[00:00:00] **Dave:** All right, Mike, Josh, welcome to Security Voices.

[00:00:04] **Josh:** Hey Dave. Thanks a lot. Thanks for having me. Pleasure to

[00:00:07] **Mike:** be here. It's

[00:00:07] **Dave:** been a long time and coming. We wanted to have both of you guys on for a while. Glad we're able to finally make it work. Josh, you're up in the peninsula now? Yeah, up in

[00:00:14] **Josh:** the Bay Area? Yeah, up in San Francisco actually, it's, it's, I know it's sunny outside, so you obviously wouldn't think it's San Francisco, but it is and this is where I am.

Yeah, spring has sprung. We're having been a good spring.

[00:00:25] **Dave:** Yeah. Yeah. I came home from the Midwest and opened up the windows today and it was like, oh my God. It's actually really, really nice. It's time to hit the beach as we, as one does in Los Angeles. And Mike, I'm assuming you're home. I see the hockey jersey up there.

Yeah,

[00:00:40] **Mike:** I'm in Denver where it's a balmy 45 degrees, but in typical fashion, the sun is out, so it looks like it will be a pretty nice day. Awesome.

[00:00:49] **Dave:** Alright, so quick intros across the board. So important that Mike has a hockey jersey up behind him because that's actually how we met. He and I met a little bit less than a year ago.

Feels like a lot longer than that. But we met about a year ago at a Colorado Avalanche game. Mike is a Avid hockey player. Hockey fan, and is that a Nordiqs jersey up behind you?

[00:01:11] **Mike:** It's actually an old Canadians jersey Carrie Price who used to used to Gold 10 for the Canadians forever.

[00:01:17] **Dave:** Outstanding. So Mike and I met, he is currently doing security things, how's that as a title?

Doing security things for Amazon Prime video where he has done many, many security things in the past. Most recently for Fastly, for regional bank out in Denver. And interestingly, like Mike is one of the people who can bridge both the practitioner side and the vendor side on security and kind of shares both perspectives.

Mike, what important thing did I miss out?

[00:01:46] **Mike:** I think you covered it pretty well. I've got back and forth in both the vendor and the customer space. Spent a lot of time on, uh, availability and resiliency across, enter both enterprise networks, financial services, big, big into DDoS protection and, and those kinds of areas.

[00:02:06] **Dave:** I always forget, you did some time with Arbor Networks with Doug and Crew, right?

[00:02:11] **Mike:** Yes. I was at ARB Networks for a number of years, both pre and post-acquisition by Donaher, which is a big kinda VC holding company. I was with another startup called Simplicity, where we actually did d n s policy technology for threat modeling and threat detection of basically command and control back in 2006, 2007.

Like I said, I've, I've spent a lot of time in both like kind of carrier space and, and FinTech has been a lot of my world, so there's always been a lot of data at scale problems that we're trying to chase, and now

[00:02:45] **Dave:** you have Mocho video and other content at scale. At Prime Video,

[00:02:51] **Mike:** we tend to have a bit, I think we were just recently put ahead of one of the number one streaming providers out there.

[00:02:59] **Dave:** Yeah, hopefully that's new episodes of the, of the Ring of Power. If it isn't, I'm gonna be deeply disappointed. Ring of Power was good, solid.

[00:03:08] **Mike:** It was something I definitely, definitely watched, created a lot of buzz. And we have a cross prime video, right? We have. We have a lot of stuff in the works. I'm a fan of the Jack Ryan series and Goliath.

If you've seen that one that was kind of filmed in your hometown, prime video's got a lot of really interesting products. It's actually, this is one of the more exciting and entertaining security roles I can say I've had in my career.

[00:03:31] **Dave:** All right. I'm hoping for a follow on to the peripheral William Gibson's agency.

Which is about sentient ai, which feels screamingly apropo given the moment. Yeah, that was really freaking good. As Gibson typically is even better than the peripheral. So fingers crossed, if you've got an in man, let 'em

[00:03:49] **Mike:** know. Sounds good. I think they don't quote me on this, but I do think they had already committed to a season two and that's public.

I'll confirm that.

[00:03:57] **Dave:** Awesome. Alright. And Josh and I met during the pandemic. We've actually never met in person of all things like you could be like eight foot and I wouldn't know. I'm actually not real.

[00:04:08] **Josh:** A mid journey. I'm a mid journey script that was generated. I'm not actually a real person. I like

[00:04:12] **Dave:** you as a mid journey script.

[00:04:14] **Josh:** Thank you. I, I think I would make an excellent mid journey script.

[00:04:17] **Dave:** Well, this isn't our first video. Josh and I, we did two things. We did kind of a security questions where we talked through things with ETL and reverse ETL and other stuff to kinda break that down for security types. And then we did a rousing coverage of a whole bunch of placebos of LAR tea.

And everything that we were drinking so that we weren't drinking during the pandemic.

[00:04:40] **Josh:** I know exactly. Cuz you and I were trendsetters when it came to like quitting alcohol basically, or cutting way back and now everybody's doing it and so Yes, exactly.

[00:04:47] **Dave:** That one's gonna age really well.

[00:04:49] **Josh:** Really well. I think definitely, no doubt, far, far long after you and I are both dead, that we'll still be, people still be referring to it.

Yeah,

[00:04:56] **Dave:** so Josh and I met through our series A investor at Kleiner, and we were trying to really wrap our heads at open Raven around folks who worked in the data world and. Bucky said, I know a guy. And ever since then, ask a smart person, a dumb question. Head of data engineering. Data science has always been Josh, and you've been an amazing advisor for all startups out there outside of those who compete with us.

Go get Josh as your advisor.

[00:05:24] **Josh:** Thank you for that, man. I appreciate that. 35 companies now, which is arguably a little too many to be honest with you. So I'm actually gonna gonna wave you off and say, no, no more. No, no more. I'm done. You're full. No more. I got plenty of people to talk to. I'm happy to do it, but it's kind of a bit much, I'm not gonna lie.

Yeah.

[00:05:39] **Dave:** So Josh, you did early data work at Google, you're ahead of data engineering for Slack. You did data engineering at Cloudera and you've done a bunch of open source, a bunch of investing and advising too. And currently you're working on Bueno Vista, right? Connected to Duck

[00:05:54] **Josh:** db. That's right. It's, it's my fun, uh, yeah, working on Duck DB stuff.

Funta is, I like to say it's my Postgres Pressu. Python proxy server. I need to figure out another way to get another P in there. So it can be like P P P P P, you know, it's my fun little hobby project that provides me with a great deal of joy. So I kind of try to imagine like, what would the next generation of like cloud data warehouses look like and how would it work and how would we integrate it into the rest of the kind of modern software ecosystem, basically is kinda what I'm thinking about.

Anyways. Tremendous fun. Yeah,

[00:06:26] **Dave:** and you advise, you invest, you work on your own projects and you'll continue to do that until it sucks, until it's boring, until you find something that pulls you back into the work world.

