SDS PODCAST EPISODE 965: FROM PHD SIDE PROJECT TO $500M ARR: WILL FALCON'S PYTORCH LIGHTNING STORY

Show Notes: http://www.superdatascience.com/965

Jon Krohn:	00:00:00	What if you could go to one place to get absolutely everything you needed to train and deploy AI models and it was all cost-effective. Welcome to the SuperDataScience podcast. I'm your host, Jon Krohn. I've got a really exciting episode for you today with the CEO of Lightning ai. They have just announced a huge merger that now means they have tens of thousands of physical GPUs. They have over $500 million in a RR. They've built a open source ecosystem including PyTorch Lightning, which has been downloaded nearly 400 million times, and that's growing quickly, just this really exciting ecosystem and products. And today with Will we get to dig into the man behind all this innovation and exactly what they've done, I hope you'll enjoy this one. This episode of Super Data Science is made possible by Dell Intel, Fabi and Cisco. Will Falcon, welcome to the SuperDataScience Podcast. It's great to have you on. How are you doing today?
Will Falcon:	00:01:05	Amazing. Thank you so much for having me. Excited to chat with you and thanks for having me on here.
Jon Krohn:	00:01:10	You're certainly experiencing some amazing times right now. It's really cool to be a part of it in some way. We've actually known each other for a number of years. And full disclosure, I am not an unbiased interviewer of you. I have a fellowship at Lightning ai, which is a really cool role that I appreciate you created for me and I've been doing for a year now and absolutely love working out of the Lightning AI office in New York. So many talented people. It really gives me a lot of energy and hopefully I'm contributing a bit back as
Will Falcon:	00:01:44	Well. We love having you, so it's been fun seeing you do your thing and I'm sure you're seeing kind of what we're doing as well.
Jon Krohn:	00:01:52	Yeah, and I think in both respects we're just getting started.

Will Falcon:	00:01:55	Yeah, well, I'm finally glad we got to work together. It has been a few years. Yeah,
Jon Krohn:	00:01:59	Yeah, exactly. And so the big news that we're here to talk about on air, and so even though I've had this fellowship at lighting AI for a year, we've been kind of holding on for a really special moment to have you on the podcast. And now the time has come because lighting AI has merged with a firm called Voltage Park, and you guys have created something really big. It's mind blowing to me. You've described it as the full stack ai, neo cloud for enterprises in Frontier Labs. It serves 400,000 developers and companies, $500 million plus in a RR, and that's been gained in under two years, which is astounding. $500 million in a RR under under two years. And now this company after merger letting AI plus Voltage Park has 35,000 GPUs making it the third largest neo cloud in the world. I guess a nice place to start on this is congrats.
Will Falcon:	00:02:56	Well, it's been, obviously these are numbers from bringing both companies together and I think it shows just how amazing each one of us were doing on our own. And when we looked at it, we said, if you brought together, I don't know, a Mac and Mac os, you'd built the best laptop in the world, it was pretty obvious when that became clear.
Jon Krohn:	00:03:20	Yeah, software plus hardware.
Will Falcon:	00:03:22	Exactly.
Jon Krohn:	00:03:22	Yeah, it's a really cool pairing. And you have actually been, so Lightning AI was first a customer of Voltage Park and then yeah, it obviously matured into something much more. How did that kind of evolve?
Will Falcon:	00:03:38	So if you're not familiar with Lightning full stack software, all the tools you need to build, train inference models, all that I'm sure we'll get into at some point. So we have a lot

of enterprise customers and developers and Frontier Labs who are, they're doing the inference with us, for example, once like Cantina, which is Sean Parker's new startup. So we do a lot of their inference there. And a lot of these players were using very single purpose tools to do that. For example, other companies where that's all they do. And so these companies came to us at first maybe for training and other things, and that eventually they kind of move into these other things that we offer because we offer all of it. And as we started to scale, we found the need to go beyond just AWS, as you all know, AWS is a very premium product, I will say.

00:04:27 And so command a high price, but it's lacking a lot of the specific AI tools that are needed because AWS is built for CPU applications actually. And so it's trying to be retrofitted back to ai. And so then you have enterprises and startups that will try to make up the gap by buying these specialized products that only do that one thing and then they stitch 'em together. And then that creates a lot of problems because it's security and enterprise, it would be firewalls and security between all these products, startups and Frontier Labs. They don't see it this way, but it creates a lot of operational overhead. So anyway, so we started looking around for where can we get better compute for our customers. It was clear that they wanted the software, but the compute prices were really expensive. So we found this thing called Neo Clouds, which I don't know what it was until a year ago.

00:05:13 It's a brand new term. And neo clouds are basically a new type of cloud that are, they're GPU first, so they have some CPUs now, but they were really designed around how do you make the best performing hardware on GPUs work? Really, really amazing for things like training models, things like Infinity Van and Vast Storage and all these different things. Whereas the traditional clouds like AWS try to roll a lot of that stuff out on their own. And so

it's not as high performance as a neo cloud because a neo cloud works directly with Nvidia to do a lot of this. So we find the slew of new clouds, there's like hundreds of these things, by the way, I didn't know this. So we partnered with the top seven or eight, we start working with them and then we start putting customers on all of these enterprise customers.

00:05:58 We're talking about people like Cisco, et cetera, and they need a special type of flexibility that I was skeptical that I could find outside of AWS. Personally, I think startups, they're okay dealing with this more. So we put the first enterprise customers and then they struggle and we struggle to get them to adopt these neo clouds for the majority of them. And then there's just this new player that we have never heard of Voltage Park, and they kind of come out of the blue and it's a interesting story how they got started. So they bring all these CPUs online. And then I think what was amazing, and probably every customer's reactions here have been how responsive they are. And so now we're a one company now, so I'm talking formerly Voltage Park,

00:06:43 They call it the white glove experience, which I think it's a great product name, but hey, we're going to do what it takes, give support as much as possible to make you successful, which is what you need as an early stage startup, right? Ultimately. So they're doing this amazing job and through that iteration process, we're able to actually win a few customers enterprises and increase retention. And so I was like, Hey, these guys are serious. They know what they're doing. And at the time, we were trying to figure out how to bring in a lot more GPUs online because Lightning, we've never sold compute. All our customers buy compute from somewhere else, and then they connect our software to that compute. It's called BYOC, bringer on cloud.

Jon Krohn: 00:07:24 And that is part of the great offering that you guys have is that kind of flexibility

00:07:28 Where that's something that I love talking about with the lighting AI product in general, is that it allows you to bring your own cloud or to have access to a range of neo clouds as well as those traditional cloud providers like AWS that you mentioned earlier. You just have a dropdown box, you can see what the pricing is in any given moment for the kind of GPUs that you want, and then you can provision them at a click of a button and have access to that in minutes. And you can have been working already, say for hours as an individual or as a team, just on A CPU, basically free compute instance. And then within minutes when you need it, switch to however many GPUs you want on whichever cloud you want.

Will Falcon: 00:08:12 Exactly. And so customers love that flexibility. But yeah, the pricing on the hardware was just a little cost prohibitive. And I think a lot of the bottleneck that we saw last year for growth, I mean the revenue growth in both companies went super high. I mean, lightning alone, we went at least 30 40 x revenue increase without the merger, and we were really just always bottlenecked by compute. We lost many deals, like millions of deals that we couldn't land because we couldn't find the compute for it. And so that was kind of the predicament we were in. And on the other side, you've got this neo clouds where the only software that they offer is Kubernetes basically, and that's it. And so you're like, Hey, I need to do inference. I need to do this development. I have training, I have model hub things, I have experiment management. There's dozens of things you need in between that we all just offer natively enlightening. And so they looked at it and they're like, oh, you guys have a full Mac os? And I was like, yeah, and you have an amazing machine, so let's get married.

