SUPER DATASCIENCE
MAKING THE COMPLEX SIMPLE

SDS PODCAST EPISODE 912: IN CASE YOU MISSED IT IN JULY 2025

Jon Krohn: 00:02 This is episode number 912, our In Case You Missed It in July episode.

00:19 Welcome back to the SuperDataScience Podcast. I'm your host, Jon Krohn. This is an in case you missed an episode that highlights the best parts of conversations we had on the show over the past month. We'll start off with a conversation I had in episode 901, in which I asked Lilith Bat-Leah, why data centric Machine learning research or DMLR has become a byword for accuracy in the field of legal tech. The impetus for having an episode when we talked about it on the train already a year ago, was this idea of data-centric machine learning. And so this is now a topic that is, it's not, this isn't just like, oh, there's some analogies here that might be relevant to your industry. Data-centric ML is relevant to every listener, anybody who's working with data, this is relevant. And so tell us about data-centric machine learning research DMLR. And my understanding is that you fell into DMLR as a result of how messy the data are in the legal space.

Lilith Bat-Leah: 01:25 Yeah, that's right. So in my first r and d role, I was really focused on algorithms and on finding the best classification algorithms for these classification tasks that we've discussed. At a certain point, I realized that the labeled data I was working with was so noisy, just had so many mislabeled instances and all of that, that it really curtailed my ability to evaluate the performance of the algorithm just because I couldn't necessarily trust my data. So that led me to be very interested in what Andrew Ng coined data-centric ai, and I ended up getting involved with a working group at ML Commons called Data Perf, where we were looking to benchmark data-centric machine learning. That ended up leading to a few different workshops that we've organized at icle and ICML data also became a NIPS paper.

02:44 And yeah, basically it turned into a whole community. So now there's A-D-M-L-R journal, there are the DMLR workshops at these conferences, and then data proof morphed into the data-centric machine learning research working group with ML Commons. So we have a lot of different things going on. We're working in partnership with Common Crawl, the foundation that curates the data sets that most LLMs have been trained on. We're partnering with them on a challenge that will result in a low resource language data set that will be publicly available. So if you're interested in joining the working group, please do get involved. Again, it's with ML Commons, you can go to that site and send the working group.

Jon Krohn: 03:35 We'll be sure to have a link to ML Commons in the show notes. And so when you say a low resource language, this is languages for which there are not many data available online. They could be rarely spoken languages or for whatever reason, languages that even if they're spoken relatively commonly, they aren't represented on the internet.

Lilith Bat-Leah: 03:57 Exactly, exactly.

Jon Krohn: 04:00 Nice. That sounds really cool. And so those acronyms that you were saying there earlier where this DMLR initiative was getting traction, so conferences like icle, ICML, nips, these are the biggest conferences that there are academic conferences that there are. And so really cool that you get such an impact there. And it's also interesting to hear the connection to Andrew Ng there because I have in my notes here somewhere, I'm kind of scrolling around in here. Yes. So at the inaugural DM LR workshop, Andrew Ng was the keynote.

Lilith Bat-Leah: 04:33 Yes. Yes, exactly. And he was involved with data perf as well. He's on that data perf paper.

Jon Krohn: 04:40 Okay. So I'm now very clear on the importance of DMLR, the traction it's getting and bigwigs like Andrew Ng being involved. Probably most of our listeners know who Andrew Ng is. He's one of the biggest names in data science period. And if you aren't already familiar with him, he was on our show in December. So episode 841, you can go back to, we'll have a link to that in the show notes as well. So yeah, so now I have a clear understanding of data-centric machine learning being very important, gaining traction, but our listeners still might not have a great understanding of what it is.

Lilith Bat-Leah: 05:18 Yeah, so the best way I can explain it is that in traditional machine learning paradigms, you're iterating on the model, you're iterating on the model architecture, on the learning algorithm, all of those sorts of pieces. And that's where you're really focused on improving performance is by iterating on the model. With data-centric machine learning, you're iterating on the data, so you're holding the model fixed and you're improving the data. You're systematically engineering better data. And then there are all these different questions. So there's the question of whether to aggregate labels or not. There's a really interesting paper dore me that looked at waiting different domains of the pile to get the best LLM pre-training performance. So it can go lots of different ways. There's another paper I'm thinking of, I can't remember the name, but they looked at selecting the best data points for training a model a priorize. So not even active learning where you're starting with the results of the model to determine which additional data points you should have labeled, but just with a dataset from scratch using linear algebra to figure out which data points are worth labeling

Jon Krohn: 06:48 From DMLR. We turn to AI benchmarks in episode 903. In it, Sinan, Osmer and I discuss approaches to circumventing the limitations of benchmarks. You right at the beginning, near or near the beginning, I was talking about benchmarks. You talked about contamination. And so what is the resolution there? This seems like a really tricky problem. How do we prevent leaks once a benchmark's been out and the answers are online? I mean, I guess one solution is to just not have answers online.

