SUPER DATASCIENCE
MAKING THE COMPLEX SIMPLE

SDS PODCAST EPISODE 769: GENERATIVE AI FOR MEDICINE, WITH PROF. ZACK LIPTON

Show Notes: http://www.superdatascience.com/769

Jon Krohn: 00:00:00 This is episode number 769 with Professor Zachary Lipton, Chief Scientific Officer at Abridge and Associate Professor of Machine Learning at Carnegie Mellon. Today's episode is brought to you by the DataConnect Conference and by Data Universe, the out-of-this-world data conference.

00:00:20 Welcome to the Super Data Science Podcast, the most listened to podcast in the data science industry. Each week, we bring you inspiring people and ideas to help you build a successful career in data science. I'm your host, Jon Krohn. Thanks for joining me today. And now let's make the complex simple.

00:00:51 Welcome back to the Super Data Science Podcast. The brilliant Professor Zack Lipton is our esteemed guest on the show today. Zack is both Chief Scientific Officer and CTO at Abridge, a generative AI company for clinical conversations that just announced $150 million of Series C venture capital investment and that's only four months after announcing a $40 million Series B. Zack is also associate professor in the machine learning department of Carnegie Mellon's Computer Science School where he directs the Approximately Correct Machine Intelligence Lab that builds robust systems for the real world. On top of all that, and he's the co-author of the great book Dive into Deep Learning, which was recently published by Cambridge University Press. Dive into Deep Learning is super popular. It's used at over 500 universities in 70 countries, and the associated GitHub repo with implementations in PyTorch, TensorFlow, NumPy/MXNet, and JAX has over 20,000 stars.

00:01:49 While the entire book is available in digital format free online, I've got a link to that for you in the show notes, I am also going to ship 10 physical copies of the book to people who comment or reshare the LinkedIn post I publish about Zack's episode on my personal LinkedIn account today. Simply mention in your comment or reshare that you'd like Zack's book. I'll hold a draw to select the 10 book winners next week. So you have until Sunday, March 31st to get involved with this book contest.

00:02:15 In terms of today's episode, despite Zack being such a deep technical expert, most of today's content will be of interest to anyone who'd like to hear about the cutting edge of generative AI applications in healthcare. In today's episode, Zack details how AI is starting to make a huge impact in medicine with detail on how he runs such a game-changing commercial R&D program, how his academic research cross-pollinates with his commercial machine learning projects, how the safety and security of clinical data are prioritized, and how his background as an exceptional jazz-sax player helped mold him into the entrepreneur and data science leader he is today. All right. Are you ready for this enlightening episode? Let's go.

00:02:59 Zack, welcome to the Super Data Science Podcast. It's awesome to have you here. Where in the world are you calling in from today?

Zachary Lipton: 00:03:06 I'm calling in from my house in Pittsburgh, Pennsylvania.

Jon Krohn: 00:03:10 Nice. We met about a month ago at the time of recording at a panel discussion hosted by Arthur in Manhattan. So Arthur is a cool, fast-growing New York startup, although I've got to say probably no one is growing as fast as your company Abridge. It's wild to see. At the time of recording, you announced last week a $150 million Series C raise, which follows just four months behind a $40 million Series B raise. That to me is almost unheard of, that kind of rapid growth. And so it's a testament to what you guys are doing at Abridge. Super exciting. You are both the Chief Technology Officer, the CTO, and the Chief Science

Officer, CSO, at Abridge. And the idea is that it's a AI platform for clinical conversation. Do you want to fill us in on kind of a high level to start on what you guys are up to at Abridge?

Zachary Lipton: 00:04:12 Sure. I suppose there's always the really broad view, which will sound crazy and maniacal. And our plan is to do anything and everything at the intersection of AI and healthcare. And then there's a really narrow view that sort of speaks to the particular use case that we've unlocked and flagship product that's hitting the market right now. At the broadest view, I think just my entire career in machine learning before I was a professor of machine learning at Carnegie Mellon, before I was a PhD student at UCSD comes from being a patient and being caught in one of the gray areas of the health system where people are not sure how to help you and where patients kind of have to dig through and figure out what's going on themselves. And I think looking around, you see things like there's diseases where you've got tens or hundreds of thousands or millions of people who are affected, and people are publishing studies based on 40 cases.

00:05:18 And just our use of information to inform care to constitute empirical foundation of healthcare is just like nothing compared to what it could be, compared to what's technologically possible today. And that was why I went into PhD, not because I was an expert in machine learning, but because I had a feeling that this is how we're going to advance medicine, or at least the way that I was maybe set up to be given that I wasn't a doctor, I hadn't gone to med school. And I think my entire research career has been sort of a combination of doing medical collaborations, stepping back to try to think what are the fundamental technologies that we need to make it all possible.

00:06:07 There's the narrow view. And the narrow view is basically there's a crisis happening right now in healthcare. That crisis is basically that we have more and more patients every year. The population's growing and the number of doctors is shrinking. And the biggest cause of this is physician burnout. So we have a bunch of decisions, all of which independently make a lot of sense. We moved doctors over from writing health records and writing notes on dead trees and putting them in filing cabinets, which seems like that can't be the way of the future, to modernizing the health system and putting everyone on digital medical records. We also said, "Well, patients should have some amount of right to see what's going on in their records. And while we're at it, other stakeholders in the health system should be able to see records, including payers."

00:07:04 And then we start going through and saying, "There's all these other things that should also be documented. And since we have the health record, why don't we add fields for this?" And they all independently make sense. If there's informed consent as part of the visit, that should be documented separately. Billing code should be registered somewhere in the record. But what it's all added up to you is a situation where these things all make sense and we kind of need and want all of them, but it's created a clerical burden on the provider where the doctor right now, for example, or other clinicians are spending two hours doing clerical scut work for every one hour of direct care. So it's sort of like, "Would you like some fries with your ketchup situation?" But mostly, they're treating the paperwork and sometimes getting to spend face time with the patient.

00:08:00 I know, it reminds me of a situation in data science and machine learning that you think like you're going into machine learning and you get to do cool modeling, cool algorithms, then you spend all of your time cleaning data.

They're there to treat the patient, but they're spending all the time mucking around in the health record. And the big insight that our founder, Shiv, had from the very beginning when he approached me back at the sort of beginning of the founding of Abridge was that sort of all of this work, upstream of all of it, the primary source of truth for almost all of these downstream, upstream of all the documentation, the rev cycle work, the diagnostics, therapeutics, the referral letters, clinical trial matching, upstream of all of that is a conversation between a doctor and patient. This is the primary source of information, the primary location of decision making, and almost everything takes place within the doctor patient conversation. If only you could sort of tap into that conversation, you get to unlock all the tooling that could reduce the pain of these downstream workflows.

00:09:08 So a realization, if you had an assistant sitting in the room just taking notes, who just broadly was medically knowledgeable, they'd be able to get you to a first draft of a note that was maybe 90, 95% of the way there. There might be a few things that the doctor had in their head or some pieces of prior context that they needed to add to the note that weren't actually vocalized during the visit, things they forgot to say out loud, but really the vast majority of the information that you need to make this work happen is context that was already given. And the treatment decisions have already been made, the referral decisions have already been made, the diagnostics have often already been spoken aloud. So much of the work here is just taking this information and sort of remixing it, reconfiguring it, pulling out what is the part that pertains to the history of present illness, what pertains to the assessment and plan, what pertains to the medication list.

00:10:02 And so we could go based on that conversation and go from there end-to-end all the way to a first draft of a note that's ready for a doctor's review. So we take a doctor, and instead of their job being to sit, and after the visit, they're kind of burnt out, they've already done this 15 times that day, and they have to first start typing up two pages of documentation from scratch, they get a running start. They get to start off with the core note is already written and their job is to check, " Are all the important pieces there?" Add anything that's missing, verify the content if they have any doubts. And we provide this end-to-end experience. Everything from the raw audio through the transcription, through the sort of sorting and structuring of content to where it belongs in the note, the abstractive summarization.

00:10:46 And then we don't see our job as ending when the draft is there, we pick up the job on the other end and provide these assistive editing tools. For example, a doctor can highlight any piece of text, like a word or a sentence or even a couple sentences, and float immediately to where in the transcript is a substantiating evidence that backs that up. And if they don't see it... It's funny, apple balloons showed up. If they don't see the evidence, then they have a basis for either doubting what they see in the note. But 99% of the time, they get right to the evidence and they could click on it and immediately even hear the audio playback from precisely that part of the conversation where that content came from.

00:11:32 So we build a level of trust because it's a level of trust that comes almost from not having to trust us, not having to take our word for it. We direct you. We have a certain stewardship for the providence of information. We take you all the way back to the source. People think that AI could have a propensity to hallucinate, but nobody's worried that microphones hallucinate and that's the essential character of the product.