[00:06:38] **Josh:** I think until it gets boring, I think is is is my current plan exactly. Till life calms down a bit, which I'm sure life will calm down any day now.

I'm sure that'll totally happen. I will find something else fun to do with myself.

[00:06:49] **Dave:** Outstanding. Alright, so we're gonna talk on this episode about. We'll get inside how data engineering, data science teams work. And then we're also gonna get inside how folks like Mike and on the data security side are learning to adapt and work with data science teams and how they can do it in a way in which we all kind of peacefully coexist.

And we also have a common friend, common enemy, common partner, and privacy. So we'll probably talk a little bit about privacy in this one too near the end and where the intersection is and how it's. You know, I wanna make sure that people understand that it's fundamentally different than security, even though security's often settled with privacy.

I think mainly because people don't what the hell to do with it many times. But we'll bump up against that. What we're not gonna do here is there's a lot of folks that are going deep, deep into things like threat modeling, generative AI and chat, G P T and prompt injection. We'll cover some of that, but instead, more of what we want to do is cover how, how to defend, how to secure, how to understand the data, all the data that goes into creating key data science initiatives like a generative ai, like a large language model, and so forth.

So, we're gonna bias us a little bit more towards there, but y'all are unencumbered. You know, as before, like if there's a thread that you want to yank on, like by all means, go for it. So this is going to be security. And data security. Data science, and we will not be able to help ourselves from talking about generative AI because it's so freaking interesting right now.

But, you know, off we go. So first, let's do this. Like Josh, we're at an insane moment where things feel like they're changing every day, every week with respect to the potential of ai. Did you see this coming? I mean, having been in this space for a really long time, did you imagine we would be here someday?

And did you kind of see it coming? Or did it just happen all of a sudden, even take you by surprise?

[00:08:45] **Josh:** That's a great question, Dave. I mean, I think the, the short answer is no. I did not see this coming. At least like to the extent that I saw this coming and figured something like this would happen Sunday. I didn't think it would be in the next decade.

You know what I mean? I tell people, like I, I have obviously like a lot of friends who've worked in the generative AI space for years, and they told me this was coming at like parties and social events and stuff. They were like, this was all they were thinking about. And to be honest with you, like Dave, I basically thought they were crazy, right?

They sounded like crazy people to me. And they weren't crazy. They had just been living in the reality we all find ourselves in well before everyone else. Right. And the most annoying thing about this for me is that they were all right and I was wrong. And so that just obviously like bothers me deeply and stuff like that.

Such as life. Such as life. I think what's hard, even for me, someone knew like I, you know, I do this stuff professionally is I know exactly. Kinda like how these large language, oh no, not exactly. I know reasonably well how these models are trained, how they work, what they're designed to predict, right? And what they're really designed to predict is just the next word.

I'm oversimplifying, right? But it just predict that given a set of words, predict the next word. That's it. And from that, Modeling process of just predict the next word and then the word after that and the word after that, so on and so on. We get this amazing stuff out of it. Like we get absolutely incredible stuff.

And so the delta between like the mental model of how this stuff works and how it's trained and what it's designed to do versus what it can do, it feels like magic. It feels like magic to everybody. And in some ways it's amazing. It's an amazing time to be alive cuz like magic exists. There is magic in the world.

Again, we don't understand how exactly this is possible and that's really, really exciting, but it's also like terrifying at the same time. It's like all these feelings all at once. And yeah, that's the kind of the, the emotional sort of like the resonance of this event happening, like right now. So

[00:10:46] **Dave:** yeah.

What do you think, I mean, you're deep inside data infrastructure, data science and so forth, tapped into that community. What brought us to the point to where even someone like yourself was able to be surprised. Was it a confluence of factors? Was it tooling and compute and just kind of time figuring out neural nets?

What were the magic kind of combination of things where we arrive at this holy shit moment, everything's different, magic is real. What do you think brought us

[00:11:16] **Josh:** here? Oh, I mean, so many things. It's kinda like I, I gotta like, I feel bad simplifying it this much, but I think we don't have a choice otherwise we could be here for hours.

You know, you have to sort of say that this starts in 2017 with the publication of a paper called Attention is All You Need, which is probably like one of the most important papers in like human history at this point. Right? It's a really, really big deal, which essentially like laid out the architecture of what is called the transformer architecture.

The transformer architecture is the neural network architecture that is the foundation for everything that's come since prior to this paper. Architectures are very complicated. They're very hard to scale. They're very, very, very compute intensive, very difficult to paralyze, very difficult to throw data at them.

All this kind of stuff, right? The transform architecture essentially stripped away everything you did not need. And left us with something that we could scale, that we could really throw an arbitrary amount of data and an arbitrary amount of compute at, and sort of see what came out, right? So that was kind of like thing one.

Thing two was we actually needed the data and we actually needed the compute in order to be able to like train these things and sort of like push them to their limits and see what they could do. And that obviously has taken some number of years, right? Like as we need GPUs to get better, they got better, we needed data, we got the data.

Like all that kind of stuff had to happen. That just takes time. That's just kinda the way it goes, right? This is not for nothing. A fairly common thing that's happened in, in sort of machine learning a couple of different times it happened with the original deep learning stuff around computer vision back in the sort of late aughts, like early, early tens roughly, where you'd have people like Lacoon who were saying, No, no.

Trust me, this neural net stuff, it'll totally work. It'll totally work. I promise. We just need more data. We just need more compute. We'll do it. And no one believed them, right? No one believed was like, yeah, sure, yawn, whatever. Like that kind of stuff. And then like eventually enough data and compute came along and it turned out yawn was right.

And that's kind of the same thing that happened here was like for years, like the transformer people saying, no, no, just wait. We need more data. We need more compute. Once we get it, it's all gonna work. And here we are. They were right and I was wrong. And so yeah, that's what's brought us here. It's pretty exciting.

[00:13:20] **Dave:** That's a great explanation and if anyone heard people typing furiously, that was Mike and I pulling off. The attention is all you need, and lemme just give people a little taste of what they're in for here. It's a barn burner. Here's from the abstract. The dominant sequence transduction methods models are based on complex, recurrent, or convolutional neural networks that include an encoder and a decoder.

The best performing models also connect the encoder and the decoder through an attention mechanism. We propose a simple network architecture. The transformer based solely on attention mechanisms dispense with recurrence and convolutions entirely. All right, Joshua, do you wanna take a stab at that?

[00:13:59] **Josh:** So, yeah, I mean, as best I can, the old convolutional and recurrent neural networks, a convolutional in particular, convolutional neural networks were fundamental to like computer vision stuff, which was the early, early use cases for deep learning and stuff like that.

And they are, again, these very dense, very just kind of convoluted looking, to be honest with you. Like neural network architectures. Like if you looked at them as like a picture, you'd be like, what? There's like, you know, lines and stuff like a, it'd be like looking at like a messy server rack. We're all old, right?

