**William**

What is that iPhone experience look like? What does that look like? And so it took us about three years to figure it out. We tried many things and ultimately we landed on studios, which we launched. It's been it's been an industry like behind the scenes for about a year, like being used by enterprises. But we made it public a few months ago. I actually do think that Studios achieves that iPhone leap where like, it is super usable, and pretty much every single tool in the world today forces you to code on your laptop, on your local machine, and then you submit to the cloud. And that's a paradigm that they've all adopted. We got rid of that paradigm. We don't let you do that. We just let you code on the cloud and stays on the cloud. You can still connect to your local IDE and code. So it still has the local experience, but everything is on the cloud.

**CRAIG**

Hi, I'm. Craig Smith and this is eye on eye. In this episode I speak with William Falcon, the creator of PyTorch lightning and founder of lightning AI. We'll introduces us to Lightning Studio, the innovative cloud-based platform that integrates various AI tools and frameworks, allowing users to train, fine, tune and deploy models at scale. We'll talks about his passion for open-source AI and democratizing access to powerful AI tools. I hope you find the conversation as useful as I do. AI might be the most important new computer technology ever. It's storming every industry and literally billions of dollars are being invested. So buckle up. The problem is that AI needs a lot of speed and processing power. So how do you compete without cost spiraling out of control, it's time to upgrade to the next generation of the cloud. Oracle Cloud Infrastructure, or OCI. OCI is a single platform for your infrastructure, database, application development and AI needs. OCI has 4 to 8 times the bandwidth of other clouds, offers one consistent price instead of variable regional pricing. And of course, nobody does data better than Oracle. So now you can train your AI models at twice the speed and less than half the cost of other clouds. If you want to do more and spend less like Uber, eight by eight and Databricks Mosaic, take a free test drive of OCI at Oracle.com slash Eye on AI. That's E Y E O N A I all run together at home. At work. We all know one person whose password challenged sticky note reminders, emailing passwords, reusing passwords, using the word password as their password.

**CRAIG**
Because data breaches affect everyone you need One password. 1password combines industry leading security with award winning design to bring private, secure and user friendly password management to everyone. Companies lose hours every day just from employees forgetting and resetting passwords. A single data breach costs millions of dollars. One password secures every sign in to save you time and money. 1password lets you switch between iPhone, Android, Mac, and PC with convenient features like autofill. For quick signups, all you have to remember is the one strong account password that protects everything else your logins, your credit cards, secure notes, or the office Wi-Fi password. 1password generates as many strong, unique passwords as you need and securely stores them in an encrypted vault that only you have access to. I use 1password and you should too. 1password is award winning. Password manager is trusted by millions of users and over 100,000 businesses. From IBM to slack, it beat out 40 other options to become Wirecutter's top pick for password managers, plus regular for third party audits and the industry's largest bug bounty. Keep 1password at the forefront of security. Right now, my listeners get a two week free trial at 1 password.com/eyeonai that's Eye on AI all run together E Y E O N A I. That's two weeks free at 1password.com/ eyeonai 1password com slash Eye on AI

**CRAIG**

Will, can you introduce yourself. Tell us about how you got to lightning AI and then we'll talk about lightning studio.

**William**

Yeah, so I'm William Falcon. I'm the creator of PyTorch Lightning and founder of lightning AI. I come to AI from Research, right, like most of us do. But before that I actually was in the US military. So I spent about six years going through Navy SEAL training in San Diego; during that I got injured. The Navy gives you a few options at that point, the option to continue and finish was not one of the options they gave me. As an officer, I only get one chance to train so that was unfortunate. But then, you know, they gave me the option to become an Intel officer, or leave and try something else, so I left. I was considering things like the CIA and other special operations, and then just kind of found my way. It was kind of the time when iPhones were coming out and iPhone apps. So I started building, just started coding, coding apps, and eventually find my way into computer science and math at Columbia. Don't ask me how but ended up doing my undergrad at Columbia University in New York. And then there, I learned that you could major computer science, right, and so I did.

**CRAIG**

Did you know a guy named Chris Wiggins?

**William**

Yeah, so I know Chris, from the Center for Data Science, I think as far as I'm from. Yeah, and so right around the time I got to Columbia it was like 2013, which is when deep learning had just taken off, like the year before that I had no idea what it was, right. But I show up to one of my classes and, and there's this Professor Tony Jebara, right, who ran Netflix and stuff for a while, on the deep learning side. And he's teaching the neural networks and I'm like, I don't know what any of this is. And, one of the demos he gives us, MNIST from Yann down at NYU, right down the street. At the time it had little carousel music and I was like, okay, I don't really know how that's useful, right? But it was interesting. And so I started learning about it. A long story short, about two, three years later, I ended up doing research in computational neuroscience and in 20, like, 16, 17, I want to say I went to NeurIPS in Montreal, back then it was like a few 100 people. So I discovered the deep learning community and somehow ended up, kind of going into that. I had taken up a job at Goldman during that time.

