# SDS PODCAST EPISODE 901: AUTOMATING LEGAL WORK WITH DATA-CENTRIC ML (FEAT. LILITH BAT-LEAH)

| | | |
|---|---|---|
| Jon Krohn: | 00:00:00 | This is episode number 901 with Lilith Bat-Leah, Senior Director of AI Labs at Epiq. Today's episode is brought to you by the Dell AI Factory with NVIDIA and by Adverity, the conversational analytics platform. |
| | 00:00:21 | Welcome to the SuperDataScience Podcast, the most listened to podcast in the data science industry. Each week, we bring you fun and inspiring people and ideas, exploring the cutting edge of machine learning, AI, and related technologies that are transforming our world for the better. I'm your host, Jon Krohn. Thanks for joining me today. And now, let's make the complex simple. |
| | 00:00:54 | Welcome back to the SuperDataScience Podcast. Today, we've got Lilith Bat-Leah, a tremendously gifted communicator of complex technical information on the show. Lilith has over a decade of experience, specializing in the application of machine learning to legal tech. She's now Senior Director of AI Labs at Epiq, a leading legal tech firm that has over 6,000 employees. |
| | 00:01:17 | She's published work on evaluation methods for the use of ML in legal discovery, as well as published research on data-centric machine learning. She's co-chair of the Data-centric Machine Learning Research Working Group, MLCommons, and has organized data centric workshops at ICML and ICLR, two of the most important AI conferences. She holds a degree from Northwestern in which she focused on statistics. |
| | 00:01:40 | Today's episode will appeal primarily to hands-on practitioners, like data scientists, AI/ML engineers, and software developers. In today's episode, Lilith details how AI is revolutionizing the legal industry by automating up to 80% of traditional discovery processes, why elusion is a critical metric that only exists in legal tech, and what it reveals about machine learning evaluation. The surprising reason why we should stop obsessing over |

model improvements and focus on something else that takes up 80% of data scientists' time.

00:02:13 And she talks about how she grew from being a temp receptionist to eventually an AI lab director by falling in love with statistics. All right. You ready for this outstanding episode? Let's go.

00:02:25 Lilith, welcome to the SuperDataScience Podcast. I'm delighted to finally get you on the show. I was talking about it with you for a while and now it's finally happening. Where are you calling in from today?

Lilith Bat-Leah: 00:02:43 Thank you so much for having me. I am in New York City.

Jon Krohn: 00:02:47 Likewise. Exactly. Both in Manhattan, though recording remotely anyway. It does make things easier. There's a lot less setup involved, if I just do remote recording sessions for people who are wondering at home why I sometimes do New York episodes remotely with guests. Though, I guess also I could be traveling. I don't know, it just the logistics easier to do things remote. So, we actually met after the Open Data Science Conference East in Boston a year ago. Was it a year ago or two years ago? Just a year?

Lilith Bat-Leah: 00:03:20 I think it was just a year ago.

Jon Krohn: 00:03:21 Just a year. And so, we were on the train back. So, the Acela, the supposed express train that is available only in this kind of northeast corridor of the US. And it isn't that fast. If people have been on express trains in Europe or Asia, you'll feel like this isn't a very fast train, but it is a really nice ride from New York to Boston or vice versa. And the only nice train ride that you can have in north, or at least in the US. Canada does actually have some nice trains, too. And I was sitting, trying to mind my own business, but behind me you were sitting. And you were going into a lot of technical detail and explaining

technical data science concepts very clearly, succinctly, in a really enjoyable way to whomever you were sitting with. And after an hour of listening to that, I popped around in my seat, because I had to find out who this person was. And it was you.

Lilith Bat-Leah: 00:04:26 Thank you. Yeah. And you can give all of that credit to the judges and attorneys that I've explained this technical concept to over the years.

Jon Krohn: 00:04:35 Yes. And so, we are going to have a legal episode here. But I think this will be interesting for anyone, because I think it's great to dig into different domains. And whether you do work in legal or legal tech yourself, there's lots of concepts that will be describing that could be transferable. So, you might think of an analogous kind of thing that you could be doing in your own industry. So, let's start off with that. So, you are the Senior Director of the AI Labs for a company called Epiq, E-P-I-Q, which is a leading legal tech company. They're a pretty big one. There's thousands of employees. I think I looked into this.

Lilith Bat-Leah: 00:05:22 Yeah. I think we're over 6,000 right now.

Jon Krohn: 00:05:25 Right. Over 6,000 employees, so it's a big legal tech company. Earlier this year, Epiq launched something called the Epiq AI Discovery Assistant, which claims to automate more than 80% of traditional eDiscovery processes and completes reviews up to 90% faster than something called TAR, Technology Assisted Review or linear review, which I'm guessing is like a human reading every word on a page linearly. So, we've got a bunch of legal tech jargon here now that I'm not really familiar with. So, tell us about linear reviews, TAR and eDiscovery. Tell us about those firms... about those terms, and then we can get into how AI can make life easier.

Lilith Bat-Leah: 00:06:09 Yeah. And I'll qualify one of those claims a little bit. It's better than traditional TAR. So, I would say that the software that we offer does support TAR workflows. And to actually describe what that is, it stands for Technology Assisted Review. And it basically describes a process whereby you use machine learning to classify documents as relevant to a litigation or not relevant to a litigation.