Jon Krohn: 00:11:59 It's amazing. And I can see how it would be such a transformative tool for a doctor to wield to allow them to get that two to one ratio of administration to patient care down and to ultimately be able to see more patients, have less burnout, also presumably improve outcomes by having such rich data from the conversation easily summarized, easily being able to look up from those conversations as opposed to having to dig through them. When a physician is using this tool in a hospital setting, do you provide them with a particular kind of hardware to be working with or it works with the hardware that they already have?

Zachary Lipton: 00:12:40 Yeah, and no special microphone required. We can record from sort of multiple sources. We could do web recording, we could record from the app interface on the desktop. The primary way that people use it is that they have an app that functions as a recorder. So on their work phone, they have the Abridge app. The app is synchronized. So another characteristic of Abridge that as much as my tendency as AI scientist is to focus entirely on the technical guts on a deep learning level, there's a tremendous amount of work that we've done in terms of our integration into the workflow and into the health system. So we have a partnership with Epic and we are integrated into the electronic health record.

00:13:28 So a patient when they pull it up will actually be able to pull their entire... Sorry, the provider, the doctor or nurse, when they pull up the app, they'll actually see their schedule for the day. And so they can click into the appropriate encounter and start recording. And at the end when they hit complete note, what's going to happen is Abridge is going to run in the background, it's already transcribed all the audio, it's going to generate the appropriate note and it's going to beam it right into the medical record for the corresponding patient against whom they were recording. And then when they open it in the health record, then they have a set of tools for editing the final note and ultimately finalizing and signing it.

Jon Krohn: 00:14:17 Two days, 50 plus presenters, unlimited opportunities to connect with a global data audience. The annual DataConnect Conference brings together industry leaders, technical experts, and entrepreneurs to discuss the latest trends technologies and innovations in data analytics, machine learning, and AI, all while amplifying diverse voices in this space. Join us in Columbus, Ohio, July 11th and 12th to hear from incredible speakers like Carly Taylor, Megan Lieu, Olivia Gambelin, and more. Can't make the trip to Ohio, join us for the second annual DataConnect Conference West hosted in Portland, Oregon on May 8th, where you'll hear from AI experts from the likes of Microsoft, Alteryx, and Women Defining AI. Save 15% on your DataConnect Conference pass by registering with the code SuperDataScience.

00:15:06 Very cool. That kind of in the flow aspect to having AI tools is essential for adoption. It's something that we've talked about on this show particularly with the GitHub COO, that's going back to episode number 730. Kyle Daigle, the COO of GitHub made... Pretty much that whole episode is about how if you want your AI tools to be adopted, they have to happen right there in the flow. So GitHub Copilot is specifically what he's talking about there, where when you are writing the code or developing a machine learning model as a data scientist, you are there in GitHub, and as you are typing, you're getting suggestions. And so this sounds similar in the sense that you've clearly worked with clinicians and figured out how you need to be integrated with their medical records and with their schedules in order to make this as painless as possible so that it's a no-brainer for them to just tap that one button and be recording.

Zachary Lipton: 00:16:06 Yeah, I think we're advancing on all these fronts simultaneously on our level of integration for the workflow on the underlying technology. And then I think at the same time, to speak to your original question, it's like we are picking the battles that make sense from a standpoint of actually improving the product. And if I had a strong reason to believe that some strange recording device with 74 microphones pointed and some special configuration would actually lead to some massive improvement and the product quality, then maybe we'd invest in saying, "You need our special hardware." I think what we found is that investing instead in our algorithms for audio processing has been just a much, much, much richer payoff, and this allows us to reach way more people in way more circumstances and allows them to get going much quicker.

Jon Krohn: 00:17:02 Do you think it's conceivable that in the future, there could be like 74 microphones in a given hospital room, and maybe it's more than just microphones, lots of different kinds of sensors and you're getting real time information much more than just conversation, but lots of data on patients? Obviously, there's privacy concerns here, but that's part of the gap that Abridge is trying to bridge, right? Is to sow trust and to show that these AI systems are private, they're secure. And so maybe there's a future someday where you have 74 different kinds of sensors that are tracking every patient in the hospital over their entire time in the hospital, maybe even after they go home as an outpatient. And you could have maybe a doctor in the loop, or maybe some point down the line, decades down the line, you don't even need a doctor in the loop at all for a lot of recommendations. I don't know. Am I going too far out there?

Zachary Lipton: 00:18:05 Look, I'm a fan of science fiction, so nothing's too out there for the purpose of conversation. What I would stress is there's no limit to the number of pieces of information to draw potentially upon, and there's no limit to the set of things to potentially do with that information. I think at the same time, what you have to realize is going deep enough to really, I'm even reluctant to say conquer a problem because I kind of had this attitude in the company. You get one moment of satisfaction when something rolls out, and then immediately when you have that mindset of like, "Nothing is good enough ever," and not in a sense of a negative sense, but it's the thing that drives us as researchers. None of us are like, "I solved AI." That's never been the outcome of a great research project. It's like a feeling of like, "I was briefly at the front and now we start all over again," and you're not special.

00:19:20 And so I think that, look, in the short run, recognition, prediction, structuring, providing hypotheses, providing drafts, this is a much safer place to be operating in than in the taking away judgment or decision making. And I don't spend a lot of time these days thinking about how do I make decisions without a doctor in the loop. I think that it's a huge part of the level of trust that we have with our partners comes from the fact that they know that we're not like children running around with knives, that we have a certain kind of sobriety about the responsibility in the way associated with different kinds of ways of participating in the system. And they know that when they talk to us, not to just like a couple of Silicon Valley guys, they know they're talking to a research academic on a shift, to a practicing cardiologist who know the gravity of these problems.

00:20:34 I think it's important at the same time to be able to bifurcate your brain a little bit, to be able to say like, "Here's where we are, here's the world we're in, the things that are immediately possible, the kinds of claims that we could responsibly make, technology we could deploy." And on the other hand, to be able to think a bit whimsically, to be able to think really blue skies and imagine. I don't think the important thing here is taking doctors out of the loop. It's not something that I aspire to, but I would aspire to, I would own to wanting to help them to make better decisions, right?

00:21:11 A first step is to say, "Let me give you the joy of practicing medicine again and let me give the patient the benefit of your attention again," right? Like allow you to practice, so we say, at the top of your license. There was a time when most of practicing medicine was seeing patients, and now we somehow came away from that. A first step is like, "Let's give doctors back the ability to practice medicine at their full potential." A second step beyond that might be, "Let us help you to improve further," right? For example, you can imagine helping doctors to introspect, right? To be able to look back at patients and say, "How are these different conversations different from each other? What's going on in these conversations? How many questions did you ask? What topics did you cover? You recommended a treatment, what are other treatments that could have plausibly been recommended or statistically might've been likely to be recommended by other doctors in your specialty? And what are links to the associated medical literature up-to-date material?"

00:22:25 This kind of flow from just immediately helping you draft a note, helping you reflect, helping you learn, there's a lot that you could do there. [inaudible 00:22:36] someone from decision-making at all, but you're equipping them to be more informed. You're helping people access better information and to access it in context where it could be helpful. And that kind of part of the dream, I see that as very realistically in focus and something we can go after without... But even there, there's a part of it that's streaming and a part of it that's like, "What is the bar that we hold ourselves to put something out in the wild?" And I think an area that we've differentiated ourselves compared to some of our competition is being a bit more forward and transparent about the methodology and statistical reasoning that we use to guide our deployment decisions and having some amount of rigor in that process.

Jon Krohn: 00:23:31 Nice. Great answer. And I didn't mean to suggest that Abridge was coming for doctors jobs. I think even when I said that, I was talking about decades in the future. You started off by talking about sci-fi, and that is more what I was thinking. But in the immediate term, this idea of helping doctors make decisions and enjoy their job, and then improve outcomes further with the kinds of data that you have allowing you to recommend when doctors are writing up their notes or maybe even in real time to be suggesting, "Based on the kinds of conditions that are being presented, this is typically what's prescribed or this is typically what's diagnosed." That sounds like it would be super useful, even just to give physicians some other ideas or to bolster maybe the direction that they were already thinking.

00:24:19 And I think it would be super cool, like you said, to be able to provide citations to medical papers that are up-to-date because, of course, nobody can be completely up-to-date on what's going on. And so to be able to see that in real time like, "Here's the latest literature, here are the latest studies." It's really cool to think that a company like Abridge could potentially even be... I don't know. Again, this is getting a bit more sci-fi again. And so take it with a grain of salt from me. And I'm sure you'll definitely say that this is sci-fi, but you could imagine that...