So we remember there, there were these things called servers and they were connected to each other, right? And you, and you know, it could be like a disaster, right? It could be like wires going every which way and all this kind of stuff, right? That's kind of what they looked like. And we thought that was what we had to do in order to make this stuff work.

Cuz it seemed like that was what we had to do to make this stuff work for the transformer. That was again, designed for computer vision. Use cases for language use cases. It took a while for folks to kind of figure out what exactly was the right architecture to do. And so we went through recurrent base systems.

We went through these things called ls, tms, like long short-term memory networks. Again, very convoluted, very complicated. Yeah. Again, I'm being very glib here, but like the attention, the transform architecture just sort of dramatically simplified that amusingly. If I can throw an aside in here, we don't really use recurrent neural networks or convolutional neural networks for vision stuff anymore.

So if you've ever used stable diffusion or mid journey or dae, those are all based on what it's called a diffusion based architecture, which is completely different. And it's really like these are the two sort of dominant architectures of, of our time is the diffusion based model for visual stuff, music, that kind of thing.

And the transformer based architecture for language arch architectural speaking is kind of like general relativity going up against quantum physics in some sense. Like they're not. Like they're very different approaches to the problem and there's a lot of people trying to apply the diffusion based model to the transformer based problem because it looks like it may be possible for diffusion to be able to do what Transformers can do now, which would be again, another absolutely seismic shift in how we build these systems and stuff like that.

So this is like a huge other thing that's going on in the research community and stuff like that

[00:15:58] **Dave:** right now. It's really interesting. I think a lot of us, I think part of the magic of Josh Wills is that you can explain these things in relevant terms to people.

[00:16:06] **Josh:** Yeah. It helps that I'm like five or six years old emotionally speaking, so like that's a kind of just resonates with me.

Yeah, exactly.

[00:16:12] **Dave:** You have a gift no matter how you downplay. You have a gift at explaining these things in terms that at least security people can understand. And I can, I'm assuming others. I haven't tried being another person, but if I was, I'm pretty sure I'd like it. I'd like the explanation as well. I think the, the interesting thing to me, Is that we didn't just happen upon the transformer model, but we happened upon the diffusion model.

At the same time, seemingly, or at least they both kind of hit this apex around the same time. It

[00:16:38] **Josh:** feels pretty close to each other in deep learning years. Pretty much Right around the same time. Yeah, exactly.

[00:16:43] **Dave:** Is there a reason for that? Is it just coincidence and kind of time

[00:16:47] **Josh:** served or, that's a great question.

That's a great question. I have no idea. To be perfectly honest with you, I have no idea why. I think it's just a lot of smart people have been thinking about and paying attention to this kind of stuff and trying to make a name for themselves in this space. I think, as we all know, that's a fantastic opportunity for innovative, amazing stuff to come out, and I, I just think that's what's happening.

Yeah,

[00:17:07] **Dave:** and it does kind of all feed off each other in an ecosystem and so on. Exactly. There's some competitive

[00:17:12] **Josh:** effects. Precisely. That's why, I mean, it's why it's, to me it's like why it's so exciting and fun. To be in San Francisco right now is like everything. It feels like the early.com or the early website, you just felt like everything is happening here.

Everything is happening at like at dinner. In restaurants and stuff and people bumping into each other, that's, that's what's so fun about being here right now.

[00:17:31] **Dave:** Mike, what's it? And of course, as always, don't say anything you're not comfortable with here, but what's it like being inside an Amazon where y'all have massive amounts of data right now and you actually, you have your own services, right?

That enable tooling and make data science easy. If not easier and so on. What's it like being inside a large organization like Amazon right now? I'm sure you have, you guys have a number of things that you're not talking about publicly, a number of things. You are, I'm assuming you're looking at this both as an opportunity for Amazon to get better, but also to create more services and so forth.

What's kind of zeitgeist on something like this inside such a massive organization

[00:18:08] **Mike:** at the highest level? Right. We're always looking. To refine how we manage and scale the data that we collect. We don't sell data to third parties, right? We keep it internal, but on on the prime video side of the house, right?

We, we do things like recommendations. So if you, if you like watching action movies, we wanna be able to recommend action movies. How do you always tailor that? How can those data sets be more timely and more accurate, right? We're doing live events. How do, is it time to notify you about sports cuz you're late a sports on without having to necessarily, without you having to go program all those rules on what you like to watch or have to pick your favorites.

How do we provide a better experience?

[00:18:51] **Dave:** I'd imagine the Alexa team is straight up losing their shit right now. Can you imagine like, every time I interact with Siri, I think of it as like, why the hell isn't this chat g p T? It seems like such a huge opportunity for the personal assistant

[00:19:07] **Josh:** tech. I dunno about y'all.

Sorry. Like, whenever I interact with g p t, with chat, G p t, like the, the 3.5 model, instead of the four model, I'm like rolling my eyes at how stupid this thing is. Like I can't tolerate how stupid like G P T 3.5 is after getting used to G P T four and it's been like a month. It's a crazy town. Anyway,

[00:19:23] **Mike:** if you go look around, there's definitely been some public comments on testing interactions of Alexa versus G P T three five and G p T four to, to show some of the differences.

I think we're a company that's working on a lot of things. I don't even necessarily have insight into, you know, we do want to provide what's best for our customers and if there's a way to make those work together, I, I don't see how we don't do it. And

[00:19:47] **Dave:** we were talking about this a little bit before we jumped on, but part of the thing that's freaking amazing about right now is, alright, so let's say the Alexa team was behind in developing a large language model and maybe didn't have like the neural network chops that they needed.

The beautiful thing of where we're arriving at is they don't have to. You know, the ability to take advantage of the technology is available to so many organizations in a way that never was before, that it's gonna democratize access to neural nets. This thing that was like, you know, arcane, esoteric science until only recently.

So it's pretty freaking amazing. I dunno. Josh, can you talk a little bit about that? Let's say you wanted to build a large language model today, and you are starting a company, Josh Wills corp. Like what would it take you? What would be your building

[00:20:38] **Josh:** blocks? So there's a lot of these, what are called foundational models, which is like your GPT three, your gp, PT four Llama is the one that came out of out of Facebook.

These are foundational models. New has several different foundational models as well. You start with one of these foundational models. You start with something that's already been pre-trained on a very large corpus, and then you're gonna do some additional, what is called fine tuning. You're gonna do some fine tuning of this model on your corpus, on your data set.

Essentially using the same encoding decoding strategy, the same embedding strategy as the underlying foundational model. You present it with additional data and you presented with additional sort of instructions and prompts and that kind of things in order to optimize it for your use case. The Databricks folks have been all over this.

They have a system called Dolly, like D O L L O Y, which is named after Dolly the Sheep. You remember like the original like genetic cloning kind of thing. So it's named after that, right? This is designed to help you do this with Spark, like starting with open source, freely available. Weights and models and foundational models, and then train them for your specific use cases.

