SDS PODCAST EPISODE 541: DATA OBSERVABILITY — WITH DR. KEVIN HU

Jon Krohn:	00:00:00	This is episode number 541 with Kevin Hu, CEO of Metaplane.
Jon Krohn:	00:00:11	Welcome to the SuperDataScience podcast, the most listened to podcast in the data science industry. Each week we bring you inspiring people and ideas to help you build a successful career in data science. I'm your host, Jon Krohn. Thanks for joining me today. And now, let's make the complex simple.
Jon Krohn:	00:00:41	Welcome back to the SuperDataScience podcast. Today's guest is the fun, well-spoken, and wildly intelligent entrepreneur, Kevin Hu. Kevin is co-founder and CEO of Metaplane, a YCombinator-backed startup and platform that observes the quality of data flows, looks for abnormalities in the data, and then reports the issues to the right people. Prior to founding Metaplane, Kevin completed a PhD in machine learning and data science from MIT, one of the world's most illustrious technical universities. In this episode, Kevin details what data observability is and how it can help any platform or company with automated data flows to identify data quality issues immediately, as well as more quickly resolve the source of the issue. He talks about his PhD research on automating data science systems using machine learning. He talks about how he identified the problem his startup, Metaplane would solve, the driving force behind his decision to launch a data science startup right out of his PhD, and his experience in YCombinator accelerating the startup.
Jon Krohn:	00:01:49	He talks about the pros and cons of an academic career relative to the startup hustle. He talks about the surprising complexity of the software tools he uses daily as a CEO, and he fills us in on what he looks for in the data engineers that he hires. Today's episode does get a little technical here and there, but I think Kevin and I were pretty careful to define technical concepts when they

came up. So today's episode should largely be appealing to anyone who's keen to learn a lot from a brilliant entrepreneur, especially if you'd like to found or grow a data science startup yourself. All right, you ready for this fun and absorbing episode? Let's go.

Jon Krohn: 00:02:35 Kevin, welcome to the SuperDataScience Podcast. It's awesome to have you here. Where in the world are you calling in from?

Kevin Hu: 00:02:42 It's awesome to be here. I'm a long time fan of the show and you're an incredible host. Usually I'm in Boston, Massachusetts. But today I'm actually in the south of France, near Marseille.

Jon Krohn: 00:02:54 Oh, wow.

Kevin Hu: 00:02:55 First time here.

Jon Krohn: 00:02:57 Wow, so you're right on the coast?

Kevin Hu: 00:03:00 Close to the coast, a little bit inland, but I'm sure we'll do some traveling around.

Jon Krohn: 00:03:06 Cool. And I guess, so we're recording in December and it'll still be winter when this episode comes out. It's warm down in Southern France at this time of year. Is it still, or is it a little bit chillier?

Kevin Hu: 00:03:20 It is definitely warmer than Boston when I left, which isn't saying that much.

Jon Krohn: 00:03:23 No.

Kevin Hu: 00:03:26 Let's just say that when I went on a hike earlier today, I had to take off my jacket, which is a good benchmark.

Jon Krohn: 00:03:31 You got lots of fun, outdoors things to do there?

Kevin Hu: 00:03:35 We were just walking around the town and suddenly we come across ancient Roman ruins right next to a cave. You keep walking, you see olive trees and it's like, "What is this place?" It really is as beautiful as people say.

Jon Krohn: 00:03:50 Yeah, it's a beautiful part of the world. All right. So Kevin, we could talk about Southern France all day, but I would love to talk about your startup Metaplane. So you co-founded it, and you're the CEO. You founded it in 2019 and it's backed by some really great organizations. So you're a YCombinator company, maybe later on if we have time, we'll talk about that YCombinator process. You're also backed by Flybridge, and the founders of a bunch of really big data startups like HubSpot and Clearbit. So clearly in just the couple of years since you founded, you're doing a lot of things right. The area that you're in is data observability. So you're building tools for anybody who's had data quality issues. So you're looking to automatically monitor data. It could be data in data warehouses. It could be data that's flowing into business intelligence dashboards. I guess you're just monitoring data flows, and then you're alerting people when things go wrong, is that right?

Kevin Hu: 00:05:04 That was an excellent pitch. I cannot do it better than that, but I can maybe tell you a bit of the story behind why this problem matters. The issue, well, imagine you're a software engineer. 10 years ago if you were building an app, you might put your rails up on an EC2Box somewhere, put a heartbeat check to make sure that it's up, call it a day. But today software engineers and DevOps engineers have amazing tools to understand the state of their infrastructure. To know how their database are performing, how their API endpoint are performing, how their servers are functioning. Incredible amounts of visibility, and we call this observability. Because software engineers can understand at any point in time the state of

the infrastructure. Unfortunately, data is about 10 years behind software, where if you go to a data team, like data teams that I've worked on and you ask honestly, "Is your data infrastructure up? Is the data what you expect?" Now frequently we aren't as confident as we would like. And ironically data teams don't have the metadata that they need to answer their own questions much of the time. So data observability tools, we strive to provide data teams as much visibility into the state of their data as software engineering teams have into the state of their software infrastructure.

Jon Krohn: 00:06:41 Interesting. So I hadn't clued in on that myself, but you just mentioned metadata there and meta is in your company named Metaplane. So I suspect there's a relationship?

Kevin Hu: 00:06:48 The relationship is in networking you have a distinction between the data plane and the control plane, and we want to be the metadata plane.

Jon Krohn: 00:06:49 Ah, cool. I thought it might be something kind of like a metaverse.

Kevin Hu: 00:07:08 Unfortunately not.

Jon Krohn: 00:07:12 When I first saw your company name, like that you were using meta in that kind of sense. But now I see it's associated with metadata, which meta is being used in that kind of metaverse sense of being something adjacent to. So I think that that's the Greek or Latin basis of that kind of meta. So data is, when you have an audio file, the information that allows you to listen to the track is the data. And then the metadata is, when was this track recorded and who recorded it? And so you're appending on lots of other metadata that allows us to have some confidence about that data?

Kevin Hu: 00:08:00 Exactly. And unfortunately we couldn't get meta.com. I don't think anyone can get it at this point, but we were there with the name before Facebook was.