**William**

So I told my manager, hey, I wanna I want to do deep learning on the trading floor and they were like, no, and I was like, fine, I'll leave. That's how I went back to school, and just did research. And then at that time, I started building like internal software for myself to be able to move faster through it. That's why I eventually open sourced which is known as PyTorch Lightning today, right. So it started super, super early. And I ended up doing my PhD down at NYU with Yann LeCun and then Kyunghyun Cho as well. Then about three, four years into it, I dropped out to start Lightning AI. PyTorch Lightning took off. I was at Facebook AI research at the time. And then you know, 1000s of companies were pinging me, they're using it, they're trying to put models in production. This is like late 2019, early 2020. And, you know, they were super encouraging. And they basically were like, hey, like, you should probably pursue this and see how it goes, and, you know, we'll support you. And they have been very supportive. So I left Facebook, and have been focused on Lightning since then. At the time, we were trying to figure out how to remove a lot of this overhead that you need to know to do deep learning, right. Back then you had to implement your own gradients and all these different things.

**William**

So PyTorch lightning solved so PyTorch for Meta, solved a lot of these problems, and then Lightning solved the scaling problem. But then you start to know Kubernetes, and Docker, and Cloud and a million other technologies. So I started trying to figure out, how do you do that? How do you it really felt like MS DOS or like a Blackberry before an iPhone, super clunky. So I was like, well, what is the iPhone experience? What does that look like? And so it took us about three years to figure it out. We tried many things. Ultimately, we landed on Studios, which we launched. It's been in industry, behind the scenes for about a year, being used by enterprises. But we made it public a few months ago. I actually do think that Studios achieves that iPhone leap. It is super usable. If you go to Twitter today, on LinkedIn, and social media, you hear people talking about it, like “revolutionary”, “it's the next thing”. That's what we wanted and that's what Lightning taught us. We want you to get that 10x experience on something.

**CRAIG**

Yeah, let's go back. Explain what Pytorch Lightning does sort of in layman's terms.

**William**

Yeah. So PyTorch is a library built by Meta, right, or Facebook AI at the time. And, what they saw is the ability to do computational graphs, right. When a model learns, you have to compute gradients, and you have to update those gradients. So it's like, how to express a math function and program in essence, right? They give you the raw tools to do that. So, it'd be me like giving you a bunch of tools to build a car, like, here's a wheel, here's this, here's an engine, but then you still have to build your own car. So people come up with ways of building their own cars. And they'll kind of look and feel the same, but they're not standard. Some people put like four wheels, five wheels, three wheels, they put them in different places. PyTorch Lightning came in as an interface, like a front end for that, that says, No, here's how you do it. Here are the standards, you put wheels here, you put this here, you do this here, and it structures your code. It's not an abstraction. It's really just structuring the code and organizing it for you. Then from that, you're able to scale and that came out of many years of research. The big deal that this unblocked was the ability to train across 1000s of machines, right? In 2019, that wasn't happening at Facebook. It was like me and two other teams who were doing this, right. I was an intern at the time and the other teams were like professional engineers, like ten of them, right, doing these things. And, I got a lot of their help to building lightning into what it is today, putting all their knowledge into the programs. So that's how I came out. So in 2020, when I left Facebook, I think the paper we published a year after, we trained that on like, one or 2000 GPUs, right. And that was in 2020. Today, people are starting to learn how to do that but we've been doing that for a long time. So it gives you that scaling ability, right? So if we're use, like, web frameworks, we basically say like, PyTorch is JavaScript and PyTorch Lightning is React, like you wouldn't build your own React library.

**CRAIG**

And then once that product had traction, you started looking around and seeing all the other tools and frameworks that you need to work with, with PyTorch Lightning, and you eventually brought it all under one umbrella, in that studio?

**William**

Exactly. So just like in Web, right— I mean, actually I’d argue that even AI is a lot harder because you have so many moving parts; but PyTorch Lightning just one piece, but you needed so many things. So lightning was the first thing to integrate other tools. We were like the first ones to integrate, like, the ability to train on different hardware. So before lightning, you didn't really code your stuff, and then go to GPUs, and then go to TPUs, today you do that, right? That's something that we pioneered in Lightning. So we started integrating different accelerators which is super valuable today because there's not enough GPUs to go around. So you kind of have to be able to try other hardware. Then we also brought in things like experiment managers, like TensorBoard, Neptune, Weights and Biases, Comet, all that stuff, right, integrated into the system. So it started growing into kind of like an operating system in essence, right, that connected a lot of things together. And that was kind of the premise behind a lot of what we did. In Studio, we basically took that farther and we just said, okay, and now we're gonna bring in your cloud infrastructure as well and then all your teams, your developers, your data scientists, your machine learning engineers, your designers, everyone can work on a single thing together from the browser but the infrastructure is at scale, right.