I mean, it's super cool stuff. It works like amazingly well. It's really remarkable how far you can get with like relatively little data, relatively little compute once you build on this sort of foundation, these foundational models, to kind of like bootstrap yourself and get yourself going.

[00:21:54] **Dave:** Yeah. Are we gonna have to suffer through people talking about the Dai Lama stack?

Is that gonna

[00:21:59] **Josh:** happen? We're gonna suffer in some way or other, Dave, it's just kind of a question of how, if you kind of wanna rank them from like, AI destroys humanity, like the terminator scenario down to the, like, we have to put up with like some annoying marketing buzzwords. I mean like, let's go for the annoying marketing buzzwords.

I don't know what's gonna happen, but I'm, I'm saying like there are worse possible outcomes here, bud.

[00:22:18] **Dave:** You know, I think one of the greatest contributions of generative AI thus far is we no longer have to hear about NFTs.

[00:22:24] **Josh:** I was grateful to not have to hear about data contracts so much anymore. To your point.

Exactly. It was nice. That sort of took over all discourse in every vertical, in every, like everywhere, all at once, which was amazing. Cleansing, I would say refreshing, yes.

[00:22:36] **Dave:** If I'm in crypto, like all of a sudden, like FTX and Sam Bankman, Frid Freud, someone at one point called him like scam banker fraud, and it just stuck in my head and I can't say his name right anymore.

[00:22:51] **Josh:** You need to reprogram your, your large language model inside of you, Dave. That sounds like you got a prompt injection there, my man. Yeah,

[00:22:56] **Dave:** yeah. But crypto can just kind of quietly go off and heal itself. Now I think, you know, we, somebody else has the center stage, so we talked about this a little bit before we, we jumped in.

But there's a lot of dialogue and justifiably so about the LLMs that are built off the interwebs, off all the data that's out there. And famously, Elon's losing his mind over all the Twitter data that's being used. Reddit's doing the same and I'm sure there'll be others as well. You're nerfing all of our data and turn it into amazing things and people aren't coming to our sites where they can do clicky things.

Like that's a big deal, a really big deal. But you look at an organization like an Amazon that has a lot of its own proprietary data. One of the interesting things about all of this, especially as a data security company, so I will, I will profess my bias upfront. Is that all of a sudden we realize data is incredibly valuable.

And I think for a lot of times we would say, oh, well, data's the new oil. It's the lifeblood of the new economy. And when people are like, oh yeah, I can kind of see that. But now you look at it and you say, oh dear God, yeah, that thing scarfed a metric ton of data and now it's spewing magic. The value of the data itself seems so much more now and.

All of a sudden we realize that we have to take care of it. And one of the things that I wanted to talk about is forget about data security for a moment. There's a foundational level of just basic hygiene of taking care of data at scale that is really freaking hard and also rather esoteric as well.

And I want to dive into that a little bit. I'm gonna put both of you on the spot for a moment. Mike, I'm gonna do this to you. First, you've been warned and then it's coming over to you, Mr. Wills, and both of your answers, I would suspect to be different because you come from different perspectives. Mike, to you, current job, other jobs.

What is data governance?

[00:24:49] **Mike:** Fundamentally, it's understanding what data your organization has and who else has access to it, who's using it, and keeping an eye on, on the behaviors that are okay and the behaviors that you can't allow. I think it's almost easier at big companies sometimes than small companies when you think of the scale of the amount of data you actually have versus the resources you actually have to chase problems.

So like a good example is, you know, smaller startups, and I've been involved in a few where you, you have, you know, three or four people that are doing DevOps and they're doing security, but they're also operating in, in 12 different countries with a bunch of d n s data or what would be, you know, search history and things like that.

Most of that stuff is going to a bunch of random third parties and. Part of your supply chain. Maybe it's going to cloud providers, maybe it's going to recommendation engines. You end up with these scenarios where you have this proliferation of data and you don't always have the visibility into how your supposed partners are actually consuming and using that data, which I think brings us back to, well, how are these L LM models gonna be applied to by the services

[00:25:58] **Dave:** we use?

Oh, that's interesting. So yeah, as a small company, you have to use a bunch of services, no doubt about it. And tracking those down and making sure that you understand who's doing what and being comfortable with that and so on. It makes a lot of sense. All right, Josh, over to you. What state of governance.

[00:26:15] **Josh:** It was like just a fantastic answer by Mike. That's exactly it, and it's sort of the simplicity of the answer in some ways belies the complexity of the problem, like just how extraordinarily difficult it is to keep track of all of this stuff and know what exactly you have, let alone what else your partners have and various services have.

And who is the effective owner of that kind of stuff. You know, at Slack this was a, a fantastically interesting problem in that we had companies data. We had their files, we had their messages, we had their, their conversations with each other. We had all of this stuff and it was unambiguously and militant clear that it was not our data, it was their data.

It was the customer's data, like capital C, capital D, it was their data. We were not in a position to do whatever we wanted with it and so on and so forth. Like Slack can't go train a large language model off of all of Slack's corpus of data. They can't do that. It's not their data. They have no business doing that right now.

If you want, I suspect, I don't know if this is happening, obviously it's been a long time since I've been there, but I, I think if you want as a service, slack can take a foundational model, apply your data to it for you and give you your own model that you can then do stuff with as a bot or whatever it else you want.

Right? But that's only if you want them to. They're not gonna just go do that for you. Right. I think to Mike's point, this stuff is tremendously, like dealing with the governance problems is just so much easier at these big companies where you can just devote the time and resources to it and so much harder at small companies and even like, especially small vendors and stuff like that.

It's just really, really hard. I always gotta call out to the, the poor folks at Drizzly who had like this massive lawsuit and I'm pretty sure the CEO. Himself was ultimately found liable, as I understand it, for their very poor data governance policies and stuff like that. Like it's just, I, I'm not gonna, you know, name names here and I obviously, like, I don't, I would be shocked if that was the only small company that was kind of in that situation.

I would sort of suspect that problem is endemic everywhere. If you went looking for it, I don't think you'd have a hard time finding more instances like that.

[00:28:16] **Dave:** Yeah. You know, it was funny, the FTC hit Drizzly and Chegg at the same time. Okay. They were like 30 days apart and I was curious if they were gonna keep doing it.

And I think some of the reason they did it is because, There's no national privacy law and the FTC can go after whoever they want. And having worked with them, they're really freaking good. You don't piss off. There's government agencies you might be able to thumb your nose at with very little repercussions.

You do not piss off the ftc. They can only solve a few problems, but if you become the problem they want to solve, like God help you, they're gonna solve you. They will indeed solve you and they will not give you anesthesia beforehand.

[00:28:55] **Josh:** It's making me think of X Corp here, Dave. It's like bold strategy there.

Cotton. Let's see if it pays off for them.