Jon Krohn: 00:08:09 Yeah, yeah, yeah.

Kevin Hu: 00:08:12 That's a great description of metadata where imagine if you're a data team at an e-commerce company, all of the other teams, the support teams, the marketing teams, they have data about their customers, their orders. And when they signed up, the data team wants to have metadata about, okay, this is how many tables exist. This is the dashboards that are being refreshed and the data that's being used by your end users. And that sort of metadata, believe it or not, is actually quite hard to come by. It's getting easier with tools like DBKEY and Snowflake, but before the recent generation of data tools, it's quite difficult.

Jon Krohn: 00:08:55 Cool. So I understand that there are some specific kinds of things that you look for in data. It sounds like there's four of them specifically. Maybe you can run us through those to help us kind of understand when data is likely to be high quality or when there might be an issue. Because I imagine most of the time it's probably some large percentage, 99 or 99.99% of data are actually fine. So you're looking for factors. And so you have these kind of four particular kinds of things that you're looking for in a data stream to say, "Hey, that one is a standout, there might be an aberration here and we should let somebody know."

Kevin Hu: 00:09:37 Totally. Going back to the idea of observability as the amount of visibility having to your data. We kind of flip it and say, "What's the minimum amount of information about my data that I need to reconstruct it, as much fidelity as I can?" And of course, ideally you just have all the data throughout all time stored in snapshots, but

most of the time that isn't feasible. So the way we like to think of it is to boil your data systems down into your four buckets. One of them is metrics, describing the internal characteristics like statistical distributions of your data. One of them is metadata, which describes more structural characteristics of your data. Like what is the schema, the number of rows, the freshness. Then we have lineage where you describe how one piece of data depends on another piece of data. For example, if A, B and C were used as inputs to a derivation for F. And then logs, which describes interactions between systems, whether it's like a transformation pipeline or end users of the data with the data itself. So with the catchy acronym, MMLL we try and describe, okay, internal and external characteristics and internal and external interactions with the data.

Jon Krohn: 00:11:11 Nice. Somehow I missed one of the M's as we went through, I got metrics, lineage and logs, and those all made sense to me. What was the other M? Oh, metadata itself?

Kevin Hu: 00:11:19 Metadata itself.

Jon Krohn: 00:11:20 Yeah, yeah. Okay, got it, got it, got it. Cool, all right. So now I have an understanding of what data observability is and specifically Metaplane's approach, do you have one or two examples that you can give us, use cases of where clients have used Metaplane and that's improved their productivity or their client's experience or something like that?

Kevin Hu: 00:11:43 Definitely. One of the beauties of working in data is the tools that you build are applicable across industries. A table is a table is a table. So some of our customers range in size from 10-person companies to 10,000 person companies and across verticals like healthcare, FinTech, e-commerce, B2B, SaaS. And to give you one example,

one of our customers, an e-commerce company around the 1,000 people. Due to the AWS outages last week had a major freshness issue where none of their tables and none of their reports were being updated. And Metaplane was actually the first to notify them, "Hey, this data is not fresh and is getting more and more stale by the hour." So we have some anomaly detection on our end that tries to take trends and seasonality into account. And to be almost like a first alert system for this team. That was like a all hands on deck code red issue. Another example is one of our B2B SaaS customers had an upstream engineering team change some of the instrumentation on their events and made it such that a lot of the product usage activity that they were sending to their sales and marketing systems were way skewed. This is very insidious because you will not be able to catch this, unless you're staring at a table all day, which I hope no one is, let Metaplane do that for you. But [crosstalk 00:13:29]-

Jon Krohn: 00:13:28 It's a kind of thing. In a lot of situations, until companies like yours have come around recently that allow us to have these systems flagging for us when there's an issue, that would be the only way. We kind of just trust that everything's in place and you don't find out until something goes really wrong downstream.

Kevin Hu: 00:13:50 And this was okay in the old world of data when data was used primarily for reporting for executives. But nowadays data is used everywhere throughout organizations, whether it's AIML and data science, to powering product experience, to even sending the data back to the customer. The stakes are so much higher. We make so many more data assets, whether they're the tables or downstream dashboards that no one can audit everything. So with a ballooning number of assets, a ballooning risk, let machines do it for you, come on.

Jon Krohn: 00:14:34 Cool. So I understand how you're observing data and then looking for anomalous events. That makes sense to me. And you have your four ways that you're looking for anomalous events. That makes sense to me. What do you do when you actually identify the issue? Is it like an automated phone call or an email, or do people kind of have their choice?

Kevin Hu: 00:14:53 The first thing we do is try to give our users as much information as possible as they need to prioritize. Not everything will be a code red issue or a P0.

Jon Krohn: 00:15:08 So it's like priorities, right. Exactly, like P0, P1. Got it, got it, got it.

Kevin Hu: 00:15:12 Exactly. So we try and have the downstream lineage to BI tools and other downstream tools to tell you, "Okay, this table is not fresh. And it's being used by these five dashboards that are used by executives every day," to give you a sense of the impact. And these are the upstream tables and sources to give you a sense of how you could possibly address the root cause. So just try and get in front of a human and give them as much information as possible. We're looking into doing some more automated remediation, but that's a pretty hairy problem.

Jon Krohn: 00:15:51 Yeah, that sounds like that would be. I mean, even if you can crack it in a small number of frequently occurring situations, which I'm sure is where you're looking, that would be a big win. Super cool. All right, so I love what you're doing, Kevin with your company, Metaplane sounds like it can make a big impact on people's data businesses. I'd love to get a bit into your background before this. So you founded this company in 2019 and we're going to get to your transition into founding it. But before we get to that, I want to discuss what you were doing first, which is a long string of evidently very successful academia, all at MIT, at the Massachusetts

Institute of Technology. So you did a bachelor's degree there in physics, and then you liked it so much in Boston that you stayed around for a masters in data visualization and machine learning. And that wasn't enough. You hadn't gotten your fill of Red Sox games yet, and you stayed on for a PhD in machine learning and your PhD research sounds super interesting. You presented it at top conferences like CHI, the Computer Human Interface conference, which I think you pronounced in a much cooler way.

Kevin Hu: 00:17:15 I like the Greek letter.