**CRAIG**

Last time we spoke, I had just spoken to Versel, which has a services or a platform where you can build a model and then access different foundation models. It's kind of like the bedrock at Amazon. But maybe more versatile and with more options, and I saw Versel is in lightning studio as one of the tools or platforms that you can access. It's a little confusing to me. You keep on getting these umbrellas that then give you access to things further down, further upstream I guess. Can you talk about this as, sort of, hierarchy of platforms, or models, or umbrellas?

**William**

Yeah, so I'm not super familiar with how Versel works but for what I know, it's mostly about web development and deploying web apps, right, like that kind of thing. That's not really the focus of Lightning, right. So what you saw in lightning is the ability to deploy web apps, you can deploy React, you can deploy Angular, Next etc. But that's not really the focus, right. The focus of Lightning is, how do you deploy models at scale? How do you how do you train models? How do you fine tune them? How do you take 10 terabytes of data and do distributed processing on it? So it's really more about the scale. What separates Lightning is this ability to take your studio and scale it to 1000 machines instantly, right? That's not something that a website ever needs. A website is happy to work on a single machine and maybe scale a little bit, but like, not at the scale that you need for deep learning. So it's a fundamentally different thing. Yeah, I think that all tools have their place, for sure. And we’re always happy to integrate with things. But definitely the main thing that we're going after is the scalable workloads, the machine learning kind of side of it, and like through that people do have to make web apps, and they have to make websites. So we give them the ability to do that pretty easily as well. But you know, our customers make tools all the time that don’t just use one.

**CRAIG**

Yeah. And when you seem to build models, you're talking about any machine learning model, are you talking about foundation models, or generative models?

**William**

So PyTorch Lightning came out before foundation models were a thing, before Gennai was a thing, before any of this, right. So PyTorch Lightning— In fact, one of the foundation models to built an AI is built using PyTorch Lightning, right. So from the very beginning, we've been at the epicenter of GenAI. So to answer your question, you know, our tools scale from the simplest models, like literally tiny, tiny models to foundation models on 1000s of GPUs, you know, over the years to our frameworks, and our tools have stood the test of time. And we know that they actually do work for pretty much any type of model today, right. And we didn't actually know that in 2021, when Lightning came out first, right. So that's been true, but yeah, it's really any type of model. And I think we definitely do like Gen AI and foundation models better. And when I say build, that's a good, that's a good point. I don't literally mean, you have to build the model. Most people just grab a model from open source, right? I don't mean that. For researchers like, fair, you're doing your PhD, like, sometimes I do mean that, and we do that. But really, the majority of people are fine tuning models, they grab something, they drop it, they put some data on or they fine tune it for a few hours. That's great. That's exactly one of the bread and butters that we do on the platform.

**CRAIG**

Does Studio then allow you to do things like connect to vector database, you know, and implement RAG and all of those things that people are doing today?

**William**

Yeah, so we have like a public collection of studios that we put out, kind of like a gallery or like a library of these, these are templates, right? In fact, just yesterday we put out a RAG one. So that one, you can grab a whole data set, do RAG with it, and there's like a UI, you can query it as well. You can integrate an open source database or your personal databases as well. And these things come up a lot, like you know, pretty much anytime there's a use case someone's like, Oh, can you do this or that? Usually either we do it or someone in the community will put it together in a few days. But the platform is very extensible, right? It's kind of like having a car and asking if you can drive to Texas or the supermarket, like, yes!

**CRAIG**

 This is open source, it's not right.

**William**

So a lot of our frameworks are open source, the platform itself is not but a lot of the pieces of the platform are open source, right. But the code that we give you, the templates are open source, like that RAG template, that code is free in there. You can use it, you can take it off Lightning if you want, right, but it already works out of the box.

**CRAIG**

Yeah, are there other things in the market that have approached this and you've gone, sort of one layer higher or broader? That people may know or could relate to?

**William**

So kind of like, remember when the iPhone came out, most things looked like blackberries and Palm Pilots, most things that were out there at the time, Razor phones, right, like that kind of thing. That's what that middle of the market looks like today. And that's because people were kind of trying to do the same thing, just slightly better, different ways, right? We said, No, we're not going to do that. In fact, we like stopped— like, I don't really care what other people are working on in general, like we don't follow what what the markets doing. We just talk to our users, hear what they want to do. And we solve it our own way, kind of like Apple, you know, they'll solve it their way however they want to do it. And, and no, so we really rethought the problem from the ground up. And so we said, like, it's got to be fundamentally different, right? And pretty much every single tool in the world today forces you to code on your laptop, on your local machine and then you submit to the cloud. And that's a paradigm that they've all adopted, we got rid of that paradigm, we don't let you do that. We just let you code on the cloud and it stays on the cloud, right? You can still connect your local IDE and code, so it still has the local experience, but everything's on the cloud. And that's a very, very different paradigm. No one else is doing that. And when we started doing that, people thought it was crazy, right? Kind of like when Apple took keyboards off phones, they thought it was crazy, right?