	00:09:12	And so that's how this kind of came about.
Jon Krohn:	00:09:14	Alright, so you've used the word inference pretty casually, and probably a lot of our listeners know what that means, but just to explain it a little bit, when you have an AI model, you first train the model, and then once you have it trained, you put it on some kind of production system like Voltage Park, GPUs or AWS GPUs or whatever for that AI model to be used at inference time in production to be able to do something for users. So this is when you open up chat GPT and you type something that's inference happening on some cloud somewhere that open AI is paying money to. And so there's lots of, now that we have this big revolution of AI in more and more places, there's more and more demand for this inference compute. And you might know the stats better than me, but it's something like 99% of all GPU usage is for inference, not for training.
Will Falcon:	00:10:09	Yeah, I mean that's right. I think I'm actually sad that the name inference stuck because inference is a math concept, which means to take a math model and infer have it infer things. So if you ever heard of in stats and things like that, they inference existed in that context. And so someone decided to call it inference, which means have the model make predictions ultimately. But there's a lot of stuff that goes into that uptime. Ultimately inference is for a developer into being a container just like a web server, except that it's receiving requests and how it handles batches of requests and streaming is a bit different. And so that became a whole thing, which I have many opinions on what that is. But yeah, one of the main things that people have to worry about there is the compute elements of it, and then people strap on all these things. So on this 99% inference thing, it's funny, I think people are missing. I think people see that and say, oh, the world is just going to be inference. Let's go back to

early two thousands. And let's say that there's a product that exists called, I don't know, S3, right?

Jon Krohn: 00:11:27 Okay,

Will Falcon: 00:11:28 What

Jon Krohn: 00:11:28 An arbitrary name.

Will Falcon: 00:11:30 Arbitrary name. So this product, people are like, wow, storage. Everyone needs storage, right? Correct. Now it's this going to be its own market. Is it going to be its own company? That seems weird, but in early two thousands, you probably think that. I think if you look at what happened over time was S3 was just a primitive RDS EC2, all those things are primitives that if you put them in a bucket and you label that bucket, what is that called? Cloud just it wasn't built yet. And that cloud mostly became a WSI think you're in this process that we're going through this building process where the thing that's in front of us right now is inference, but there are other primitives like vector dbs, like training infrastructure, like Kubernetes, like agents, and there's dozens of other primitives that would be created that have not yet been created. And not all of those are going to be their own companies. I argue they shouldn't be. All we're doing is that we're in the process of creating something called a cloud. And that cloud, I believe we're the first ones to actually have that today, which is a new type of AI cloud,

Jon Krohn: 00:12:39 This full stack ai, neo

Will Falcon: 00:12:41 Cloud. Yeah, just put all those things, all those little pieces into a bucket and label that now that's an AI cloud. And we have the first one of that.

Jon Krohn:	00:12:49	So something that I've said to you before, I've come into your office and said, how do you explain all of this functionality? There's so many different things that Lightning AI Studios can do. And so I guess that's something that we should talk a little bit about is just even this idea of, so we talked about the Lightning AI company, the Voltage Park company. So we're aware that lighting AI is the software, voltage park is the underlying hardware like Voltage Park. They are literally standing up data centers and GPU centers and there's people physically screwing and screws
Will Falcon:	00:13:22	And
Jon Krohn:	00:13:22	Running cables and making,
Will Falcon:	00:13:23	We're standing up our seventh data center today just for context cursor, for example. We build their training cluster.
Jon Krohn:	00:13:32	There you go. That's super cool.
	00:13:34	And so that gives us a sense, probably as much as we really need to know for what Voltage Park, how that works for this kind of audience, for AI practitioners, data scientists, that kind of listener. But the Lightning AI studio part of it, which is the product that Lightning AI has been building for years, and that you described that 30 40 x growth in the past year. Tell us about that lightning AI studio journey. You've described it as being kind of analogous to what AWS was, but designed for ai, so all of these bells and whistles together. But if I'm a user, if I'm listening right now, what's my experience as I type landing AI into Google and start using it for the first time?
Will Falcon:	00:14:18	Yeah, I mean the products evolved, right? I would say our first product, waspy to sliding, which we'll talk about in a

minute, open source for training models, any kind of model including LLMs. And then the problem became, okay, this is cool, and I can train at scale. So at the time, this number, I want to come back to this, you said 99% of all workloads are inference. I was training models of Facebook AI in 2019, and if you ask anyone, what's the number of GPUs that everyone uses? Everyone has said one 99% of workloads are one GPU, but why is it because people only want one GPU or because it's hard to do multi GPU. So even the Facebook cluster was massively underutilized. And then I rolled out Pieter Lightning and suddenly people started training on multiple GPUs at Facebook, which was a very good team.

00:15:09 And then it kind of rolled out and eventually they trained most of their models like that, and then it kind of spilled out of Facebook and went to other companies. And that's how Piper Sliding was born. But you look at stats now, and it's not that I don't think anyone would say 99% of workloads or was single GPU, right? It's just that people didn't have the tooling to do that. So I argue for inference, it's not that 99% of the workloads are going to be inference is training is very hard and people don't yet have the tooling to do that. We do today on Lightning. So if you're struggling with training, you should go to Lightning. But so the studio is designed to basically be a, people are starting to need this today. So people are using cloud code, right? Well, what's the problem with cloud code on your laptop could delete everything. So people are starting to try to find these cloud environments. That's what a studio is.

Jon Krohn: 00:15:59 You're also limited in, you can't be training models in cloud code on your local machine.

Will Falcon: 00:16:04 Exactly. So Lightning gives you the studio. So Lightning has many products. One of them is studios. The studio is a cloud development environment sandbox, if you must

call it that. That's persistent. That acts like your laptop. So you could go and run cloud code on there and leave it running overnight and it'll do something for you and you don't have any kind of problem that is going to delete your laptop or anything. And you can have 20 of these running at the same time all on different GPUs and CPUs and things.

Jon Krohn: 00:16:30 And it doesn't matter if your preferred environment is vs code or Jupyter Notebooks.

Will Falcon: 00:16:34 Well, that's the idea of how you connect to the environment.

Jon Krohn: 00:16:36 Oh, right.

Will Falcon: 00:16:36 Yeah. So I want to separate the environment itself from how you code in that environment.

Jon Krohn: 00:16:41 I see

Will Falcon: 00:16:42 We provide a cloud kind of VS code interface, and we have Vibe coding in there as well. So you can describe things, but if you prefer to use Cursor locally, go for it. And if you prefer to use cloud code, go for it. We don't care how you're connecting to that thing. The point is, you're getting a cloud environment that can be shared. If you're an enterprise, it can be audited. For example, you probably don't want your developers to have local files that are customer sensitive on your laptop, but on there you can have them, right?

Jon Krohn: 00:17:09 Right. Yeah, it's a very flexible environment. It's allowing you to be there in the web interface. And that's kind of the default screen. So go to lightning.ai, you can create a free account. We'll have something in the show notes for people to be able to skip the queue and get access to some number of free monthly credits. So check out that

link to get access to Lightning AI right now. And then from in there you are in this VS code experience that you described as a default. And from there you have access to everything that you imagine you could need to train and deploy AI models. I feel like we could spend literally hours, and I have seen you demo for literally hours on all of the functionality.

Will Falcon: 00:17:58 I dunno if people know this, but lighting AI itself today is built on studios. Every single developer at the company today codes in studios, whether you using Golan or Python, whether you're training models to an inference or coding a web app, everyone does it on studios today because it's easy to reproduce, it's easy to onboard new people, but they will probably code from the local id, but it's connected to this remote thing. And then I think the other thing to notice is the studios themselves, like I said, are persistent. You can SSH into them, and you don't have to use a web interface either. There's a whole command line version of it where you can just start a new studio and just open it up and it's like you open a new terminal, that type, that's remote terminal, and now you can do whatever you want. So there's that developer experience as well.

Jon Krohn: 00:18:44 For sure. Yeah. So whether you're more comfortable and it's kind of sometimes easier to get started maybe right off the bat, especially maybe if you haven't been coding in a little while, you're coming back to it, you're like, oh, I've heard that all this tooling makes training and deploying AI models easier than ever before. Maybe you just get started within the web app itself, but if you are already used to doing things all the time in Cursor or VS code or Jupyter Notebooks or whatever environment you prefer, you can very easily at the click of a button connect that or yeah, just like you said, a terminal, any of those kinds of experiences connect to this remote studios instance. And so you get all of the security, all of the flexibility

associated with studios and you can, something else that's really cool is when you're working as a team, just like working on a Google doc together, you can be collaborating in the same environment, pass something off to someone else working in another time zone. And if you work in an enterprise or whatever organization you're in, everyone there is happy with the security of that. Everything's happening in their controlled world environment.