Jon Krohn: 00:17:17 Right, right, right. And of course, KDD, which is a big data discovery conference, a data mining conference. And it isn't just in academia that you made some ripples, because I understand that your research was featured in popular and very well respected publications, like The Economist, the New York Times and Wired. So Kevin, first of all, why did you stay at MIT so long? And then, I mean, actually, I can kind of answer your question for you because it is one of the most venerable institutions in the world for technical work. So I suppose if I had the opportunity to continue staying at MIT, I would too. But yeah, so let us know a bit about that and what your experience was like, and then specifically what your PhD research was all about?

Kevin Hu: 00:18:04 That makes one of us, because I didn't plan on staying in academia, I didn't plan on going to grad school at all. For me it was all about having the best teacher I could. And I consider myself so lucky to have had the privilege of learning from some of the best at MIT. One of them being my PhD advisor, Cesar Hidalgo, who is originally my undergrad research advisor. And he is really an incredible man who invested in his students in trying to cultivate them, not only as technical researchers, but researchers who can identify interesting problems and tell the story

that's needed to really make their message resonate with other people. And I started, like you mentioned in physics, and this is actually how I got into the data world, which going back in time a little bit, one of the most intimidating things about studying physics at MIT is that there are some very, very brilliant students.

Jon Krohn: 00:19:20 I bet.

Kevin Hu: 00:19:21 Like physics, Math Olympiad, gold medalists. On the log scale they are way down there. But everyone has to take an experimental lab course called J lab. And this lab is known as the gauntlet course where it takes 30, 40 hours a week. And in that course you have to replicate a Nobel Prize winning experiment every two weeks.

Jon Krohn: 00:19:48 Wow.

Kevin Hu: 00:19:48 One week you do the experiment. Very cool. Whether it's like counting [inaudible 00:19:53], or trying to approximate the speed of light. It's a very fascinating course where one week experiment, one week analyzing the data. And I realized at the time, everyone does experiments at the same pace. The people who had the hardest time in the course, including some of those brilliant classmates that I had, who had the hardest time where the people weren't able to analyze the data. They didn't know MATLAB, they didn't know statistical analysis. They couldn't present the results in a meaningful way. They ended up having to burn the all nighters. So with that happening and my sister getting her PhD in neuroscience at the time, she like-

Jon Krohn: 00:20:40 Oh, no kidding.

Kevin Hu: 00:20:42 Yeah. She studied fish, cichlids at Stanford.

Jon Krohn: 00:20:45 No kidding. That's cool.

Kevin Hu: 00:20:47 Yeah, she's awesome. She spent five years collecting data. And at the very last year she was like, "Kevin, can you help me analyze some of this?" And I was like, "How insane is it that some of the brightest minds in our world are bottleneck because they can't analyze data when they can otherwise articulate the problem?"

Jon Krohn: 00:21:07 Yeah. You know what, Kevin, I had a really similar experience. So doing my PhD in Oxford, I decided right at the beginning to specialize in machine learning. So I also, like your sister I did a neuroscience PhD, and lots of my colleagues did very specific lab-based things like putting a recording electrode into a ferret's brain and monitoring its activity as it runs around or growing some kind of cell in a Petri dish. And I was like, "Well, those aren't super transferrable skills." If I become an expert at machine learning, at managing large amounts of data, identifying patterns in data, I was like, "Even if I don't stay in academia and decide that I want to be growing cells in a Petri dish forever, this is going to be a highly transferable skill." And exactly like what you're describing, I was surrounded by people whom I looked up to so much. These amazing researchers who would come by my desk and be like, "Look at this plot. I've got more, this one is higher than the other one," I can see that. But I don't know what statistical test I can use to prove of that. And so I got a lot of papers out of my PhD and a bunch of them I did like a couple of days' work on the back of years of somebody's research. Just like you're describing with your sister where they spent years designing an experiment, collecting the data. And then at the end they're like, "Ah, I think there's this interaction between these two things. And I don't know how to model that." And I'm like, "I can do that."

Kevin Hu: 00:22:54 It sounds like you made the right call, both with machine learning as a subject and the transferable technical skills.

I'm wondering, why do you think this skillset of being able to reason statistically and code is so rare amongst people who obviously can do it, can obviously learn how?

Jon Krohn: 00:23:18 Yeah. I think it's just, it's random, people, you get exposed to different things at different ages. And then, maybe when your sister sees at the end of a neuroscience PhD that her little brother, I guess, can come along and solve these problems and do it in such a short amount of time. She's like, "Oh, that's cool. I should spend a bit of time learning about that." And so maybe like you, I was extremely fortunate that even in high school I had amazing computer science teachers and math teachers that I just loved what was happening in those classes. And I found myself kind of tinkering around with writing code on my own time outside of projects. And so just always kind of had this, it always just, it seemed natural to me and it seemed fun to me. And so throughout my bachelor's degree, my master's degree, I just, even though those were, my bachelor's and my master's were in science disciplines and my PhD, the whole time I was self-teaching, like you mentioned MATLAB, how could I be doing this? How could I be doing some analysis here with MATLAB as opposed to in an Excel spreadsheet or with SPSS, which seemed clunky.

Kevin Hu: 00:24:33 Yeah, I definitely see early exposure being a huge factor, not only in programming, but also in math, for example. Different countries, different cultures having different proclivities towards math and part of it being just getting the self-confidence sometime.

Jon Krohn: 00:24:53 Exactly.

Kevin Hu: 00:24:53 If I can do this as a kid, I can do it as an adult, but I've never done it as a kid. I'm going to do this for the first time when I'm 30, it's tough.

Jon Krohn: 00:25:03 Yeah, something that I've said on the show before, but probably haven't said enough is that I know there are listeners out there who are listening to this podcast because they're thinking about taking their first steps in data science. And parts of it, like maybe it's the math, or maybe it's writing code seem intimidating because you haven't done much of it before or any of it before. And maybe you don't know other people who you can see and kind of just learn off of and ask questions. And so Kevin and I are here to let you know that math stuff, coding stuff, it's just like learning anything else.

Kevin Hu: 00:25:45 Totally. And if you're listening to Jon and being curious and intentional with your learning, you've already done the hardest part.