[00:29:02] **Dave:** Yeah, so they, they hit drizzly and they did, they held the CEO accountable and they said, if you screw up your data governance, your data security, we will hold you accountable. And I think wherever that poor soul goes, in his future job, he has covenants around him and things that he has to do from a data security perspective and with Chegg.

I honestly didn't understand why they went after drizzly, but Chegg, it was a reaction to all the growth of educational services that proliferated during the pandemic and all of the sensitive student data that people had. And I think they made an example out of Chegg to say, if you're in ed tech and you're holding sensitive student data, particularly you know for young people, you sure as hell better take good care of it.

Otherwise we're going to come crashing down upon you with the wrath of God himself. So that one I understood a little better. I don't know what drizzly did to piss off the ftc.

[00:29:57] **Josh:** Got it. Nb same. No idea.

[00:29:59] **Mike:** What would be interesting Josh, and it'd be interesting to get some of your thoughts on this too, cuz when you think of data governance and how it's evolved over the last like 20 years, I would say like data security's been really big probably the last seven to 10 years.

But you know, I know we've all been in this industry longer than that and I think of, you know, back to the time in. Early, I wanna say it was 2002. I'm doing security architecture for a d p and we're building the payroll systems. At the time, I think we were doing three out of five paychecks in the us, one out of five worldwide.

And none of this stuff was online yet. Right? We were one of the first teams to build this architecture, start putting payroll on the internet so you could go download your, you know, your W two s and, and your pay stubs and update direct deposits and all your beneficiary information. And, and what I would say is like when I think of a lot of the compliance things, forget.

Internal security policy for amendment. But you think of all the compliance things. It was, it was Sarbanes Oxley cuz it was after Enron and that was the big governance thing everybody was worried about cuz executives could go to jail for, for Sarbanes Oxley. But you look at a lot of the privacy laws and data regulations that exist today and are you required to notify on a breach and all those things.

Those weren't around when we actually had a lot of very, very sensitive data starting to get put on the Nat.

[00:31:23] **Josh:** That's a hundred percent right. Yeah, without a doubt, man. Totally. For what it's worth, I was, I was a Google before I was a Cloudera and I did data analysis kind of things there, and I was always very annoyed with the folks who controlled the logs, which is Google's crown jewels in terms of a data source like the sawmill and the saw all and stuff like that, who were absolutely the most crazy militant data privacy and data security people you will ever meet in your entire lives.

They went out of their way to make my life miserable in terms of making it difficult for me to do what I considered interesting and fun analyses they considered like harmful and threatening. In retrospect, they were probably right. I was pretty young back then. They were right. And so, I mean, you're exactly right.

The regulations have like come like tremendous way and stuff like that. I, I was not for nothing. Very annoyed as a Google employee. These people existed, but also very grateful as a Google user that these people existed cuz they went outta my way to protect my data. Well beyond what the letter of the law, especially at the time.

[00:32:21] **Dave:** So since you're talking, Josh, I'll ask you this one first. Who's responsible for data governance? If you could pick one group that either typically is, or, or I'll ask it this way. Who do you think should be responsible for data governance? How should it be structured?

[00:32:36] **Josh:** Obviously I'm, I've got my own biases here, man.

I do think it should be the head of data or, or the, the head of the data team that is ultimately. Responsible for it. I mean, I guess like when we say responsible, do we mean like accountable in the sense of like the c e O of drizzly level of accountable, or like what do we mean

[00:32:52] **Dave:** here? I think if we're gonna look at data as the essential lifeblood of the business, as the thing that's feeding, like the L L M, I don't think it's wrong that the c E O ultimately is accountable for data governance.

But having said that, on a day-to-day basis, the CEO understandably isn't gonna be the one doing the things. Who's the one who should, who do you think should be holding the bulk of the responsibility, or how would you structure it? Maybe that's the better question of how would you structure a data governance given your

[00:33:21] **Josh:** druthers?

I mean, I guess I would again, make it ultimately the responsibility of the head of data, like obviously, or whatever the sort of like most senior data person in the organization is. Again, this is not surprising. I'm a data person. I'm gonna want to give the responsibility of the data person, right? To me, it's like generally I do not feel like, say the CISO or whatever has the sort of.

In my highly biased opinion, nuanced understanding of like how the data's generated and like what its sort of impact can be on the business and stuff like that, that the head of data does. That's why I put it there cuz I think of data as like data is both like very clearly an asset and also like a tremendous liability, right?

If it's just one or the other. If you purely see your data as a liability, give it to the ciso, like put the CISO in charge of it and just collect as little data as humanly possible problem solved. If it's an asset and it's not a liability at all, then you know, just kinda like give it to, I don't know, whoever the head of product.

The junior data analyst you just hired, it doesn't matter. And let them just go crazy. Let them build all kinds of crazy machine learning models and stuff like that, right? If it's both and you need a nuanced kind of understanding of these things, then I think that to me is like that's the job of the head of data is to have that, again, super biased, but that's, that's me, man.

Yeah.

[00:34:34] **Dave:** It makes sense. Mike, what's your take?

[00:34:37] **Mike:** I'm agreeing with Josh here, where you need somebody that understands the data and the business and the regulations and controls. I think part of the challenge is even for the head of data that often people think about. Governance is just not losing the data, but there's a large part of governance that's about making sure the data is used appropriately in, in line with regulations and in the best interest of the customer and the business.

There's things that aren't part of, I would say, the crown jewels, like employee data that you tend to have or confidential like emails or source code or all. All those kinds of things that are, security has a role. Legal has a role. I do see value in like ahead of data that's likely reporting to A C I O that could balance the asset versus liability risks of the data.

I would say that. Privacy and security are, are probably more aligned than data governance and security, even though they're all, they're all important. One of the other things that I often see as a gap, especially in medium sized companies, is because of all the data breach stuff and the impacts of loss, there's been a bit of an over rotation towards just pure confidentiality and protecting data and not as much focus on sort of the resiliency and the integrity of it as well.

So, is your data good? Is your data clean? Is it available when you need it, how you need it? You look at like the string of things like ransomware attacks, where there was a TV provider that's been offline for, you know, I think a month or two back. It was public, it was dish networks. It's not fully clear what happened, but like that's the kind of thing where people couldn't log into their accounts and they couldn't use their services.

Is that a security problem? Is that a data governance problem? I think we're still waiting to see how, how all that kind of comes out. Cool.

[00:36:37] **Josh:** Yeah.

[00:36:37] **Dave:** So this is something that I think you and I have talked about before, Josh, that interests me. I'm curious as to why this failed, and especially, you know, you see things like what's happening with generative ai and again, looking at it now, it all seems so obvious, but there's a bunch of stuff that seems really freaking obvious that never comes to pass, you know, which is part of the reason in tech.

I think you become skeptical over time and you kind of take a wait and see approach on many things. Like what are those would be? Data ops and the premise behind data ops. Was that as we work with more data, as data becomes this essential fuel for data science and the rest of it, we need to normalize the practices of how we work with data at scale.