Will Falcon: 00:19:47 And it can run on-prem on your cloud. So it's fully secure with all the guardrails enterprises need. No, but yeah, I mean look, I think it's really a collaboration tool, first and foremost, a development environment tool, I guess second. And then there's a gap now where you have normal developers, not normal, but non-AI developers who are maybe data scientists who dunno how to train a model, who dunno how to do in France. You can come on there and we have our vibe coding and you can choose the skill you want and you can say, Hey, I want to fine tune a model or even I want to do data science and it'll actually help you vibe data science and vibe, ml and Vibe, analyze a notebook or whatever. So it'll complete sales for you. It'll do all of that there. But you happen to be in the cloud now and you've got cloud resources and it can submit jobs for you. It can spin up a cluster of a thousand GPUs. So basically the vibe coating helps you do the mops, I guess is the right way to say it.

Jon Krohn: 00:20:43 Yeah, it's something certainly worth being excited about and yeah, that's why I am so honored to be associated with the company and all the great things you guys are doing. Let's talk a bit more about this Neo cloud. Before moving on to PyTorch Lightning, there's been press specifically, I have this McKinsey article that I'll link to in the show notes for people. I think you're already familiar with it. And in it claims that neo clouds so kind of bare metal as a service, the economics are fragile and they risk

repeating mistakes that happened when AWS was starting to become popular. There were lots of companies that were trying to be like AWS, but smaller over leveraged players were acquired sidelined or forced into niche roles. And so in a Forbes article about the merger with Voltage Park that we have also in the show notes for you, you told Forbes why you shopped around for Neo Clouds, despite the fragile economics and why in particular you chose Voltage Park. And it sounds to me from reading this McKinsey article, all of the things that they're concerned about with neo clouds through this merger of lightning AI in Voltage Park, you avoid all of those issues.

Will Falcon: 00:21:59 Exactly. So they don't have any debt, which is amazing, and that allows us to do a lot of things that other clouds can't do. The CEO who is running Voltage Park, who's now on our board, he's a finance guy, so ex finance guy. And so he's very good at making profitable businesses and being rigorous about that. So I'm excited to partner with them on how do we keep scaling that? And both our views are that if you want to have profitability in the future, you have to check them amount of leverage that you have in the business. And this means loans and so on. Not all the new clouds had the kind of start that we had. And so they didn't have that luxury, and so they had to take out loans. And a lot of those loans depend on your customer, not defaulting. It's a bit like a mortgage crisis really, where if your customers default, and by the way, most of the customers are startups in those companies.

00:22:57 In our company we have some, the cursors and so on, but these are high quality companies like growing super fast. The vast majority of our customers are enterprises with great credit. And so we don't really have that issue, but it takes a lot to sell a new cloud to an enterprise, you need a lot of security, blah, blah, blah, blah, blah. Things that we've been doing for six years now lightning. And so we

don't really have that problem. And so really the thing that we are focused on is how do we make enterprises successful and give them an alternative to the traditional clouds and also work with them. Because a lot of enterprises use all the clouds, including us, they use AWS, they work with us. It just depends on the different use cases. So we still partner with those clouds. I think we're more like a hybrid cloud than anything else.

00:23:42 And we just happen to have our own supply now of data centers. So just the fundamentals are vastly different. And a lot of these companies went public. A lot of the credits that they need to get loans to build these data centers are based on their market caps because when you're public company, what's your credit? It's actually your market cap. And so what's happened in the last six months is all the new cloud prices have dropped by like 60%, maybe more. And so then now the creditors are like, wait, how's your market cap doing? And so if this continues to happen and it's a little too low, then they're going to get calls on these loans. And then we're to actually, it's literally equivalent, not equivalent, but it is very similar to a mortgage crisis situation where the banks were over leveraged and the Federal Reserve had to bail them out at some point.

00:24:33 And so that probably happens here. We can speculate who that would be, but there are a lot of people who'd probably be interested in doing that. But yeah, it's a new industry and it needs a bit of a TARP program or support from big players to help the ecosystem be successful ultimately. So yeah, that's where I think we're at, but I think we're very differentiated in that we have real revenue that's growing extremely fast, extremely high margins, and we don't really have a lot of debt. We'll probably take on debt in the next few years, but very controlled. And we have enterprise customers.

Jon Krohn: 00:25:10 Speaking of prices and margins, something that I'd like to dig get into a bit here is, you mentioned earlier in the episode how AWS can be very expensive for training or inference for ai, and how do you manage to maintain these good margins that you just cited more recently while also offering these great deals? So there's a recent Lightning AI press release that highlights a claim that you're cutting AI costs by 70% while offering a free tier and global access to everyone from undergrads to Fortune 100 companies. How do you manage to square all of that? At the same time?

Will Falcon: 00:25:48 A lot of smart financial engineering, I guess. And we just have a lot of enterprise customers, and so they're the ones who are kind of spending the most money with us because they have enterprise contracts and so on. And I could cut developers off a hundred percent, it would probably be more profitable for us, but I was a grad student at some point, and I don't think I could have done what I did without having access to this kind of compute. So I've given personally a lot to the community. I've given a lot of open source pipes are sliding. I've spent a lot of years working on open source sinks. So I'm an open source person by nature and try to give back as much as I can. I still have to build a business and make it profitable and take it IPO, right? So within those constraints, I try to do what I can. And I think the free tier is something that I'm super proud of our team for doing. And yeah, I want to support the next generation of builders of undergrads, grad students and help 'em figure out how to build the next bites or sliding or the next open ai. A hundred percent.

Jon Krohn: 00:26:52 Yeah. So speaking of giving back and your grad experience and PyTorch Lightning, let's talk about that now. That was is, you mentioned already earlier in this episode how PyTorch lighting is kind of what led to the lighting AI studios product and now this merger with

Voltage Park, let's talk about PyTorch Lighting. It's been downloaded over 300 million times, I believe is the last

Will Falcon: 00:27:12 Close to 400 now.

Jon Krohn: 00:27:13 Yeah, maybe by the time this episode is out. And yeah, you developed that while you were a PhD student working at NYU and you had some pretty well-known PhD supervisors. So Yanika and Kung y Cho, who also, I mean, he's been cited hundreds of thousands of times.

Will Falcon: 00:27:32 He's the next yon for sure,

Jon Krohn: 00:27:33 And he's got papers like neural machine translation that are huge and that everyone in AI needs to know. So yeah, tell us a bit about that experience. Maybe you can even back up a bit and give us a bit of what led you to that program, founding PyTorch Lightning and the big impact that I know that it had at Facebook and beyond.

Will Falcon: 00:27:52 So I actually started working on PyTorch Lightning probably around 2015 before I was a grad student. I was a grad student in 2018, a few years later. So I was an undergrad at Columbia and I was at a neuroscience lab, and this is when auto encoders and gans were a big thing, and we were learning to train models that could decode neuro activity from the brain. So it was a partnership with Stanford. So at Stanford, they had monkeys, they would collect neural activity from the retina. So if the monkey died, they would cut out the retina and then they would show images to it, ImageNet specifically. And so you had a way to measure what was happening in the retina, and then you had what was shown. So you had an X and a Y, and at Columbia, we were doing the computational work. And so we would get that data and then we would try to figure out what is a model that can take those raw neural activities, which are

pretty much zeros and once and some simulations, and basically decode the image that the I saw.

00:28:53 And so that's called Gen AI today. That was not called Gen AI back then, right? Because imagining an image back then we were using something called auto encoders, and then gans, which people I think are still using, I'm still very bullish in auto encoders. And so in 2015, you didn't have TensorFlow, you didn't have PyTorch, you only had fi. And so the first version of PyTorch sliding wasn't even based on PyTorch. It was actually on the, and it was around how do we try many ideas at once without rewriting the code all the time, which is why I was forced to write that.