**CRAIG**

Yeah, I would imagine, though, that there are latency issues if you're working in the cloud. I just— this is way before machine learning, but I used to have to do the— or, expenses at the New York Times, you were working on a remote platform. So everything you input, there would be like, a little incredibly irritating delay, and then, that sort of thing. So how do you overcome the latency issue?

**William**

We’ve spent a lot of time making that go away, right. And that's why I also said, you can also code from your laptop, so there's zero latency. When you connect your local VS code, your IDE, it's exactly the same experience. You know, the market has gone through this many times, and I've been in the other side, too. I've been skeptical about things, right. But, have you ever tried Figma for example, for designing?

**William** 18:01

So Figma is like a browser design tool, right. It competes with Photoshop, and those kinds of tools, right? Now, when Figma came out, you know, you ask any designer, and they're gonna be like, Wait, like design on the browser? Why? Like, it's not powerful enough, I have my Photoshop on my laptop, right? Adobe bought figma for $20 billion. Why? Right? Because that is the future, and most things are moving that way, you just have to be very good about how you do it. And our team has spent a lot of time trying to solve these issues. And I think we've landed there, now it's gonna get better. 100%, right, the first iPhone and the current iPhone are very different. But we have to take the industry and leap forward that way. And we have to say it can be done. And we're going all in on it because we believe it is the future. And most people are, I would say about three years behind now because of that, right? But we have a lot of knowledge on how to keep improving things as well.

**CRAIG**

Yeah, and then you integrate all of the, like copilot or I mean, GitHub Copilot, or Codex, or all these other tools that people use while they're coding. So you're building by coding or you're building by putting together different elements that are pre written?

**William**

Yeah, exactly, an iPhone has iPhone apps, right? And you need apps to do your work, right? You need a way to write, you need a way to send emails, calculator, flashlight, etc. When you're developing AI, you need apps, right? So to us everything is an app. So you want to train a model, there's an app for that, you want to fine tune the model, there's an app for that, you want to deploy a model, there's a fun app for that, you want to do RAG, there's an app for that, right? So every one of these things has an app embedded in it. Depending on what you're doing, it may be a full code experience. So like all you want to do is code and develop there’s literally an IDE to do that. But if you don't want to do that, there might be another UI. Some of these embedding things, there's no code. You’re just like watching plots and dots on the screen, right? It's kind of like no code. So, just like your Mac, sometimes you write into an app, and sometimes you don't, you click into it, right?

**CRAIG**

So once you have this built, the studio built, and it's got PyTorch Lightning and all of these other apps, are you continuing? — I guess the unique thing is that this is all being done in the cloud, or at least if you're doing it locally, it's uploading to the cloud. I mean, then you have all these apps that you can use, but what are the other features of Lightning Studio that are not covered by an app? Or is it really an orchestration layer to like, manage all these different tools?

**William**

So there are features that are not apps, specifically, right? So the first thing is like, you can create one studio for every task. So you know, let's say you want to train a model, you'll create a studio just to do that. And then you'll create another studio, maybe to do RAG, and then you'll create another studio maybe to do, I don't know, fine tuning, and another studio to serve. You can have infinite of these. In essence, what you're doing is you're taking your laptop— imagine if anytime you want to start a new project, you could just have a brand new laptop, that's what you have, right? And you can store that laptop, and then when you want to work on that you just open the laptop, and it's ready to go. So that's kind of the first thing, I think it's the most useful. I program in probably three or four things, you know, front end, back end, machine learning, data science, different times. When I want to context switch, I just go turn on that studio and I'm there. I don't have to set anything up. So it's a lot faster. That's, I think, probably one of the biggest features for everyone. Things just work out of the box, then I can share that with you. So let's say you today, you're like, Hey, how did you find him that model? Can you share that with me also? Sure, duplicate my studio, and you literally take a carbon copy of my laptop, and it's done, right? So, the community creates these and they share these. You asked me about the RAG, we didn't create the RAG someone else did. And now you can go take their RAG up and deploy yourself in five seconds, right? And you had to do zero work, the RAG studio. So that's one, you have the ability to work together. So like, let's say you and I are coding and maybe you're learning to code, you send me a link, just like in Google Docs, and I go and type with you and we just code together, right? So I can teach people how to code, I can help junior engineers onboard easier, a lot of things we can do there. You have the ability to measure cost in real time. So you can say, I don't want to spend more than $100 on this thing. And you put a budget of 100. And we will measure the cloud costs to a second. Ask any person in an enterprise today, any company, how much they spend on anything machine learning, they can tell you. You ask us? I can say, yes, on that thing we spend $50,000, $271, that's it. No question asked, right. It's all there. And then there's a bunch of team management stuff, like who can see what when, or what levels of the org, who was an admin, who isn't, what data can they see, what can they not see, right? The ability to connect to different data sources like Snowflake and Databricks data, and then bring in data from S3, upload your own data, right? So all of that gets ingested. So yeah, Lightning, in essence, is kind of like an operating system. It just becomes the center that everything kind of comes together. And then you can create studios that are specialized to specific tasks that you need.