Much like when we moved to the cloud, when we moved to agile development, we had to normalize the operational flows around it. In the famous book, there was the Phoenix Project, which illustrated why DevOps is so important and so on, and there was a movement a little while ago that I want to unpack for the briefest moments here, but I want to touch upon it where it said, look, just like DevOps.

Normalized all the behaviors, created the teams DevOps teams, SREs the whole shebang and makes things work in the crazy world of the cloud in a healthy fashion, a productive, healthy, efficient fashion. We're going to need data ops on the data side, but I don't live in the world that you do, Josh, but I keep an eye on it and it seems to have been like a flash in the pan and people don't talk about it, but it seems like that would be one of the teams that would be responsible for data governance or at least play a huge role in it.

What's happened with data ops?

[00:38:15] **Josh:** That's a great question, Dave. I was kind of a skeptic of the data ops stuff. I'm not trying to like dance on data ops grave or anything like that, like data ops is still very much a thing and there's still people trying to pursue it. The reason I feel like DevOps works and DevOps resonated with people is fundamentally that shifting kind of control and responsibility for like the deployment and monitoring and, and all this kind of stuff of software to engineering teams is good for engineering teams.

It helps them ship features, it helps 'em do what they want to do, which is push code, right? Same thing with like whatever SEC Ops or DevSecOps or SEC dev ops or I'm not, there's too many goddamn ops, right? All that kind of stuff, right? Same kind of thing. It's like if pushing this stuff left let's me ship faster, what I want to be doing anyway, which is features for my customers, then like, let's do it.

La Gungho, whatever I need to do. Tools. Great. Fantastic. The problem with data ops is that if I'm just a random feature team working on like a random feature, me taking responsibility for the data exhaust, my system generates. Doesn't help me ship features. Now to Mike's point, if you're shipping a recommendation engine, then absolutely you care a lot about your data ops.

You a hundred percent care about your data ops. That's a huge deal because that data powers your recommendation engine and that's your product. Same for like fraud or spam or any of these kinds of teams that are very, very data intensive. They've been practicing work this way for years because they had to, because the data powers their product.

But again, if I'm just a random team and I'm not like response, obviously responsible for some machine learning model, there's no incentive for me to do the data pipeline stuff. There's no incentive for me to really care about the quality of the data at all because it doesn't help me ship features.

That's it. That's the whole thing. If it helps me ship features, great. I'll do it. Does it not help me ship features? Then like I just don't care. I'm never going to care. This is the only thing I'm here to do. That's my take on

[00:40:12] **Dave:** it here. It makes sense, and especially given the climate we're in right now where everybody's making cuts or at least you know whether you need it or not, there's a general prevailing perception that everyone should be cutting back, trimming down, getting rid of excess.

And to your point, like if it isn't directly contributing to the core thing that needs to get done, it's on the chopping block. Particularly if it's hygiene oriented. And a lot of this stuff, I feel like we're talking about hygiene. At the end of the day, if it's not helpful, it generally doesn't get done.

This is the way of the world, man. How normalized in your world, Mike, and I don't wanna put you on the spot too much here, because I think that it'd be easy for me to ask this in a way that you couldn't answer, but how normalized our kind of data governance practice is. And I know there's been a number of articles on Amazon, but Amazon is massive.

I mean, it's like a hundred thousand people now, if not more. How many businesses? I gotta imagine that data governance and data hygiene is very specific to the area. So maybe if you can talk to your area a little

[00:41:16] **Mike:** bit. Sure. I think data governance is a big deal, right? I think whether you're talking Amazon or anywhere else, right?

It's important to your customers, it's important to your business. The amount of regulations out there has a ton of influence on what you're allowed to do with your data, how you're allowed to use your data, how things are supposed to get tracked. I think one of the hardest things to solve is when you actually have a real mature governance organization and you know all the things you need to be doing, making sure that those are actually applied consistently everywhere, all the time, in the right places, right?

Like just the scale of, well, is everything properly tokenized? Is everything minimized as much as it can be? Honor, all the controls are in place. How well are you tracking your things going to third parties? When you think about the importance of governance, I think it's key to have. You've gotta have that structure in place and you have to communicate it in an organization of any size, especially in the DevOps world or the DevSecOps world, where you have a team that's kind of siloed, but they only have the one problem or the set problems that they're, they're consistently working on.

You know, if you wanna let a team ship features fast, they need that governance structure to do it

[00:42:32] **Dave:** safely. Let's move on from governance to security. I'll never forget a conversation that you and I had Josh about security. Actually several of 'em, because I'll confess, I didn't really understand security and what it meant to the data team, and you told me that upfront.

I think we were, we might have been your first security advisory.

[00:42:51] **Josh:** You are, you're my, my first, and for some reason only, no other security company has found me useful. So I don't know if that's. Good thing for you, like I'm your secret weapon, or like, actually, I'm just basically useless. It's one or the other.

[00:43:04] **Dave:** Well, you've been insanely useful for us, so maybe we're letting the cat out of the bag with security voices

[00:43:09] **Josh:** here. Again, I'm not taking any more advisees, so I don't think there's any, I'm no risk here, so I think it's okay. All right.

[00:43:14] **Dave:** But it was interesting to hear what you worried about versus what I thought you would've been worried about.

What do you, as a person who works in data engineering and data science, what do you worry about from a security perspective? Let's talk about just baseline, you know, kind of going back to slack days and cloud errands, so forth. I think it will be interesting, let's open up the aperture after this and talk about what we worry about in the future with generative AI and so forth.

But let's, let's continue to keep it at kind of a foundational level

[00:43:45] **Josh:** for now. I think my great overriding nightmare of fear was always kind of twofold. Dave, as I think back on it, one was inadvertently collecting. Logging or whatever, a bunch of P I i data or security sensitive data, some places did not belong into kind of our own internal systems that we use for observability and for analytics and stuff like that, right?

That was like worry number one was that this data would show up in my systems at all, ever was a persistent source of stress for me. As you and I know, the easiest thing to do as a data person is just like, Say, send me whatever you want. Just send me just arbitrary J s o payloads, just, you know, fire it off.

I'll take care of it for you. Don't worry about it. Right. And that is just the easiest way to guarantee that a whole bunch of pii, very sensitive data is gonna inadvertently end up in your system a hundred percent. Right. So that was always like stress one. And then of course, like stress two was that somehow by hook or by crook, somebody external to the system, some external threat, be it government or otherwise, would gain access to our data systems, both of the business risk.

And then if you combine that with the first one, the p i I risk. Then it's like a way, way worse deal. Then it's like a, you're, you're on the front, but like a lot of data people, you know, security people. My nightmare is to be on the front page of the New York Times or the Washington Post because of a breach.

That's the worst possible thing that can happen. To me, addressing sort of those two threats is like foundational and it's like up there with like, you know, breathing oxygen and drinking water. In terms of things I worry about.