00:24:55 You were talking from the very beginning about dozens of subjects being in a medical study when there are tens of thousands or hundreds of thousands of patients with that condition. And so it's conceivable that a tool like Abridge could someday be in real time saying, "There are 100,000 people in the United States currently afflicted with this condition. And of those, 20,000 were given this prescription and there was this outcome." That kind of thing becomes possible when these data are being collected and when models are being run over them. I'll give you a chance to comment, but I know I'm getting out of control.

Zachary Lipton: 00:25:39 No, I think that broadly, sort of everything is possible. And I think that that's sort of the magic of this company. We like to talk about it as a T-shaped company where you could think of like the stem of the T would be like, "If that's all you had, that would be like one of these other companies in the space that are more like just sort of pure commercialization efforts." The tip of the spear, that's the thing that has shot off and become a real full-fledged go-to-market operation, a commercialized technology, something that's out there. But the base of the T is the more foundational R&D. If all you had was the base, you'd be a research lab. And if all you had was to stem, you'd be a kind of pure commercialization effort of like, "Let's just kind of grab the tools and wrap them in a software package and try to hit the market hard."

00:26:56 And I think we see that there is a special connection between these because we nail one problem and really do right by our customers. We're able to reach a level of quality that other people wouldn't be able to have if they didn't have the same caliber of scientists working, the same kind of technical acumen in the company. But also, we're able to learn from that data, we're able to come back and potentially... Whether you think of it as a new product line or you think of it as expanding the offering, we're able to develop new technology. And some of that effort in the kind of foundational R&D level is pursuing ideas that might be transformative in 10 years, but some of it is pursuing ideas that could be relevant, could be reshaping our current offering on the scale of weeks to months. So I think that's...

00:28:02 I think the thing that's been special and that is really important to me to keep is that somehow we have this thing of, we have the foundational science happening, we have the commercialization happening, and we have all of them kind of informing each other. And we've been able to do it in a way where it's not like a siloed company. Because this is, I think, the failure mode that I've seen so many times at companies, is that you wind up in a place where research is this completely segregated out unit and the definition of a good manager in research is someone who keeps anyone associated with a product from coming anywhere near you. And it's like that kind of company, where it's like your main job is just maybe to not touch the company data, to not really think about their business problems, to publish some papers. The value prop is you make the company look good, there's reputational benefit, they get to be like a leader in AI, but there's this weird gap that forms of friction.

00:29:11 Oftentimes, there's even a weird elitism or hierarchy of the product people think the research org folks are out of touch, the research people think the product people are not... whatever, whatever. And obviously, I think this challenge, it's easier to do it when you're a smaller company. So, I don't mean to, by any means, be like, "I figured it all out and Meta hasn't," or something. It's much harder on a larger scale. But to create a certain kind of energy where every single scientist feels directly dialed into the mission of the company, is directly aware of what's happening, and deriving insights and inspiration from what we're seeing in the field and is even [inaudible 00:29:58] wired.

00:29:58 And we've been very clever about how we design the processes around how feedback is collected and where it's channeled so that everyone's wired into it. It's also a lot easier to get excited about it when the form of feedback is you're getting notes from the field, from clinicians. It's written feedback. Often it's personal. It's insightful. Sometimes they're even designing features with us, because they're like, "Oh, I would've liked it if the app did this, this, this." And sometimes it's like, "Oh, why weren't we doing that yesterday?" And other ideas, you're looking at it and you're like, "Well, that's a little bit sci-fi, but maybe you should take it seriously." And other ideas are kind of wild. But I think that given that... is you can get excited about it in a way that I think is hard to get excited about, like improving click-through rates or something like that. And that feeling of impact.

00:30:58 It's a strong selection criteria. We have to select on talent and skill and grit and hustle. But we select really hard on... if someone's sort of like, "I could just as easily be doing this work for a hedge fund," or something. It's sort of probably not a great fit for Abridge. And I think because we've selected on a certain kind of mission alignment, that's the glue that holds together the kind of more foundational research, the near-term research, the sort of translation of models into a live product, the app engineers, the customer success folks, the go-to-market team. It's strangely cohesive. Every other organization I've been at, you're like... there's this sort of like, "Engineers keep the business people away from me." And I think here we have this kind of energy that is made possible by a kind of belief in what we're doing.

Jon Krohn: 00:31:55 This episode is brought to you by Data Universe, coming to New York's North Javits Center on April 10th and 11th. I, myself, will be at Data Universe providing a hands-on generative AI tutorial, but the conference has something for everyone. Data Universe brings it all together, helping you find clarity in the chaos of today's data and AI revolution, uncover the leading strategies for AI transformation and the cutting-edge technologies reshaping business and society today. Data professionals, business people, and ecosystem partners, regardless of where you're at in your journey, there's outstanding content and connections you won't want to miss out on at Data Universe. Learn more at datauniverse2024.com.

00:32:33 I feel like people who want to have a belief in what they're doing is increasingly common. It seems like the younger someone is, the more compelled they are to be having their life work, their employment, be into something like Abridge is doing, where you are making a tangible medical impact. I could just be super biased because that's how I feel, and I left a career as a trader to hedge fund to get into more medical applications of data science. But it seems like that is the kind of thing... When I think about data scientists I know from the community from New York, and I think about, if somebody was to reach out to me and say, "Hey, we're looking for a data scientist for our hedge fund, do you know anyone great?" It's a relatively small list of people that I can think of that are like... That is really what they love doing, finance. But when you have something like Abridge, I'm like, "I think almost everyone I know who's a data scientist would love to be working on these kinds of problems." So yeah, super, super cool. To kind of recap based on our research, some of the cool things that we found that you can do today that are not science fiction... Though, some of them sound like they could be. If we'd been doing this podcast two years ago, you'd be like, "How are they doing that?" And so, it's things like using ambient listening and generative AI to reduce the cognitive burden of clinical documentation, as we've been talking about, reducing burnout, as well as enabling clinicians to spend less time with computers and more with patients. Industry-leading automatic speech recognition that's specifically designed for healthcare applications can accurately transcribe speech in challenging environments. For example, when there's background noise or when multiple people are speaking.

00:34:24 It supports over 14 languages including handling people switching rapidly between languages. For example, when there's an interpreter mediated conversation, which must happen all the time in hospitals around the US and around the world. And then there's a development of an in-house large language model, LLM. Allows greater customization and responsible use features such as transparency, for example, links to transcript or audio like you described in your workflow there, and evidence extraction for a verification process. So this sounds like incredible stuff and it is really at the cutting edge. Since we're a data science podcast and we've now talked about, kind of, how the tool works in the flow of work, maybe now is a great time to dig into the cool stuff that you're doing on the LLM side, on the data science side?

Zachary Lipton: 00:35:15 Sure. Yeah, where to begin?

Jon Krohn: 00:35:20 Yeah. Well, I guess, with... I don't know, the coolest stuff that you feel like you're working on now, that the company's working on now, I guess, that you can talk about on air, would be awesome. The kinds of projects that you're excited to get to work on in your workday.

Zachary Lipton: 00:35:43 Sure. There's so many aspects of it that are really exciting. I guess... I don't know if the best way to go is... It's hard to order by excitement. I guess, one way to order, is to order by the pipeline, maybe? Like, what's there?

Jon Krohn: 00:36:01 Yes. That's a good idea.

Zachary Lipton: 00:36:03 I think I'm [inaudible 00:36:04]. So I had a funny route to the company. I was an advisor for a while, and when it became clear, I was like, "This is our moment to do this." I dove about a year and a half ago, that was when I became chief scientist. That's kind of why I have too many titles, is my role kept changing. And it was a few months later, I stepped into the fuller CTO role. When we were thinking about how do we go from cool toy, cool proof of concept, to a technology that people can't live without, a technology that sets new standards for quality, for robustness, for the sheer range of situations of which we could deploy.

00:36:56 One kind of weird thing to get around is, there's a tendency to... One thing that's wild about the system is how many parts of the apparatus have to chain together to get to the full experience of the product. There's speech recognition, there's the alignment model that sort of gets the timestamps right of exactly what [inaudible 00:37:21] names are said, of what times. There's diarisation, there's the content structuring, there's the abstractive summarization layer, like the different LLM things that are going on in there. There's the guardrails, then there's through the editor, through the linked evidence, and all these components are kind of coupled in a tricky way, which is that like, you make a change upstream, you have a distribution shift on the data downstream.

00:37:46 And so there's a natural tendency, like inertia or friction or laziness or something makes you want to... What you want to do is you want to... Or like the tendency. Maybe not what you want to do, but like what laziness makes you want to do, is to kind of build a component and freeze it, or make a commitment. [inaudible 00:38:07] going to use extra ASR freeze, then we're going to iterate on something else, freeze, then we're going to iterate on something else, freeze.