Jon Krohn: 00:06:41 And a litigation is just somebody suing someone else, I guess.

Lilith Bat-Leah: 00:06:45 Yeah. So, the way I explain discovery for people outside of the legal industry is that basically anytime two companies sue each other, they have to exchange anything and everything that might be considered evidence in the case. So, what this ends up looking like is piles and piles, maybe hundreds of thousands, even millions of documents, emails, Word docs, Excels, tweets, text messages, anything and everything, tons of unstructured data that might be relevant to the litigation. And then attorneys have to go through all of that and determine what is going to be produced to the other side, what they're legally obligated to produce to the other side, because it might be evidence. So, that is eDiscovery in a nutshell.

Jon Krohn: 00:07:33 What makes it eDiscovery?

Lilith Bat-Leah: 00:07:35 So, way back when, and you might be familiar with those TV shows where the lawyers have the boxes of documents, that's traditional discovery. But now, all of the business records, all the data, as it was maintained in the ordinary course of business, is electronic. So, in the early aughts around then, I think we started calling it eDiscovery, rather than discovery. But now, pretty much all of document discovery is eDiscovery for the most part.

Jon Krohn: 00:08:06 Got you.

Lilith Bat-Leah: 00:08:06 There are exceptions where some asbestos case where you have to go back to the paper documents, and scan them in, and then review them.

Jon Krohn: 00:08:18 This kind of reminds me of watching those old TV shows. And you'd see it seems like it's a deliberate strategy to flood your opponent with as many documents as possible to bog them down and increase their fees, that kind of thing.

Lilith Bat-Leah: 00:08:34 Yeah. So, that's considered a bad faith approach these days, if you try to just overwhelm your opposing counsel with documents that aren't actually responsive to their RFPs. And that's where precision matters a lot, if you're using a certain version of Technology Assisted Review. So, so now that we have all these great machine learning tools for eDiscovery, it's much easier for opposing counsel to uncover the needles in the haystack that they might be looking for and can point to evidence that really matters to them.

Jon Krohn: 00:09:09 Nice. Very cool. So, now I think we have an understanding of the territory. So, tell us about Epiq AI's discovery or Epiq's AI Discovery Assistant. Tell us about how that's different and how it accelerates, again, the claim that you qualified appropriately, but then we get this 80%. It automates 80% of discovery and completes reviews up to 90% faster than linear review, other approaches.

Lilith Bat-Leah: 00:09:44 So, traditional TAR technology, basically it's a classifier with active learning. And depending on the prevalence of documents that you actually care about in your overall population, you'll use one of two different active learning workflows. And before I keep going, can I assume that your audience will be familiar with active learning?

Jon Krohn: 00:10:09 I would love to hear a bit more about it.

Lilith Bat-Leah: 00:10:11 Excellent. So, active learning is just a way to select data in a more efficient way for training your classifier. And there are generally two popular approaches to it in eDiscovery. So, if you have really low prevalence, you're probably better off using relevance feedback, so you're going to have human annotators label the documents that are already most likely to be considered relevant by the model. So, you're going to use that and then you're going to iteratively retrain the model several times in order to improve performance. And that's in a low prevalence situation.

00:10:59 If you have more balanced classes that is a more equal proportion of relevant and irrelevant documents, then you're going to want to use uncertainty sampling, where you're looking at the entropy of each data point, and having human annotators label the documents that the model is most unsure about in order to improve performance. So, those are the two flavors of active learning that we tend to use in this space.

Jon Krohn: 00:11:29 Very cool. That's exactly the kind of clear, technical explanation that I heard on that train ride. Fantastic. Thanks, Lilith. So, we kind of took a little bit of an excursion to talk about active learning, but you were filling us in on Epiq's AI Discovery Assistant.

Lilith Bat-Leah: 00:11:47 And I am very excited to talk about Epiq tools. But again, with traditional TAR, it's basically traditional long text classification. Anything from a random forest algorithm to an SBM, to logistic regression is pretty popular. You can use any of these algorithms or some ensemble learning and arrive at your classifications, along with that active learning component of things. What's very cool about Epiq AI Discovery Assistant is that it uses more traditional methods for a long text classification, but it also leverages LLMs.

| | | |
|---|---|---|
| | 00:12:28 | So, you get a head start on the documents that you start training on by using retrieval augmented generation to find the documents most likely to be relevant to whatever it is that you care about, whatever issue the attorneys might've specified, and kickstart things there. And then both your human language instructions and your labeled examples are going to go into training the best classifier possible. So, it takes input from both example data and from natural language instruction. |
| Jon Krohn: | 00:13:08 | So, you need a classifier for basically every case, a separate classifier. |
| Lilith Bat-Leah: | 00:13:14 | Yeah. Sometimes many, many classifiers in one case. It depends on how many different things they care about classifying. So, generally, you'll always have a responsiveness, assuming that it's in preparation for a production to opposing. You'll always have what's called a responsiveness model, basically a relevance model. Is it relevant to any of the issues in the case? But then you might also have classifiers for things like privilege, whether the document is protected by attorney-client privilege and therefore it's not mandatory to disclose it. And then confidentiality, potentially, and then sorts of issues that the attorneys working on the case might care about. |
| Jon Krohn: | 00:13:54 | So, does this mean that big law firms typically have data scientists on hand, or do they rely completely on tools like Epiq AI Discovery Assistant to allow these classifiers to be trained in a fully automated way without some kind of technical expertise, like a data scientist being involved? |
| Lilith Bat-Leah: | 00:14:17 | Yeah. They mostly rely on these tools. Maybe that's changing, but I would say very few law firms have data scientists who are involved in the discovery component of the practice. So, they do rely on these tools. With that said, you do need to have some expertise, some domain |