Sinan Ozdemir : 07:23 Tell that to the internet. Well, because the thing if a benchmark literally comes with the answers, that's the whole point of the benchmark is you're supposed to know the right answer. So the same place where you get the questions for the benchmark also has the answers to the benchmarks where you can validate that it's correct. So it's impossible to not have the answers, not on the internet.

Jon Krohn: 07:45 You have something like it could be like Kaggle. Exactly,

Sinan Ozdemir : 07:50 You could, but then who owns it?

Jon Krohn: 07:52 Who owns?

Sinan Ozdemir : 07:53 Who owns someone has to own it. Well, someone has to because if it's going to be hidden from everybody else, someone now is in charge of holding those answers. So who is it?

Jon Krohn: 08:03 The developer I guess, in kind of the same way. So okay, here's an interesting idea. So what about a solution like Chatbot Arena? Where in chatbot Arena, there's no correct answer necessarily. So it's run by Berkeley, the LM cis lab, if I'm remembering correctly, I think it's Joey Gonzalez's lab. And so Joey Gonzalez has actually been on this show talking about it. If I can find that episode quickly. Yes. Episode 707, you can hear from the Berkeley professor that it was in his lab that this chatbot arena was devised. And so in the chatbot arena, it's different from benchmarks in the sense that you don't have a specific set of questions and answers. You pit two LLMs against each other and you as a human evaluator of arena, you don't know which two you're seeing output from, but you pick one as better than the other. And so first of all, I'd love to hear thoughts on the arena, but the reason why I'm bringing the arena up is that in that situation, I mean, so you're talking about ownership, you could have a similar kind of thing where for a benchmark where somebody creates a training set like humanities last exam, you could have a holdout answer set. And yeah, I mean some like a university like Berkeley could be administering it. People submit, people submit their responses, and then they get a grade back.

Sinan Ozdemir : 09:34 Yeah, a few things. I'm a fan of the arena in general. The idea of blind judging from a human for me is one of the best ways to really get a good sense of an LLMs usability. Now, a couple things caveats there. If I'm just a lay person talking to a chat bot, to your point, I'm not coming in with structured questions. I'm just going to pick the one I like the most. And that might come down to which one's talking the way I like it to talk, which kind of leads to the whole S of fancy thing, right? OpenAI said, well, we rely too much on people's thumbs up and thumbs down, and that's what got us in trouble. The yellow marina is pretty much a thumbs up and a thumbs down. It's all we're really doing is saying, I like that better. I'm not telling you why, just because it cursed once and I thought that was cool.

10:23 We have no idea. And sure, over a at scale, when you aggregate these, you'll get a much more stable answer. But again, at this point, we're just judging preference as opposed to knowledge and again, without that structured dataset. Now also, I think you mentioned this, there is no answer to any questions on the arena. You are just shown response like you are not coming in with a question, you are just shown answers, and it's up to the human to decide which one is correct. So whoever is judging it behind the scenes, how are they doing it? Are they paying a human being to read each one and actually comparing it to the right answer? Or are they going the LLM as a judge route where they're saying, well, we have yet another LLM who has given a reference answer and this answer, and it's asked to say, how closely does it compare?

11:16 We don't know. And again, a lot of it just comes back to what actually is the right way to judge this? Who has the right to judge whether or not the AI was correct or not? That's a big question. And again, that's why we have benchmarks is that is our current proxy to that question, which is, well, if we all agree that Pablo Picasso painted this thing, and that's one of the answers they can pick from, it's on the right track to knowing general world knowledge. But if it just comes down to which one do you like talking to better, like an arena would be, you're going to miss a lot of the actual important pieces of information you're trying to get out of that LOM. I'll say one more thing. It's funny you brought up the arena. That's actually one of the allegations from law before, again, total allegations. But one of the separate allegations from law before was they released a tested or a trained to test model specifically for the arena that was different than the LAMA four. We all got in the end, again, total allegation, but those rumors start bubbling up when people notice discrepancies and who's to say those discrepancies are correct. They're all just our own interpretations and our own expectations maybe not being met by what we were shown. There's no way to prove this.

