EPISODE 04

[INTRODUCTION]

[0:00:00.3] JM: A hedge fund is a collection of investors that make bets on the future. The hedge refers to the fact that the investors often try to diversify their strategies so that the direction of their bets are less correlated, and they can be successful in a variety of future scenarios. Engineering-focused hedge funds have used what might be called machine learning for a longtime to predict what happens in the future. Numerai is a hedge fund that crowd-sources its investment strategies by allowing anyone to train models against Numerai's data.

A model that succeeds in a simulated environment will be adopted by Numerai and used within its real money portfolio. The engineers who create the models are rewarded in proportion to how well the models perform. Xander Dunn is a software engineer at Numerai, and in this episode he explains what a hedge fund is, why the traditional strategies are not optimal, and how Numerai creates the right incentive structures to crowd-source market intelligence.

This interview was fun and thought-provoking and actually quite difficult, because I was having trouble totally understanding what the long-term vision of Numerai could be if it succeeds. Numerai is one of those companies that makes me very excited about the future, and I hope to have other people from Numerai on the show to discuss the business further.

[SPONSOR MESSAGE]

[0:01:29.9] JM: To understand how your application is performing, you need visibility in your database. VividCortex provides database monitoring for MySQL, Postgres, Redis, MongoDB, and Amazon Aurora. Database uptime, efficiency, and performance can all be measured using VividCortex. Don't let your database be a black box. Drill down into the metrics of your database with one second granularity.

Database monitoring allows engineering teams to solve problems faster and ship better code. VividCortex uses patented algorithms to analyze and surface relevant insights so users can be proactive and fix performance problems before customers are impacted.

If you have a database that you would like to monitor more closely, check out vividcortex.com/sedaily to learn more. GitHub, DigitalOcean, and Yelp all use VividCortex to understand database performance. At vividcortex.com/sedaily, you can learn more about how VividCortex works.

Thanks to VividCortex for being a new sponsor of Software Engineering Daily, and check it out at vividcortex.com/sedaily.

[INTERVIEW]

[0:02:45.4] JM: Xander Dunn is a software engineer at Numerai. Xander, welcome to Software Engineering Daily.

[0:02:50.8] XD: Thanks very much.

[0:02:51.3] JM: Numerai is a new kind of hedge fund, and I want to work our way towards discussing Numerai with some simpler discussions. First of all, what is a hedge fund?

[0:03:02.8] XD: Very simply, a hedge fund is a collection of money that allocates those funds into the stock market, typically. You have some alternative investments; hedge funds that invest in cryptocurrencies. Typically, you've got hedge funds that invest in the U.S. Stock Market or the worldwide stock markets; Japan, et cetera.

[0:03:27.7] JM: What does the term hedge refer to?

[0:03:30.5] XD: Actually, hedging bets. In terms of actually creating risk portfolios and making decisions based on those risk assessments. It's basically, "I have some idea of what the market is going to do, and I'm going to hedge based on that idea. I'm going to make certain balances to have a good risk portfolio." Those kinds of things.

[0:04:03.8] JM: What's the financial structure of a hedge fund? You have investors. You have people who are actually placing bets. Where are the incentives in a typical hedge fund?

[0:04:14.4] XD: I think I see what you're asking. Actually, a hedge fund will have LPs, limited partners, and the limited partners are external to the hedge fund, and they have nothing to do with what the fund actually does. The limited partners will give some amount of money to a fund, and then the fund will be completely responsible for actually allocating that.

[0:04:40.3] JM: Okay. Let's talk about the market even more broadly. There is some large percentage of the market that is automated. You could call it bots. You could call it automated trading. Even the humans that are trading are mostly using automated orders. Can you give me a sense for what kind of software hedge fund and trading desks are using?

[0:05:03.3] XD: Yeah, that's a good question. I suppose it depends very much on what kind of trading you're doing. There are funds that trade on very long time horizons; that's six months to five years sort of trading. That trading might not even be automated, because there's really no need to it, such a long term position. Then, you have high-speed trading, which is taking place on nanoseconds. Obviously, that's very much automated. Those are probably mostly custom software suites written by the funds themselves.

Of course, there are tons of software suites that typical hedge fund use. They have software that interfaces with — What's the term? Nasdaq, for example.