Jon Krohn: 00:25:54 Yeah. And then it's, we're at an amazing time. Things just keep getting easier and easier and easier to get started with learning these kinds of things. And maybe something I don't mention enough is that I maintain a big list of resources at jonkrohn.com/resources for getting started in data science. So lots of free websites, books, all kinds of things just to get started in general with data science or with specific topics like machine learning or the deep learning branch of machine learning. So yeah, if you're looking to get going, just find something that looks kind of fun and there's lots of easy ways to get into this field.

Kevin Hu: 00:26:38 This is actually what I've been trying to work on for most of my research career. My grad school research was entirely around trying to make data analysis easier. And at the time when we started, to give you a very scope down example, if you're given a CSV, a spreadsheet, how can you recommend interesting "analysis or visualizations" so that data analysis turns into a search and selection problem more than a specification problem? Because it's the specification which is a huge barrier. And by that I mean, type down in code what you want to do,

as opposed to trying to reason either graphically or with natural language about what you want to do, and then letting the machine kind of figure out the codification of that for you. And first we started with the rule-based systems, which only took us far. Then we, me and my colleagues, were some of the first to apply deep neural nets to large data sets, trying to recommend data visualizations. That project was called [inaudible 00:27:57], to produce a type detection model. So if you get a column of latitudes and longitudes that we don't just say that it's afloat, but we say, given the distribution, we think that this is a latitude or a zip code or a name, that sort of thing.

Jon Krohn: 00:28:17 Wow. That's cool. Wait, this is your PhD research?

Kevin Hu: 00:28:19 That was my PhD research.

Jon Krohn: 00:28:22 Wow. That's super cool. And I can see why I go out picked up by popular outlets, like The Economist, because that's pretty fascinating.

Kevin Hu: 00:28:30 Yeah, we were super lucky to work on these topics and for my advisor to give me really the latitude and support to work on these topics as well.

Jon Krohn: 00:28:43 Amazing. And so, all right, so now I understand a bit more about your PhD research. So how did you go from that concept, this kind of automation or streamlining suggestions for data analysis or data visualization typing, how did you go from that to then the Metaplane idea, and then talk about having the courage to do things. We're talking about having the courage just to get into data science or to learn math or to learn some code. But you made the plunge from academia, from doing a PhD right into founding your own startup. So I realize I'm kind of asking two main questions there. And so if we can put a pin in one or the other, but how did your focus switch

from the PhD research you were doing to the data observability focus that Metaplane has today? And then the follow-up question about how making a startup about it, did you have in the same kind of way that you and I were describing with, at an early age having mentors to get us into math and programming, did you have the same kind of thing in startups?

Kevin Hu: 00:30:01 There's a lot of talk in the software world about 10X engineers. Engineers who are just prodigiously more productive. And I don't know if that's entirely true, but someone told me once that the easiest way to be a 10X engineer is to help 10 people be two times better. And I've always viewed myself as someone who, I'm definitely not the sharpest tool on the shed, but I can help you sharpen your tool. And if we can build tools that can help people like your colleagues, my sister, or really anyone in the world that works with data, which is many, many people nowadays do their jobs better, then that's an impact that I would be happy to have. And that's how I felt throughout my entire PhD where a lot of it was impact-driven. How can I build tools that help more and more practitioners out there? And in trying to commercialize the technology that we built, we did try and build a tool around automated data analysis. It had legs, but I think it was a little bit too early, but we very quickly discovered that the real barrier to working with data within most companies is not the data analysis piece. It is still there for sure, but it's actually making sure that the data is in a good shape.

Jon Krohn: 00:31:29 Yeah, that's cool. And that makes a lot of sense to me and I can totally see how that would happen. And this is one of those great examples of how much value there is just in starting on something. So I think this happens a lot out with startups and we hear this idea of startups pivoting after they've already been founded, where you go

to tackle one problem, you think you've got this problem that you're going to be able to get clients for. And then you discover pretty early that what your clients really need is something else. And so this is an example of just, if you want to have your own startup or be creating a product that can be getting consumers, then just start something. You don't need to spend forever ideating because opportunities will present themselves kind of like it has for you.

Kevin Hu: 00:32:19 That is one of the huge differences between a PhD and startup life for me. I'm very curious how it's been for you, and also why YC has been helpful. Like of course there's a lot of overlap between the two, such as dealing with uncertainty over a grueling period of time with an uncertain outcome is something that isn't common between research and startups. But besides that, there's a lot of differences. For example, not having an advisor in the startup world, like one person that can guide you along the whole way. Startups are also highly collaborative, whereas research I would say for many, many people is actually quite solitary, right? But in startups you have peers, co-founders, customers, you're talking with people a lot. And interestingly, as much as research is of course motivated by the truth, much of the time, the work that researchers work on is driven a lot by peer approval. That's [crosstalk 00:33:30]-

Jon Krohn: 00:33:31 That's actually, that's something, going into a PhD I was really excited about uncovering the truth. And one of the things for me personally why I didn't end up staying in academia was that I realized that people's motivations were actually quite different. It is about, how many citations can you get on papers? What conference can you get accepted at? And that makes sense. I mean, because that allows you to get bigger grants. It allows you to have bigger impact. But yeah, it's amazing how much, I think

from the outside for people, for listeners out there who haven't done a PhD, who haven't gotten deep into academia, it's very easy to see it as kind of like this kind of pure investigation, but really it's a lot of chasing, oh, like what's getting a lot of citations. What's an angle that I can spin on that to also get a lot of citations, which then kind of indirectly ends up springing up some truth along the way?

Kevin Hu: 00:34:35 Yet another case where frequently the proxy you use to measure what really matters, ends up becoming the goal.

Jon Krohn: 00:34:46 Right, right, right, right, exactly.

Kevin Hu: 00:34:46 Right.

Jon Krohn: 00:34:48 That's right.

Kevin Hu: 00:34:48 Yeah. That's a whole another conversation, another hour for us to talk about the impact of prestige seeking mechanism status in academia. Ironically startups as much as there is a capitalist motive, there is also a source of truth, which is the customer need.

Jon Krohn: 00:35:09 Totally.