[00:45:16] **Dave:** Mike, what would you add change to that? I think one of 'em it sounds like is gonna be availability, which is interesting cuz I, I would've expected to hear Josh say that.

But you've worked tightly with the business and, and seen things from many perspectives in your years. What would you add on top of that? What are your primary concerns and maybe what are a few of the non-obvious things that you're concerned about either at a p v or elsewhere?

[00:45:39] **Mike:** There's a few things. One of 'em is right, you don't wanna lose data.

I always joke from the sort of the security thing that, like the security triad, they talk about confidentiality, integrity, and availability, right? Like, don't let your data get corrupted, don't lose your data and like kind of don't go down. I've always phrased it as like, don't lose data, don't go down, don't get sued.

That's kind of how I've always tried to sort of, you know, simplify the conversation. I think there's a lot of regulatory requirements that are constantly changing, which is forcing the double down on data governance to make sure companies stay in line with that. You look at like just in the US. You know, there's arguably 20 to 30 privacy laws in flight that organizations need to be, you know, cognizant of.

And some of these regulations are actually forcing your teams to come back and actually collect more data or tag different data differently, or go back to rebuild systems, to, to keep track of things in a matter that, you know, you weren't thinking about originally. I think Josh brought up child data or teen data.

Earlier in the kind of scenario as well, I just have users. I wasn't really tracking on them by age, but now there's regulations and expectations where I need to go figure out everybody's age so I can stay in compliance. So now I have a bunch of data with ages attached to it. These are problems as you have, as your company matures and you have services, uh, that are years and years old and you're not, you didn't have that split in your data upfront.

You've gotta go back and do that. I think the other big piece, and I've, I've, I think I've brought this up once or twice, is third party or supply chain security. The amount of things that can happen, it's in your control cuz you've made a decision to use them or partner with them, but it's outta your control whether, you know their governance is as effective and their security controls are as effective as they say they say they are.

Right? And this can lead to the corruption of data. This can lead to the loss of data and you can still, in theory, be liable. These are all things that really, really concern me.

[00:47:42] **Dave:** All right, so I'm gonna ask you an entirely unfair question. In your mental model for risk at a P V or even elsewhere, if you could divide it into percentages, how much do you worry about your own organization losing a grip on your own data and going down, losing it, getting sued, your triad there?

Versus your partners? Is it kind of a, a clean 50 50 split? Is it more a 60 40? Like how much do you worry about your own versus what other people are doing with your data?

[00:48:14] **Mike:** I think this is where you actually, and this is where a good governance and risk framework come in, cuz you've gotta prioritize it.

Like you can't have everything equal everywhere. I think there's certain vendors I'm more worried about. Than I am about internal controls. And then there's internal controls. I'm, I'm more worried about, like, I think of, I was running security for a, a FinTech and we're starting to build out this security program.

We're starting to figure out like, you know, what services are people using where customer data going or might be going or employee data and Right. This is a, you know, a 200 person startup, 300 person startup, you know, that had close to, you could argue 203rd parties that were using between like the calendars of the world and Eventbrite and Constant Contact for mailing lists and a bunch of things.

You're not necessarily top of mind and you may or may not have a good, good harness on 'em, C r M systems, like the list can go on and on. I think you need a way of mapping your challenges against a framework, uh, and going after the ones you think are gonna be the biggest risk based on what they're providing.

[00:49:26] **Dave:** And do you create your own framework for it, or is there like a, a standardized framework that you lean on?

[00:49:32] **Mike:** I think it's both, right? I think there's some foundations that always like always hold true, whether you're using Sands or NIST or anything else, right? Like go build an asset list of all your data and where it lives.

You can ask most organizations that and they will struggle. And whether you're asking about data or number of computers they have or number of locations they have, getting organizations to produce a pretty comprehensive asset list of everything everywhere tends to be a pretty tough ask.

[00:50:01] **Dave:** Yep. We heard that consistently enough.

We started a company for it. Let's shift gears a little bit and let's, I think we've resisted the temptation to go full on like generative ai, but let's indulge that for a moment. Given everything that's happening now, given the emergence of widespread LLMs and generative AI being talked about and used in so many places, things changing by the week, if not by the day, I'll start with Josh.

What do you worry about today with respect to data security, data privacy, that's different than before? Like how does it change the mental model for

[00:50:38] **Josh:** you? Oh, I worry a lot lately about how to keep certain information out of LLMs. Forever. That's, that's what I've been spending a lot of my time thinking about is like how there are certain digital assets that very clearly need to be accessed by.

People need to be used by people or important and stuff like that, but need to never, ever, ever, ever, ever make their way into the weights and the embeddings of a large language model. How the fuck are we gonna do this? How are we gonna do this? This is what keeps me up at night these days, is like, it's not just what these things can do.

It's like, how do we keep certain things out of their hands? Forever and ever and ever. That's what I worry about these days.

[00:51:24] **Dave:** Any particular data? Is there one that kind of preoccupies you more than others?

[00:51:29] **Josh:** Oh, like a sort of like a single source of data or like

[00:51:31] **Dave:** what do you mean? Are there data types more that preoccupy you more than others?

[00:51:35] **Josh:** I mean, I think, you know, national security information, like various kinds of classified information's, classified assets, I think is what is top of mind for me right now. For reasons without a doubt.

[00:51:45] **Dave:** Well, I mean, we just had the Ukraine links too, so I mean, it's right on top of That's

[00:51:50] **Josh:** exactly. And like, will those leaks end up in a large language model in the Nazis future?

Uh, yeah, they absolutely will. Like, without a doubt, without a question, and it's like, yeah, it's, it's precisely that kind of stuff, Dave. How do we keep that stuff out?

[00:52:01] **Dave:** So arguably data privacy,

[00:52:03] **Josh:** arguably, arguably, but just with like with a blast radius, that's a lot greater than like any individual person having their identity stolen or something like that.

That's what I worry about. Yeah.

[00:52:13] **Dave:** National Secrets, incredibly sensitive data that now can pop up just at someone's whim. Who knows how to write the right prompt

[00:52:21] **Josh:** in a discord or something. And again, once it's in the model. It never goes away. It's with us forever. It's not like we can go like surgically extract it.

Right? We can't like zero out a bunch of vectors and have it go away. It doesn't work that way. That's the problem. There's

[00:52:35] **Dave:** no data subject access request or r tbf, right, to be forgotten yet for LLMs.

[00:52:41] **Josh:** Precisely, precisely. It's a huge

[00:52:43] **Dave:** deal. Mike, how does it change your model? And, and this can be broadly too.

I'm not sure that this is your biggest concern at a p v at the moment. How does it change your model? The, all of the things that are happening with generative AI right now and, and just AI broadly.

[00:52:59] **Mike:** I think, uh, like some of the real interesting things, and I think this is much more general security, but there's, people have been using micro G P T and you can go search, you know, the innerwebs and you'll find it to hack other systems.

