing and mapping it very fast to that conversation. That’s really where the unique sauce comes in. We don’t need to record the conversation [inaudible 0:10:39.8]. Analyzing it, capturing the events, and we will store the events. That could be like data that could be saved in the CRM or so forth, or summary, but it’s not necessarily the actual conversation.

[0:10:51.6] SC: Okay. I get that you’re not required to record the data in order to do the things that you’re trying to do, but a lot of companies aren’t recording data because they need to to deliver their service. They’re recording their data because they use it to train and make their algorithms smarter.

Kind of flipping it on the other side, if you’re taking a hard position instead we’re not going to record any data, do you miss an opportunity to produce a better product. Is that why?

[0:11:22.3] KS: To make it clear. There’s the recording of the audio which we’re not focused on. Where we are focused on is we are storing the metadata, which is the mapping piece. That we are maintaining and that we do need to know. Basically, from that standpoint we can then see where our engine is getting false-negatives, false-positives and so forth. We can see where we are mapping the right data or the wrong data so forth. For us, that’s why I keep using this analogy of we’re a real time mapping engine or matching engine.

Let’s use the analogy, if you go to whatever you might use for an online trading system and you go, “I want to buy Apple,” and it gets you Google. You’re not going to be happy with it, right? That’s what we’re looking to ensure we don’t do that. It’s different than recording someone’s conversations, storing those conversations.

Where a lot of times they’re focusing on the analytics and storing that [inaudible 0:12:21.3], which then becomes a — Actually, there’s a legal aspect to it. You have to tell people, “This conversation is being recorded and monitored.” It’s a law across most states. Because we are not recording or monitoring those conversations, it’s just being analyzed in-line and done in a way effectively, we don’t need to even say that [inaudible 0:12:40.6].

[0:12:42.4] SC: You take in the voice, and then are you converting that to text? Then your metadata is based on a text transcription of what the person said?

[0:12:53.7] KS: Then we're identifying keywords or key phrases and then we're building — Basically, our chief scientist had written some papers on this. He was most recently a professor at an institute at Budapest where he wrote some papers around the concept of attribute variable matrices and doing a real time.

Basically, the idea would be in a conversation is not always a linear process of finding and achieving when an event is achieved. When I say an event, it means when an invent is triggered. If you think about it, if I go to buy a train ticket and I say, "Okay. I want to buy a train ticket." Okay, I can't buy the ticket yet. The person is asking where are you going and what data, what time. Now, I can start maybe buy the train tickets. That triggers now finally gets me how many people and so forth, right?

In the same way, we are doing the same thing where we basically set up this matrices who says, "Okay. Let's identify from that text, we can start filming out their matrices and is there an event triggered?" Once that even gets triggered, it could be one keyword. It could be a bunch of keywords that we need to hear.

For example, it could be on a trading floor, it could be how many share are traded. You can't maybe trigger there, but if someone said, "What do you think of IBM?" Then said, "How many shares traded?" Well, now we know for IBM, right? Or someone says, "How many shares of IBM traded." We can then find that.

That in that vein, we don't need to record the audio or save the text, we just need to capture the face that someone knows and how many shares and what stock. We can capture in any order and then when it's triggered, we map it to the data so that we know where to find it.

[0:14:38.4] SC: I guess the personal assistant type of use case that we discussed is kind of a clear one. Are there other use cases or what's the initial target customer and use case that you're going after for this?

[0:14:54.8] KS: Probably, as I said before, there is that knowledge worker general that you said, we're seeing interest from the financial services vertical, financial advisers, traders, where there's a need for having that information. Sales people having to get information, close that sale and having that data, managing that information and so forth. Yeah, basically.

Any mark, if you think about a real estate, someone's calling up. They can't manage all the properties in the MLS listings, right? If someone says, "What property is available in the west village here in New York?" They're not going to know. We could push intelligently what those properties are based on some criteria that person says. You can now deliver that information intelligently.

One area we're seeing, we're sort of working with a data provider and exploring is around attorneys. They're on the phone all the time. They're in conversation all the time. They're constantly trying to get information on precedents and other legal facts that require so much search and data. The more we can push that to the person as they're having a conversation. Well, now you don't have to have that asynchronous process of, "Let me get back to you and let's set up another call."

Think about the time it takes to set up another call, get on the call, that process of a few minutes of talking, of, "Okay. Chat, chat, chat. Okay, now let's get to it." Then, "Oh. I don't know that answer. I'll write that down. Get back to you." That constant process. We are trying to streamline that where you can now take maybe two, three, four calls and you can make it into one, maybe two.

Basically, now, significantly reducing the amount of time that people need to be on those calls and searching and you're basically making people much more efficient, productive. As you sort of said, hopefully, a dream of being super intelligent or smarter than otherwise.

[0:16:43.7] SC: We've talked a little bit about the frontend of this process. How are you handling the kind of knowledge retrieval aspect of it on the backend. It strikes me that there are a number of interesting challenges there that you have to overcome.