Jon Krohn: 00:29:28 I didn't know that. What did you call it? It wasn't called PyTorch writing,

Will Falcon: 00:29:30 It was just called research lib. Just my research lib

Jon Krohn: 00:29:34 Got to name this thing.

Will Falcon: 00:29:36 No, I didn't really care. I was just the only one using it. And a few people in my lab, and I remember this postdoc at the time, Scott Linderman, I think is his name, so he's now a professor at Stanford. He's a statistician, best statistician I work with ever. And he was like, Hey, will, you should formalize this into some sort of, he called it a harness so we can use it in other projects.

00:30:02 And I was like, okay, yeah, sure. And so I kind of tracked it out and then we started working on other projects. And so I was like, oh, very quickly I can just rerun the same stuff. All I need to change is the computational part of it, which is the forward and backward, which is where the science was for us at the time. And then TensorFlow came out, I rewrote it in TensorFlow before, actually, this is before I spoke to Scott. So I rewrote in TensorFlow

because the paper we're working on, we needed to publish in new's deadline was like March, this is November. And every time we went through ImageNet, it was 21 days for one epoch, and the auto code needed hundreds of epochs just to see if that one worked. So you can imagine that we're not going to complete this. I think it was laptops. So I went out to my PI and I was like, Hey, we need to buy GPUs. And so then I bought 4 24 GPUs, something like that, no, yeah, 16 GPUs.

Jon Krohn: 00:31:04 That was probably around the era of 10 80 tis.

Will Falcon: 00:31:06 So I bought 10 80 Ts. I bought a bunch of 10 80. I actually have the box, I'm probably bring it to the office, but it's cool. I still use it for gaming and stuff. So I built these four boxes from scratch. So when I went to Voltage Park to the data centers, I was like, ah, so this is how you do it at scale. I mean, they're experts at this, but I think building those machines definitely gives you a good background. And I actually posted on my GitHub the instructions. So if you guys go to my GitHub, you'll see exactly everything, all the parts and how you do it or whatever. This is eight years ago at this point. So I built these machines with four GPUs on them so we could try one idea per GPU. So that gave us some boost, but it wasn't as fast as I needed. So then roughly at that time, tension flow came out and one of the things they were talking about was multi GPUs. And I was like, oh, okay, cool. Let me try that. So then I rewrote all our code into TensorFlow, and then I actually did get it to work on four GPUs. And then we were able to train that model in maybe an hour from 21 days,

Jon Krohn: 00:32:03 21 days to an hour.

Will Falcon: 00:32:04 And so then I was amazing. And then we actually did it eventually published a paper to new's, my first paper ever, and also my last paper to new's. So I got super

lucky. I was not the first author on there, but the team was incredible. And that was as an undergrad. And so then fast forward, I started my next project with Scott, and then he's like, Hey, rewrite this. And then at the time PyTorch came out and then we looked at it and it just looked more mathematical and we were doing more probabilistic models at the time, so I needed to actually look at a function and then map it to the code. And so it was easier to just write the literal math on there and then model the probability solutions as a model. So like P of X given some parameters, some data, that data was a model.

00:32:49 And so you could take all be stats, probability and just put models around them. So that's what we did. And then, yeah, we started training this model, that paper. I don't think we publish anything. I think I screwed something up. I dunno, maybe, I dunno if the data had data, it was weird data. So in neuroscience, you have to know if the data's good. I dunno if that was great data. I think he did eventually publish something about it, but not with me. And so then that was it. We were using, this ate many years later, I started a company sold it. And in that company we used PyTorch separately. We didn't use PyTorch Lightning because we weren't training models, we were just doing what people call inference now. So I was running large scale inference across 60,000 customers. This is low income students we're helping them figure out how to pay for college.

00:33:32 So it was all in text message. And it was some of the first production systems I've ever, I think I've seen put out there. And we gave a talk at data-driven NYC. So we gave a talk there, and you guys can watch my talk and I called it bot powered Humans. Basically, I was trying to figure out how to use AI to augment the humans. And so these were people that we hired. We built a little UI for them where they would have this conversation with the

students and then that UI would suggest what to say next. But the way it would do it is it would treat you as a game. We use reinforcement learning now RL is super hot, but back then, so I treat the conversation as a game where your message is a step in that game. And then my message is a step in that game. And the goal of the game is to get you as much money as possible for college. And so the reward function is optimizing for your financial aid. And then the games are the words that I say to get you there. And so that's what the model was trained on, and it would take your whole

Jon Krohn: 00:34:33 Context, like convincing people to give you money,

Will Falcon: 00:34:34 No convincing, asking the right questions to help us get them the most money for college.

Jon Krohn: 00:34:39 I see.

Will Falcon: 00:34:40 So the model is like, hey, whatever my function says, I don't really know your SAT yet. And if I get it, I can figure out my policy better and see if I can get better. So it would ask you these questions. And we didn't have an alarms back then, so actually it was more of a ranking. So we had a list of questions and it would choose what's the best one to ask with some probability, and then ask that question and then answer and then play this game back and forth. We ended up getting a bunch of people, about $60 million worth of financial aid with this.

Jon Krohn: 00:35:11 Oh my God.

Will Falcon: 00:35:11 Yeah. Well, we sent inner city kids to Princeton.

Jon Krohn: 00:35:16 That company is called NextGen Vest, is that right? Correct. Yeah. And you sold that to CommonBond?

Will Falcon: 00:35:21 Yeah, so they acquired us. So we're doing this, this is like 20 17, 20 18 or something. So think about it now. We were doing reinforcer learning with models that I guess are called language models now. And it was all inference and production. This is people's problems today, and we're doing this and we have a whole company based on this, so I need real time inference. It's got to be fast and this and that. So then these guys come along and they're like, this is insane. So they basically offered to acquire the company, and I wasn't the CEO, my co-founder was so she's like, let's do it. And I was a CTO and I was like, Hey, cool, whatever you guys want. I'm do whatever you guys want. And then I had an offer to join NYU to basically do a PhD. So in this whole process, my undergrad PI was like, you should consider a PhD at some point. And I went to public school in Florida. I'm from South America. I was like never even considered going to college in the first place. So my PI is like, you should do a PhD. I was like, you're insane. And I was like, fine, I'll apply. So I applied and I get in at all these places randomly, and then one of the just apply to work with

Jon Krohn: 00:36:25 Beyond the camp

Will Falcon: 00:36:26 Seems like a good choice. These people were big, but they weren't. I mean, they're very big now, but then they were big to us, so they still were taking students. And I ran into Jan randomly at an event and he saw a paper, the neural decoding thing. And I think that crew of Yasha, banjo, Hinton and Jan think a lot about what does neuroscience say about things? And so they got interested and they invited me to the lab meetings. And that's actually when I met Khan. He was sitting there and I was like, Hey, what do you do here? I dunno who he was. He's like, oh, research and stuff. He's very humble. And then the whole meeting, I was like, dude, you're the guy in those papers. And then every paper right after that was he was on it. And so I was like, Hey, yeah, I'll apply.

And I asked him, I'm like, would you take me? He's like, I don't even know you. And I was like, all right, fine. And then over the year, I think I kept going to lab meetings and got to know them. And then I did apply and then actually got into his lab and then Joshua's lab.

Jon Krohn: 00:37:21 So wait, let's talk about this for a second. So you're going to lab meetings, but you're not a lot of students, you're just showing

Will Falcon: 00:37:26 Up.

Jon Krohn: 00:37:28 So basically you exited NextGen Vest, and so you're kind

Will Falcon: 00:37:31 Of like, no, I'm still at NextGen Vest at this time, and I'm still going to Columbia. I'm just taking smaller, I'm doing fewer classes, doing the company thing. And so that was kind of it. And so I'm doing all of these kind of three things at once. And then the acquisition offer comes in and that kind matches PhD applications. I get in with Yasha, Bengio and Mila, and then with Jan and then Columbia, and then a bunch of other places. And so I go to Montreal and I'm

Jon Krohn: 00:38:00 Congrats. I did not know that. That's wild.