[0:05:56.9] JM: Right. Exchanges.

[0:05:57.7] XD: Yeah, exchange. Thank you. Yes. Different pages that interface with exchanges to place trades or to go through various brokers, et cetera.

[0:06:07.1] JM: Right. Machine learning has crept into Wall Street overtime. I think machine learning has been used before it was even called machine learning at places like Jane Street, or other trading firms through Wall Street. How prevalent is machine learning in Wall Street? By the way, I totally understand, you are not a financial expert. For context for people who are listening who don't know Xander, he's a software engineer. He's got a background in software engineering. He's never worked in finance. I'm asking you this as somebody who may not have any insight to this, but I would love to know what's your perspective for how trading companies use machine learning.

[0:06:49.7] XD: Yeah. That's very accurate as far as my finance background; nonexistent. It's my understanding that most funds are not doing anything like machine learning. You do have quantitative finance, which is probably the closest that most funds come to doing machine learning. I think you have some funds like Two Sigma, or Renaissance Technologies that probably are doing some machine learning. Those are the kinds of funds you'll actually see with boots and nibs. They're really interested in that kind of talent, but those are really, really rare for hedge fund.

For the most part, hedge fund are not using machine learning at all. Actually, I think what most of them do is what I call voodoo trading. It's kind of humans making bad decisions based off of very little data, because humans actually can't put very much data in their brains.

[0:07:54.0] JM: Okay. We'll get to that question later on. I have some follow up questions about that. A company that's doing machine learning or, actually, basically, any type of trading, they're typically building financial models. A model is a vision for the way that the world is and where the world might be going. Describe what goes into a financial model.

[0:08:19.0] XD: Yeah. There's a huge spectrum for what could go into a financial model. Closer to our end of the spectrum, you can have something like scraping quarterly earnings reports, and then doing some kind of analysis based on how frequently does something appear in the report, or what do the numbers look like in the report. You could make — On the one hand, you could make a report, a human digestible report and making trades based off of that. On the other hand, you could frame it in a machine learning problem format where you actually predict the market directly from that information, whether it's word frequencies, or number trends.

I guess the simplest financial model would be take stock price data and just see if you can find trends. That's, of course, where random walk theory comes into play and you find out that doesn't work very well. That's kind of the basic idea of a financial model. A super basic financial model would be, I think, companies that — I think companies that are out to IPO are going to be worth a lot. That's a financial model. It's a stupid one, but that's an example of something super simple you could do to allocate funds.

[0:09:44.9] JM: Does a model just present a version of the world, or does it also make trades?

[0:09:50.2] XD: I would say a model simply presents a version of the world. Yeah, there are various degrees of how you connect fact to making trades. You could actually include in your model-making various things like risk balancing, or you could not, and then your model is just saying, "I think this will go up and this will go down, then it's somebody else's job to balance risk." I don't want to have all of my funds allocated in a single market. I don't want everything I buy to be in agriculture. That's very risky. You want to balance it across, "Oh! I have some tech and some agriculture, some in Japan, some in the U.S."

You could structure your model such that whatever comes out of it is exactly what you will be placing on the exchange, or there can be various degrees of human decision making between the model and the actual exchange.

[SPONSOR MESSAGE]

[0:10:54.7] JM: You are building a data-intensive application. Maybe it involves data visualization, a recommendation engine, or multiple data sources. These applications often require data warehousing, glue code, lots of iteration, and lots of frustration.

The Exaptive Studio is a rapid application development studio optimized for data projects. It minimizes the code required to build data-rich web applications and maximizes your time spent on your expertise. Go to exaptive.com/sedaily to get a free account today. That's exaptive.com/sedaily.

The Exaptive Studio provides a visual environment for using back end algorithmic and frontend component. Use the open source technologies you already use, but without having to modify the code, unless you want to, of course. Access a k-means clustering algorithm without knowing R, or use complex visualizations even if you don't know D3.

Spend your energy on the part that you know well and less time on the other stuff. Build faster and create better. Go to exaptive.com/sedaily for a free account. Thanks to Exaptive for being a new sponsor of Software Engineering Daily. It's a pleasure to have you onboard as a new sponsor.