00:38:16 And then you kind of wind up in this world. And the danger of that world is that you find yourself... almost everything you're doing is in post-processing at some point because you've committed to this, committed to this, committed to this, committed to this, and then you're like in bandaid world. Now everything's about the bandaids. And then it's like the first giant process thing that we've been excited about is like, how do we design an [inaudible 00:38:38] that can move simultaneously and improving the speech recognition, improving the content structuring, improving the abstractive summarization, improving the guardrails, improving the link evidence. So there's both... I think a lot of companies think about scientific practice in terms of, you are training model and there is some code that does the training and there's an objective function and there's an eval loop. This is where this science happens, in this box, and each component has its own little science box.

00:39:09 And I think I've realized that we have to zoom out and have a view of scientific practice that involves end-to-end evaluations of entire systems. That has to wrap our heads around, how do we do evaluation in a world where there's not a single agreed upon scaler? There's no analog of accuracy and... Speech recognition might have the closest because you have word error rate. LLM part of the equation, you have many, many different ways that you try to capture different aspects of accuracy. But at the end of the day, any automated metric you have is a partial view and you need to wrap all of these in human loop eval with clinician experts to basically cover yourself for all the unknown unknowns.

00:39:57 And so, I guess, the first thing I've been excited about in this role is, as an academic, usually you just don't have the people to simultaneously be building better speech and making noise robust speech, and making speech deal with medical terms that are emerging in the vocabulary that weren't there during the cutoff date. And also improving the content filtering and the LLMs and the guardrails and handling multilingual, and doing all these things at the same time. But you can only do one of those things, so you never get to have these systems level view. But we're able to step backwards. So it's like on one hand we have research programs in each of these components of the system, we also have evolved a kind of statistical practice around how we do end-to-end evaluations and on what criteria they need to be triggered. And then on top of that, how we do in-vivo evaluations, how we do staged rollouts, how we do A/B testing when appropriate.

00:40:54 So having the opportunity to think... As an academic, I've thought for a long time about problems of distribution shift and problems of model monitoring. And usually, from a very academic perspective, let me drop a theoretical model of like, what is it that's allowed to shift and what's not allowed to shift? Because if I don't make some commitment, then the problem's impossible at [inaudible 00:41:12]. And then once I make that commitment, what would be a statistically efficient algorithm to detect the shift, to quantify it, to correct for it? But that work doesn't quite prepare you for the messiness of reality and the extent to which you ultimately need a human loop eval to which you need in-vivo eval.

00:41:31 So the first thing that's been really exciting has been designing on that scope, thinking of that whole thing. The entire product is the province of science, and it is part of what attracted me to step into a more producty role as opposed to just being like... or the product manager will tell me what to build and then we'll work on the modeling, is I think so much of what's intellectually exciting is how you manage the entirety of the system as a living, breathing organism. Then we could go into each of those components and there's just so much red meat there. Speech recognition, I did that before. One of the things that's really challenging is you've got speakers, many languages being spoken, sometimes simultaneously, or rapid switching, like code switching, and kind of polyglot conversations. You're going from one language to another.

00:42:23 You also have just a wide variety of noise conditions. You have noisy backgrounds. You have people who are starting to use [inaudible 00:42:30] like emergency department where there's all kinds of chaos going on and background beeping and machines. You have people using it in outpatient settings. You have people with interpreters in the room, people with parents in the room. But then this other problem is staying current because things aren't fixed. So before we took control of our own destiny with speech, we were relying on third-party speech recognition. And the problem we kept running into is the available systems have no coverage at all. We're talking about 4% accuracy, or something like this, on novel medical terms that were introduced to the vocabulary after the cutoff date for their data sets. So if you go on... I won't say which... you go on some cloud provider speech recognition and you say COVID-19, it gets transcribed as COBRA-19.

00:43:23 And you say like... There's all these new drugs, semaglutides that are really popular for... they're originally designed for diabetes management but have become popular weight loss treatments. And you'd be shocked at how many different ways you could misspell Wegovy or Ozempic. Ozempic could become Olympic. So figuring out how to deal with these kinds of problems where... And we do all kinds of things to get around this. A ton of it... I'm thinking about what I could say without running it by... a broader consensus for what I could share, but I'd say that there's sort of a couple sides to how we're able to do so well on these kinds of terms. And one is the way we're dialed in on feedback.

00:44:19 We have really precise feedback coming in. Our doctors are dialed in and because we're able to process and act on feedback about what's being misspelled where so fast, they often feel like when they type in their feedback forms that they're typing to the AI, to the point that if they are like, "Oh, this thing... I saw this thing at noon," and then if it's not fixed by 2:00 PM, they're like, "Didn't you here? Did you get my last message?" And then the next day they'll check in and it's fixed and they're like, "Ah, all right. Thank you." So it's not exactly that the Abridged brain is reading the comments and self-improving exactly, but we do have... we've designed pretty sophisticated internal business processes for capturing that information, for acting on it really quickly. And we have [inaudible 00:45:17] that facilitate that.

00:45:19 On the other side, we're able to do things like pull all the new drug names, new disease names. We're able to find out what they are to use them in context, to ultimately create a view for our model where it's as if they had been around and they are encountering them in the course of training. So one problem is that if you have a word that has never occurred before in a dataset, it's like you're doing... Think of it in like Bayesian terms, like you're inferring a posterior, but you've assigned zero prior to this being a word. So the likelihood that it's a word conditionally but is also zero because the prior is zero.

00:45:58 And so a big part of what's important is it's not even like the data being perfect, but it's being able to get a model that's otherwise amazing to have some context for having seen these words before. And we're able to do a lot on that front. So all this work makes ASR much better. The LLM space is a really exciting one. And it's one where there's a really complex interaction between like, what can we do with RLLMs? What are capabilities that are already really strong in general-purpose LLMs? What are the places where the real path to progress is learning from in-vivo, sort of, human feedback, whether it's in the form of doctor's edits or Likert scale ratings versus where we're able to move in other ways, use other kinds of engineering techniques. And there, I'd say that at a high level, our first commitment, always, is to ship the best possible product.

00:47:15 And that means that the exact... It's not like we're married to one method and that will be what powers every single component of the system, it's like we will shift the best top to bottom product that we possibly can at any point in time, and we will continue to iterate on all this. And I kind of call this like... my philosophy on it, when I evangelize with the company, it's like, you need to earn the right to be elegant. If there's something... You need to earn the right to be elegant, you need to earn the right to be proprietary. You get to ship your model when your model is the best thing in the world, that's the best thing in our fingertips. And if it's not, then you don't get to.

00:47:57 And that sets a culture because you can only have one first commitment. Your first commitment is either the quality or your first commitment is to, like, "I'm prideful about this thing that I built." And it can't be both because sometimes they'll come into conflict. And I think by making everything quality first, it set the tone and it set a standard that everything... When did we switch from third party speech recognition to our speech recognition? When it was undeniably better, faster, cheaper, more accurate, better out of sample, better on new medication domain, and not a moment before. And I think that that's been the kind of attitude that we have. There's something... I think there's overall a workflow emerging that's really interesting, which is that I think there's this... Whenever you're going after a new use case or a new behavior, there's this product discovery phase where you're developing a new feature, but it's not even like... It's not like, "Oh, just train on the data." There is no data. There is no ground truth.

00:49:09 You're going after something where you're really the mediator between many stakeholders, like helping to try to tease out, like, "What would doing this task really well look like?" And the data that even exists doesn't fully live up to that. Even if the doctors are amazing or whoever's the other stakeholder, like a case manager, because they haven't had the time to be able to make perfectly polished notes. They're falling short of their own standards because you've given them 15 minutes to write two pages. Imagine what your prose would look like in that context. So we're often in a situation in the LLM space where we're able to take full advantage of the rich toolkit that's provided by general-purpose LLMs to get in the neighborhood, get in the ballpark of a behavior, be able to iterate with stakeholders and find out, "Are we even..." Get the [...] pointed in the right direction.

00:50:04 And then once you are, there's so many ways to learn from there because you're learning from feedback, you're learning from edits. And then this becomes this engineering project that moves in a different direction and often is best served by proprietary models. But I think there's these incredible strengths of controlling the model top to bottom because of all the different granular control that you have over, what's the learning objective, what's the training data, how are you steering that model? And at the same time, there's this incredible generation shifting ability to get off the ground that is provided by some of the general-purpose LLMs. And by having, sort of, that full spectrum of competencies, we're able to do... kind of have the best of all worlds. And that's sort of the playground that we're working with.