expertise, and some familiarity with basic evaluation metrics in order to make sure that you're using the tool in a defensible manner. And we are trying to build in as much of that as possible, build in the expertise, build in all the metrics and intuitive explanations of them. But I would say at this point it's still ideal to have that domain expertise and a little bit of familiarity with evaluation metrics.

Jon Krohn: 00:15:11 Got you. So, perhaps a law firm might work with Epiq, not only to get access to a tool, but also to leverage expertise for people like yourself.

Lilith Bat-Leah: 00:15:23 Exactly. Yeah. We have an amazing team that helps clients with specific matters and helps them achieve whatever it is they're looking to achieve for that particular case. And that can entail building dozens of models for one single case.

Jon Krohn: 00:15:41 When the stakes are so high... As big law firms, when you're talking about hundreds of thousands or millions of documents, obviously these are going to end up being very expensive cases. You're talking at least millions of dollars and is probably very often in these kinds of litigation situations, tens, hundreds of millions of dollars, billions of dollars on the line one way or another for the defendant or the plaintiff. Does that happen in litigation? Do you have a defendant and a plaintiff in litigation?

Lilith Bat-Leah: 00:16:12 Yeah. Yes, absolutely.

Jon Krohn: 00:16:16 This episode of SuperDataScience is brought to you by the Dell AI Factory with NVIDIA, helping you fast-track your AI adoption - from the desktop to the data center. The Dell AI Factory with NVIDIA provides a simple development launch pad that allows you to: perform local prototyping in a safe and secure environment. Next, develop and prepare to scale by rapidly building AI and

data workflows with container-based microservices. Then, deploy and optimize in the enterprise with a scalable infrastructure framework. Visit www.Dell.com/superdatascience to learn more. That's Dell.com/superdatascience.

00:16:56 And so, in that kind of situation, the stakes are very high. So, how do you balance speed and automation, which are so important, with the legal field's high standards for defensibility, a word you just used, and due diligence?

Lilith Bat-Leah: 00:17:12 Those are great questions. So, one of the fun things about working in the legal industry is that these standard evaluation metrics, we call them precision, generally, sometimes get negotiated with opposing counsel or some governmental body. So, it's a one time where your evaluation metrics and being rigorous in your evaluation processes, reasonably rigorous in your evaluation processes, really, really matter. You get to argue about the margin of error and all sorts of things like that. And as a data scientist, you do have to be able to explain what that really means, and what the consequences of it might mean to attorneys and sometimes judges.

00:18:00 But every case is a little bit different. Defensibility boils down to what that particular attorney is comfortable defending. And there are proportionality considerations and undue burden considerations that go into it. So, for example, if you have a really, really low prevalence relevance tag, if you're looking for a subset of the documents, that's really rare, just sampling enough documents in order to be able to evaluate it could become overly burdensome potentially. And then we have this metric that I've never come across outside of eDiscovery. We call it elusion, where we're just sampling the subset of documents predicted not relevant, and we have the human ground truth labels for all the relevant documents. So, from those two metrics, we can then

estimate an interval for recall. And that's an interesting case, and the defensibility around that is debated.

00:19:14 I'm a proponent of it, because we don't use any of these metrics to evaluate what we call linear review, which is just humans with eyes on everything. And if we're just going to assume that that is the gold standard, that all of those labels are in fact correct, which we kind of know they probably aren't, then why should we hold machine learning workflows to a higher standard? We should be able to accept that those labels are the gold standard. So, lots of interesting areas of debate, lots of different angles. And again, it just depends on the case, and who's requesting what, and how onerous it's going to be for the producing party to appease opposing. And all of that goes into a "defensible workflow."

Jon Krohn: 00:20:14 All right. Right. Let's talk a little bit more about this elusion term that seems to be unique to legal tech. So, for our listeners, it's not an illusion like a magic trick with an I. It's like elude, E-L-U-D-E. Elusion, E-L-U-S-I-O-N. So, it's kind of like this idea of deception or avoiding detection, I suppose, because it's not deliberate deception.

Lilith Bat-Leah: 00:20:47 Yeah. Those are the documents that have eluded you.

Jon Krohn: 00:20:51 Right, exactly. And so, then why is that different from a machine learning metric that would be equivalent? I often like getting a little two by two table in front of me to make sure I'm not butchering this, but that would be a false negative?

Lilith Bat-Leah: 00:21:12 Correct. Yes, yes. False negative out of false negatives and true negatives. Exactly.