Now you're gonna have machines actually tacking machines. It's gonna be interesting to see how that plays out and evolves. What does that mean for like individuals and users? I think it gets even more interesting cuz I think about, like, I've spent a lot of time dealing with things like bot attacks for people trying to buy Xboxes.

So if you tried to get an Xbox or a PlayStation over C O V I D, it was near impossible and it was all these scalpers that had, you know, these essentially bots that are run and retooling at a. Semi fast rate to basically snap up inventory, whether it's concert tickets or PlayStations, et cetera, and that's been a cat and mouse game going on forever.

Like Ticketmaster getting called in, you know, publicly over Taylor Swift concert tickets. The potential for these machines to evolve faster than the defenses have an opportunity to make these systems worse. I think Elon's even, you know, supported this where he is like, Hey, one of the reasons I want everybody to pay for Twitter is that actually cost of paying is a control that will actually block some of these automated systems if they actually have to write a check.

Because all our tools, whether it's CAPA or trying the fingerprint or whatever the case may be, these LLMs are smart enough to evade them and have reasonable answers that you can't tell human from machine. So there's potential for scarcity in, in, you know, higher end items

[00:54:40] **Dave:** if you ask chat. G p t is, I did what the major risks are to LLMs.

One of the things it mentions is model poisoning and model theft and model inversion and so on. How big a risk are these? Like I understand them as the theoretical risks. The things that the both of you mentioned are the things that concern me way more. But I'm curious, you know, particularly maybe a question for you, Josh, is how real are these threats?

Are they, you know, kind of staunchly in the land of theory today, but coming to us soon? Do you know of any examples of these? How do you reason about it? I

[00:55:17] **Josh:** mean, people have done fairly lightweight versions of this to kind of demonstrate that it's possible in like non-threatening ways and stuff like that.

And that's kind of the current, that's the current sort of state of affairs here. I am not personally aware of anyone doing. Intentional sort of attacks or stuff like that. But I mean that's just because no one knew this was possible before. I'm sure that literally as we speak, people are working on attacks for exactly these kinds of cases and stuff like that.

Again, it's, it's one of those things where like, because we don't fully understand how the magic works, we don't understand how it's possible to attack these things and how you would do this very sort of task. Right. And that is sort of the challenge here is like, what we don't know we're gonna find out in the next 12 months, we're gonna find out.

But I don't know what way, shape, or form it's gonna look like. Honestly, I don't, and I think anyone who does is either lying or is actually like actively doing it right now. Like one or the other.

[00:56:10] **Dave:** Yeah. And going back to what we said before, there's a lot of theoretical attacks that never materialize into anything.

Massive. They don't. And even ransomware. I remember being at, being at Symantec security response back when Symantec was a thing. It was 2006, 2007, and we saw ransomware back then. Ransomware came out. But it wasn't a thing because we didn't have Bitcoin, we didn't have crypto in order to, you know, provide for easy money laundering.

And I think a lot of these concerns that people have a generative AI right now, they could take a very long time to manifest. And we look at 'em and say, oh, dear God, there's a prototype. There's it happening. Much like with ransomware, it could take 15 years. 10, 12, 15 years to where the supporting technologies and the adoption bring it to the point to where it's actually a predominant concern.

[00:57:06] **Mike:** I would say that I think the old's gonna become new again. So some of these things like model poisoning and, and such, right? Like. These kinds of things have been going on in a while for a while In non l l m sort of vectors, people are gonna figure out what it takes to, you know, repackage 'em and retool 'em to inject them.

Like I think, you know, one of the things that was kind of big for a bit was deception and security. So you'd have these honeypots that you know of vulnerabilities and they'd, they'd adapt to whatever methods the attacks were doing to basically profile the kill chain and respond. I could see some large sites looking for scrapers and deliberately returning bad data to break a model as a defense mechanism.

What does that get right? And then how does that play into what are known to be use cases? So like airlines and hotels are known for doing dynamic pricing based on who you are and. Where you're searching from. Will these kinds of, you know, injections be used to game or exploit the system in some way? I think to Josh's point, I don't know how they're gonna materialize, but I think there's, there's enough sort of proven science that I expect these to come back in, in some shape or form.

[00:58:21] **Josh:** It's like SEO techniques. It starts out with like the worst actors basically, and gradually filters out to everybody else. Without a doubt. I think Mike's right, history repeats itself without a doubt.

[00:58:30] **Dave:** So to wrap up, give me one, um, kind of indication of what you're reading, what you're doing to stay on top of it.

Like what are your data sources in, in order to understand this world and kind of further your education. Could be a podcast, could be what you're reading, could be a book. What are you guys reading to stay on top of it? For me,

[00:58:49] **Josh:** it's, it's a, a guy named Sean Wang sw. He goes by S W Y X, SWX, he's a, he's done development stuff, developer relations stuff across data, stuff across front end stuff, and now mostly AI stuff.

And he is, Absolutely fantastic in terms of like folks who explain things really well, so that. Lay people like myself can understand them. He's, he's just one of the best. So I follow him in what he does pretty avidly. His podcast and his his newsletter are my main kind of go-to source for like what's absolutely happening right now.

[00:59:18] **Dave:** Excellent. Love that. What's the name of his podcast? I don't

[00:59:21] **Josh:** actually know. sw.io I think is his website. S W Y x.io and everything on there is fantastic.

[00:59:28] **Dave:** Mike,

[00:59:28] **Mike:** how about you? To keep my understanding with a lot of the L l M stuff to date, most of it's been following forums, things like Reddit specifically and a few others, to see how people are experimenting with things like micro G P T and some of the others.

I'm particularly interested in how this is actually being applied much more than trying to pull apart the, the actual learning models and training models. Myself, I'm trying to understand the applicability better.

[00:59:58] **Dave:** Yeah. I mean, there's so many angles we didn't even begin to touch upon here. Everything from like corporate policy towards responsible use of ai.

Towards, you know, prompt injection. I mean, there's just, there's a myriad of angles here as they are with any dramatic sea change. I'll throw in there like, there's a few books just to, for people who want a bigger kind of more theoretical resource. I read Homos by you've all Noah Harari and he Reasons about a fair bit of this and what it means for society.

And if you want something fun, the William Gibson Novel Agency, the Second in The Jackpot Trilogy is really fun as he's just, I mean, he is the guy who coined the term cyberspace. He knows a few things. His ability to articulate the future in a way that's super engaging is fun. And my, my next up here is Stuart Russell's human compatible ai and the Problem of control, which, uh, my buddy Tim from Tesion has told me is outstanding.

So for those of you who like to, to dig into a book, those are a few things that I've personally found pretty interesting.

[01:01:00] **Josh:** Awesome. That's a good, I'm gonna buy that book right now. Thank

[01:01:03] **Dave:** you so much. All right. Book club. Here we go. Gentlemen, this is awesome. Thank you so much for the time. This is great.

Thanks a lot y'all. Thank you.