00:51:00 Now, linked evidence is an entire other beast for how we accomplish that. But that's, I think, the amazing thing, is that as a data scientist, as a machine learning scientist, a fear in going into a startup is maybe they're talking about AI, but once you get deep into the weeds, or once you finally pivot one, two, three, four, five times and find your PMF, maybe what you're doing, it's like machine learning is not the important thing. And they're like, "What are you doing?" And I just feel incredibly lucky to be in a place where we are building something that I believe in from a mission standpoint, but also where, as a pragmatic or as a application-oriented scientist, the richness of the problems every day is kind of like being a kid in a candy store.

Jon Krohn: 00:51:58 Starting on Wednesday, April 4th, I'll be offering my Machine Learning Foundations curriculum live online via a series of 14 training sessions within the O'Reilly platform. Linear Algebra, Calculus, Probability, Statistics and Computer Science will all be covered. The curriculum provides all the foundational mathematical knowledge you need to understand contemporary machine learning applications, including deep learning, LLMs and A.I. in general. The first three sessions are available for registration now, we've got the links in the show notes for you and these three sessions will cover all of the essential Linear Algebra you need for ML. Linear Algebra Level 1 will be on April 4th, Level 2 will be on April 17th, and Level 3 will be on May 8th. If you don't already have access to O'Reilly, you can get a free 30-day trial via our special code, which is also in the show notes.

00:52:45 Two things from that amazing section describing all of the data science things, the key things that excites you about what you're up to at Abridge. Two things that I really want to highlight are this idea of being committed to providing the best quality models. So it's so easy for an AI company to say, "Okay, we're an AI company. We're going to roll out AI models even if they're not the best. Even if there are existing LLMs out there or cloud provider models or third party models that could be used instead." So I'd love the comment that you made about earning the right to be elegant, earning the right to have a proprietary model in production only when that model is definitively better in terms of capabilities, cost, et cetera. I think that's a really great guiding principle to have in your company.

00:53:39 The other thing that I wanted to highlight from everything you just said, that I thought was one of the funniest things, was consumers' expectations when you are an AI company. So the physician in the hospital network that has bought onto Abridge, that physician knows that Abridge is an AI company. And so it's so funny to me from a data scientist perspective, and I'm sure from many of our listeners' perspective, that you could be chatting with the AI and the AI is understanding what you're saying and adapting its code and model weights to deal with the situation that the physician is complaining about in these messages. "Why haven't you done anything about this?"

Zachary Lipton: 00:54:21 Yeah. It's wild. It's crazy and it's not crazy, right? Because you're like on one hand... day one, we had this feedback channel. I think we were maybe smart about setting up in a way where really directly informed product development, but also the only process, day one, for ingesting it was... also, because day one, you've got the 10 users at your first pilot or something, is hawkishly reading every single comment that comes through. And I still read almost every comment. I'll read thousands of them. I'm addicted to it. I'll drink from the fire hose. But at some point you get into a world where it's like, this is no longer a set of particles, this is a wave. And at that point it makes sense to... in addition to the fact that... I think you can never stop reading the prose because there's nothing like what that does.

00:55:28 Again, it's the sensory apparatus you'll build on top of a comment stream will detect the things that you know are issues and the things that you know are axes of the conversation. They won't tell you what are the things that you never thought to develop a sensor for. So we still read, but at the same time, for all those things that we know about, we are able to start making a certain kind of programmatic sense of the comments. So you get to the point where in addition to the fact that we're reading constantly the comms, there's also the ability to start turning this stream of data into a real-time kind of indicator of what's going on.

00:56:13 And at that point, you're like, it's not necessarily that there is this one monolithic artificial brain called the AI that is at one serving the note and reading the comments, but you have an AI system that is reading the comments and an AI system that's sharing the notes, and the various processes by which these things can be coupled through the process of model training are... That is not... it's not so... it's not conceptually removed from their picture as to be like... It becomes less and less goofy. Like, the idea that the AI is reading it. At the same time, I think something that we have to be very careful about is I think a lot of people might not appreciate the ... When we're in academia, I think it's very popular to think about online problems. And you're just like, "Oh, a data comes in and then stuff happens, and then the model's updated and then the next data comes in, and the updated model does something on whatever," and then everyone who's actually tried to train a complicated model knows how, just like when you're making updates to models, all bets are off. Anything can happen. When quality rests on a certain kind of rigor in evaluation, you can't just be willy-nilly making updates to weights at whatever cadence you want. And so you wind up in a situation where there are all these signals coming in and they are influencing behavior, but it's on a different kind of cadence.

00:57:50 I think OpenAI has been a great pioneer here. And I really admire how creative they've been here. What they're doing and this whole discipline that they've pioneered in the space with ChatGPT and learning from preference feedback, it is something that at once on a technical level looks nothing like reinforcement learning. You're not like, "I got a state, I chose an action, I received a reward, I transitioned to another state, and let me do some kind of Bellman update." It looks nothing like that, but the spirit of what they're doing is very much like reinforcement learning. If you're not thinking as a technician, but you're thinking as a philosopher, the core differentiator between reinforcement learning and supervised learning is this, am I learning from exemplars or am I learning from evaluative feedback? And the crazy thing about evaluative feedback is it allows you to continue to improve in a world where nobody knows what perfect looks like. It allows you to follow the reward signal.

00:58:53 And in a sense, what OpenAI has figured out is that, hey, you could build these absolutely wild systems, these absolutely wild action spaces and learn from evaluative feedback, but it's not on the cadence of every update you get "I'm improving". There's these longer windows of collecting data, a much longer process in the background of figuring out how to learn from it to get to the next checkpoint, of deploying that, of collecting more information, of people finding out, do they really find it better, do they transition? Does that next checkpoint not actually gain traction and they got to rewind and start from somewhere else? It doesn't have maybe the smoothness of academic RL, but it has the spirit of it and it is doing it on a scale that probably nobody ... When has someone built a system like that, that impacted so many people and that learned from the evaluative feedback so effectively?

Jon Krohn: 00:59:53 All right, Zack, so as you've been talking about all of the pieces, all of these models that need to be interconnected, ASR, LLMs, making sure that everything's happening ethically and how different parts of the system can be changing and you need to be making sure that there's not negative impacts downstream, that got me thinking about how your tech stack might work. And I know that some of that is going to be proprietary sauce that you can't divulge. But maybe there's something that you can share with us about what platforms you're using, maybe what kind of programming language makes sense when you have all these pieces working together, and maybe the orchestration of all of these pieces.

Zachary Lipton: 01:00:41 Sure, I can get into some of these and probably not others, to the same degree. At top level, we're a GCP shop, so we work with Google Cloud. That's where all of our production ecosystem lives. We also have our own data center where we do a lot of model training, and that's one source of challenge, I think, in today's ecosystem. The economics still really favor owning hardware and at the same time being harmonious with a cloud environment, is an interesting engineering challenge.

01:01:29 When it comes to making all the pieces work together, it's tricky. I think it's one of these things where there's this line that you're always walking that I think startups have a tendency sometimes to ... You could fall victim on either side. You could over-engineer too quickly and you wind up spending almost all of your time like building infrastructure that will never be used, or you're on the other side, which is you are getting expediently to technology. And then as it starts to be used, as you start to scale up, you start seeing how the pieces are fitting together and where the pain points emerge. And then you have to start making some longer term infrastructure commitments.

01:02:26 I'd say that overall, we're probably, for a startup of our size and age, a pretty mature company. But there's still a continued process of re-engineering, when I think you go 10X or 100X in usage. So I think a lot of the pieces ... And also, it's a technology that's living and breathing. And decisions that make sense when you are, say, only doing note generation are not necessarily the right decisions in a world where you're simultaneously handling speech recognition and note generation and also have modules doing post-processing and doing linked evidence, then, as all these pieces start coming together. There's endless orthodoxies and systems people have around both how to structure an organization and correspondingly how to structure software.

01:03:31 At a high level, there is a little bit of a distinction between the application, which is very stateful, and machine learning, which for the most part aims to be relatively stateless. On the machine learning side, I think almost everyone relly heavily on Python programming. That's something, again, that could evolve. I don't think that model training taking place in Python commits us to have everything happening at the model serving layer or at the API layer or at the orchestration layer necessarily dialed into Python. But it's kind of like the natural place. I think almost any AI-powered startup will find themselves at a similar age.

01:04:28 Within the model serving, we make heavy use of certain technologies that make our models really fast, that help us to handle a lot of the intricacies of dynamically batching during serving time and managing load and sharing load across GPU instances and auto-scaling GPU instances. So some of the technologies that we rely really heavily on are our model compilation. And in particular, we also make a lot of use of NVIDIA's TensorRT-LLM. And we also make a lot of use of the Triton model serving framework. And this was a commitment that we made a while ago and has paid off pretty handsomely for us.

01:05:23 So if you naively go and you start trying to build speech recognition system based on the best open-source tools, maybe tweaking them to your purposes and stand up some model serving, kind of like, "I got some L4s and I'm running those," for something like a 30-second chunk of audio, you might be encountering processing times in the ballpark of 15, 20 seconds. And we're down to a median of something closer to one or two seconds.