Kevin Hu: 00:35:10 It doesn't matter how charismatic you are, how beautiful the story is or the technology, it doesn't solve a customer problem, it is not useful. And that's where, like you said, pivots are almost inevitable.

Jon Krohn: 00:35:27 And to kind of, to circle back on something. It is interesting now that you mentioned it, I haven't thought about it this way before, but the real-time economy and being able to get sales is actually much more the truth, the reality, than a lot of the truth that you're uncovering with science. And I hadn't noticed that before, but with science you can, those kinds of things that you

mentioned, like charisma, presentation style, writing capability can end up being a much bigger driver of success than the actual truth uncovering. And you can't get away with that in commerce, anyway.

Kevin Hu: 00:36:09 One good proof point to, back at the argument is, if you're a scientist and you want to know about prize, that is an amazing predictor of future citations. Regardless of how high quality those papers are, but in the startup world, if you exited one company, you sold it or it went public, it's actually not that great of a predictor of the success of your second company.

Jon Krohn: 00:36:33 Interesting.

Kevin Hu: 00:36:34 So there's almost this independence that goes on that at least removes the fact of the shine, let's just call it that.

Jon Krohn: 00:36:46 Right, right, right. So here's another example related to Nobel Prizes that I learned about recently. So the Nobel Prize committee, at some interval they publish historical decisions. So it's nothing recent. But in the last year, for example, they published a bunch of decisions that they made in the mid-20th century and earlier. And so the kinds of people that were on the Nobel Prize committee were people like Albert Einstein, who were these extremely well known, themselves already Nobel Prize winners. And The Economist did a really interesting graphic in a short article. They have this graphic detail that's always the last page. Well, the last page is always an obituary. But the page before that is what they call graphic detail and it's really interesting graphic. And so they did a visual of Nobel Prize winners and how they might have been nominated for a decade, but they didn't win the Nobel prize until someone like Albert Einstein, a really prominent person came along and supported the nomination. And then all of a sudden everybody changed their mind and everyone was like, "Yeah, yeah, yeah,

yeah. That person is really good. We should give an award." That's a perfect example of how it isn't to do with the truth. But it's like, it's kind of just Albert Einstein's opinion really that's driving that's getting the Noble Prizes.

Kevin Hu: 00:38:11 Yeah. The brightest, most rational minds in the world. Subject to such herd behavior. It's really, I got to look at that graphic.

Jon Krohn: 00:38:24 Yeah, it's good. I'll try to find it and put it in the show notes for everyone. That's going to be a fun one to try to figure out.

Kevin Hu: 00:38:32 In fairness to them, if Albert Einstein told me to do something, I would, if he told me to jump, I would ask how high.

Jon Krohn: 00:38:41 I've got these anti-gravity these anti, these relativity bending boots that I'd like you to put on, Kevin. It's totally going to work.

Kevin Hu: 00:38:51 I'll do that.

Jon Krohn: 00:38:52 "Why don't you do yourself, Albert?" And, "I'm too old."

Kevin Hu: 00:38:56 It's just shoot me up, beam me up.

Jon Krohn: 00:39:01 Nice. All right. How did we get here? All right, so we were talking about differences between academia and startups. You were talking about 10X-ing, by being a 10X engineer by two X-ing 10 people. Yeah, and we'd kind of identified the flip from what you're researching in your PhD and how that uncovered that there was this data observability problem. Okay. But then let's go from there. So you've identified, there's this data observability issue, you hypothesize that people would be willing to pay for solutions. Then what, what did you do next?

Kevin Hu: 00:39:41 Then we are in the stage that we are in now where we try to grow like crazy.

Jon Krohn: 00:39:47 But I mean, so what about the ... I mean, you had YCombinator in there, you had, I mean, you [crosstalk 00:39:56] proof of concept or something. How did that unfold? You didn't just jump to now being funded by these amazing people and having amazing clients. There's a little bit in between surely.

Kevin Hu: 00:40:07 YCombinator was an amazing experience. I did skip a few steps there. One of them was getting the support of amazing angel investors, who I like in, if you played Warcraft or RoomScape, when you leave the tutorial you're wearing cloth clothing, you're level one. And then you encounter people who are level 99 and they drop an item that is useless to them, but is the most valuable thing in the world to you. And that's how I view the advice from some of our angels where they can dispense that all day. But it's so hard won. And if you're open minded about it and it receives, it comes to you at the right time, it could really change the trajectory of your company.

Kevin Hu: 00:40:54 And YCombinator was an example where you can get that kind of advice really distilled in one point in time. And they injected some urgency behind your companies every week for three weeks. They ask, "What is your KPI? How are you tracking against that? And what's your goal next week?" And one of the best parts I would say is related to your comment before about pivoting is hearing from the founders of successful startups like Airbnb and Segment, for example, about how they build their companies. And the answer is that there's no one right way. The Airbnb founders, the Coinbase founders, they had a vision and they executed on it, straight through to the end from the garage days to IPO. Other founders like the Segment or Amplitude founders, when they went through YCombinator, they were building completely different

types of companies. I think they may have both been ed tech companies and now they're highly successful public companies. Amplitude's public segment was bought for a very large sum.

Kevin Hu: 00:42:09 So I think in both cases it's knowing the founders had an amazing taste for when they were onto something. And then when they should put the pedal to the metal. For us, it was similar. We did pivot around quite a bit and it was painful for sure. It was very painful. We were lucky to have that support of amazing investors and colleagues. So that finally, when we're solving a problem, which we feel has some meat to it, data observability, we know that the product we have solves a real customer problem and we know how to reach those customers. And the goal is, how do we bring data observability to 10,000 companies as fast as possible?

Jon Krohn: 00:42:55 Cool, that is amazing. So yeah, it sounds like YCombinator was a hugely valuable experience. So did you apply while you were still in your PhD or was that shortly after?

Kevin Hu: 00:43:09 I've applied four times.

Jon Krohn: 00:43:12 No kidding.

Kevin Hu: 00:43:14 At first I thought that was a lot, but I've talked the people who have applied [inaudible 00:43:18] to 10 times.

Jon Krohn: 00:43:20 Right, right, right.

Kevin Hu: 00:43:21 So it was, shortly after the PhD. And finally, I think we were on something and we're committed full time to it and we were able to get in.