Will Falcon: 00:38:01 Wow. And actually, the guy who I was going to do my PhD with, Mila became the VP of chat, GPT. He's one of the chat GPT creators, right?

Jon Krohn: 00:38:11 Ilias?

Will Falcon: 00:38:12 No, no, Liam Feta.

Jon Krohn: 00:38:13 Oh, sorry. He was from University

Will Falcon: 00:38:14 China. Yeah, Liam just started Periodic Labs. You should check out his company. So that lab was insane. I mean, SKR I think did spend some time there. I dunno how

they're related, Hinton, whatever, the Canadian mafia, these guys are amazing.

Jon Krohn: 00:38:26 There was a complete brain fraud on my side. So Giver was in Hinton's lab for his PhD, and they had that big 2012 Alex net paper with Alex.

Will Falcon: 00:38:34 But make no mistake, I mean, at the time Mila was, I think the top lab I would say, and NYU is pretty much up there, but it was a newer lab, at least to me it was a much smaller lab. And so I was like, I would love to go to Mila, but it's freezing. And I'm from South America and I'm already in New York. And n NYU is amazing. NYU felt like a newer lab to me, but they weren't. They've been around for a while. It's just a lot smaller. Mila's, a factory of researchers at the time, they already had hundreds of people. And so I was like, Hey, I'm just going to make this easy and go to NYU. And so I started there and that coincides with acquisition. And so then they offered me a bunch of stuff to join this acquiring company, and I'm like, no, I'm good.

00:39:15 I'm going to start this PhD with these guys like world class. And so I leave and I started the PhD at NYU, and then I dust stuff PyTorch Lightning, and I started doing my research with it. And then I think Jans being a big proponent of open source. And so he's kind of the reason why Facebook open source, so many things, applied Torch included. People don't know this, but a lot of projects were incubated at NYU things think SK Learn had a huge, huge help from NYU. People were working on it there. Adam Passkey who wrote the original PyTorch was found by a professor at NYU to work on that project. I think mpi, there's a lot of projects. There's osa. I think it's like a sound thing. So a lot of these Project N, NYU U has a weird history of doing this. I didn't join for this. I didn't know I was going to create an open source project, but I think they encouraged that a lot. And so a lot of these

things have been teed from there. So I commend, they basically recommend that I open sources so that the other people in the lab can use it somehow moving through my research faster than everyone.

00:40:20 So that's cool. My advisors are mostly focused on the research. So Khan's like, Hey, whatever, as long as you do your research. They're like, we don't really care. I'm like, okay, fine. So then this gets open source. And then I joined Facebook many months later, and when I joined, I have to name this thing so that other people use it. And so I was like, well, makes me move super fast through research. So I guess lightning. And then I just looked at, I don't know, a lightning bolt and NYU color, and I was like, there it is. And it was like 30 seconds, right?

Jon Krohn: 00:40:52 Wow.

Will Falcon: 00:40:52 And then that was it. And then I put it on the Read Me, and then I started my internship at fair. And then

Jon Krohn: 00:40:59 Facebook AI research. And there's also, there's a big, you talk about all these collaborations or this atmosphere, this environment of supporting open source that NYU has. They also famously have this very strong relationship with Facebook now meta. So Facebook AI research, they brought in, it was around 2013 or 14, the yaun became Chief AI scientist at Facebook. Well,

Will Falcon: 00:41:24 He founded Fair with I think a few other people, but he's one of the fair founders.

Jon Krohn: 00:41:31 And so there was a big revolving door between PhD students at NYU and people working in this prestigious AI lab fair at Facebook.

Will Falcon: 00:41:41 Yeah, it's incredible. I mean, we're in New York right now. If you go down to NYU, the Facebook Fair Office used to be on seven 70 Broadway, which is Cooper.

Jon Krohn: 00:41:52 Cooper Square.

Will Falcon: 00:41:53 Yeah. One block away from N NYU U. And so you have lab meetings and you're like, is it at n NYU or Facebook? And you just, it's the same distance. You just walk to one or the other. So you all have Facebook badges. It's literally there's kind of no friction because the Facebook clusters and data, or they're not Facebook internal. And so there's no customer that or anything there. It's all academic stuff. And so you have collaborations going on all the time, but the professors are dual. They're working at NYU and they're also working at Facebook now. That's fair. During this heyday, 20 18, 20 19, I think most fair people will tell you that was the peak time. Maybe I missed a year before that, but fair at the time was only a hundred people worldwide, maybe 150 if that, including admins and things. You had people like Dawe who was the chief scientist, a hugging face after he left Fair, now he's got his own company. You had people like Jason Weston who created a lot of the machine translation things. And the thing was so interesting, I showed up to my desk, this is, I remember 14th floor I want to say, and I sit down by the window. There's me and another desk here. I don't remember who was here, maybe another intern, but to my right, a Sumit who's leading PyTorch at the time,

Jon Krohn: 00:43:07 Tala.

Will Falcon: 00:43:08 And then to his right is the whole PyTorch team, which was seven people at the time. And then behind me is the leadership team. And then behind them is the team who wrote the first distributed training models in the world ever, who then they eventually became character ai. They left Facebook and started character ai, and then now

they're at thinking machines. And then you had PyTorch, these guys and then me, I guess, and I was probably the first one to get funding and then start a company. And so then I think that let people know that, hey, you can probably get funding to do a lot of these things as well. There might've been someone before me, but

Jon Krohn: 00:43:47 That was Grid AI at the time was the name. And so how did that happen? How did you, you're in this great ecosystem, both with this commercial relationship with Facebook, but also all the open source stuff that's encouraged by Facebook, by NYU. What were the conversations that led you to found a company? Was it because you'd already done it with NextGen Invest?

Will Falcon: 00:44:10 So I did not want to start a company, first of all. I think if you've done startups, they're hard. And I just come from a grind. And so I was very happy to become an academic for a bit read. I mean, it was amazing. I was reading books and doing research. I love science. I just love thinking about things going into really hard, impossible problems and going as deep as I can. So if everyone's thinking, ever should I do a PhD or not, that's what I would ask you. Do you love science so much that that's what you want to do? If all you're doing is for a career step, you're going to hate your life two years into it. But I loved it. It was amazing. And so I did not want to leave. I mean, the problem is Pieto lighting took off, and I wasn't trying to make that a thing. It just took off and people started using it. And so then VCs started pinging me. They started emailing me being like, Hey, we see your open source project. It's working this and that. And I was like, Hey, I'm good. I just sold a company. I just want to focus on research. They're

Jon Krohn: 00:45:08 Like, I'm more interested.

Will Falcon:	00:45:09	Yeah, exactly. And now, at the time I think about it, all these things have names today, but what I was doing back then was pre-training world models today that has a name. Back then, we didn't call it that, but I was training me pre-training models on four 8,000 GPUs, single person doing all the things that the pre-training folks are doing, it, open AI and whatever. How should you scale the thing? What gradients should you use? What's a distributed strategy? How do you prevent faults? And what if the loss function is this? So most of our research at fair, at least for me, was more theoretical. So it's more like how do you change the math function? So every model has a loss function that optimizes something. And so it's more like, how do you change the math function? Do you add a regularized here? Do you use this one or this one?
	00:45:52	And so I think what I learned was cool, and I have lost that skill a little bit. I think I need to get it back, is when you leave undergrad and you've done math, right, you're really good at reading math. It'd be like learning a language and you're really good at hearing it, but speaking, it is something you have to practice and speaking. It is more like, let's say you wanted me to build a model to model this world here that we're in right now. I'd have to figure out how to model the wall and what does gravity look like and how does the lighting hit and this and that. And that's a math equation. So being able to translate that to math, that was a skill that we developed a lot, and that was what FAIR was really good at.
Jon Krohn:	00:46:30	Oh, really?
Will Falcon:	00:46:30	Yeah.
Jon Krohn:	00:46:31	That is a cool skill.