[0:16:57.4] KS: Yeah. The first thing we're doing is, in terms of our core product, we are offering it to — We're initially focused on enterprises and we're offering that as a virtual machine. We offer — Again, this goes back to some of the IP and technology I have built in pervious companies, which is sort of like a very much a publish-subscribed real-time architecture. I'm sort of — One of the things I've always loved about building companies is you find the sort of the gap between two areas and you sort of focus on there, because if you focus on just machine learning and so forth, you obviously have very big players that have a lot of expertise. You focus on, say, publish-subscribed in real-time where you have Tipco and IBM that are very big players in there. But if you get in the gaps, you can sort of build a very exciting company in the wedge.

What we're building is sort of like a wedge company that's in between those markets which is saying, "Hey, we're building a published subscribe event based technology. It's using machine learning and a lot of the capabilities around there." What we're doing there is to — We're leveraging what we call like a high performance caching engine where companies can open data sockets from any database or system to our product and we basically have both an in-memory cache as well as a cache that could sit in [inaudible 0:18:18.9] companies would be able to access data in real-time and we've got to show it.

It's not a major integration process. Basically, it's just opening up a data sucker to our caching edge, and then we would pull a lot of this between this speech recognition, the NLP and the cache all learning a network.

[0:18:37.5] SC: On this backend side of things, what are some of the data sources that you envision your early customers connecting to? Is there like a top three list of their email servers or something else, or do you expect it to be more specialized, maybe proprietary systems or databases?

[0:18:57.0] KS: Yeah. Great question. What we've done for building the initial prototype is we did the Dropbox, Google Drive, Gmail, Yahoo Finance to be able to show that. Right now, you could have a conversation on our product and you ask for the price of Amazon, it shows up. You ask for an email on X, it shows up.

We're also working on exchange integration. What we're doing is there are certain integrations where they're more broad across many enterprises and doing a very tight custom integration where we can find information very fast. Exchange integration gives us both enterprise emails as well as a calendar information. That's awesome.

If you say, “Hey —” When I give you an example; when are you available? Boom! Who was on that call? Who’s going to be on that call? You can bring in all that information and show that person. You don’t have to sit there and fidget your email, your calendars, whatever. A lot of enterprises now are using things like bots and Dropbox. That would be a more general type of integration that we would offer.

Where we see, for our example, for financial services. Conversation with Reuters and market data, on Bloomberg, sources there. Same thing with legal, like Reuters or Lexix Nexis are major players. Salesforce would be an integration on the sales side. MLS would be an integration for real estate. That’s sort of the — The integrations are really not the focus. Really, it gets to being — Make sure to get the language corpus for each one so you’re getting good matches of data.

[0:20:25.7] SC: Have you tackled that challenge?

[0:20:29.0] KS: Like I said, we’re focused on initially a couple of verticals. We’re not going to target the entire marketplace and we have customer interest in both side. We’re actually in the very close releasing an enterprise version of produce here within a couple of weeks, so we’re excited about that.

Then we have — These are fortune 50 type companies that are going to be demoing the product to and go from there effectively. Then in terms of — The goal is to expand the capabilities to a little more the cognitive side of things where we can — This is sort of why we joined the NYU’s program, so we can work closely with some of the professors in doing some of the things around building out our API so we could do a little more, not have to be so limited to building at each vertical, but you can do quite a more of a broader solution for multiple type of verticals.

[0:21:22.4] SC: Okay. Cool. What’s next for the company?

[0:21:26.3] KS: As I said, we’re excited about releasing this product here and getting these customers teed up. It’s one of these things where it’s fun to show the product, because it’s sort of magical. It’s also — You want to move faster, but you have to obviously go sort of a certain course to sort of achieve those results.

Basically, it’s first completing this next, what we call the beta of the enterprise version. Get these customers using the product and using it and then scale from it.

[0:21:56.8] SC: Great. I really appreciate you spending some time with me. It sounds like you guys are doing really interesting things and I’m looking forward to keeping up with you as the company evolves.

[0:22:06.0] KS: Thank you. I appreciate the time as well.

[0:22:07.7] SC: All right. Thank you.

[END OF INTERVIEW]

[0:22:12.8] SC: All right everyone, that’s our show for today. Thanks so much for listening and for your continued feedback and support. For more information on Kul, Second Mind, or any of the topics covered in this episode, head on over to twimlai.com/talk/65.

To follow along with the NYU Future Labs AI Summit Series, which we’ll be piping to your favorite podcatcher all week, visit twimlai.com/ainexuslab2. Of course, you can send along your feedback or questions via Twitter to @twimlai or @samcharrington or leave a comment or write on the show notes page.

Thanks again to NYU Future Lab for their sponsorship of the show. For more information on the AI NexusLab program, visit futurelabs.nyc. Of course, thanks again for listening, and catch you next time.

[END]