**CRAIG**

Yeah, and you were saying that without Lightning studio, people would have to set all this up themselves. Can you walk me through a use case, to build something, and the steps that you would take, and then how that is different with Lightning Studio?

**William**

Yeah, so let's say mixed trail, I think is one of the best models out there, came out, and this is the mixture of experts.And let's say you want to deploy that model, right? And for you to do that you could have an option. You can either go to one of these API companies and then hit a button and then they'll deploy it for you. But you are kind of subject to them. What if they go down? Like, OpenAI goes down all the time, right? So what happens? Is your app going to stop working?  Your service is done, right. So do you really want to give that up? I don't, I'm not sure, right. So that's your first option. Second option is you go find the code in open source and then you download it to your laptop and then you like set it up and try to get it to work, then you try to find a cloud machine somewhere, and then you go to that cloud machine and you got to set it all up again, and then maybe it doesn't work. And you spend like a week or two doing this. It's extremely hard, right? On Lightning, you go to Studios, there's the gallery there with all the templates, you find the MOE template, you click on it, you press duplicate, takes about 35 seconds, it's up and running, and the model’s already deployed and everything there, and you have the code fully. So, you can delete the code if you want. But it's fully transparent to you. So you have full control over it. You can hack it, or you can leave it alone and leave it how it is, right.

**CRAIG**

AI might be the most important new computer technology ever. It's storming every industry and literally billions of dollars are being invested. So buckle up. The problem is that AI needs a lot of speed and processing power. So how do you compete without cost spiraling out of control, it's time to upgrade to the next generation of the cloud. Oracle Cloud Infrastructure, or OCI. OCI is a single platform for your infrastructure, database, application development and AI needs. OCI has 4 to 8 times the bandwidth of other clouds, offers one consistent price instead of variable regional pricing. And of course, nobody does data better than Oracle. So now you can train your AI models at twice the speed and less than half the cost of other clouds. If you want to do more and spend less like Uber, eight by eight and Databricks Mosaic, take a free test drive of OCI at Oracle.com slash Eye on AI. That's E Y E O N A I all run together. At home. At work. We all know one person whose password challenged sticky note reminders, emailing passwords, reusing passwords, using the word password as their password.

**CRAIG**
Because data breaches affect everyone you need One password. 1password combines industry leading security with award winning design to bring private, secure and user friendly password management to everyone. Companies lose hours every day just from employees forgetting and resetting passwords. A single data breach costs millions of dollars. One password secures every sign in to save you time and money. 1password lets you switch between iPhone, Android, Mac, and PC with convenient features like autofill. For quick signups, all you have to remember is the one strong account password that protects everything else your logins, your credit cards, secure notes, or the office Wi-Fi password. 1password generates as many strong, unique passwords as you need and securely stores them in an encrypted vault that only you have access to. I use 1password and you should too. 1password is award winning. Password manager is trusted by millions of users and over 100,000 businesses. From IBM to slack, it beat out 40 other options to become Wirecutter's top pick for password managers, plus regular for third party audits and the industry's largest bug bounty. Keep 1password at the forefront of security. Right now, my listeners get a two week free trial at 1 password.com/eyeonai that's Eye on AI all run together E Y E O N A I. That's two weeks free at 1password.com/eyeonai 1password dot com slash Eye on AI

**CRAIG**

Yeah, and you said that there's been a very strong response, how big is that? Well, first of all, how do you charge?

**William**

Yeah, so we have four different tiers. So there's a free tier, so everyone, when they sign up, they get a free tier, they automatically can run a studio for free anytime of the day, right, CPU studio. They literally could code on the cloud all day long. They can make infinite studios and they can only run one at a time. So if you want to run multiple, then you're going to use credits, right? So you have to buy credits. You pay as you go as you need them. We also give you 15 credits, right. But 15 credits means you can basically run about 22 GPU hours per month, which is a lot, right? Like, I don't think most people need that. So you have the 15 free credits. So in essence, you're gonna get one free running Studio and 22 GPU hours every month that you can use for free. And then when you need more advanced features, or you need more hours, you buy more credits or you upgrade to the different tiers, right. If you're a company, there's stuff around security and the way that you work with people that are in the different tiers, as well. What we wanted to do is get back to the open source community. I do think that there's a big accessibility problem. I'm from South America, right. And if you tell me that I can get a GPU in Venezuela, like, it's not going to happen but I can through Lightning. And so now you've allowed people in developing countries to like, participate in this as well and do something, which is cool. And, and then after that, you have the ability to integrate all the open source stuff as well, right. So it's, I think, the best of both worlds.