Jon Krohn: 00:21:20 Right. We actually, candidly for our listeners, I just took a second to do some research and pull out that it seems

like kind of a generic term for this, for elusion. So, false negatives divided by false negatives and true negatives can be called false omission rate in machine learning in general. But I guess, it's a bit of a mouthful. Elusion sounds nicer. It sounds like it's a simple word. I like it a lot.

Lilith Bat-Leah: 00:21:54 And I don't know who to credit with that term. It did kind of pop out of nowhere, so I wish I could tell you more. But I did figure out how to get an interval for recall based on the elusion rate. The problem with the elusion rate, and it's a very legitimate problem, is that people will take an elusion sample and just decide that, "Hey, yeah, it's low, that's good," without thinking about the starting prevalence. So, people will say, "Oh, if the elusion is under 5%, then it's good." But that's not good if your prevalence was under 5% to begin with, then that doesn't tell you anything. So, with this standard workflow, now I can talk about... I hate these terms, but there's TAR 1.0 and TAR 2.0. And they basically are heuristics for different workflows and there's different permutations. There's different ways of getting to whatever model you're using to serve up documents and different stopping points for training.

00:23:14 But at the end of the day, it boils down to TAR 1 being a workflow where you produce documents that have only been classified by your classifier and not necessarily been looked at by human attorneys. Whereas, TAR 2.0 heuristically describes a workflow where you're looking at every predicted relevant document before it goes out the door. So, in this TAR 2 workflow and this workflow where you are having humans actually annotate every predicted relevant document, then again, now you have that known quantity. You do know how many actual relevant documents there, where you don't have to estimate that from a recall precision curve or confusion matrix. And then you can estimate what the interval for recall is based

on the interval for elusion. And you hear me go on and on about intervals. I am obsessive about focusing on confidence interval and not point estimate.

Jon Krohn: 00:24:22 Exactly. I actually had some questions for you later on in the episode about that, but we might as well get into it right now. I mean, I can guess, but please tell us why you prefer going in ranges, why you prefer providing information in ranges, as opposed to point estimates.

Lilith Bat-Leah: 00:24:42 Yeah. So, the short answer is that the coolest thing about statistics is that you get to measure your uncertainty. So, why wouldn't you do that? Why wouldn't you measure your uncertainty? But the more serious answer is that... I mean, we are dealing with uncertainties. You shouldn't assume that a point estimate is truly representative of the parameter that you're trying to estimate. You really should think about those confidence intervals, because then you can feel pretty good about knowing that it's going to be somewhere within that range. And you're taking the uncertainty into account. So, it's easy to fixate on a point estimate. But I've said before that a point estimate without sample size, without confidence intervals, is basically lying with statistics. You have no idea what the actual claim is there.

Jon Krohn: 00:25:43 Right. And so, how do people who aren't trained in statistics react when you provide them with confidence intervals, as opposed to point estimates? Do you ever get kind of a confusion or backlash on that?

Lilith Bat-Leah: 00:25:57 Yeah, yeah. So, it depends on who I'm working with. If it's an attorney or a judge, I try to just demonstrate. I have even just a quick calculator in Excel. I'll show them how varying certain things affects those estimates and try to give them an intuitive understanding of it. If it's a consultant and I'm trying to give them a more intuitive understanding of it, I'll have them randomly sample half

of the documents in a certain population and label them something like documents I care about. And then I have them sample with 90% confidence, at least 10 times, so that they can see that, hey, on average, one out of 10 times the point estimate that I'm estimating from the sample is not within the range that I've estimated. And I think that builds up an intuitive understanding of confidence intervals.

Jon Krohn: 00:27:05 Yes, yes, yes. The old law of large numbers sounding familiar here. I will eventually create YouTube content on these concepts, if I haven't already. I can't remember where I am. I've been creating this mathematical foundations content. And I was really good about putting it on YouTube for a couple of years, up until three years ago. And I know that somewhere in there, I do have a law of large numbers video, but I think I might not have released it yet. So, it is coming eventually someday. Nice. In the meantime, people can look it up. But basically, it's like the more data that you sample, the tighter your ranges will tend to be, your estimates will tend to be. You begin to get a better picture of reality without having to look at every single data point.

Lilith Bat-Leah: 00:28:01 That's right. And I was actually just showing colleagues today, that when your sample is large enough, the intervals for 95% confidence level versus 99% confidence level tend to converge. So, once your sample is sufficiently large, it doesn't really matter whether you're estimating something at 95 or 99% confidence.

Jon Krohn: 00:28:26 Right. Right. Right. That makes a lot of sense. Nice. All right. So, we've now learned a lot about law, about legal tech. Have we gotten into yet...? Yeah, I guess we have. We've gotten into the Epiq AI Discovery Assistant as well, because you explained it had things built into it like retrieval augmented generation that allowed it to outperform Technology Assisted Review 1.0 or 2.0

Lilith Bat-Leah: 00:28:48 Traditional technology. Those are more workflow terms than technical terms. So, I hate the terms, because they confuse people so much. But it's been what the industry has been using for quite a while now.

Jon Krohn: 00:29:08 Nice. All right. And so, the topic that I actually thought might be the topic that we talked about the entire episode, but it ended up being that there were so many interesting things to go into around legal tech AI, that I wanted to have the conversation that we just had. But 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... 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: 00:30:07 Yeah. That's right. So, in my first R&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 label 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 MLCommons called DataPerf, where we were looking to benchmark data-centric machine learning.