Jon Krohn: 12:42 Finding the right way to judge a system is a huge question and is one that I am sure will continue to perplex and challenge us over the years. In episode 905, my guest, Dr. Sebastian Gamon and I continued this conversation specifically regarding ways to select models for the domains we might be working in. There's a second paper that you also recently published. So your first author on a paper that was submitted to archive in April called Understanding and Mitigating Risks of Generative AI in Financial Services. So mostly so far in this episode we've been talking about generally how models fair into Reg, but in that paper it's related to risk of gen, ai, finance. You emphasize that most foundation models are not trained on finance specific corporate bodies of knowledge. So what are the limitations this creates for LMS in general, but particularly for rag? And I'm assuming that this same kind of sentiment, you looked at it with finance specifically because Bloomberg as a financial services company largely, but do you think that the same kind of limitation would apply in other sectors as well?

Sebastian Gehrm...: 13:58 Yeah, absolutely. So yeah, I gave a little bit of a teaser of this paper earlier in an answer as well. And the way that we wrote our paper very much should be seen as a case study finance here or financial services in particular. Capital markets and asset management is the case study that we use to make the point that we really need to think about risk and risk taxonomies and risk management in our domain in what we are trying to build. And as you say, we made the point, yeah, models are not necessarily trained on financial domains. We see that both in helpfulness and the homelessness angle. Often complex financial tasks are not being able to be sufficiently handled by large language models by themselves. But also in our paper, we make the point that even safeguards that are dedicated models or systems to provide these kind of first paths like is this safe?

14:54 Is this unsafe judgment? They're also not trained on financial services. And if you use them out of the box and say, look, I use Lama Guard, I use Shield Gmma, I use

Aris, I'm safe now. Right? You are protected against a particular view of safety that is very much grounded in categories that are relevant to broad populations, to things like chatbots that help you do productivity day-to-day tasks. The typical applications that you would see in those AI productivity tools, no matter which one you use, they all have similar mechanisms, but those are not necessarily the same risks that we are under in financial services. Those are not the same obligations that companies, organizations in healthcare are under or law or any other highly domain specific knowledge intensive domain that has a lot of specific regulation jurisdiction, specific regulation considerations about whether just refusing to answer or giving disclaimers is enough or whether questions should be blocked altogether. And there's just this difference of view that can be calculated in a single model that a provider can give that very much is focused on a different use case.

Jon Krohn: 16:06 Nice. Yeah. So I don't know, do you have guidance for us if we're trying to select an LLM for a particular use case? What do you recommend we do? I mean, practically, how can we move forward with all the information that you provided in this episode in selecting an LLM for a particular use case for a particular domain, particularly if we want to be applying it in rag situations?

Sebastian Gehrm...: 16:32 Yeah, so in our paper, we also have a list of best practices and recommendations that we have for especially for knowledge intensive domains and regulation heavy domains. Not necessarily everything has to be followed if you're building something for a much broader general population, but especially for these kinds of domains, all I can do is spray my mantra, evaluate the system in the context that it's deployed in. If you are building something for healthcare, well you better evaluate in the context of healthcare. If you are building in the context of financial services, you better evaluate with subject matter experts in financial services. And specifically on the safety angle. Our paper makes a couple of suggestions here. There are very good starting points. There are taxonomies such as the n NIST risk management framework for ai. There are other industry collaborations going ongoing. There's ML commons. Those all provide more general purpose taxonomies, but just taking 'em as a starting point and then from there, adjusting them to your domain can often save a lot of time.

17:39 And especially if you're a large organization with a compliance or risk department, it'll help them also understand how one can classify and then categorize these kinds of risks. Risks. Another recommendation we make is to organize red teaming events or any other kind of red teaming. Red teaming in this case is this practice that had to start in the Cold War where you have users trying to be malicious. So we get people in the same room and we say, look for the next couple hours, try and break the system, try and play evil. Here are some instructions on how to do this. And then afterwards we can look, how often was this actually broken? How often did the system give financial advice? How often did it refuse? And from there we can quantify the risk surface since we were talking, we were earlier about this unknown risk surface.