[INTERVIEW CONTINUED]

[0:12:25.1] JM: The picture you're painting is hedge fund, typical hedge fund, have a bunch of people work for them. Maybe those people all work off the same model and maybe they all make their own models, then they make some human decisions based off of those models. They probably have some strategic sort of framework for how they're positioning a trade, maybe some combination of options, and stock, and futures, or something. What's wrong with that approach? What do you take — I sense that you have some disagreement with that being the best approach.

[0:12:58.5] XD: Yeah, very much. I very much disagree with that approach, and the market doesn't like it either. Hedge funds have been slaughtered this past year. If you just Google hedge fund, all of the news reports are about how awful hedge funds did this past year.

I disagree with this position, because I think it's largely a crapshoot. I think there's a huge survivorship bias and the few funds we see that are doing well are just kind of the long tail of some Monte Carlo simulation. I guess, in really simple terms, I think this is wrong, because the market is nothing but numbers. It's literally just massive amounts of numbers. Whether those numbers are crop yields, or Twitter sentiment, or stock prices, or whatever, and humans are just really bad at numbers. How many numbers can you fit in your head at once and visualize them, or understand them? The answer is not very many. That's exactly what machine learning is really good at; finding trends in massive amounts of data.

For the same reason that now is the perfect time to be doing image recognition. It's the perfect time to be applying it to large sets of data related to the stock market. We've been — I guess our capability to execute machine learning models has been greatly increased because of the hardware we have now. Machine learning research has really taken off with Google Brain, DeepMind, oPENai. There's an immense amount of progress in it, and I think it's exactly this sort of thing that's really good at understanding the market.

[0:14:51.7] JM: We're seeing this tipping point in certain games. Poker has got into the point where machines beat humans, same thing with Go obviously. I'm not sure what the status is on the centaur chess sort of thing, because there is a question for a while, at least, and maybe this is still a question, is the centaur, which is a human working with a computer, better than the computer working by itself. My perspective from the AlphaGo thing, it sounds like that, basically, the computer working alone has surpassed the centaur. I think you'd agree with me that that's an inevitability. Once you can refine a problem enough, but the market seems like a maximally complex problem.

I guess what I'm asking is; do we have enough refinement at this point where we can say, "Yes. The completely automated solution beats the human working together with the automated solution," which is basically what we have in most of these hedge funds that you're talking about, people that are making models and then they're like thinking through it and then they're making up iron condor or whatever and saying, "Okay. We're doing the iron condor, and that's our strategy for this." Are we there yet, where we can do a completely automated solution?

[0:16:12.3] XD: Yeah, certainly. The ultimate answer to that question is what are your returns? What are your returns and over what time period? That's the ultimate answer to your question? If my claim is correct, then that will beat the market. If my claim is incorrect over a long enough time period, it will not beat the market.

I suppose our track record is very small, simply because we're a new company. We've been trading for just over a year. I can't point to that as conclusive evidence yet, but there are a lot of things you can do in the intermediary. You can, for example, train machine learning models on time periods, or portions of data and then actually execute it across time periods or portions of data that have actually happened, but it's never seen before.

[0:17:10.7] JM: Back testing.

[0:17:11.4] XD: Yes, exactly. Back testing. That's one useful data point in knowing, "Does this actually perform?"

[0:17:21.3] JM: Let's get into what Numerai is so we could have some more far-fledged conversations. What is Numerai?

[0:17:28.6] XD: Yeah. Numerai is a hedge fund, and our fundamental thesis is that we can take datasets, and make those data available to anyone in the world, and that we will get from anonymous data scientists around the world really good predictions from those datasets. These data scientists compete on our tournament on our website. There's a leaderboard and it shows you the current rankings of all these data scientists, thousands of them around the world. We create a metamodel, an ensemble, of all of these predictions. Actually, we use the output of that metamodel to place the trades for our fund. They're actually competing to control the funds in our — The money in our hedge fund. They are paid based on the quality of their predictions. The people with the best predictions will be paid the most. It's this feedback loop of them producing better predictions, us doing better on the market. We can pay them more, et cetera.