00:31:12 That ended up leading to a few different workshops that we've organized at ICLR and ICML. DataPerf also became a NeurIPS paper. And basically, it turned into a whole

community. So, now there's a DMLR Journal, there are the DMLR workshops at these conferences. And then DataPerf morphed into the Data-centric Machine Learning Research Working Group with MLCommons. So, we have a lot of different things going on. We're working in partnership with Common Crawl, the foundation that curates the datasets that most LLMs have been trained on. We're partnering with them on a challenge that will result in a low resource language dataset that will be publicly available. So, if you're interested in joining the working group, please do get involved. Again, it's with MLCommons. You can go to that site and sign up for the working group.

Jon Krohn: 00:32:17 We'll be sure to have a link to MLCommons 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: 00:32:40 Exactly. Exactly.

Jon Krohn: 00:32:41 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 ICLR, ICML, NeurIPS, 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 he... I have in my notes here somewhere. I'm kind of scrolling around in here. So, at the inaugural DMLR workshop, Andrew Ng was the keynote.

Lilith Bat-Leah: 00:33:15 Yes, yes, exactly. And he was involved with DataPerf as well. He's on that DataPerf paper.

Jon Krohn: 00:33:22 This episode is sponsored by Adverity, an integrated data platform for connecting, managing, and using your data at scale. Imagine being able to ask your data a question, just like you would a colleague, and getting an answer instantly. No more digging through dashboards, waiting on reports, or dealing with complex BI tools. Just the insights you need - right when you need them. With Adverity's AI-powered Data Conversations, marketers will finally talk to their data in plain English. Get instant answers, make smarter decisions, collaborate more easily—and cut reporting time in half. What questions will you ask? To learn more, check out the show notes or visit www.adverity.com.

00:34:06 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, 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: 00:34:45 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,

DoReMi, that looked at weighting different domains of the pile to get the best LLM pre-training performance.

00:35:41 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 priori, 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: 00:36:14 Right. Right. Right. Right. So, the idea here is that... And so, I think this contrast with the idea of what we mostly end up doing as data scientists, as machine learning engineers, as AI engineers, where we're trying to change our model weights in order to get the best result for whatever situation we're in. With data-centric machine learning, the idea is that you could actually potentially keep your model weights the same. And you make adjustments to the data themselves, in terms of how much you have or the composition of those data, or how you sample from the data. And so, basically, you're concerned, you're focused on data. They become central to the way that you develop your machine learning models and ultimately provide results.

Lilith Bat-Leah: 00:37:02 Yeah. That is a much better way of explaining it than [inaudible].

Jon Krohn: 00:37:06 I doubt it's better. I doubt it's better. It's just different because you explained it very, very well indeed. Seriously, you are gifted at explaining this stuff. Nice. So, in a paper written by the DMLR community members. And so, it's a paper called DMLR: Data-centric Machine Learning Research-Past, Present and Future. I'll have a link to that paper in the show notes. And I think you were a co-author on this paper, am I right?

Lilith Bat-Leah: 00:37:38 Yeah.

Jon Krohn: 00:37:38 Yes, you are. In fact, you're the third author on this paper amongst a couple dozen. And in that paper it quotes that everyone wants to do the model work, not the data work. So, what mindset shifts or incentives do you think are necessary to elevate the perceived value of data-centric contributions in the ML community? So, that's more than enough questions.

Lilith Bat-Leah: 00:38:05 Yeah. That's a great question. So, one of the major challenges when DMLR was getting off the ground, was that there were no really prestigious archival venues for this kind of work. So, that's starting to be addressed with the Datasets and Benchmarks track at NeurIPS and then launching the DMLR Journal. Which by the way, it's the newest sibling journal to the JMLR Journal, which has some street cred. So, finding or establishing these high-impact prestigious venues for publishing this kind of work, I think that goes a long way toward encouraging more of the data-centric work. But we still have a long way to go. I mean, it is true. I think 80% of most data science projects are way more about data cleaning, and data engineering and all of that.

00:39:11 But we really focus on that 20% that's iterating on the models. But we don't look at that as the fun, exciting part. So, I think we do need to just bring our engineering mindsets. How can we systematically improve data? How can it be a task that goes beyond just annotating, finding better ways to annotate the data? All of those things have to happen for it to, I think, gain even more traction than it has.

Jon Krohn: 00:39:43 Yeah. Once you put it in that kind of stark term, we've probably had 100 guests on the podcast confirm that 80/20. Around 80% of a real world data science project is spent on data cleaning and 20% is actually on model

building. And it's so interesting that when you think about that ratio, how little there is published on that 80%. DMLR should be most of it.

Lilith Bat-Leah: 00:40:13 Right. Well, I agree with you on that one.

Jon Krohn: 00:40:18 This is me just completely riffing and I'd love to hear what you think about this. But I guess what ends up happening perhaps is that people might feel when they're doing that work that the problems that they're encountering are unique to their particular dataset. Maybe ideas don't come to mind for them that generalize well across many domains or even within their subject matter that they're a expert in. What do you think about that? What are some of the big trends or big themes that you see in DMLR that apply broadly into a large range of circumstances?