Will Falcon: 00:46:32 Exactly. And so then that's what we train models with it. Oh, here's a model, train it. No, it's like what is the training algorithm itself look like? And if you want to have better representations of the world, how do you accomplish that? Do you make things come together or separate? Do you embed images or not? And if so, how do you do it? Do you use this type of embedding or this type of function? And what's a similarity between them and what if you add a penalize and regular and this? So most of pre-training today is actually not that. It's more like the engineering of pre-training. We were not just doing that, but we were also doing the math. And that's fair. That's why Fair was what these are math people who are also really good engineers and they can solve these problems together. And so that was called pre-training.

00:47:13 And so yeah, I mean, I think the VCs were kind of onto something, I guess, very early. And they asked me to come in for meetings, and then I flew out on a Thursday, and then by Monday I had 10 term sheets, and I kind of got out of hand. And so I was like, well, I guess I'm starting a, and so I called my advisors and I was like, Hey guys, so I got all this money that they're offering me, should I do this? And they're like, well, your research is going super slow. So probably no. I mean, I've been telling them, I was like, Hey, listen, I kept getting dragged into coding every night. And so I was like, research, run, submit. My training runs. Okay, it's running. And then I'm like, okay, go merch. This pull request from Piper Sliding. And then it was stuff that I wasn't doing.

00:47:53 People were like, oh, can I have this function for RNN, the way you do a loss function for an RNN back prop Through time, I was like, I'm not using back prop through time, but now I have to implement it for you. So I started working for people. I was like, this is terrible. I almost shut down Piper sliding in September. I was like, this is distracting. And so my advisors were already kind of like,

Hey, your research is a bit slow. And I was transparent about it. I was like, Hey guys, I just keep getting dragged into this.

Jon Krohn: 00:48:19 There's a funny irony here, which I don't know if you've noticed this before, but you telling me this kind of history that you ended up doing the PyTorch Lightning Open Source project because you were doing your research much more quickly than everyone else. And they were like, you've got to open source this thing so everyone else can take advantage of that. But then once you did that, it ironically slowed down your research because all of a sudden it became so popular. You have tons of people asking you to be reviewing prs and fixing bugs and adding functionality. So yeah, there's an interesting irony there.

Will Falcon: 00:48:54 Well, ironically also, I didn't want to code. That's why I wrote in the first place, so I could focus on the math, not the way the distribution happens or the training algorithm. And then people just ask for features that ended up being their engineer. And I was like, well, that's kind of what I was trying to avoid in the first place. So it's fine. I think ultimately it's great because I ended up putting so many things into this that were very specialized knowledge. That was a Facebook AI at the time, because SMU was helping the character I folks were helping, they weren't a character back then. So we had all these amazing people who are just world-class engineers helping me do all of this. And so I learned a lot. I put it in there, and as a benefit, the whole world was able to benefit. And now, yeah, 400 million downloads later. The reason you can train multi GPU U and multi-node and fine tune LMS and pre-trained LLMs is because we spent a lot of time at fair putting that into Pieto Lightning, and then all the contributors that came after I tried to add it up the other day. I think we've had about 500,000 hours of engineering time into Pieto Lightning. If your

pre-training code can be better than that, good luck. Right?

Jon Krohn: 00:49:54 It is a really great tool. And that's how I became aware of you in the first place. You were doing talks around the New York meetups, around PyTorch lighting. So I became aware of you. I started using it. It became really obvious to me that it was invaluable for speeding up, especially things like multi GPU training. It was something that was an impossible nightmare in almost any circumstance to get going before PyTorch lighting. And then that made it easy. And so I actually, I think it's four years ago now, I did a half day or a full day training at the Open Data Science conference east in Boston, and I got that edited and put it on YouTube. So I'll have a link to that in the show notes, which is kind of like an introduction to PyTorch Lighting. I had some other libraries in there as well, but that was one of the key libraries for helping people get off the ground with training their own models and deploying as well. Inference is a big part of

Will Falcon: 00:50:45 As

00:50:45 Well. I mean, I think there's a common misconception that Piper sliding is not for gen AI that's factually wrong. Literally, I was training world models before they were called world models. People have trained LLMs. All of NEMO from Nvidia is built using Piper Lightning. AWS has pre-trained LLMs using Piper Lightning. LinkedIn recently published an article six months ago how they train a hundred billion parameter LLM using Piper Lightning. I think that at some point if you are wanting to mess around with the distributed mechanism, which is DDP or FSP, and you have your own way of sharding, gradients and whatever, then the piper sliding can get frustrating because remember it was built at a time when that's not the thing we did. We focused on the math. So then we added the flexibility to have people plug in their

own strategies into this. And then we also created something called lighting fabric, which kind of gives you more less managed pipe PyTorch because PyTorch sliding is basically managed PyTorch, less managed PyTorch, but still you get all the beauty of standardizing the tools, having the flexibility, having different training strategies, but it's a bit more hackable.

00:51:54 So as you get in deeper into a Frontier Labs situation, then probably the journey is like you outgrow PyTorch Lightning because you need more control. You then go to a lightning fabric and then if you really, really need more control than that, then you go to raw PyTorch. But then in raw PyTorch, you need to know exactly what you're doing. You can mess things up. So as a result, I think only a few of the Frontier Labs, we'll use PyTorch directly. The vast majority of enterprises and startups use PyTorch Lightning. And then sometimes you've got a PhD who's like, oh, I can do it better. And then they try PyTorch and that's fine, but then their manager really should find out and they're like, Hey, but standardize with the rest of the team. But no, it's used for everything today. Yeah,

Jon Krohn: 00:52:35 And you kind of gave us a sense there of the other open source projects. So obviously we started this episode by talking a lot about Lightning AI studios, which is a product, a web-based product that you can use as an individual or as an enterprise at basically any size. But in addition to PyTorch Lightning, there are, you touched on some other open source projects out of lightning like fabric. I think there's others, there's a compiler.

Will Falcon: 00:53:02 Yeah, I mean we've open source pieces of the stack that you need for inference. We found Piper sliding pipe sliding is for model training and fine tuning. You can use it to do forward passes. There's a mode where you just freezes the model and it'll do that. But for an inference like production inference, the forward pass is only one

ingredient. It actually needs to be more of a system. Maybe you connect the RAG DB and this and that, and so people will use something like fast AI to do that, but then you have to implement your own batching and streaming and this and that. And so then we created something called Lit Serve, which is Piper sliding for inference basically where you can grab multiple models if you want, you can put them all together, you have full control, but you get automatically all the features like batching, streaming, security, all the things you'd need to put an actual inference server together.

00:53:50 But you have full control over what the models are. I think to my knowledge, that's kind of the only tool that exists out there today. There's other stuff, like I said, fast AI and things like that where you can write it yourself. And then things like VLLM and SG Lang are more like if you were to take lits serve and write a very opinionated, very optimized LLM serving engine specifically with batching streaming like KV caching, and you started with LITS serve to do that, and you put all those pieces perfectly together, you would end up with something like VLLM. So it's more like a high level tool than anything else. So if all you want to do is just hit run on ACL I, then lid serve is not that. But if you've got multiple models and you need to build your own inference engine, then that's what lids serve lets you do. And so tools evolve over the years and we found that there were certain gaps that we needed to fill in the market. All of lighting today, all our inference runs on lits serve, and we have so many customers run on lits serve today that are consumer products with millions of users and it's perfectly scalable.

Jon Krohn: 00:54:49 What do you think are the big gaps that you're going to need to fill in next? Obviously you have this big ecosystem of open source tools and commercial products that allow us as AI engineers, data scientists, people who

just want to make AI work in the real world tons easier. What else is missing? What else is next?

Will Falcon: 00:55:07 So we had two gaps. One gap was having enough GPUs for everyone, which I think we solve now. And then the second gap is a talent gap, which is a data scientist who wants to do inference or fine tune or something. And they're used to more like SK learn MPI type stuff. So that's what we have our vibe coding agents on our studios for now. I'll bring those problems there, describe what you want and that'll just do it for you. And so we want to close the talent gap with the vibe coding part.