[0:18:43.7] JM: Okay. Let's say I'm a new trader to the platform. I want to start competing in a tournament and I want to ramp up my credibility, so that I can have more of the portfolio allocated towards me. What is the onboarding system, and what kind of code am I writing. Give me a picture of what a trader that is participating in a tournament trying to get on boarded. What do they do?

[0:19:13.4] XD: Yeah, that's a good question. I wouldn't actually even call them traders, because they're not really making any trading decisions, not directly anyway. The onboarding procedure would actually be super simple. You go to the website, you click signup, same as any website. Okay, now you have an account. You download the data. Anyone can download the data. You can download the data. Even if you don't make an account, you can download the data. You open up the data, it's just a CSV file, just a simple text file with all of the values. Then, you conduct some machine learning on it.

I think most of our users will use Python, or R. Simplest example; do a logistic regression. We give out two sets; we give a training set and a tournament set. The training set has targets, so you train your logistic regression to predict those targets. Then, you test, you predict using your trained logistic regression on the tournament set. In those values that you have predicted are what you actually submit to us on our website. You just upload your text file of values. You have a prediction for each row of our data.

[0:20:26.3] JM: This CSV that I download to get started, does this get updated?

[0:20:32.3] XD: Yes, on a weekly basis. Our tournaments are one week long. Every Wednesday, actually, we give out a new dataset.

[0:20:43.7] JM: I see. The training set is what are you building your model based on. Then, the tournament set is what are you building your model based on. Then, the tournament set is data, basically, "from the same pool", but it's a larger dataset. By testing against that — Testing their model that they trained against the training dataset, on that tournament dataset. You get to make sure, "Okay. Did they over-fit certain things to the training dataset?" If they did, they're probably not going to perform well in the tournament. If they made a model that's very intelligent, then it's going to extrapolate to the tournament dataset.

[0:21:19.6] XD: Yeah, that's a great point. Generalization is very important. We kind of have a built in back test. A portion of the tournament set, we actually know the answers to. It's actually already happened. We can grade them approximately based on their performance on that set of data, because we know the answers. We haven't given the answers out, so they don't know them, but we do. We test them based on that, and that gives us an initial idea of how it might perform.

Then, there's also some portion of the tournament set that we do not know the answers to. That's actually live data, but that's happening on the stock market. These values, these predictions, are what we actually trade based on.

[0:22:05.8] JM: When you said you know the answer to it —

[0:22:09.0] XD: Sorry. The targets. We actually know — Let's give a simple example.

[0:22:15.3] JM: Sure. Please.

[0:22:16.1] XD: One row of the dataset could be Apple. We'll have a bunch of features, a bunch of values for Apple. Then, we'll have a target. This target is saying, "Apple, at this time, was a zero, or a one. It was a good buy, or not a good buy."

Then, those are the answers, when I say, "We have the answers." We have those targets. The targets in the training set are things that we have the answers for, that we actually know that's what happened on the stock market. The same is true for some portion of the tournament set, even though we don't give that out.

[0:22:59.4] JM: The thing that people should take away from this, whether or not they want to know in machine learning — Machine learning [0:23:05.6] in detail is you get started, you've got a playground to play in to build your model. Then, you can test that model in an actual tournament. If you do well in the tournament, you're going to rise in the leaderboard, and the leaderboard dictates how much of the Numerai funds are allocated to your strategy. Is that correct?

[0:23:31.1] XD: Yeah, that's quite accurate.

[0:23:32.4] JM: Okay. The core principles of Numerai include open participation and anonymity. Can you explain those two principles?

[0:23:41.8] XD: We're the most open hedge fund in human history. Unlike any other hedge fund ever, all of our data are public. You are not going to be able to get data from any other hedge fund in the world. It's a huge problem. If Two Sigma were to release the datasets that they're using, then they would lose their advantage, and other hedge fund would be able to trade based on their data.

We don't have this problem, because we obfuscate our data in a manner that preserves structure, so the machine learning potential is preserved. I can't tell what it is in the real world. It doesn't relate to the actual stock market. I wouldn't be able to place trades on it. We can place trades on it, because we know what every row corresponds to, but you don't know that. You wouldn't be able to place trades based on our data.