01:05:56 So yeah, taking infrastructure seriously really matters, at least once you have real users. And I think that's something that also requires someone operating at the level of designing an organization to recognize that very often ... You might think of R&D being the long-term play and everything else, like software, as being pretty quick. But a lot of infrastructure bets actually require a bit of patience on the scale of it could be a month, it could be a year. And that's, I think, a big challenge when you're growing is figuring out when is the right time to make those bets and to make them and commit to them, especially when you're talking about really ramping up on new technologies that you're introducing to the organization.

01:06:59 So I think we made a commitment to get really serious about the quality and robustness and efficiency of our speech serving. And there's still so much more we could do, but the full fruition of that was over a long period of time. But on the other end of it, now that we're having the kind of meteoric growth that we are, thank God that we took the time to make those sorts of investments.

Jon Krohn: 01:07:29 In addition to these kinds of production runtime considerations that you have to be making when you have a model being used in real time by clinicians by patients, another consideration that I know is something that is top of mind for Abridge based on all the research we've done, is ensuring that there's a really high level of reliability and quality, that especially in a clinical setting, you need to be sure that you're not making mistakes. So you already gave some examples earlier of Ozempic, you got to make sure that doesn't get turned into Olympic. Is there anything else that you'd like to add on how you ensure that in a regulated industry like healthcare, the AI technologies that you're deploying will be safe and well received?

Zachary Lipton: 01:08:20 There's so many aspects. I would start even before you get to the AI, of just operating a software company in the medical space that is handling sensitive information. Whether you are of the new generation of AI companies like us, or you're a traditional software company that's doing something in the payments processing space or something like that, or even a traditional scribing company, it's like humans in the loop working out of call centers, if you are somehow taking custody of sensitive information, you can't just be like a couple kids who are like, "Hey, I whipped together a web interface."

01:09:07 I think you have to get really serious about platform and infrastructure and security at a stage where a social network company of your size wouldn't necessarily be taking that nearly as seriously, because this is a kind of situation where you need to wake up every single day feeling the weight of that responsibility, that there can be no data breach, there can be no data contamination. And you need to have strong leaders who are the authorities when it comes to building robust and secure software, who are empowered to put their foot down and set really high standards for ... Because you don't have the luxury of let's grow up eventually and then do it. As soon as you are the trusted provider for a major medical enterprise, you have to already be, day one, taking security seriously.

01:10:08 And obviously, I am in a position as someone overseeing the entire organization that ultimately the buck stops with me and I have a certain responsibility as the head of the builder organization, but I'm also ... my background is AI research. And that's why our VP of platform engineering, Steve, is extremely responsible, extremely experienced at operating secure software in regulated environments, and has the authority to set standards, and the platform team can say ... Whatever it is that you're trying to do, you need to do some data annotation, you need to test the DID software, anything that comes into contact with production data is extremely locked down.

01:10:59 And you just have to take that with a level of sobriety and a level of respect that otherwise might seem discorded with the startup. When it comes to how you handle data, you cannot be like a move fast and break things company. You have to operate in a different kind of way. That's sort of at the level of integrity of the software. How do we have the sort of tests that make sure that things like data contamination are of approximately 0% probability? How do you make sure that data breaches are exceedingly unlikely?

01:11:42 Then there's the other side of what you brought up, which is quality. How do I make sure the quality that I'm putting out is great? And I would put another dimension to this. Part of it is quality, but also, part of it is, what is the function that you are serving within the healthcare environment? If we were going and taking automated decisions and acting completely autonomously, you'd be entering a whole different category of activity. If we were going and automatically placing medication orders without a doctor's review, that would be a very different sort of action from acting more like the world's smartest auto-complete.

01:12:23 And it's exceedingly important that we can't be running around like Elon sort of saying full self-driving, full autonomy, in this kind of pitching ahead of our skis, pitching ahead of the technology, because in our industry ... I mean, one, I think it's just not aligned with my values, personally, and our values and the leadership team throughout the company, but it's also in our industry, that would be a non-starter. So I think taking every kind of action to ensure that human agency remains firmly in place.

01:13:05 And this is a challenge, because people could have a tendency, if technology starts working really well, to over-rely on it. And this is why we've invested so much more than other folks in not just, how do we get to a draft that looks good or even a draft that looks great, but how do we invest in making it easy and frictionless for someone to stay in the driver's seat and really for the doctor to stay involved and to fact-check work and get from that draft to the final note?

01:13:38 So respecting and empowering people to remain high agency in that home stretch is really important. We see it ourselves, because we do a lot of human in the loop reviews of notes. And if you're just looking at one note versus a different note for the same encounter, it's extremely difficult to know which is ... because you could look at them and say, "Okay, I like the style of this one better than that. It flows better." But what if it's all [...]? How do you actually know? Because the most important thing to be evaluating for besides style and composition is faithfulness.

01:14:26 And if you have to go and fish through a 40-minute conversation full of ums and ahs and ohs and okays, it's extremely time-consuming to get to the right evidence. And also, you might not even be able to Apple F to get there, or Control F, because it might be very different language. They might've said heart attack in the transcript and they might've said myocardial infarction in the note. So actually getting to where is the evidence quickly is something that supports human agency because it cuts down the amount of time that it takes to exercise that agency.

01:15:05 Then there's quality control. How do we actually measure quality? And I think this comes from working really closely with all of our partners to identify a kind of ... you can never say exhaustive, but a thorough ontology of all the known categories of problems, and whenever possible, to have corresponded automated metrics that track them. On top of that, we have a clinician in the loop internal eval process, that whenever we push a significant component to any model, we're actually evaluating across a large number of conversations head to head of a prior system versus a new system. And these are blinded trials, so you don't know, it's not like, "Oh, the new system's always on the left and the old system's always on the right."

01:15:53 So we have a whole framework that we've developed where actually clinicians in the loop are evaluating an old system versus a new system. And they have the ability to check off specific categories of problems, but they also are registering preference information for which system's better than the other than other. And we sort of implemented a sequential hypothesis testing framework in the background that is sort of operating where it's controlling for false discovery. So there's danger of if you do a hypothesis test every single time someone registers a preference, then you might at some point be like, "We have a winner," because you just tipped over the threshold to be able to be like, "Oh, if I hadn't done all these previous tests, if all I was doing was this one test at this one sample size, I can make a statistically strong conclusion that one system's better than the other." So you have to do some kind of false discovery control that usually involves widening your concept intervals by something like a login factor.

01:16:45 So we do all of these checks in the background, such that by the time we conclude an internal eval, we have a really strong level of statistical evidence that the new system going out is usually not even a little bit but massively preferred compared to the other one, over a reasonably large corpus of conversations. Then that's not the, okay, let's deploy it to everyone. That is the, okay, we can begin the testing framework that goes on in the wild. And there's sort of a couple aspects of this, and one is a staged deployment, and the other is AB testing.

01:17:28 And so there's concentric circles of cohorts, that the innermost circle is a set of folks who are kind of like early adopter types. We've been working with them, we have daily contact with them sometimes. These are folks who they were eager to use technology when it really sounded science fiction to people. They are specially trained, they know to have a certain level of vigilance. They're in daily contact sometimes during ... if we're making a change to product with our folks, either our product folks or customer success folks who are ... We're getting a lot of back channel.

01:18:04 And so we're able to get ahead of any kind of thing, especially, I don't know, let's say we reorder the problem list to be more concordant with expectations for different specialties. And we've made sure not only to get overall feedback and to track overall star ratings, things like that, and to track the written feedback, but also to maybe even have specifically folks who are in that specialty and get their feedback and make sure we actually improved in the desired access of performance. And only after we've passed these kinds of levels of rigor are we then pushing for wider availability. And I think that's this challenge of on the one hand you'd love to just be able to be like, "Oh, let's just be updating all the time."

01:18:57 And on one hand you should be able to be improving really rapidly, and we are. At the same time, what it takes to go from, "My experiment in the ML sandbox looks good," to, "This is what everybody is using in the wild," is this complicated process. And I think where we continue to innovate is, how do we develop the sorts of processes where there's so much system around these things that we're able to make improvements and kick off these steps? And there are some things that have to be bottlenecks, like anything that involves an internal clinician in the loop eval, that we're able to maybe bundle up updates in the right way that we know and we're able to define an intervention appropriately such that the thing that is being shipped has been exhaustively and rigorously compared to the thing people are currently getting.