Lilith Bat-Leah: 00:41:01 Yeah. So, I'll go back to DataPerf, because we were aiming to establish this benchmarks pushed model-centric machine learning pretty far. So, we were hoping that we could push data-centric ML further along by establishing benchmarks there. To be honest, I don't know how far we really made it, but it was an interesting endeavor and we focused on a few different types of tasks. So, one was data selection. So, from a very large pool of data, how do you select the subset of data to train the highest-performing model? And we did that in both the speech and the vision domains. So, that was one challenge and benchmark. Then we had a data debugging challenge where participants were encouraged to find the mislabeled data points, the mislabeled instances in a dataset, and correct the labels or exclude them from training.

00:42:16 So, I think that has pretty broad application. Anytime you're doing supervised learning, if you have mislabeled data, then that's going to be pretty practical. And then we

also did a data valuation challenge. So, how do you value each piece of data? Not all data are equal when you're training a model. Some have much more impact than others. So, we looked at that. And that's a whole really interesting area of data-centric machine learning research that I didn't know anything about until I joined DMLR. But there are all these different ways to estimate the value of certain data points. And that might become increasingly important as we try to figure out how to compensate people for all the data that we're using to train all these models. And then we had a red teaming challenge called Adversarial Nibbler, where [inaudible]. You know the reference?

Jon Krohn: 00:43:38 Is it Futurama?

Lilith Bat-Leah: 00:43:39 Yeah. Yeah.

Jon Krohn: 00:43:42 That's funny. I didn't actually think that there would be a reference until you asked for one. But fortunately, I have seen quite a few episodes of Futurama.

Lilith Bat-Leah: 00:43:50 Cool. Well, I did not come up with the name. I can't take credit there. But the main objective of that challenge was to find benign-sounding prompts that generated unsafe images. So, for example, a child sleeping in red paint sounds benign, but generates an image that looks horrific. So, the challenge was all about finding these pairs, these text image pairs for use and then helping to make these models more robust and all of that.

Jon Krohn: 00:44:24 Wow, what a visual, a child sleeping in red paint. Yeah, that is interesting. In just a little red paint puddle that just happens to be on the floor. I'll have links in the show notes to dataperf.org, which I'm guessing stands for data perfect, maybe.

Lilith Bat-Leah: 00:44:49 Data performance.

Jon Krohn: 00:44:50 Data performance, of course.

Lilith Bat-Leah: 00:44:54 But that site is super outdated, just MLCommons. And then dynabench.org is the platform where we host all of these challenges.

Jon Krohn: 00:45:02 Dynabench, that's like dynamic bench?

Lilith Bat-Leah: 00:45:06 Yeah. Yeah. So, that's a platform that we've used to facilitate a lot of these data-centric challenges. And that's still maintained by MLCommons. And if you're interested in that, that same DMLR working group that I mentioned before, we maintain Dynabench and continue to host challenges on Dynabench.

Jon Krohn: 00:45:29 Nice. And then I'll also have a link to your paper. So I already mentioned the DMLR Past, Present and Future paper. We'll also have a link in the show notes to your DataPerf paper, which is on benchmarks for data-centric AI development. And that one, you're just a couple commas away from Andrew Ng in the authors of that paper there. Cool. So, those are resources that people can dig into deeply, if they have more interest in data-centric machine learning, which probably all of us should, given that 80%.

Lilith Bat-Leah: 00:46:02 There is probably a lot of value to people sharing domain specific solutions, because it might inspire people to find some new domain specific solution for their domain. And actually, one of the future workshops we're considering is an applications research-focused DMLR workshop. Because oftentimes, at these academic conferences, applications research gets looked down on a little bit. And we do think that there is more need to ground everything in really practical use cases. And we're sure that there's going to be a lot of really interesting research that is domain specific that different people can learn from. So,

that is something that we're hoping to undertake in the future.

Jon Krohn: 00:46:55 Very nice. And not only would it be great for people to be publishing more on the kinds of situations that they get to with their specific domain, much in the same way that us at the beginning of this episode talking about legal tech, AI applications, people can have analogous ideas come up for their industry. And not only that, but you could end up having... I totally see the idea of how benchmarks and competition have led us to having such a model-centric approach to machine learning. And so, things like DataPerf, where you have benchmarks, where you have competitions and people can be trying to get the best results, how that can drive more and more data-centric ML adoption, it's a brilliant initiative.

Lilith Bat-Leah: 00:47:47 Yeah. And at the same time, I think we can be critical of it, too, because there is a critique that the intense focus on benchmark performance doesn't necessarily translate to real world impact in the way that we would expect. So, there's definitely a balance to be found there.

Jon Krohn: 00:48:07 Nicely said, as you have done throughout the episode. All right. So, before I let you go, we've gone through the most exciting technical things that you're working on today. But you have an interesting background that I'd like to ask you at least one question about to get into. So, just going over your LinkedIn profile, it looks like you had a pretty interesting journey, where there was a point where you were an administrative assistant at the beginning of your career. And it looks like you kind of grew through legal roles, increasing seniority within legal firms, and then got into data science as well. And now you are a data science leader. So, I think this is an interesting journey and I'd love to hear just a bit about what happened.