Jon Krohn: 00:55:34 Cool. I like that. I think I'm going to start to wrap up the technical questions here, but there is one part of your life, your earlier life that I still want to dig into because as much as all of these things that you've described in this episode across working for people like Kung y Cho, and Yaka, and getting the opportunity to work with people like Joshua Bengio working with, I

Will Falcon: 00:56:01 Never worked with him, but

Jon Krohn: 00:56:01 No getting opportunities, having top VCs fund your startup, now you're at this $500 million plus R run rate and under two years serving 400,000 people. And so we've talked about this kind of later part of your journey, but the beginning part is also really interesting before Columbia, which is the earliest that we've gone to so far. So if I understand this correctly, and this was something I didn't even know until we started doing the research for this episode, is you came from Caracas, Venezuela at 13 years old and it came to the US and didn't even speak English. And I mean, tell us about that experience and I don't know, maybe somehow the way learning or being in that unknown environment or are there any kind of links to that immigration story to what's allowed you to be this incredibly productive human later in life?

Will Falcon: 00:57:05 Yeah, I mean, I did not learn any English in Venezuela. We had English classes, but in Latin America, this was status thing. If you know English, it's like, oh, you're so fancy. And so I didn't believe my teachers knew English. I think they probably did, but I was like, nah. I was like, you no way. So I just ignored. I was like, I'm not going to learn whatever you're teaching here. And so anyways, we win the green card lottery. We moved to the US and I remember showing up, I don't think it was at immigration, but a few days after. And they force you to take this test and I've never seen a Scantron because that's not something that we have in Latin America. And so I'm like, what is this thing? And so then I have to go,

Jon Krohn: 00:57:49 So that's where you have a pencil and you have to fill in a circle, have A, B, C, and you pick which one is the right answer.

Will Falcon: 00:57:55 So give me an English test, just a standard write stuff, and then math test. And so on the English, I'm like, I don't understand anything. They're like, you still have to answer the questions. So I just randomly picked, I think I just drew a diagonal across all of them. And then the math is universal. So I wrote that and I think that's what kept me from going back one grade. You probably would've been held back. So I came straight, I think it was fifth or sixth grade, sixth grade I think. And so then if you're someone like that, you enter this program called ESL, right? English as a second age. And so this is in Virginia. And if I had started in Miami, I probably wouldn't speak English honestly, but in Virginia you have to learn. And so you're so out of place there and you're put into all these weird situations where the American kids get things that you don't get.

00:58:47 And so I was very annoyed at that. And so I was like, I need to get out of this thing quickly. And so I made it my life mission to learn English as quickly as possible. I

think I was out of ESL in four or five months. It's just a two year program. I was just very annoyed at it, and I was like, I don't want to be pushed to the side of this thing. And so that's kind of how I got into it. And I think if you speak to a lot of immigrants, there's this need to want to simulate quickly so that you don't stick out. I think people change your names. You see this a lot in Indian culture or Asian culture. They'll want to change your names to simulate more. Luckily, my name is William, so I was great, but I need to figure out the English thing and not have an accent at the same time. So I did that. And so that's kind of what helped me, I guess. I mean, I don't know. The only trade that I could think of is if it annoys me a lot, I want to solve it quickly, and the cloud annoyed me a lot, so I'm trying to solve it quickly.

00:59:40 And training models at scale annoyed me a lot, so I'm trying to solve it quickly. Yeah, cool.

Jon Krohn: 00:59:46 And then one last tidbit is I guess in between learning, I need to learn English in between learning English and Columbia, you were training, you were a, lemme try to get this right from memory. You were a US Navy officer and you were in the SEAL training program, but then a medical injury meant you never finished the SEAL training program. But yeah, I dunno if there's anything, I mean that's also kind of another probably given everything else that we've said in this episode, people probably didn't expect me to be now talking about you as a Navy SEAL or

Will Falcon: 01:00:21 Trying to be one. No, I mean for context, I was not a Navy seal, but again, public school in Florida, you're never thinking, will I go to an Ivy League and become a programmer? You literally dunno anyone who does that. And so my best career path that I thought was super interesting was, well, maybe I can do special operations and be in very tactical things that require being honest

and having the decision to do the right thing when it matters. And then I'm like, Hey, I'm going to join. And I find out about the seals and of course I tell everyone about it and they're like, oh, you're never going to do that. And I'm like, okay, well let's wait and see.

Jon Krohn: 01:01:01 Right, exactly.

Will Falcon: 01:01:02 So I trained most of my high school to do this, and then I get selected to go into SEAL training as an officer. Getting selected to go to SEAL training as an officer is an incredibly difficult thing to do. That year that I applied, there were about 12,000 applicants. They selected about 20.

Jon Krohn: 01:01:20 Wow.

Will Falcon: 01:01:21 Because it's not, there's this crazy physical test you have to do. You have to

Jon Krohn: 01:01:26 And a really big math test.

Will Falcon: 01:01:28 No, but I mean you're competing against Harvard grads. My roommate, one of my roommates was a Stanford quarterback. So not only do you have to be top physical shape, but you have to have proven something. And I was a terrible high school student, so I don't really know what they saw other than I ran pretty fast and I was really good at Pullups and did all these things and crushed the numbers. And then I did a lot of community service. I worked a lot. I met a lot of seals and tried to help the community before I was there. And I think I got lucky. I think one of the guys who wrote me a letter eventually became the commander of SEAL Team six. So I think at the time he already had a lot of clouts, and so I think his letter carried me through and got me into SEAL training.

	01:02:10	And the reason they're so selective there is because every SEAL class is about 300 people in training, and you have only maybe 12 officers. And if an officer quits during training, you usually bring 50 people, not 50, 20 people with you. You have to understand these are basically collegiate athletes at this point, very, very good at everything. And then you've got some 18-year-old looking up to you and then suddenly if you can't do the run and if you can't do the thing, they're like, how could I ever do this? And so they'll usually quit. And so they select you out. Everyone knows SEAL Training has probably an 80% attrition rate, so about 80% of people who go through it don't make it. Not everyone quits. Some people get medically discharged or whatever. And the officers, if you only look at officers, it's actually about an 80% rate that makes it through.
Jon Krohn:	01:03:05	I see.
Will Falcon:	01:03:05	So it's only 20%.
Jon Krohn:	01:03:06	I see.
Will Falcon:	01:03:07	And most of that is injuries or some sort of training violation. Officers don't really quit. They've been so selected at that point. And so I go through training with that expectation. I'm 21 years old, and the first time I go through SEAL training, I get to basically beginning of hell week. And I get to that point, this is two months in already, six weeks at this point. So that class started with about 300 people. We get to the beginning of hell week with about a hundred people. So it was already, most people are gone and they look at me and they do these scans right before you go into hell week. And then they're like, Hey, you have pneumonia? And I was like, well, I can't hide that. You try to hide all the injuries during training. You don't want to get medically rolled out. And so I'm like, well, clearly I can't hide from an x-ray. So

they're like, Hey, you have pneumonia, I have to start again. And I'm like, I just went through this thing that almost killed me

01:04:02 And you want me to go through it again? And I'm like, all right, fine, whatever. So I go and it's like, I dunno, six months to heal up to clear pneumonia, then get back into training, get back in shape. So I do that, and then I go back in and then I go through training and I go through hell week, but I come out of hell week with injuries. So in my class, I think we had about 300 people going into training. And then we got to hell week, probably around a hundred hell week starts on Sunday. By Tuesday we had maybe 35 people left.

01:04:35 Within two days you kind of watched that most of the class.

01:04:38 So then 36 or 37, so there were two guys in there. So of my class, I think I was the second person to get rolled out of there. So 35 made it, and then for me, I get rolled for these injuries. So I've been basically running, I got sick right before I got, what do you call it, VGE. So food poisoning on Sunday. So I'm not able to keep food down and I'm going through this thing. So I weighed like 200 pounds going into this. By the time I got through hell week, it was like 160 pounds. And so around Wednesday, Thursday ish, Wednesday night roughly, they sent me to medical and you have to Thursday to fully graduate. And so at medical, they're like, Hey, you've got, I was throwing up. I have a bunch of issues. And so they scam me, they're like, Hey, you're good to go. They thought it was an appendix or something, and then they send me back, but it's been more than eight hours. And there's a policy that if you are out of hell week for more than eight hours, you have to start again.