01:19:55 And so I think it's all those things, all at once. That's probably a much longer answer than you wanted, but all the traditional stuff that you have to do just to be operating in the medical space responsibly, it's also all these ways that we assess the quality of the ML system itself. It's all the things that we do to keep humans in the loop so that we maintain, I think, that appropriate place and the agency structure of the process, where the doctor remains in the driver's seat. And then of course, it's not just relegating eval to offline eval, but thinking of eval as a never-ending pursuit that includes offline eval but extends into Invivo monitoring and staged deployments and a mature experimental platform.

Jon Krohn: 01:20:59 Yeah. It was a long answer, but it was also a great answer. And it's interesting to me, it seems like there's a hint of irony to me that so much of what you do now in your role at Abridge is ensuring correctness and quality, when your lab name at CMU is the Approximately Correct Machine Intelligence Lab. And so I think your research there is focused on domain adaptation and robustness. I don't know if there's something that you want to quickly add on. This will be probably just the only minute or two in the entire episode about your ... This is obviously a huge part of your life. I don't know if you want to add something in there.

Zachary Lipton: 01:21:42 Yeah, the name of my lab, the Approximately Correct Machine Intelligence Lab, it's kind of a play on a few things. And the most obvious machine learning like haha pun, is that the sort of dominant theoretical framework for thinking about machine learning algorithms and providing statistical guarantees are probably approximately correct. You're almost never in a system or a situation where you can just say the system's perfect, but what you want to be able to say is, "Oh, if I have at least so many samples, then with probability, at least something, I am at least so close to the best model in class."

01:22:22 For us, the term Approximately Correct is kind of another meaning that's layered onto it that maybe speaks more to the flavor research that I've always been attracted to as an academic, which is there's that obvious, almost banal sense of which machine learning is never perfect, because it's statistical in nature. Just like if you ask somebody what is the average height of people in the population, you never get the exact height. You just get closer and closer to the truth with higher and higher probability.

01:22:55 But there's something deeper, which is that the ML problem as conceived as a academic thing to go after with statistical tools, it always bears a approximate relationship to the real world desiderata that you're after, like our goal is, say, eliminate physician burnout, or our goal is improve care. But then what we end up doing is then, sure, we maybe work from that to something more specific, like help people take more accurate notes more efficiently. But then if you work, say, in an academic sense, you go all the way back to what would be in an ML paper, very often it's something like, "Given a data set consisting of some number of transcripts and associated notes, produce a model that scores well on some unseen test set in terms of rouge score," which might have some relation to performing well as actually a doctor would experience, but also leaves out all sorts of aspects of the real world problem.

01:24:06 And then of course, you're usually, in an ML paper in an academic setting, evaluating on a particular test set, whereas in reality, you want not only to chase a much more expressive set of desiderata, but you're also trying to perform well in changing conditions, in a world where new people are going to use the product in new ways, where their behavior is going to change via interaction with it, where new terms and medication names are entering the lexicon all the time. And so the name Approximately Correct, it's sort of like an expression of humility, I think, of all these ways in which the narrow academic treatment of a machine learning problem can become disengaged from the thing itself.

01:25:03 My academic agenda was largely around and still is around, has been about trying to come up... pick some phenomena, some axis along which this kind of divergence between what people are actually doing and how we think about it as modelers become disengaging to try to bridge those gaps, try to come up with whether it's some theoretical framing or some more realistic empirical treatment or some just conceptual problem setting. To take some of those aspects of the real world and pulls it into the province of what we think about as machine learning scientists. So that's been my impulse as an academic to think how can we model a distribution shift problem? When would such a problem be well specified? What assumptions would license which statistical methods and what guarantees could you make about them? Could you reduce them to practice and actually get even under the slightly restrictive assumption, create a semisynthetic setting where you actually get these algorithms to work on high dimensional deep learning type data? So that's been my impulse as an academic.

01:26:16 Also on the more moral and political side, when we talk about fairness, when we talk about transparency to try to bridge between what actually are people calling for in the real world? How are these concepts understood, for example, in the context of the law or in the context of political philosophy versus the kinds of naive things that maybe we've been doing in the technical machine learning discipline. And to try to sit in the middle and to try to pull the technical work closer to some of that real world messiness.

01:26:54 I think that in some ways as an academic, I've always been drawn to these things. It's the ways that the real world is messy that are unaccounted for. But then you go look at most machine learning research operations within a large company and you say who's the person who's taking those problems seriously? Who's the person who's actually getting the phone call when things are going wrong and it's impacting a customer?

01:27:26 Who's the person who's really seeing all of those? They're ultimately business metrics that are tracking in vivo performance. Who is that at a company? And very often it's not the machine learning scientist. Very often the machine learning scientist, they got a data set, they built a model, they threw it over the fence and someone else took it and implemented it. It's very often it's someone whose title is GM or someone whose title is product manager, who's actually the owner of all of that, all those things that correspond to that rich messy domain.

01:28:05 I think my impulse as an academic has been to try to take some of that and internalize it into how we think of machine learning as scientists. My impulse as an entrepreneur has been also to recognize that actually those are the problems you care about and that's where you're serious about excellence then you can't have the attitude of product is a pejorative or product manager is your enemy. You have to be no, I want to own this. I want to be in the trenches with the product managers and with the customer. I want to actually shadow doctors in the hospital and I actually want to read all their feedback. So I think that the name speaks to this kind of thing, the separation that normally takes place between ML modeling in the real world. Maybe the name is approximate, but the corresponding impulse is to try to pull these things closer together.

Jon Krohn: 01:29:09 I like that idea that it's humility, it's an expression that these things are never going to be perfect, but that we can invest a lot of time and effort in getting very close, having these things be probably approximately correct. So final question for you at least on your background before we get into my, I already told you we always end with what's your favorite book. But just before that one is you wrote a book called Dive into Deep Learning, which is a really popular open source book. People can also order a physical copy if they want to from Cambridge University Press. I'll have a link to the book in the show notes. It's very simple. URLD2L like the number 2D, 2L.ai and great, great, great book.

01:30:00 So obviously you are also big into helping people learn themselves about deep learning, which is the technique that underlies all these kinds of things that we've been talking about in this episode particularly the things like large language models. You don't hear the words deep learning spoken as much as you did five years ago, 10 years ago, but now it is kind of just taken for granted it is this underlying foundation to make all of these magical applications happen. So fun question that came up from our research. So in addition to being a computer scientist, you're also a talented musician. So you're a tenor saxophonist and you studied with the famous jazz musician, Branford Marsalis. So Branford said that musicians should learn Giant Steps by John Coltrane, which is often seen as a rite of passage for jazz musicians. It tests their ability to navigate its complex core progressions, for example.

01:30:56 In machine learning, is there a giant steps kind of equivalent that our listeners should be pursuing to be able to show their chops as machine learning practitioners?

Zachary Lipton: 01:31:13 So I guess you really did your homework, found out who my saxophone teacher was. Yeah, Branford. Branford is amazing. Branford is kind of a legend. For those folks who don't know, Branford lived in my town where I grew up and being an aspiring jazz musician at the time, he was kind of like a god. He was hosting The Night Show Band long before John Batiste or folks like that had late night gigs. He had been on the road with Sting for a while and him and his brother, Wynton, were the biggest names in jazz music I think of that generation. Branford was a tough teacher so I think he taught me a lot of things. I'll spin back on one thing is that he does have a lot of videos talking about Giant Steps, but he has a different perspective.

01:32:13 If you go through those videos and watch them all the way through, he's often chastising the young jazz musician for being preoccupied with shredding over Giant Steps versus maybe really knowing the language of the blues, really having the time feel of the rhythm of jazz music, of really playing swing. He has a certain kind of... It's almost like the way someone might chastise a certain kind of engineer for getting lost in what is technically cool, but not having the meat and potatoes of what matters of making music. That kind of really speaking the language of really doing the things that matter.

01:33:01 So I think that in some ways I know Branford's impulse is sort of to direct people into not just being a so-called jazz nerd, but to having a certain contact with what's important. I think that that maybe partly informs my impulse. On one hand I like to grow technically, but I do have a kind of gravitation to what I think is at the heart of the problem, which may or may not have anything to do with a flashy technique. I think he also just taught me a lot. Branford was on top of the world, he didn't need to be taking students. There's a certain character of recognizing when someone takes his music seriously. When I was 13, it was the most important thing in the world for me, probably for another 10 years it was. Maybe even somehow deep down it still is. But Branford, he wouldn't charge me. So he had no obligation to butter me up or tell me that I'm talented or tell me that I'm good.

01:34:22 There was a culture and that was you go take a lesson with Branford because you want to know. He would tell you [...] as it was and it was brutal. You're a 14-year-old kid being told just how badly you play. But there's also something real when you earn just a little bit of respect and you know it's earned. There is, I think in that generation of jazz musicians, a certain kind of culture of a certain respect for the craft and a certain culture of straight talk, a kind of candor that I think is rare to find in most workplaces. I think that part of, when I think about how we set the bar and how we define what it means to really perform this craft, I think some of that I think-

Jon Krohn: 01:35:40 The craft of machine learning.