Yeah, unlike any other hedge fund, anyone can download our data. Then, in terms of anonymity, we don't collect any information about our data scientists. It can be anyone. You can use Tor to participate in our tournament. It's just — You have a username, password, and an email address. Those can be anything.

[0:25:03.2] JM: Give me more clarity on where these datasets come from.

[0:25:06.6] XD: We don't actually talk about where our data come from.

[0:25:10.9] JM: Oh, okay. You get the data from somewhere, or maybe that's like your special sauce. Aggregating — Can you tell me about what the contents of the data are. Is it the "price of something" in a given time, or a quote that you're getting somewhere, or you're aggregating a quote from across different exchanges?

[0:25:31.8] XD: Yeah. I think, actually, exactly what is the underlying data isn't the right question.

[0:25:40.1] JM: Oh, okay.

[0:25:40.6] XD: I think the right way to think about it is it could be any kind of data. This could be tweets that have been turned into vectors. This could be — Sure, it could be fundamentals. It could be something we scraped off of quarterly earnings reports. It could be something from standard stock prices. The basic idea is that you take any kind of dataset and you attempt to correlate it to the future of the stock market. Is it machine learnable? Do we get valuable predictions from data scientists?

[0:26:30.2] JM: Now, I'm seeing the scope of this company and how big it is, because you could predict pretty much anything.

[0:26:36.7] XD: Exactly. Yes. We're not specific to any kind of data. We have no loyalty to any particular kind of dataset. Indeed, our mission is definitely to take anything we can and actually release it to make correlations to the stock market.

[0:26:54.9] JM: Numerai has its own cryptocurrency, which is Numeraire, and this is used to show confidence in how well predictions will do. Explain what this is.

[0:27:09.4] XD: Numerai is really exciting for us. Currently, as I was describing data scientists are rewarded based on their performance on this kind of back test, this portion of our tournament set that we already know the targets to. That's not really what we care about. It's now what the hedge fund cares about Numerai. We really care about predictions that will perform well on the future, on the livestock market.

There's this kind of incentive misalignment at the moment. They have incentive — The data scientists have an incentive to over-fit to our private dataset, and we actually want them to have incentive to generalize to the future. We are creating this mechanism such that this incentive is completely aligned.

Basically, what I can do once I've submitted predictions to the website, I can stake Numerair to indicate my confidence of my predictions to perform, to generalize, to the livestock market.

[0:28:27.1] JM: Wow!

[0:28:27.9] XD: Yeah. Then, these Numerair will actually be held in a contract completely outside of Numerai's control. At the end of some time period, one month, three months, something like that, your Numerai will actually be either destroyed if you were wrong, you did not perform, you did not generalize to the livestock market, or your Numerair will be returned. Then, also, Numerai will give you some additional payout, some USD payout. Currently, we pay USD in Bitcoin. The basic idea is that your Numerair are not destroyed, and you get more money.

[0:29:11.2] JM: You get more money, you don't get more Numerair.

[0:29:13.1] XD: Correct. You actually get Numerair by competing in our weekly tournament, which is still based on this back test principle. You could think of it as two tournaments running in parallel. One is this weekly tournament for a back test, and one is this three-month long tournament for live test.

[0:29:35.6] JM: Is there a resell value on that Numerair?

[0:29:37.7] XD: Yeah, definitely. We expect that there will be some secondary market value for Numerair, just as you can exchange dollars for Bitcoin on exchanges. The same thing will inevitably happen for Numerai, just as ether is on secondary markets, et cetera. That's something we'll be able to facilitate, but it will likely happen.

Having that dollar value, you actually know the value of each Numerair you hold. You can actually understand the risk that you're taking.

[0:30:14.9] JM: Right. Risk of ruin.

[0:30:16.3] XD: Yes, exactly. You can actually reason, "Okay. I'm super confident about my predictions. I know based on how I created my model, how I created my predictions, I know that it will win in the livestock market." Then, you know that you should place a lot of USD value on that confidence. You should place a lot of Numerair on that confidence. Then, you expect the value of Numerair to increase as the value of Numerai's fund increases, because Numerai will be able to payout more money and it's this feedback loop.