Jon Krohn: 00:43:32 Super cool. And so that was in the winter of 2020. So I guess everything was remote?

Kevin Hu: 00:43:40 We were the notorious COVID batch where COVID was started. Everyone was in person at the time. And all previous batches you fly to the Bay Area and meet every week in Mountain View. And now nowadays it's fully remote and COVID started getting very, very serious in the world before then, but in California, at least around March. So everything was accelerated, whether it's the famous demo day where you done to investors, the wrap up of the program, it was sped up at the end.

Jon Krohn: 00:44:15 So now it's still remote? All YCombinator classes are still remote?

Kevin Hu: 00:44:20 I think so.

Jon Krohn: 00:44:21 Huh. Interesting. Yeah, I wonder if they'll ever get back to it. It's so interesting. It's just like the transition to remote working. There's obvious cons, but there's also really great pros. It's nice that you could be in Southern France and in YCombinator, but then the con is the, you're not going to bump into people in the hallway and have that conversation that could be really valuable, that first client or some new pivot idea that comes about organically. So there's pros and cons to being remote.

Kevin Hu: 00:44:54 I think with, many of us coming into remote work were pessimistic, I was. And I think it's another case of technological change where we, the downsides are very concrete, but the upsides are harder to imagine.

Jon Krohn: 00:45:10 And also to some extent harder to quantify, I'm sure there are industries, even software to some extent when it's like when you're a really big software company and everything's defined as tickets, I guess you could monitor. And in fact, definitely you could, I know you could, because I've seen COVID studies about this of people's productivity before COVID when everyone was in the office and then during COVID when everybody was

remote. In a remote condition, a lot of different companies can measure this increase in improvement. But I think for managers like me where we're doing data science R&D and some problems are tricky and they're legitimately tricky, but that also means that somebody could be just kind of telling you that it's really tricky, and in practice, they're just not trying that hard at it. And it's impossible for me to know what the situation is. So there's a lot of managers in that kind of scenario in the world. And I think that this idea of remote working was scary because we were like, "Well, how am I going to keep an eye on how much people are working, and really know?" And at least what we've seen with our company is, people are delivering like never before. And it shows in how much we're able to produce in a given sprint, for example. It's, so it does work and I'm sure absolutely YCombinator works remotely as well. So yeah, yeah, yeah. Go ahead.

Kevin Hu: 00:46:38 You know, we're both at high growth, early age stage startups. I think something that I didn't expect to be a boon of remote work is the ability to hop on a call with anyone anytime. Before to do customer discovery, take someone out for coffee, meet them at their office, suddenly half your day is gone.

Jon Krohn: 00:46:59 Exactly.

Kevin Hu: 00:47:00 But now, if you want to do early sales, you can do 16 calls in a day. I do not recommend it, but you could do 16 calls in a day. And that can really accelerate things.

Jon Krohn: 00:47:09 Yep, I've heard that from a number of other founders who have been able to do their series arrays by Zoom and exactly what you just said. Being able to get so many more meetings in a day, and even actually friends of mine who are at big banks like Goldman Sachs, who would be going out and they'd have to be flying around, in their case it was their job pre-pandemic to everyday of the

week be flying to wherever the investors are. And okay yeah, they're based in New York, a bunch of investors in New York. So maybe you get to spend two days a week in New York meeting with them, but then the rest of the time you're flying off to Miami and you're flying off to the West Coast. And in those kinds of situations, in those kinds of jobs where you're seeing how much support there is for a new, a stock that's being issued, you might be doing two flights a day where you just, you fly into the airport in Seattle, have a meeting, and then you fly to San Francisco for another meeting in the same day. And they don't have to do that anymore.

Kevin Hu: 00:48:20 Which doesn't even touch on the environmental impact of taking those meetings. But just like, it's 2021. Even if it wasn't for COVID, it isn't the rational thing to do.

Jon Krohn: 00:48:36 Yeah. Yeah.

Kevin Hu: 00:48:37 I got that there is a face-to-face element in those high stakes, wear your best suit, firm handshake meetings, but for everything else, like our conversation right now, would it be done better in person, I'm not sure, it would be fun.

Jon Krohn: 00:48:52 Yeah.

Kevin Hu: 00:48:53 But this is also fun.

Jon Krohn: 00:48:55 Yeah, yeah. And I was reading an interesting article yesterday that opened my mind to this concept, because I think I had this idea that things would go back to normal, meaning pre-pandemic, to the kinds of habits that we had, the kinds of lives that we lived then. And this article opened my eyes to the reality that it's never going to be like it was before. That how we travel, how we work, is going to be changed forever and trends that were happening slowly, like remote working or shopping

online, the pandemic accelerated those things. But now that I'm constantly ordering my groceries online, which I didn't do before, I'm not going to go back to it. It's so much more convenient.

Kevin Hu: 00:49:50 Totally. It's an irreversible process, like you said. And ordering groceries is such a good example where not forever, but at least for the past couple of years it was always an option. It's like, why didn't we do it? And I think it's one of those cases where they say about startups is true, where you need to have a solution that's 10 times better than the existing status quo. The problem is that 10 times is actually very hard to get to. But you only need a slight improvement to make it so that it's painful to go back. Ordering groceries online, I wouldn't say it's 10 times better than going in person, maybe it's like two times better, but once you've tasted it two times, I'm not going back.

Jon Krohn: 00:50:37 Yeah, exactly. Amazing discussion. I'm really enjoying this. So going back to your remote work now. I know from the bio that you sent me, that you spend a lot of time in a SQL editor. So specifically you wrote to me that outside of the SQL editor, you enjoy reading all kinds of fiction, biking around Boston and cooking. And that really cool caught me that you said that you spend a lot of time in a SQL editor. I was thinking to myself, "I wonder how many CEOs spend a lot of their day in a SQL editor?" So tell us about why you end up using it so much and how it's useful to you as a CEO of a data company?

Kevin Hu: 00:51:24 SQL in addition to English is probably the most useful language that a data practitioner could acquire. I know that in the last decade there have been many attempts to kill SQL, but SQL is still standing and it is still king. We use it all the time to sync our customers' databases and to query metadata from those databases. But we also use it on our own data. So I find myself with either TablePlus

or DataGrip open all of the time. And honestly, a lot of the questions we have around sales or product analytics, I find it's much easier just to have a SQL editor open and being able to express the question in SQL. There are no code tools, but if you know SQL, it pays off dividends.