Zachary Lipton: 01:35:42 Yeah, exactly. And even machine learning, now, people just sort of think oh, it's a job. It's a high paying job. It exists. There's lots of them. But also when I came into machine learning, there was also a feeling of you had very few routes from machine learning grad school to get to continue to do machine learning. There were very few jobs in industry and there were very few jobs in academia and most likely you came out, it wasn't now I'm entitled. Where's my machine learning job where I get to do cutting edge AI research all day? It wasn't an expectation. I think that there's something, I don't know, something special that music is even more extreme.

01:36:25 Music is like there's 20 people who get to play saxophone for a living and everyone else, you either have to teach or you have a day gig. Granted I didn't necessarily become one of the 10 great saxophone players in the world, but there's an attitude, a mindset of when you think of this is what it is to do creative work, this is what it is to earn the right to do creative work, that there's something special in that that exists in a field where you can't take anything for granted. I think that coming from that into software, into machine learning, it definitely colored how I view what it means to create an organization that operates at a level that feels like it's world class.

Jon Krohn: 01:37:18 Nice. You really tied that together well. The high bar at Abridge, perhaps thanks to Branford in some ways. Fantastic. This has been an amazing interview. Zack, as you know, before I let you go, I've got to ask you for book recommendations. It sounds like you might have a bunch for us. I don't know if you have time to give us a bunch because we're already running overtime now, but whatever you'd like, the floor is yours.

Zachary Lipton: 01:37:44 I'll give a few. One book recommendation is a book called Recoding America by Jen Pahlka. So I think it's probably not the only person to put this high on my list. But Jen had served in Office of the CTO to the United States, and I think she co-founded Code for America. She was around during a lot of the efforts, for example, to rescue healthcare.gov. But she goes through so much of how the disconnect between passing policy and having an idea of what you want to happen versus the execution of that policy, which these days in the modern era takes the form of software. You get to live through in this book all these ways that the best intentions of policy makers get translated into completely broken and backward systems because of the way the institutions operate.

01:39:03 There's so many books out there about this kind of, I don't know, whatever you want to call it, tech business person, self-help kind of genre of The Lean Startup, books about Agile, books about how a tech organization should operate that are all kind of talking about, I don't know, evangelizing some system that's some response to the old backwards waterfall development. But there's something so refreshing about Jen Pahlka's book that you're not just getting the same cliche perspective. You're getting maybe the most powerful version of it because you're seeing policy initiatives that are some of the most important movements in society towards universal health coverage that are falling on the rocks of implementation.

01:40:01 You get a piece of how policy works, but you also get a piece of how software development works, a bit of how design works, a bit of how our institutions fail. A bit about interaction between folks and the software kind of world who have one way of operating and large slow moving institutions that come from a completely different time and era. Where the way you build something is you write out a 50-page requirement doc and hold someone to every word of it, even if it turns out through the process of implementation that you realize half of the document was wrong or not really aligned to what we wanted to do in the first place. So I think it's this very special way to come in contact with that material that feels like it hits two birds with one stone and maybe is kind of less cliche and kind of refreshing. That's one book that I love that I read recently.

01:41:02 I'm a major literary sci-fi nerd, so I'd be remiss if I didn't recommend a science fiction novel. And one of my favorite kind of unsung writer is Peter Watts. I think he published under the Tor Imprint for a while. He has so many great books, but one in particular that was the one that introduced me to his writing, it's called Blindsight, and it's got a little bit of everything. It's kind of wild. It's got a bit of aliens, it's got a bit of vampires, and it's got a bit of a meditation on the relationship between consciousness and intelligence. I think maybe in today's era is suddenly so much of the philosophical energy of our time is around ChatGPT or something that's building towards, is it really intelligent? Is it really unconscious? Is it really conscious or does it possess consciousness? In what sense are those or are those not the same question? It's just a enjoyable read. It's relevant.

01:42:12 And then maybe as a third and final book recommendation on the music side, since you call out my background there. There's a book that I finally read recently that was just breathtaking and it's called Kansas City Lightning. This is a book about the sort of early life and rise towards the artist that he ultimately became. It's a biography of the early part of the life of Charlie Parker. It's told by the late great Stanley Crouch who was a long time a musician, a critic, provocateur writer for the Village Voice.

01:42:58 I feel like so much of the story of jazz music and how it came about is often told just in terms of style. These people played like this. And then the be boppers started playing lines that sound like that. You hear about someone like Duke Ellington as they arrived and you hear about them as kind of a great artistic force, but divorced of all of the insane social environment that someone like that entered and had to succeed in.

01:43:29 He wasn't just there were people who played in a different style and now he was introducing all this harmonic sophistication, he was coming into playing venues that were normally programming, exceeding racist programming and had to go and step up and create in that environment a whole new way of perceiving Black art and creativity in this country. I think among other things, it's a window into the birth of a truly transcendent artist in Charlie Parker.

01:44:02 It's also a window into what real hardship looks like, and I think for a lot of us on our startup journeys and on our research journeys, it's very easy for us to think our life so hard and to get mired in a certain kind of, I don't know. I don't want to sound too hard, but a certain kind of whether it's self pity or a certain kind of indulgent mindset.

01:44:35 I think to just teleport back to the experience of folks coming up completely inventing modern musical vocabulary and the way we think about across the globe improvised music in an environment where you're growing up in a segregated United States. You are having to cut your teeth playing at brothels and just a completely different perspective on the environment in which folks can still create truly transcendent creative work. So I think that is just as a creative person and as a citizen and as a music lover, I think one of the greatest.

Jon Krohn: 01:45:25 I wonder how many machine learning practitioners we have out there listening to the show today who are cutting their teeth in a brothel. So fascinating recommendation, Zack, and a fascinating episode overall. So much rich content in there for us to sink our teeth into. Zack, if people want to follow you after the episode, what's the best way to do that?

Zachary Lipton: 01:45:48 I guess best way to connect is I'm still for the time being on Twitter, so I guess @Zacharylipton. I have been derelict in posting new content, but I still run the approximately correct blog and hope to put more cycles in on it soon.

Jon Krohn: 01:46:13 Nice. All right. Thanks so much for the time, Zack, and for running over. Being so generous with your super in demand time at your fast growing startup, your flourishing lab. Thank you so much. And we'll catch up with you again in the future.

Zachary Lipton: 01:46:26 Awesome. Thanks so much.

Jon Krohn: 01:46:33 What a fascinating guest succeeding and accomplishing fascinating, impactful work. In today's episode, Zack filled us in on how Abridge uses a pipeline that includes several proprietary generative AI models to transcribe, summarize and highlight clinical conversations safely into a high degree of accuracy and security. He also talked about the GCP Python and NVIDIA tech stack they use at Abridge to develop and deliver clinical machine learning reliably at scale.

01:46:58 He talked about the etymology and focus of his approximately correct machine intelligence lab at Carnegie Mellon and how receiving instruction from jazz legend Branford Marsalis helped shape Zack to have the exacting professional standards he has as an academic and machine learning leader.

01:47:14 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 Zack's social media profiles, as well as my own @superdatascience.com/769.

01:47:27 And if you'd like to engage with me in person as opposed to just through social media, I'd love to meet you in real life at the Open Data Science Conference, ODSC East, which will be held in Boston from April 23rd to 25th. I'll be doing two half-day tutorials. One will introduce deep learning with hands-on demos in PyTorch and TensorFlow. The other will be on fine tuning, deploying, and commercializing with open source large language models featuring the hugging face transformers and PyTorch lightning libraries. In addition to these two formal events, I'll also just be hanging around, grabbing beers, chatting with folks. It'd be so fun to meet you there.

01:48:03 Alrighty. Thanks to my colleagues at Nebula for supporting me while I create content like this Super Data Science episode for you. And thanks of course to Ivana, Mario, Natalie, Serg, Sylvia, Zara, and Kirill on the Super Data science team for producing another enlightening episode for us Today. For enabling that super team to create this free podcast for you, we are deeply grateful to our sponsors. Please consider supporting this show by checking out our sponsors links, which are in the show notes.

01:48:29 And of course, if you yourself are interested in sponsoring an episode, you can get the details on how at jonkrohn.com/podcast. Other ways you can help us out include sharing the episode with someone that you think would like it, reviewing the episode on your favorite podcasting platform, subscribing if you're not already. But most importantly, of course, we just hope you'll keep on listening.

01:48:52 So grateful to have you listening. I hope I can continue to make episodes you'd love for years and years to come. Until next time, keep on rocking it out there and I'm looking forward to enjoying another round of the Super Data Science Podcast with you very soon.