**CRAIG**

Yeah, and how big is the community now, including freemium users?

**William**

So the PyTorch Lightning community is about 1 to 2 million people across the world, that's on the open source side as well. So the platform just became public like, two months ago and we already have about 50,000 people on there, right. So it's growing extremely fast. Some people might end up in a waitlist if we can't verify you, for whatever reason. So we're trying to get people off the waitlist as much as we can. I think probably about 20 or 30,000 of them are still on the waitlist. So we have to go through and like, do that. But you know, we're moving as fast as we can. And we're a little overwhelmed at the moment but overall it's been really, really positive and overwhelming, right? It's kind of what we saw when we first launched PyTorch Lightning in the very beginning, it just went viral on its own. Because, you know, people appreciate good design, they appreciate usability, especially in a world where like, you know, I mean, when I came into AI and I saw the way that people were working, I was like, How? It just really bothered me and, like, I hate using clunky things, right. And so I just wanted to make something that was extremely easy for the world to use so that they could focus on science, they could focus on the models, they could focus on the business problem, instead of like learning Kubernetes, and Docker, and this and that, right? Even most hardcore engineers, when they get on they're like, Oh, my God, thank God! Like yes, it was intellectually curious to know these things but it really was so annoying for years that I don't want to deal with it.

**CRAIG**

Yeah, you mentioned Weights & Biases, I had a conversation with the Weights & Biases guys and that's, that's on the platform. They do things like monitoring. A lot of the ML ops stuff, which now is becoming LLM ops, and maybe it'll be LMM ops at some point. Do you do any of that? Version Control monitoring after deployment, things like that, or again, you're letting people like Weights and Biases. do that through the Studio?

**William**

Yeah, so they're apps, right, they’re partnered with us. So, we don't provide that capability. We provide a few apps that do those things. So on the platform, people use TensorBoard, which is open source. They do use Weights & Biases. What else did they use? They use things like Comet, they use MLflow. But yeah, those are all integrated and people can use them out of the box.

**CRAIG**

Yeah, so last time we spoke it seemed that you were about to release something new on the platform, or am I mistaken? I talk to so many people.

**William**

Yeah, I mean, we've released a lot of stuff already. I mean, we have many cool things that are coming in the next few months for sure but right now people are just getting started to even try Studios.

**CRAIG**

Right, and how big is your team?

**William**

Yeah, very small. So about 40 people, right, which is cool. Remember, I grew up in special operations, right? SEAL teams are very small. You have very, very talented people who just really worked together well and internally, the company structure is how the SEAL teams are structured actually, because that's kind of the management style that I know. So everyone moves aggressively fast. Everyone is a team player, very collaborative. And we hire the best people in the world. We love to train them and, you know, I think one thing that I do is teach them how to work really, really well as a team, right, which is something that I thought was crazy how the civilian world doesn't really have that. There's a lot of individuals working on a team, but not an actual team working together.

**CRAIG**

The Generative AI and world models are coming on, this space is moving very quickly. How do you keep the studio relevant as the research moves, and then that research becomes adopted?

**William**

Well, remember, that was the goal of the Studio, is to build an integration platform, right? So inherently, it is designed to do that. So actually, if you don't use Studio, you probably will fall behind. I'll give you an example. Just this week, there were two models released, there was one from Alan AI, and there was another one, right, he Facebook one, the Code LLaMA. On the day that the Code LLaMA model was released, within an hour or two, we had a studio out, a template already ready to go with it. Then when the other one was released from Alan AI, within an hour or two, we had another one out. So like, we can get you that latest AI stuff, literally within a day, like worst case scenario, no matter what it is, RAG, you know, new database, something crazy comes out, we can always integrate it, because it is built for integrations first, at the end of the day. And that's that that's the fundamental problem that most platforms have, is their point solutions. They just do the one thing. They're like, a calculator, and then when a flashlight comes out, they're trying to stick a flashlight in a calculator and you're like, No, it doesn't really work, right?

**CRAIG**

Yeah, actually, that's interesting, because it also gives people a central source of knowledge. So, they don't have to be reading about every new tool or new technique that comes out, it'll appear in Lightning Studio. Is that right?

**CRAIG**

And I would imagine there's documentation, or tutorials, or something to help people understand when something new appears?

**William**

Yes, and we'll do the work to; vet something is real or not, most of us are researchers, most of us come from academia, like, you know, PhDs and that's something we have related. So we can pretty quickly know when something's hype or not, and we will vet things. We'll experiment with everything but we’ll decide what actually sticks around or not. So, you know, you can kind of defer that work to us as a company.

**CRAIG**

you also are in a position to see what people are using the platform for. Does that give you visibility into where the markets going, what's popular? What's hot? That sort of thing?