Jon Krohn: 00:52:26 Yeah. So maybe this analogy isn't going to end up being right and you and correct me where I'm wrong on it. But you often hear about CEOs being, especially in a small company like yours, where you're focused on growth right now, you're kind of the head of sales, and by having access to SQL and being able to look at your customer data really quickly, being able to look at data from your product really quickly, on your own, without depending on anybody else that probably it's kind of being a sales engineer. Like you're this advanced sales practitioner that has access to sophisticated data tools in order to be able to do your job better than a sales person that would have to rely on somebody else. And so instead of being in a few minutes able to run a query, that salesperson who doesn't know how to do the SQL query has to rely on some analytics team or BI team to report to them. And it might take days or longer, and they might not get exactly what they were looking for.

Kevin Hu: 00:53:27 That's, I never thought of myself as a sales engineer, I might just change my LinkedIn title to that. Because that is a perfect description of what I do on a day-to-day basis. It's I think being able to make decisions based on business data, of course, the most important decisions won't have data to back it up. But at least many micro decisions, I would say, do have data to back it up and being able to retrieve that quickly is a huge boon.

Jon Krohn: 00:53:56 Very cool. Are there any other tools that you use on a daily basis?

Kevin Hu: 00:54:04 My handy dandy Jupyter Notebook.

Jon Krohn: 00:54:06 Nice.

Kevin Hu: 00:54:07 I basically have localhost open all the time. I would, you know how it is, anything that you cannot express in SQL without Python, got your pandas, NumPy, Altair. I love Altair for data visualization [crosstalk 00:54:28]-

Jon Krohn: 00:54:28 Altair, I hadn't even heard of that.

Kevin Hu: 00:54:30 Yeah. It's a Python wrapper around a tool called Vega, which was produced by one of my thesis readers and many colleagues at Stanford University in Washington where to make a visualization, instead of being procedural like in D three of, do this and then do that, it's a declarative approach. So I say, "Make me a bar chart with this variable on the X-axis, this variable on the Y-axis." And I found it's a very organic and rapid approach to exploring in a sense visually.

Jon Krohn: 00:55:09 Amazing. I love when I learn about new hugely valuable tools like this, for me. I'm sure it's valuable to listeners, but I get so much from asking these questions too. I can't wait to check this out later. Altair.

Kevin Hu: 00:55:23 Altair, like the, I believe it's a planet. I mean, sorry, not a planet, it's a star.

Jon Krohn: 00:55:29 Oh cool. Oh, I see. Yeah, I was spelling it completely wrong, but we will definitely get that in the show notes. That sounds great. I can't wait for people to learn about that. I can't wait to learn about that. So yeah, using Jupyter Notebooks for anything that you can't be doing in SQL and Altair for visualization in Python has proved particularly useful for you. I love that declarative way of creating plots. And then when you are using SQL, you love working with TablePlus and DataGrip. I'll be sure to get all of those in the show notes. Very cool. All right, so you can't be doing all of the engineering, including the

sales engineering on your own, Kevin. You have a growing company. I know that you're hiring software engineers and data engineers right now. What do you look for in the people that you hire?

Kevin Hu: 00:56:22 We look for intentionality. We at the company and me personally try to do things in an intentional way, not to the point that you're paralyzed by analysis, but to the point of being powerful about things, whether it's [inaudible 00:56:45] you produce or how you treat other people. We also look for people who prioritize actions over thoughts and words, even though those are important, we're ultimately trying to be action driven. And we just try and have fun at our company, right? We spend too many hours at work to not get some enjoyment out of it. At least to try to get some enjoyment out of it.

Jon Krohn: 00:57:10 I do miss the ping pong table of working in-person.

Kevin Hu: 00:57:15 When you're playing it, not when you're trying to focus.

Jon Krohn: 00:57:18 That's true. That is a good point. I might be the unruly distraction. I am, it turns out the reason why the data scientists on my team are so much more productive now that we're in a remote environment is they don't have to deal with me playing ping pong next to them all day.

Kevin Hu: 00:57:33 Just trying to not destroy people with your spins.

Jon Krohn: 00:57:38 I wish I was better. There's not a huge amount of spinning, but I can reliably hit the ball. Anyway, we don't need to get into that. Okay, so intentionality is something you look for in people you hire, that's awesome. And also to be able to have a bit of fun, that sounds like a really nice work environment. So what do you look for? What's different that you look for in a data engineer versus a software engineer?

Kevin Hu: 00:58:03 That is a great question and something that's changing all the time. To go back to the very beginning of our conversation, we feel like data is years, maybe a decade behind software. And the tools that software engineers have is significantly better than the tools that data teams have. But I believe that's also true when it comes to unfortunately like best practices and just kind of the culture around engineering. Where software engineering has so many well established best practices, whether it is incidents response playbooks, test-driven development, CICD, and only now are we starting to get some of that goodness into the data world. So I think that a lot of the technical aspects and non-technical practices are different between software and data engineering. That said, they're equally important. And software might be a bit ahead in terms of its adoption across companies in the world. But very soon data will be used across verticals, across sides of the companies. We see that in our customers, and honestly, data is a production system, right? When data goes down, the company can sometimes grind to a halt and we aren't treating it that way today.

Jon Krohn: 00:59:38 Really great explanation there. Nice that you tied the program back from the beginning in describing how tools are ahead for software engineers relative to data engineers.

Kevin Hu: 00:59:49 That's my job as a sales engineer.

Jon Krohn: 00:59:54 Nice. So I only have one last question for you, which is, do you have a book recommendation for us?

Kevin Hu: 01:00:02 I do. And I can't stop talking about this book. Shout out to Matt Housley at Ternary Data for recommending it to us when we grabbed coffee in Boston, it's called, Newton and the Counterfeiter. So Isaac Newton, physicist, inventor of calculus, believe it or not, spent the last 30

years of his life not as a scientist, but as master at The Royal Mint.

Jon Krohn: 01:00:29 Huh?

