accenture

AI LEADERS PODCAST: DATA-CENTRIC AI

VIDEO TRANSCRIPT

ANDREW NG: What has been emerging is more systematic ways to engineer the data.

SANJEEV VOHRA: Hello, everyone and welcome to this episode of AI Leaders Podcast. I'm Sanjeev Vohra and I'm Global Lead for Accenture's Applied Intelligent Practice. And I'm your host today. Well, today is a special day and we have a special guest to interact on an emerging topic in the field of artificial intelligence. So let me take an opportunity to introduce our special guest, Andrew Ng.

I'm sure most of you know of him. He has been the Founder and Lead of Google Brain, adjunct professor of computer science at Stanford University, chief scientist at Baidu, Founder of Coursera, deeplearning.ai, and Landing AI and he's been doing a lot of work in the last few years on AI and the evolution of AI. And we will be discussing about something that he shared last year, which is Data-centric AI.

Andrew, thanks for joining in. Thanks for your time. I'm sure that people will - this session will be useful for a lot of us. But before I deep dive and dig in, are you up for a quick warm up questions?

ANDREW NG: Sounds great. Let's go for it.

SANJEEV VOHRA: So tell me, what's your favorite genre of music?

ANDREW NG: I'm going to sound like a nerd. I actually listen to a lot classical music. Right now, one of the pieces I love is Pachelbel's Canon. I'm trying to figure out how to play a piano adapted version of Pachelbel, and I'm not doing very well yet.

SANJEEV VOHRA: Andrew, let's goes to my next question. I mean I think I know the answer, but maybe it's good for audience to know. But are you a night owl or early bird?

ANDREW NG: Night owl. I know that I've have early bird friends that emit a feeling of moral superiority for their ability to wake up at 5 am. Unfortunately, I'm not one of them. When I was younger, the computers tended to be more free at night. And so, you just had to – staying up at night meant you could run lots of compute intensive jobs when everyone else is asleep. So your jobs could run faster. So I find a lot of my friends in tech all wound up being night owls because we've got into habit of staying up late to run our compute jobs when everyone else wasn't using the servers. And so, our jobs would run faster.

SANJEEV VOHRA: Let's move to the content then, Andrew. And I thought and if I see that, you know, we haven't always been talking about data being the food for AI, data being the source for AI, without AI there's no data. So when I heard you first time talking about Data-centric AI, the question was what's new about it? So, Andrew, can you elaborate what do you mean when you say Data-centric approach and how do you think about that approach, how it's different than traditional approaches?

ANDREW NG: Yeah, so Data-centric AI is an emerging discipline of systematically entering the data needed to build a successful AI system. I think we've known for a long time that data is important for AI. You know, there's big data and people have been fiddling with the data and improving for a lot of applications. But what has been emerging is more systematic ways to engineer the data. So rather than, for example, in a lot of computer vision or machine vision tasks, sometimes the data is mislabeled. And what used to happen is we would count on the skill or lack of some data scientist or some machine engineer to maybe find a problem with the data and maybe clean it up or maybe not. And what has been emerging is more systematic tools that can tell you this part of your data is good, this part is problematic, so that instead of counting on the skill or lack of individual really good engineers, we can come up with software tools and maybe checklists or ways of approaching the problem, so that lots of people can successfully engineer the data for the AI systems

. And what's new is a lot of academic research in AI has focused on building models, inventing new neural network models. All the stuff you see in the news is someone invented a new piece of software or a new AI model. But with the data centric approach, I and many of my friends have been calling for more attention not just on the software, but on improving the data, because for a lot of applications, there's an open source machine learning model that work just fine. It's going to be even more productive to engineer the data that you feed into the software.

So that's been the shift in the mindset of a number of people, even if it's something that practitioners have been doing a little bit by hand for a long time.

SANJEEV VOHRA: And are you expecting people to use some sort of other technology or skills to kind of spend more time and in cleaning and curating the data sets? Is that the approach?

ANDREW NG: Yes, and I hope we will get there. I think we're in the early phases of coming up with making those skills systematic. And then also mix the building tools. Maybe to make an analogy, about ten years ago, 10, 15 years ago, started the rise of deep learning. And at the start, a lot of people were a small number of people that do it to build new networks. And then after a while, more of us started publishing papers and than more people could do it by implementing your networks and C++, but it was still very error prone. But the principles became better known, and then eventually there came tools like TensorFlow and PyTorch that made it much easier for a lot of people to apply deep learning models.

I think the evolution of Data-centric AI development is an earlier phase, with more and more people publishing papers that got Europe's workshop, we ran in 2021. You know, the principles are becoming a little bit more widespread, and they're just a small handful of teams, including the in AI, but a small handful of teams, they try to invent tools that then a lot of people could use to make the Data-centric approach more systematic. And I think that for many faceted applications, entering the data is just

going to be a more practical way to get to good performance.

SANJEEV VOHRA: And I think a lot of people, obviously, I mean the two problems that I see in the industry when we deal with the problems around data, right, and I just wanted to get your reaction to that. The first one was basically, in some cases, there isn't much data anyway. When we go to core of things to the problem, the data doesn't exist to that extent that is needed to create an approach or to create the right models for that data set. So I just want to ask you, is this approach also considers element of synthetic data creation as well? Like here we don't have data, I mean are you giving any guidance around that piece and considering creating data sets.

more time now quantifying data then actually just working on the best model they can produce?

Do you see the shift of knowledge and any comments on what sort of skills they need to double up or not?

ANDREW NG: Yeah. Yeah. It's been interesting. Over the last decade, with the rise of deep learning, the field's gotten more complex. So if you like, today, understanding the say beyond models, that is still important. So people should - to build a cutting-edge model, you do need to know whether the cutting edge in their architecture is and pick the right one relatively well.