Lilith Bat-Leah: 00:49:03 Yeah. So, I fell into eDiscovery as an admin assistant. Basically, as a temp receptionist, actually. And that was how I started my career. I was still finishing my undergrad at the time. And then at the same time that I was getting really familiar with eDiscovery and developing my domain expertise there, I fell in love with statistics. I took my first stats course and I got an A in it. And I didn't feel like I understood how or why I got an A, because I didn't understand... I mean, I could calculate the correct answers, but I didn't have this intuitive understanding for why they were the correct answers. So, I figured, okay, let me take more stats courses. And I took all the ones that made sense for me at the time. I took econometrics, psychometrics, various finance courses with portfolio theory. That's where I learned PCA.

00:50:06 Yeah. So, I took all these applied stats courses and I kept getting A's. But after each one, I had no idea how I was deserving of an A when I still felt like I didn't understand the material at all. So, finally, I asked the chair of the statistics department at Northwestern, if I could take his probability and stochastic processes course without any of the prerequisites. And I wrote him saying, "Okay, I know this is going to sound crazy, but here's why I think I can do it." And I'll never forget his reply. He wrote, "Dear Lilith, anything is possible. But of course, I would have serious reservations about letting you enroll without any of the prerequisites at all. Write me back in a year. Let's see if you really have picked up calculus before I consider this seriously."

00:51:00 So, I did. I crammed. I crammed for a year. I used MIT OpenCourseWare and Khan Academy and everything out there to just learn calculus on my own, a little bit of linear algebra. And then I came back to him and I said, "Okay, well, I didn't get as far as I wanted to, but I think I still want to take your course." So, he said, "Go ahead." He sent me the textbook. It was a PDF. It was the first

real math textbook I'd ever come across. It was just no images or anything, just coding problems. That's how I learned how to code. And math problems. And I crammed and I got an A on the finals, and then I finally felt like I understood statistics. Then since then, it's just been a lot of self-education, and diving really deep into all the different flavors of confidence intervals you can use. Really understanding what probability coverage means from that angle and just nerding out on the stuff I find most interesting.

Jon Krohn: 00:52:10 Very interesting indeed. That was an even more exciting story than I was anticipating. And it's interesting that I mentioned, because I don't talk about that often anymore, my machine learning foundations curriculum, but it's covering a lot of those subjects, linear algebra, calculus, probability theory and statistics. And we go in that order so that hopefully by the time we get to the statistics part, you're able to understand based on the fundamental building blocks that are lying at what's going on, as opposed to just being able to get an A by following the examples. Not by rote, that's not exactly it, but I guess by being able to apply the abstractions, as opposed to understand the underlying fundamentals.

00:52:59 And it's kind of interesting. I guess you were very excited and you said, "I fell in love with statistics." And it's interesting, because in a machine learning foundations curriculum, I don't really need to include statistics. Many people would argue it's not essential. But I also love statistics and it ends up being useful in so many ways, particularly around exactly what you described earlier in this episode, around being able to think of results that we have as being over a range, as having a confidence interval, as opposed to being a point estimate. And it's through statistics that I feel like I have a really good understanding of what those confidence intervals are.

Lilith Bat-Leah: 00:53:43 Yeah. And I think you're not... if you don't understand statistics, I don't think you're able to properly evaluate the performance of the models that you're building. So, you might be able to build the model without statistics. But I think especially in this era of black box models, it's so important to be able to actually evaluate the performance of them.

Jon Krohn: 00:54:06 And that is exactly... that ended up being the focus. When I would try to come up with relevant examples during the statistics section. And a lot of the time it was in exactly what you're describing, about evaluating different models. And being able to not just run the model once, a stochastic model once, one way, and a second time, another way. It'd be like, well, I'm done, it did better the second time, and therefore the second model is better. You should be running that model a bunch of times in both the A case and the B case, get a distribution of results and be comparing those. And then if you have a statistically significant result, and that is actually something statistical significance came up in our research of you. So, Serg Masís, our researcher, pulled up some quotes from you around how awful these kinds of ideas of a 95% confidence interval, having that as law. I don't know if you want to go into that at all, that kind of perspective.

Lilith Bat-Leah: 00:55:04 Sure, sure. You mean just being fixated on it being a 95% confidence?

Jon Krohn: 00:55:11 Yeah. If an alpha of 0.5 being the significance threshold, arbitrarily from the early 20th century, which particularly today, when we have very large datasets. When we had datasets, when our sample sizes were eight, 16 in each group, that arbitrary confidence thresholds of... And you can correct me if I don't say this exactly right. I'll do my best here. But that if you ran the experiment 20 times, you would anticipate with a 0.5 alpha, that one of those

20 times you would get a significant result by chance alone. And this is a century old idea from the Asia Fisher and Pearson statistics. And so, the idea there is that you'll kind of accept that you'll end up getting a significant result by chance alone one out of 20 times. And that's kind of tolerable, but it is completely arbitrary. And then today, when you have thousands or millions or billions of samples, you're going to get a significant result every single time at that kind of threshold.