Kevin Hu: 01:00:29 The same mint that creates coins.

Jon Krohn: 01:00:32 Right.

Kevin Hu: 01:00:33 And I don't want to spoil too much, but do you know why our coins have ridges along the sides?

Jon Krohn: 01:00:42 Well, I'm now going to just speculatively guess based on the title of this book, that somehow that makes it harder to counterfeit.

Kevin Hu: 01:00:50 It's highly related where you would, without those ridges, it became much easier to shave off the edges of coins. That there are actually three kind of monetary crises going on at the time. Counterfeiting was a huge one. Another one was shaving off the edges of these coins. And yet another one was-

Jon Krohn: 01:01:13 Why would you shave the edge off a coin?

Kevin Hu: 01:01:16 Because back then they were made of silver and gold. So they were actually valuable shavings. The coins were actually made of silver and believe it or not, the value of the metal of the coins was sometimes higher than the value of the coins themselves.

Jon Krohn: 01:01:36 Did you know that that's the case with pennies today in most cases?

Kevin Hu: 01:01:41 Really?

Jon Krohn: 01:01:41 That a proper penny is worth more than penny and it's one of the arguments why, so a lot of countries have

gotten rid of pennies. My home country, Canada, you cannot get a penny anymore. Everything just it's rounded to five cent increments. So if you calculate the tax and it comes out to 83 cents, it just rounds down to 80 cents, because creating pennies is a fool's errand. We're creating these assets that are worth more than a penny. And I think, I don't think they have, I could be wrong. I can't remember if they still have pennies in Europe. But I know that in the U.S. we still have pennies and there's a movement to get rid of them because of this.

Kevin Hu: 01:02:28 Wow, funny, that is news to me. So I guess some of the struggles that Isaac Newton had to deal with still affect us today. And what I love about the book, related to your audience is when he came to The Royal Mint and tried to optimize the production processes, because England was running out of money, he took a very data-driven approach to the process.

Jon Krohn: 01:02:54 Cool.

Kevin Hu: 01:02:54 Isaac Newton was a smart guy, who would have thought?

Jon Krohn: 01:02:59 Yeah, yeah, yeah, yeah, yeah. He certainly was. And it sounds like a really interesting book from an interesting part of his life that I didn't know anything about. You hear a lot about his early years at Cambridge and that infamous apple story, but yeah, didn't know anything about this in his later life.

Kevin Hu: 01:03:16 He-

Jon Krohn: 01:03:20 It sounds like a good instance of somebody who was quite brilliant and well known in their day, who sounds like based on this job they might have also been rewarded nicely for that, which sometimes you hear the opposite. People like Nikola Tesla, who really struggled through

despite being so brilliant. Anyway, you were about to say something and I spoke over you?

Kevin Hu: 01:03:40 No, you're exactly right. I think he did get a cut of, the master of The Mint I believe did get a cut of the amount of money that was printed. So I think Isaac Newton, he wasn't that bad off as a professor, the most famous professor at the time, but he definitely died a rich man.

Jon Krohn: 01:04:01 There you go. I guess it's better than dying a poor man. So yeah, you don't get to enjoy your fame posthumously or your wealth posthumously. All right, awesome. That is a really cool book recommendation, Kevin. So you clearly are a brilliant guy. I have loved the way you have communicated so many interesting concepts so clearly today, I'm sure there are listeners who feel the same way out there. How can they follow you or connect with you to hear your latest thoughts?

Kevin Hu: 01:04:32 You can find out more about Metaplane at metaplane.dev or twitter.com/Metaplane. For me, I'm /kevinzenghu, Kevin Zeng Hu. There's a lot of Kevin Hu's out there, you got to fit the middle name in. And you can email me too at kevin@metaplane.dev, whether you're interested in data observability or you're looking to get into data science and have any questions there, I would just love to chat. Doesn't have to be about the company at all.

Jon Krohn: 01:05:07 Cool. Thank you for making that offer to our guests and I hope some take advantage of it. Kevin, it's been so wonderful having you on the show, and I hope we'll get to catch up with you again on the show in the future.

Kevin Hu: 01:05:19 Such a pleasure, take care, Jon.

Jon Krohn: 01:05:20 What a wonderful scholar and a gentleman Kevin is. I thoroughly enjoyed getting to meet him over the course of filming today's episode. In it, Kevin filled us in on the four

ways Metaplane looks for abnormalities in data flows, namely metrics, metadata, lineage, and logs. He talked about the super cool MIT junior lab or JLab for short, that requires physics majors to successfully replicate a Nobel Prize winning experiment every fortnight, and how being competent with data analysis makes tackling these experiments much easier. He talked about how we can be a legendary 10X engineer most easily by 2X-ing, 10 other engineers, either with guidance or software solutions. He introduced us to the TablePlus and DataGrip tools he uses on a daily basis to make SQL queries as the sales engineer in chief, AKA CEO of his company, as well as the Altair declarative visualization library he uses in Python to more intuitively create graphics from data.

Jon Krohn: 01:06:28 And he also told us about the intentionality he looks for in the software engineers and data scientists he hires. As always, you can get all the show notes and including the transcript for this episode, the video recording, any materials mentioned on the show, the URLs for Kevin's social media profiles, as well as my own social media profiles at www.superdatascience.com/541. That's superdatascience.com/541. If you enjoy this episode, I'd greatly appreciate it if you left a review on your favorite podcasting app or on the SuperDataScience YouTube channel. I also encourage you to let me know your thoughts on this episode directly by adding me on LinkedIn or Twitter, and then tagging me in a post about it. Your feedback is invaluable for helping us shape future episodes of the show.

Jon Krohn: 01:07:15 Finally, we've prepared something new for you. If you'd like to check out a detailed spreadsheet of all of the book recommendations we've had in the 500 plus episodes of this podcast, you can make your way to superdatascience.com/books. Thanks to our podcast manager Ivana for dutifully maintaining this epic

directory and now publishing it online for everyone. All right, thank you to Ivana herself, Mario, Jaime, JP and Kirill on the SuperDataScience team for managing and producing another awesome episode for us today. Keep on rocking it out there folks, and I'm looking forward to enjoying another round of the SuperDataScience podcast with you very soon.