[0:30:59.3] JM: Makes sense. Numerair is going to be deployed to the Ethereum blockchain. Ethereum is built with smart contracts in mind, and the basis for the way this currency works is the smart contracts defined by what Numerai is saying. Basically, the smart contract is these properties about how Numerai is wagered? How you are rewarded based on how much Numerai you've wagered?

I did all these shows about Ethereum, and people were talking about smart contracts. I was like, "Okay. What is a smart contract used for?" There was some conversation around, "Oh! You can use a smart contract to standardize how you would build an Uber-like service on the blockchain." I was, "Okay. I still don't quite understand how that works."

I think I'm understanding it better, thanks to your description of Numerair. Give me an idea of why it is on the Ethereum blockchain.

[0:32:02.1] XD: Yes, that's a great question. It's on the Ethereum blockchain because it facilitates creating some trustless interactions, and it facilitates creating something with value that is not under our control. It is not under Numerai's control. Once these tokens, these Numerair tokens, are in the possession of some data scientist, we cannot take them back. It's also very important that the value of these tokens is tied to Numerai.

We could do the same thing with Bitcoin in terms of giving out tokens that we can't take back, but what we cannot do with Bitcoin is create a network effect where the value of that token is the value of Numerai. There's this huge incentive for everyone who holds Numerair to make the value of Numerai go up. You want the fund to do better. That's extraordinarily powerful for Numerai, because you couldn't do that with Bitcoin, because you can't affect the price of Bitcoin. Unless you're some massive portion of Bitcoin's outstanding value, you're not going to be able to do that. Whereas with Numerair, it is in particular to Numerai. It is for Numerai's fund, and it is paid out based entirely on how well Numerai's fund is doing.

[0:33:33.4] JM: Right. Got it. Help me understand why you need to build that on the Ethereum blockchain as a smart contract rather than building your own cryptocurrency.

[0:33:42.3] XD: Oh, I see. From the ground up.

[0:33:44.9] JM: Yeah.

[0:33:45.4] XD: That's just hard.

[0:33:48.7] JM: Is this what Ethereum was ultimately built to do, is facilitate this kind of — "Build your own cryptocurrency." "We're a platform for cyrtocu —" I never heard it be called a platform for cryptocurrency. I heard it be called a platform for smart contracts, but you're basically saying a smart contract is — You could draw an equivalence between smart contracts and currencies.

[0:34:09.4] XD: Yes, you can. Smart contracts are just programs, nothing more, nothing less. You can imagine an incredibly — You could have a super simple smart contract that every time you send ether to it, it would just send back that ether. That's a smart contract. It's a super duper simple program.

[0:34:31.5] JM: Is that a currency too?

[0:34:33.5] XD: No. I wouldn't describe that as a token, or a currency. That's would just be — Yeah, that's just a trivial, "Oh! I got something okay. Send it back."

[0:34:45.1] JM: Right.

[0:34:46.6] XD: I guess, a currency as a smart contract. They're actually called tokens in terms of smart contracts. A token written as a smart contract is actually — You basically create your own tracking system. You have this integer, and this integer is all of the Numerais. This is all of your tokens. Then, based on whatever logic you put it in, it can be extraordinarily complicated. It could be super simple. These Numerair are sent to addresses, or they're destroyed, or they're — Whatever you want.

I guess a super simple token, a smart contract, would be just mint — Let say we have a fixed one million and every week I send 10,000 of those to a particular address. That's a token on the Ethereum blockchain. There's no intelligence there. There's nothing that makes that token interesting, but that's an example of a super simple test token as a smart contract.

[0:35:51.9] JM: Ethereum, here, is just serving as the shared ledger that is backing this state, basically.

[0:36:02.4] XD: Yes. I view Ethereum as infrastructure. I don't want to have to create my own distributed ledger, my own logic. It would be extremely difficult to replicate what Ethereum has done on our own. Instead, we can simply piggyback on it. We don't have to create our own blockchain. We can simply use their blockchain, and it executes on every node in the world that is executing Ethereum nodes. It's basically an infrastructure that we can use without having to create it all ourselves.

[0:36:38.6] JM: What are the hard engineering problems you're working on right now?