**William**

Yeah, I mean, look, we obviously don't look at people's code or anything like that, right. So I think, what I can tell is when we get new customers who like are excited to use a platform that kind of generally tell us what they're working on? Not the details, but like, what they're trying to do. And yeah, I mean, we see trends, I will say that, you know, I come from research and research is leading enterprise usually. So even people who come in, they're already like six months behind. So they're talking to me about things that were out many months ago. And, and it's rare that I find a new team who's like right on it with like, the thing from that week, who’s like thinking about it. And honestly, as a business, you shouldn't be operating at the bleeding edge. You shouldn't be a little bit behind. But researchers are always kind of trying the new stuff and we have researchers on the platform because it's one of the best tools for R&D and, like, iterative work. And so we see a lot of stuff that they do and we collaborate with them on papers as well. And then those papers get published. So we're like at the bleeding edge with them a lot of times as well. But you know, there's no news there. Its a lot of the same stuff, fine tuning, pre training, deploying, right? It's a lot of the same. And then like whether use RAG or not, it's a detail usually, like all these tools that come out, they’re details to a lot of the ways that these things are done.

**CRAIG**

I had a conversation this morning with a guy who was talking about using multiple models, and then having a consensus layer that, sort of, reconciles contradictions within the outputs from several models and comes up with an average or with a— can that kind of thing be done through Lightning?

**William**

Yeah, so that's, there are few terms for that, but that's probably mixture of experts (MoE) is he's referring to. And yeah, so it's just about training different models, and then somehow aggregating their outputs, you can either average time or you can have another model do consensus for all those models. So that mixtural model, the mixtural MoE, that's exactly what it does. It has eight models and it does that. You know, when people talk about Chat GPT and they compare with open source, it's actually an unfair comparison because Chat GPT behind the scenes, it's likely like 30 models that are all collaborating to do stuff, right? And then you compare that with, like, one model? Well, it's not true, you got to compare it to a collection of models, right? So you can do that. And it is actually probably one of the only platforms if not the only platform that can build systems because every other platform in the world today can only do like, one thing, like a model, right? But they can't connect things, whereas lightning can keep systems kind of going and have many of them working together, right? So the RAG example that we had, it's a good example, where on one studio you deploy the model and then on a separate studio, you have the RAG system completely isolated and they're talking to each other, right.

**CRAIG**

Oh, yeah, that's fascinating. How are you funded? I mean, was this costly to build?

**William**

Very expensive. So yeah, we've raised about 70 million so far, from Index Venture or Spin Capital, Coatue and a bunch of amazing angels. Yeah, like, really, really great angels and, you know, I would say, most enterprises that we work with have spent, probably even 100 million easily, to build some version of this internally. And then their headcount is crazy to maintain these things, right. So if you want to build something like this, you probably have to invest about $100 million, and like two or three years to try to do it. And most companies have tried that. And today, if they tried it, they hate it, all of them and they're trying to get rid of it somehow, right. And so they're looking to us to replace that. Even the world's largest banks for example, that we work with, they've invested a lot of money, so much. It's extremely difficult to build. The thing is, you look at it, it's pretty easy, right? It looks looks simple, but the iPhone looks simple. Right?

**CRAIG**

Yeah, Wow, and so Studio came out, it was publicly available in December? And your community's growing. What, what's next? I mean, where does this go? Is it just a matter of keeping up with all the tools that are appearing on the market?

**William**

Yeah, I mean, we want studio to be the way that you develop and code, that's it. I think we're successful, when you’ll get so annoyed at having to set things up on your laptop, on your local machine. You know, you're like, “Oh, I gotta start a work project”, and you'd rather just go to Lightning, turn on a studio, and you know, you're done, instead of trying to mess around with your local machine. We're getting there already with a lot of people but hopefully by the end of the year this is the standard for everyone in the world, they understand it. Now it will take probably more years for that to reach everyone but like, how long did it take the market to convince BlackBerry users that keyboards were not a good idea, right? It took a while but today we know that that opened up a lot of doors, right. So when there are paradigm shifts that occur, it takes a while for the market to catch up. But in the long term, it is the better thing.

**CRAIG**

Yeah, and you have some major enterprises with a lot of developers using the Studio?

**William**

Yeah, we have quite a few customers all the way from small startups to huge enterprises. Probably one of the largest ones that we have, it's like a top bank, top two, three bank and they have at least four or 5000 data scientists internally that do ML and they have 1000s of use cases. And so we're currently undergoing, like, a full platform deployment there as well, which is cool to see because it is a massive scale, just in that one customer.

**CRAIG**

Yeah, and this is a little off topic but I talked to somebody frequently about unit testing; which is kind of something people don't talk a lot about, it takes a lot of time. Does this support unit testing tools, automated unit testing, or things like that?

**William**

Yeah, so we actually run, so PyTorch Lightning, all our open source projects that live on GitHub, they have rigorous unit tests and they have CI/CDs because a lot of companies depend on them. So we run probably PyTorch Lightning, on any given day, we run, I don't know, 5000 tests per pull requests and there are probably a few dozen pull requests. So we're spinning up 1000s of machines a day just to test open source frameworks. Recently, we started switching to using studios for a lot of those. So we're even getting off those platforms, because it's a lot easier and faster for us to do that, and cheaper.