Having said that, what I see is that for a lot of the applications, there will be some open source, new or natural architecture that you can get for free. It's soon to be licensed and work just fine. And what that means is finding the right open source model, then tuning it, which is important. You've got to understand that piece, but that doesn't take as long and now we can spend more of the time. so what you said, Sanjeev, of qualifying the data, fixing the data, generating more data, that results in the best AI model.

And then probably one high level, one thing that you have chatted a lot about over the last year or many, many months, still is matching these AI models to the business application and closing that gap. That's the real remains a struggle.

SANJEEV VOHRA: Yep, well, that makes sense. Completely makes sense. And I think now what's happening is that some people and some of our leaders in the industry, they obviously are really more and getting more and more understanding the large language models and what's happening in GPD3 and likes of that and they started now looking at making this time to understand more about how they can use that effectively for business applications or business models in the business way, how we can get value from there and investment in that, but what is your view on that? And given the tools that we just mentioned, does it make sense to make a larger investment that those tool sets and those technologies? Or is it important just for us to understand what the need? I mean what sort of recommendation for people as to and how they should approach their investment profile and the roadmap for creating their AI board in the companies?

ANDREW NG: I think for many businesses with a reasonable but not very large AI budget, I would tend to focus on starting the use cases and thinking through the applications and the ROI. I think the large language model is a great technology. I have deep respect for the work that open AI idea and like GBC and various competitors and alternatives thereof of these large angles. It's a wonderful technology. Definitely encourage anyone doing basic research to look at it. Hopefully, lots of people will keep on investing in that and move the whole thing forward.

I think it could be a great technology for some applications. I see teams using GTV to adapt it to make chatbots who cover things or summarize documents. The fixed (inaudible), but to most business-oriented teams. I'd be cautious about making a big investment in technology, of a very specific technology, prior to fleshing out the use case and the ROI. But I do think the use case and ROI could be there for some businesses. So maybe I think you and I probably have a sort of our philosophy on this, I think really exciting technology. But that for most business-oriented teams, you know, starting the use case, thinking through the ROI. That could be a more promising approach to making a bigger investment.

SANJEEV VOHRA: That's good. And I think I go back to the questions that I may have asked earlier. But let me

just rephrase it. These people have to go to this Data-centric approach, which means they should invest, invest time, disproportionate, let's say, investment in the area of making sure the data is rich and done well, so that AI can actually create that level of performance and output that they're looking at. Is there a framework or is there an approach that you would suggest how to actually build this capability? Like is there a method to the madness of how to do this?

ANDREW NG: Yeah. So I would say common advice to chief AI officers or chief data officers is start small, run a relatively small pilot project, put one thing into production, rather than trying to build something grand right away. I know that many boards will say, give me a whatever your plan, then I'll approve the budget, then you can go and do it. But I find that for most organizations, building up the organizational capability happens better by doing smaller projects, so that the team can start to train, hire, build capability, and only with growing capability to then if that becomes more able to think through how to keep on growing this ambition.

So I've seen more companies fail by starting to big than fail by starting too small and by doing one or twoprojects that often helps you make better decisions as well about what further investments to build. So rather than spending X tens of millions of dollars building the data infrastructure, hoping it will be useful for something in the future. Instead, I think if you find even a small use case and get that done, then that actually tells you what's the most fruitful way to invest to improve your data infrastructure and improve AI team and improve these capabilities, and then to gradually step up the size of the investments as well as the capabilities.

SANJEEV VOHRA: Thanks. And, Andrew, this is great. I think this was very useful session. I just want to ask you one more question. What are you working on right now which is exciting? Would you be open to sharing with all of us?

ANDREW NG: Some ideas about I'll share about. I met AI teams in a couple of the large consumer internet software companies. I think that was and remains the exciting thing to do. But what excites me a lot now is how to taking out all of the other industries, because it turns out the recipe for AI adoption in consumer software internet companies, which have a hundred million or maybe a billion users, is quite different from what is needed in other industries. In consumer software internet, because you have so many users using a homogenous, relative homogenous product, you can build one monolithic AI system, like a web search engine or speech recognition system that serves a lot of users and the economics for that itself. But when you go to other industries such as manufacturing, where it's not about one monolithic inspection system, it's about for each of 10,000 factories, each of which makes something different. How do you help each one of them build the custom AI system or in healthcare? Every hospital has a slightly different way of encoding medical reports. So rather than a monolithic AI system, I think every healthcare system or maybe even every hospital will need their own AI system for building their own, you know, for leading their own healthcare records.

So the thing that excites me is how we build tools to enable the customers to do their own customization. And I think that the Data-centric and AI technology, which Landing AI is working on, I think that would be a key part, maybe the key part in the recipe because of easier for the IT staff and manufacturing and furthermore, healthcare, IT or something else to engineer the data in a way that lets them express its own main knowledge to then help them get the custom AI system that they need.

So, I think Data-centric AI, it sounds simple maybe, but I think it's such a very hot technology. So many days I find myself debating the engineers, really find nuances on how to move that technology.

SANJEEV VOHRA: Thanks very much, Andrew. It was great pleasure talking to you. And I think we share a very common perspective on this topic, as we really want to get to a place where the technology can be used for business application getting value. And I think this approach that you just mentioned is much more pragmatic approach of taking the step forward in using AI and just not fall to huge investments in both technology itself without looking at the business problem and the data, which is a food for AI, as we always have talked about.

So, I think this is a very, very good discussion. I love discussing with you. I love hearing your views again, and I hope you enjoyed it and stay tuned and please do not forget to subscribe. Thank you very much for your time.