18:28 Well just measure it and then you have it. So that's kind of the main takeaway that we have. We give pretty specific advice for how to go about this and how to set up risk management frameworks. And all this needs to go hand in hand. Also with, again this evaluate in the context that's applied, make sure you invest a lot in evaluation. Don't just take the word of the large negative model providers that their benchmark scores are going to translate into all the downstream applications. And if you follow that advice, you're going to have a system that is in the end much more trustworthy, reliable, robust, and you're going to have users that are going to keep using it rather than trying it twice, getting really bad answers both times and never touching it again.

Jon Krohn: 19:09 So the best way to know if you've got a system that's fit for purposes to test, test and test, again, this is how we can measure AI behavior and ensure the model we choose does what we want it to do. Determining human behavior is also a critical topic for AI practitioners are we so predictable? Dr. Zohar Broffman thinks we may be, and he has the research to back him up. He explains human predetermination and desire. In this clip from episode 907, something very interesting that Benjamin Labey and countless others have shown is that you can have a neurological, the neural basis of some conscious idea that you have happens hundreds of milliseconds before you have the conscious thought. And this is a very disturbing thing to think about because most of us go around through the day with this illusion that you have some kind of control over what thoughts come into your head or what action you take next. But in fact, what these experiments show is that there's, you become aware of a decision after that decision has already been made subconsciously in your brain. And so yeah, I don't know. There's plenty to dig into there, but maybe talk to us about this a bit more and then maybe tie it into your belief in AI's ability to anticipate user behavior.

Zohar Bronfman: 20:37 So I think we were talking about once you get exposed to something in the realm of neuroscience and ai, you lose sleep. This was probably the biggest sleep deprivation I had because it's mind blowing, right? If we think about it deep, we might end up in a rabbit hole of mi, just an agent carrying my neurons or something like that, which I don't think is very easy to dispo by the way, in all honesty. So it's mind blowing. It's obviously a lot to digest. But yes, by the way, those experiments happened first sometimes during the eighties. Since then, it's been replicated and reproduced and in different settings and in different environments, in different technologies, in different animals. So many times that I don't think it's anymore even just an open question, it's a truism it's given. Now obviously there's room for interpretation, but the fact that there are brain processes that are directly causally related to decisions we make and that we don't have access, we don't have conscious access to those processes, I think is already completely agreed upon.

21:50 Obviously you can ask the questions of how elaborate these processes are, whether as something reaches consciousness, it can override or kind of change some of these or veto some of these processes and so on and so forth. But the fact that this happen is hard fact now, it means that much of what we are doing as humans is predetermined by things that have nothing to do with our immediate desires. So you can put someone in FMRI, like a functional MRI that basically shows the blood in your brain and which areas are active and you can tell 10 minutes or 15 minutes and they drive in a car simulator. You can know 10 or 15 minutes in advance before they reach the junction, whether they're going to turn left or right, and it has many contributors to it. Maybe it's a question of your stronger side, maybe it's something that happened in the morning, maybe your neck was sore and it's harder for you to look to the right.

22:57 By the way, it's my case at the moment. So there are many different ways that can contribute to these unconscious processes that you end up affecting your decision or your action. But what it also means, and this is something again that is quite known for many years, it means as a consumer, your behavior is also affected by many things that you are not aware of. And it means that as a business that sells to consumers, you can probably know much in advance of your specific customer's behaviors before the event takes place. So you can predict the purchases that customer is going to make, the conversions or lack of those lifetime value, best products, churn and so on and so forth. And that ability to make those predictions based on their historical behavior is, like I said earlier, the biggest level we know in the industry for transforming businesses.

24:09 So I'm basically saying if you collect data about your consumers as a business, there's a good chance you can start making predictions about their behavior in the future and you can optimize their experience. You can optimize your processes, you can basically just make the most out of those precious interactions the consumers have with your business. My personal, and this is what PK is all about, my personal mission. I want to bring these capabilities to as many small and mid-sized businesses as possible because they also deserve unquote that remarkable technology that basically tells you what people are going to do even before they know what they're going to do. And that's why we've invested so much in connecting M'S data and machine learning together in one nice package.