**William**

I'll tell you some other interesting use cases, because people always hack with new tools, right? So coding interviews, so we've actually started doing our coding interviews in Studios, which is cool. So like, if you interview for us, and you're an engineer, you'll get on a studio and you’ll code with me, right? And it's probably the only tool you can do that with for machine learning because like, what other tool can you just run GPUs on and code together? You can't, so if you're testing a machine learning engineer, who's joining your company, I mean, this is kind of the only way to get them to actually do real world work on that interview to see how they work. So you can do things like, you set up a model and you break it and you see how they debug it and understand the GPU profiling, and all these different things, right. So that's something cool that people started doing, not just us, a few other customers. Hackathons is a big one as well. So turns out, you know, if you want to set up a hackathon, you can go set up a bunch of studios for everyone and then they go off, and you give them credits, and it's good. Classes, so right now there's, I won't say which one but you can probably guess, big professor from NYU, teaching a big deep learning class, using Studio in that class. And, you know, it's obviously from our lab. And so it's cool to see how they're doing it, and then they're a bunch of other universities as well, that are doing it. But imagine, as an undergrad, Columbia, when I was there, they didn’t really have a compute cluster. I don't know if they still do, they might, a little one, and you're learning computer science and you have to do everything locally. It's like, super slow and hard right? Now, professors can just be like, here's a studio with your homework on it, like, go solve it there, and they could just immediately get started. So I think it'll accelerate education a lot as well.

**William**

I'll take a minute just to say, about open source, like we've been behind open source forever. So most people today use a lot of libraries out there. And they use things like a trainer, or they use different interfaces that they code with. And, you know, a lot of those ideas came from PyTorch Lightning in 2019. And we introduced this to the world. And I think like, what's that's done, is really standardize the way people do AI. And I think it's great because like, you know, it's not just us, a lot of other companies as well have been pushing really hard and getting open source to be to be big and I really do want to ask everyone who's listening to support open source not just us, but every company who's doing it because I think we're at a critical junction where the last thing you want is to have one or two companies like own key IP of models or something like that. It's like not having an Apple and then having IBM be the only one you can get computers from, like, that would be crazy right? And so I think the world has to come together and keep AI open source and continue to support things. And I think Meta and my old lab at Facebook, they're probably doing the most out of anyone, which is great; but I think companies should not be scared of that and they should understand that it's actually better for their business. Go look at Meta stock today, like, probably Trip Forex since they started doing AI and working on open source as well. It's a net benefit for the world.

**CRAIG**

AI might be the most important new computer technology ever. It's storming every industry and literally billions of dollars are being invested. So buckle up. The problem is that AI needs a lot of speed and processing power. So how do you compete without cost spiraling out of control, it's time to upgrade to the next generation of the cloud. Oracle Cloud Infrastructure, or OCI. OCI is a single platform for your infrastructure, database, application development and AI needs. OCI has 4 to 8 times the bandwidth of other clouds, offers one consistent price instead of variable regional pricing. And of course, nobody does data better than Oracle. So now you can train your AI models at twice the speed and less than half the cost of other clouds. If you want to do more and spend less like Uber, eight by eight and Databricks Mosaic, take a free test drive of OCI at Oracle.com slash Eye on AI. That's E Y E O N A I all run together at home. At work. We all know one person whose password challenged sticky note reminders, emailing passwords, reusing passwords, using the word password as their password.

**CRAIG**
Because data breaches affect everyone you need One password. 1password combines industry leading security with award winning design to bring private, secure and user friendly password management to everyone. Companies lose hours every day just from employees forgetting and resetting passwords. A single data breach costs millions of dollars. One password secures every sign in to save you time and money. 1password lets you switch between iPhone, Android, Mac, and PC with convenient features like autofill. For quick signups, all you have to remember is the one strong account password that protects everything else your logins, your credit cards, secure notes, or the office Wi-Fi password. 1password generates as many strong, unique passwords as you need and securely stores them in an encrypted vault that only you have access to. I use 1password and you should too. 1password is award winning. Password manager is trusted by millions of users and over 100,000 businesses. From IBM to slack, it beat out 40 other options to become Wirecutter's top pick for password managers, plus regular for third party audits and the industry's largest bug bounty. Keep 1password at the forefront of security. Right now, my listeners get a two week free trial at 1 password.com/eyeonai that's Eye on AI all run together E Y E O N A I. That's two weeks free at 1password.com/ eyeonai 1password com slash Eye on AI

**CRAIG**

That's it for this episode. I want to thank Wil for his time. If you want to read a transcript of today's conversation you can find one, as always on our website EYE hyphen on A.I. In the meantime, remember the singularity may not be near, but AI is changing your world, so pay attention.