01:05:37 So I'm like, okay. So I go back and then I'm delirious. At this point, it's been multiple days, and this is Thursday morning, and usually they'll give you a pass on Thursdays if you're there long enough. And so they didn't give me a pass. So then they send me back to the beginning, but they gave me that second chance, which is an interesting thing. As an officer, you only get one chance usually. So then they're like, Hey, look, go back to one of the SEAL teams, heal up and then we'll let you come back and finish. So this is kind of the deal that we made. So I go back, the team is not called SRT one, but back then it was called SA one support Activity one. So I'm in there and I'm really in a supporting role there. So I'm helping the team do whatever I can.

01:06:15 I can't go on deployments, I can't do a lot. So I'm mostly just showing up, training, helping the command as much as I can. I speak Arabic, Spanish, I'm trying to help it much. It's annoying place to be because you're not a seal, but you're in training. And then all the seals know you're in training. So every time they see you, they're hit the surf or something. You're like, ah, man, this sucks. So you're never fully out of it. So you're in there doing this thing. And I was in there for six or eight months, then they start sending me, so after Bud Seal training, they pull you through something called SQT, second part of SEAL training. That's actually where you earn your try that. And so my commander, I'm like, Hey, I've been helping out. Can you hook me up? And so they actually sent me through some of the SQT schools. So I go and I do, there's this amazing guy, everyone knows him, Jocko and Leaf Bain. So Leaf teaches my class called jsi, and I'm in that JY class. That class became a book called Extreme Ownership, if you ever read it.

Jon Krohn: 01:07:13 I recently bought it.

Will Falcon: 01:07:14 Yeah. So that's how

Jon Krohn: 01:07:17 On your recommendation.

Will Falcon: 01:07:17 Yeah, exactly. And so they didn't have this back then, but it is what they teach Junior Seal officers on how to lead smallest special operations team. And so you learn everything from mission planning. You do a bunch of training missions, and it's pretty cool. And so that's kind of where the birth of my leadership really comes from. I mean, in training, I was in charge of my class a bunch of times, so you learn a lot there, 300 people at 21 years old, but there's really where you learn the tactics and the stuff that Jocko talks about. And then Jocko came and gave a talk. I remember, and this is right before he retired, it was probably one of the most intense talks I've ever heard. He's a very intense guy in the real life. I'm very fortunate to have gone through a lot of the early lessons.

01:07:57 I think how week teaches you a lot about resilience, about grit, leading teams, very small teams as a 21-year-old over there teaches you a lot. Showing up to my first SEAL team with seals were 30, 40 years ahead of me. And having to figure out how to manage that, it's very different. So you learn a lot of these leadership lessons early. And then I think the grit and the ability to kind of do anything is what then later allowed me to go from public high school kid to Ivy League student successfully doing it. I'm pretty sure if I hadn't gone through this, I wouldn't have been able to do that. And so overall, I think it's great. A lot of my buddies are still in guys I went through training with. They've had amazing careers and it's been really interesting watching them evolve and see how things have shaped. And so now a lot of guys are leaving and they're about hitting the 20 year mark as we all join around 2008. And I am always trying to figure out how I can help 'em do their next thing. And so hopefully I'll for any team guys watching this, if you need funding or you need a VC connection, let me know. But yeah, I

mean, I'm very excited to continue supporting the community as well. Yeah,

Jon Krohn: 01:09:04 Really cool. I didn't know almost anything that you just told me about your history with the BUDS program or anything about it really. So thank you for that. It does seem like there's a little bit of a common thread that if listeners want to be able to achieve something that is extremely difficult to get invited to do, it sounds like, whether it's a top AI lab or the seals, a key thing is volunteering, hanging around the right people, not getting paid, but just learning the right people and eventually an invite can be forthcoming. Obviously a lot of hard work.

Will Falcon: 01:09:41 I think there's never an invite. There's you going out to get it

Jon Krohn: 01:09:46 In

Will Falcon: 01:09:46 Whatever way that looks.

01:09:48 And today it's a little bit harder, I would say. I mean, I don't envy the PhD students have to apply today. There's always a joke in every PhD class where, wow, where we could not get in the next year. You're always like, oh, I could not have gotten in this year. It's true. I mean, you need 20, you basically have to be a professor now to get in. But no, you have to just go out there and get it. And I think startups are very similar. It takes grit, it takes dedication, and just knowing what you want.

Jon Krohn: 01:10:15 Well, I think contributing to open source projects like the many that PyTorch Lightning has led to that whole ecosystem in GitHub, people contributing to those, making a big impact, writing papers, publishing blog posts, maybe making instructional YouTube videos or something. Sure. All that stuff helps for getting into PhD programs or becoming a Navy seal.

Will Falcon:	01:10:39	Yeah, I mean, it takes Mike Kiss a bit of luck, right? For sure. But I think, yeah, if I look at the guys who became seals, I think everyone has very similar traits. Sometimes people get lucky, sometimes people don't. Right. But no, ultimately it's that drive of I'm never going to quit. That's it.
Jon Krohn:	01:10:57	So I always ask my guests for book recommendation at the end. Is your book recommendation extreme ownership?
Will Falcon:	01:11:01	A hundred percent.
Jon Krohn:	01:11:02	Nice. Well, that was easy. And for folks who want to be able to continue to follow your story, the Lightning AI story, where are the best places on social media to do that?
Will Falcon:	01:11:14	Yeah, my Twitter, I guess. William Falcon at Sign, William Falcon or Will Falcon, I can't remember one of those. And then the Lightning ai, Twitter, and then our LinkedIn. I think that's maybe it. Where else do we have this? Yeah, I think those are it.
Jon Krohn:	01:11:27	Yeah, we'll have all those in the show notes and anything else? Well, your podcast, I guess. Yeah, hopefully people are aware of that. Nice. Yeah, thanks for mentioning that Will. And yeah, thanks so much for coming on the show. It's been great to have you on. Maybe we can check in a year or two and see how things are coming along.
Will Falcon:	01:11:46	Yeah, once we're close to being IPO or at IPO, then we'll talk about that during. Sounds
Jon Krohn:	01:11:50	Great. We'll talk about the IPO story.
Will Falcon:	01:11:51	Perfect. Cheers. Well, thank you. Awesome. Thank you.

Jon Krohn: 01:11:57 What an episode today with the extraordinary Will Falcon in it, he covered how neo clouds are GPU first cloud providers offering higher performance for AI workloads relative to traditional clouds like AWS, which were built for CPU applications. He talked about how lighting AI merged with Voltage Park to create the third largest neo cloud in the world, with over 35,000 GPUs, over $500 million in a RR and a full stack AI platform serving over 400,000 developers. He talked about how Lightning AI got started and how it was based on the open source project Pi Torch Lightning, which started as a personal research tool in 2015, but has now been downloaded nearly 400 million times. He also talked about how his leadership philosophy has been shaped by Navy SEAL training. 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 Wills, social media profiles, as well as my own@superdatascience.com slash 9 6 5.

01:13:01 Alright, that is it. Thanks to everyone on this SuperDataScience podcast team, our podcast manager, Sonja Brajovic, media editor, Mario Pombo, our partnerships team Natalie Ziajski, our researcher, Serg Masís writer, Dr. Zara Karschay, and our founder Kirill Eremenko. Thanks to all of them for producing another stellar episode for us today for enabling that super team to create this free podcast for you. We are deeply grateful to you and to our sponsors, and you can support the show by checking out our sponsors links, which are in the show notes. And if you'd ever like to sponsor an episode yourself, you can find out how to do that at Jon Krohn.com/podcast. Otherwise, help us out by sharing this episode with folks who would like to have it shared with them. Review it on your favorite podcasting platform or on YouTube, subscribe if you're not already a subscriber. But most importantly, just keep on tuning in. I'm so grateful to have you listening, and I hope I can

continue to make episodes you love for years and years to come. Till next time, keep on rocking it out there, and I'm looking forward to enjoying another round of the Super Data Science Podcast with you very soon.