[0:36:41.5] XD: Yeah, we have a lot of those. I suppose one definition of hard is mathematically hard. Actually, is there a solution to this? Those are the hard problems that I find really interesting. We have a lot of encryption and machine learning related problems internally that are very hard problems. For instance, it's very important that the way we present our data to the world is done in a manner that not only preserves all of the important information that makes it machine learnable, but also that it would be very difficult for, an adversary, or some other hedge fund to discover what it actually is.

[0:37:31.3] JM: This is homomorphic encryption, right?

[0:37:33.4] XD: Yes, exactly. You're preserving some homomorphism in your data, preserve some structure that you're interested in, but completely obfuscate what the original data were.

[0:37:46.0] JM: This is like cutting edge field, right? This is like nobody — There's a lot of work around it, but not great answers, I guess, right? Is that accurate?

[0:37:57.2] XD: Yeah.

[0:37:59.3] JM: It was a state of the art of homomorphic encryption, and where are we bottlenecked right now?

[0:38:04.7] XD: Right. We have fully homomorphic encryption, which is to say, you can preserve all structure in the data. As a matter of fact, we even have pretty efficient fully homomorphic encryption. Most people are mistaken in thinking that it's inefficient.

For example, Microsoft published a paper just last year where they fully homomorphically encrypted the MNIST dataset. It's your standard simple image dataset that's used as a benchmark in AI. They fully homomorphically encrypted this image dataset, and then successfully trained and tested a deep network on this encrypted data. It's your standard 99% accuracy on the MNIST dataset.

The difficulty with this, actually, the hard part here, the remaining thing for this being practical for something like Numerai's use is that you actually have to be aware of the encryption used on the dataset to be able to interface with the data. If I took this fully homomorphically encrypted data and just ran it through any old net in TensorFlow, or any popular deep learning library, you would learn absolutely nothing. It's just garbage. It's just noise. What you actually have to do is write your neural net, structure your neural net with knowledge of the polynomials in the homomorphic encryption.

That's just really impractical for us, because you go from the number of data scientists in the world who are capable of performing complex machine learning on any dataset is quite small. Then, if you've add on to that the requirement of, "Oh! You also have to know how to write all of your algorithms from scratch with knowledge of homomorphic encryption," then you're down to a handful of experts.

That's just not practical to us. That's kind of the state of the art in homomorphic encryption research. It also feels like it's a big stagnated, and that they're haven't been really any interesting breakthroughs recently in homomorphic encryption. You have this whole spectrum of encryption. On the one end of it, you have obfuscation, which isn't really encryption. It's just do a random projection, or something, that just obliterates the values into some new distribution. You're going to preserve structure but it's not really encryption. There are no guarantees. It's just some new distribution. On the far other end of that, you have fully homomorphic encryption. You also have partial homomorphisms. You can actually preserve specific features in data without actually having that requirement of knowing detailed information about the encryption that was used.

[0:41:16.9] JM: The anonymity being so important to the way that Numerai functions is so interesting to me, because when you look at finance — I worked in finance very briefly, five months. I've worked at a trading place. If I think about a leaderboard — Typically, a leaderboard — If you take a bunch of finance people, especially, they're not going to be want to be anonymous. They're doing this, because they want people to know that they're making volumes and volumes of money.

The impression I get with Numerai is this is catering to a different type of person that wants to play the stock market game. This is a type of person that's just in it because it's an interesting problem. What's the bigger vision here? Because I see anonymity — That is something that caters — If you have a completely anonymous leaderboard. That's something that caters to a very different type of personality than is historically found on Wall Street. Why is anonymity so crucial to what you are building?

[0:42:20.7] XD: First of all, I'd say you don't have to be anonymous. If someone on our leaderboard who's doing very well, like [inaudible 0:42:28.5]. If he wanted to be publicly know, he could very easily do so. You could prove that this account belongs to you, and everyone can see how much that account has made on the website.

It's basically just not something we're forcing. We're not forcing us to know who is participating, and that's simply because it's not an integral component. The integral component is people who are really good at machine learning who want to either, A; solve hard problems, or both A and B; get paid for it. The long term vision for us, the vision, is that we actually don't think the world needs another hedge fund. We think the world needs one hedge fund controlled by every AI.

[0:43:20.5] JM: Right.