Lilith Bat-Leah: 00:56:35 Yeah. So, the way that I've described it... And it is one of those things that's really hard to explain in plain English. But with a confidence interval, if you sampled this population an infinite number of times, you would expect that one out of 20 times the point estimate that you arrive at through your sample is not going to be within the estimated interval if your confidence level is 95%. So, like I mentioned earlier, whether it's 95% or 99% confidence, at a certain point, those converge. The intervals for those will converge if you have a sufficiently large sample size. But by that I mean a huge sample size. You need to be in the millions for them to really start to converge. And otherwise, it's just smaller and smaller differences between the intervals as you increase your sample size.

00:57:45 But if your sample size is very small, if your sample size is, as you mentioned, eight or 10, then there's actually a pretty huge difference between the interval that you got using 95% confidence and 99% confidence. And sometimes I think people just need to think about the question that they're trying to answer. So, how important is it for me to be right about my interval? So, you're basically trying to answer the question, how likely is it that my inference is correct? And if your inference is as conservative, you have that larger interval, then you're in a better place to be correct more often.

| | | |
|---|---|---|
| | 00:58:39 | Even though your uncertainty is wider, your interval is wider, you're still going to be correct about it being in that interval. Whereas, if you're really focused on that 95% confidence level and you have this really small sample size, then yeah, you're at a higher risk of just estimating something wrong, inferring something wrong from your statistic. I don't know if that made sense. |
| Jon Krohn: | 00:59:15 | No. That was pretty good. It is tricky and I followed it along there. Nice. All right. So, this has been a fascinating conversation. I knew it would be. You did not disappoint. For people who want to follow you after the episode and get more of your insights, how should they do that? |
| Lilith Bat-Leah: | 00:59:35 | LinkedIn is the best place for me. |
| Jon Krohn: | 00:59:37 | Nice. As it is for most guests these days. Actually, I don't know if I've said this explicitly on air before, but I've stopped tweeting. And I don't really check X anymore at all. For me, social media has migrated completely to LinkedIn at this point. Nice. And I missed my penultimate question there, so it's becoming the ultimate one, which is, do you have a book recommendation for us, Lilith, before we let you go? |
| Lilith Bat-Leah: | 01:00:07 | So, I have to give the recommendation to read the DMLR Journal for that one. That's just a easy, convenient answer. If it's okay, I also want to give a shout-out to dmlr.ai, which is the website where we post the latest about our workshops at these various conferences. And there's a link to the DMLR discord if people are interested in following both the journal and the workshops. |
| Jon Krohn: | 01:00:38 | Fantastic. Yeah, great resources there. I'll be sure to have dmlr.ai in the show notes. Thank you so much, Lilith. This has been great. We'll have to catch up with you again in a few years when everyone's talking about data-centric |

machine learning and all that we're worried about, instead of all these model benchmarks.

Lilith Bat-Leah: 01:00:56 And it may have had its moment already. I think for a minute people were talking about data-centric AI, but it never made it to the peak of inflated expectations or what have you. It kind of fell off the radar, but I'm still passionate about it.

Jon Krohn: 01:01:12 Yeah. Maybe we're still approaching it.

Lilith Bat-Leah: 01:01:13 I hope so.

Jon Krohn: 01:01:15 Thanks, Lilith.

Lilith Bat-Leah: 01:01:16 Thank you.

Jon Krohn: 01:01:23 Such a great episode. In it, Lilith Bat-Leah covered how when companies sue each other, they often exchange millions of documents as potential evidence, and how Epiq's AI Discovery Assistant uses LLMs and retrieval augmented generation to classify these documents as relevant or irrelevant up to 90% faster than traditional methods. She talked about how legal tech's elusion rate measures false negatives amongst predicted non-relevant documents. She talked about how while 80% of data science work involves data cleaning, most research focuses on the 20% spent on models. She talked about how the DMLR movement, the data-centric machine learning research movement backed by Andrew Ng and major conferences like ICLR, ICLR and NeurIPS aims to flip this by systematically improving data quality rather than just iterating on models. And she talked about how in legal settings where millions or billions of dollars are at stake, confidence intervals matter more than point estimates, because understanding uncertainty is crucial when your evaluation metrics can be dissected in court.

01:02:26 As always, you can get all the show notes including the transcript for this episode, the video recording, any materials mentioned on the show, the URLs for Lilith's social media profiles, as well as my own at superdatascience.com/901. Thanks, of course, to everyone on the Super Data Science Podcast team, our podcast manager, Sonja Brajovic, media editor, Mario Pombo, our partnerships team, Nathan Daly and Natalie Ziajski, our researcher, Serg Masís, writer Dr. Zara Karschay, and yes, of course, our founder, Kirill Eremenko.

01:02:54 Thanks to all of them for producing another outstanding episode for us today, for enabling that super team to create this free podcast for you. We are deeply grateful to our sponsors. You can support the show by checking out our sponsors' links, which are in the show notes. And if you yourself are interested in sponsoring an episode, you can head to johnkrohn.com/podcast to find out how. Otherwise, share the episode with people who'd like to listen to it as well. Review it on wherever you listen to it, subscribe. But most importantly, just keep on tuning in. I'm so grateful to have you listening. And hope I can continue to make episodes you love for years and years to come. Till next time, keep on rocking it out there. And I'm looking forward to enjoying another round of the SuperDataScience Podcast with you very soon.