Jon Krohn: 25:10 Having ai, knowing your next move might sound farfetched, but given how much of our activities take place online, AI tools may soon be able to detect behavioral patterns we might not even consciously recognize ourselves. We have to note though that AI doesn't make intuitive decisions the way we do. Instead, it relies on patterns and probabilities is this set to change. My final guest in this episode of In Case You Missed It, is Microsoft researchers, Dr. Robert Ness. In episode 909, Dr. Ness investigates how we can start to build causal AI models that go beyond this standard correlation based patterns and probabilities that we're used to machine learning models generally detecting.

25:53 I'm a data scientist at a gaming company and I want to figure out whether the users, whether users of this game, when they tend to engage in more side quests, does that cause them to spend more money on in-game assets that they can be buying?

26:12 And so there are potentially confounding variables out there, like things like being a member of a guild that you mentioned there. And so if we weren't collecting that guild data, we'd have to have more assumptions. We'd have to basically make the assumption that being in a guild doesn't matter or not. And so it seems like, so this has made clear that there's a lot of assumptions, more thinking potentially about your problem that you need to do if you're engaged in causal ai. So that's a great thing to understand about this. But to kind of get into the nuts and bolts, your book does a great job of using PyTorch code using examples to make causal AI or causality in general, which is often a very theoretical, difficult to understand topic because of all of your examples and use of code in the book, it makes understanding causal AI more intuitive. And so let's say that you were the data scientist, Robert, at this gaming company. What Python tools would you use then to then do causal ai? How would you model this in order to come up with a causal conclusion?

Robert Ness: 27:27 So one of the things that I had mentioned that I was trying to do with my book was to separate out the abstractions that have to do with statistics and computing, right? Scale it up algorithmic complexity from the causality. And what's cool about the libraries that we have today is that they can actually help us if we're able to separate those abstractions, then we get to focus on one thing while leaving the nuts and bolts to be handled essentially by the library. To some extent, you saw this, you mentioned you interviewed somebody who talked about Stan, and what's cool about Stan is that that inference algorithm, Hamilton and Carlo is, I mean you can go in there and understand it's not, well it's physics, but it's not rocket science. I'm like, well, it kind of is rocket science, but you can still kind of just specify your model, specify whether the parameters, what's the model, and as long as you satisfy certain set of requirements, I think namely that's all of these things have to be continuous, that the inference will kind of just work for you without you having to go to implement your own inference algorithm.

28:48 It's the same thing here where if you can, as long as you can kind specify your causal assumptions in some cases in the form of a graph for example, then you can rely on say, graphical causal inference to the inference algorithms and from say, probo graphical models or to handle the inference there for you if you're implementing it in pie tours, plenty of PyTorch examples in the books, in the book, as long as you can exactly incorporate your causal assumptions in the structure of the model in your algorithm and write a basic inference algorithm that has a differentiable loss function, then PyTorch is going to handle all of the nuts and bolts of the inference for you. Right? That's kind of why we invented PyTorch to say, well, if I can differentiate it, then if I can get a gradient, then I can just turn it into an inference problem there.

30:05 And so I do have examples there in PyTorch that are saying, okay, well let's not worry about whether or not we need to use linear regression here, or propensity scores or double machine learning or in instrumental. These are all different types of statistical methods for doing the inference you want and you can learn all these things. Great. And there's great books for that. I think I name a couple off the top of my head, but you can also say, let's work with some libraries that are just going to handle that stuff for us under the hood and kind of treat it as either an objective function to be optimized or as configuration in a configuration parameter and some model specification. And then focus on our ability to think causally and write that thinking down in the form of a model and focus on the actual domain that we're modeling as opposed to all the inference stuff that we need to do to get that to work. And so to answer your question, I talk a lot in the book about using probabilistic models like modeling with libraries like pyro, as well as some more conventional tools like the DOI DOI from the broader pi y suite, which is a big collection of causal inference libraries.

31:41 And so even in Doy, right, there are different types of statistical techniques that you can use to estimate a causal effect, but at the end of the day, you're thinking more about what are your modeling assumptions and can you answer the question given your assumptions and your data? And then if you can, you want to get to an answer and then all of the various statistical approaches you can take to a ride at that answer, given your assumptions and your data, you can kind of just toggle between them to see which has giving you more stable results, for example.

Jon Krohn: 32:14 Alright, that's it for today's in case you missed an episode to be sure not to miss any of our exciting upcoming episodes. Subscribe to this podcast if you haven't already. But most importantly, I hope you'll just keep on listening. Until next time, keep on or rocking it out there. And I'm looking forward to enjoying another round of the SuperDataScience Podcast with you very soon.