[0:43:21.2] XD: This is a platform for AI. AIs are our end customer. We really see a future where efficiency of the market is greatly increased, because you have some amazing AIs that are making amazing predictions.

[0:43:40.1] JM: Because, then, you basically have people — If you take this to the extreme where Numerai is the only game in town, and you usually have data scientists ensembling their models into making bets on the market, and that market is, again, entirely defined by Numerai. Then, the only thing they're really betting on is the pure stochasticism that is really hard to define. They're basically betting on physics, or weather, or genetic mutations that we don't understand, right? Is that accurate? Because that takes so much of the determinism — The missed opportunities for determine — I think of a market inefficiency as a missed opportunity to act on something that is deterministic.

[0:44:29.4] XD: Yeah, I think that's completely accurate. There are a lot of those things. I think some of the examples you gave are accurate. The weather is largely deterministic. Even if that's what it comes down to, maybe that's something you can. As a matter of fact, that is something you can place trades based on.

You're right. When AI are considering every conceivable piece of information and making the right decisions based on all of that information, then, we have a market that is way more efficient, and that's better for everyone. The entire world wins. That's very much our role.

[SPONSOR MESSAGE]

[0:45:20.9] JM: Good customer relationships define the success of your business. Zendesk helps you build better mobile apps and retain users. With Zendesk mobile SDKs, you can bring native in-app support to your app quickly and easily. If a user discovers a bug in your app, that user can view help content and start a conversation with your support team without leaving your app.

The conversations go into Zendesk and can automatically include information about the user's app information, device information, usage history, and more. Best of all, this is included with Zendesk for no extra charge. Use the out of the box iOS UI to get up and running quickly, or build your own UI and work with the SDK API providers. Keep your customers happy with Zendesk.

Software Engineering Daily listeners can use promo code sedaily for $177 off. Thanks to Zendesk for supporting Software Engineering Daily, and you can check out zendesk.com/sedaily to support Software Engineering Daily and get $177 off your Zendesk.

[INTERVIEW CONTINUED]

[0:46:42.6] JM: It's a big vision. It's going to take a while to work on. In the midst — We're talking about a company that is taking advantage of all the deep learning tools out there, the open source deep leaning tools, because these all data scientists are using these sort of things. We're talking about a company that takes advantage of all the infrastructures that is plugged in to the stock market, which is this really effective hive mind we have. We're also talking about a world where cryptocurrencies are increasingly stable, and Numerai is leveraging the fact that Ethereum has reached some stability. You're actually building a smart contract on top of it that does something.

When Ethereum first came out, people were talking about, "Oh, he's built a decentralized Napster platform, or something," and there were all these theories around stuff. This is a tangible thing that you're building on top of it. This is a broad question, but what are some of the other things that we're going to see as a result of this confluence of really big changes that are happening in our world? Give me your most far-flung, crazy ideas that you wouldn't tell other people.

[0:47:56.5] XD: It depends on what kind of time horizon we're thinking of. 10 years is a really long time for me. That's what I call long term.

[0:48:02.9] JM: Okay. Let's do 10 years. We're both still alive then. Most of the listeners will still be alive in 10 years. I like that one.

[0:48:08.9] XD: Right. Okay. In 10 years. I think AIs are going to be controlling an immense number of things in 10 years, definitely cars, definitely a lot of the stock market. I think one of the challenges is actually going to be understanding how human skills fit in with something that is so much better at us at so many things. It's, in every way, a good thing, because humans shouldn't be doing mundane tasks. The markets should be efficient. There's no loss for anyone if the market is efficient.

On the ultimate horizon is, I think, humans becoming completely obsolete. Either evolving out of what we currently are, or seizing to exist, because there's something better.

[0:49:06.6] JM: Do you think that's in the 10 —

[0:49:08.1] XD: No. That's not 10 years.

[0:49:08.6] JM: You don't think — That's 20, 50 —

[0:49:11.7] XD: That's really difficult to say.

[0:49:12.9] JM: Yeah, a pretty big question mark.

[0:49:15.0] XD: Yeah. It's definitely not 10 years, because that's where — Way before that, we have to worry about safety. Now is not the time to worry about safety. That's something I very much disagree with.

[0:49:25.6] JM: Oh, really?

[0:49:26.3] XD: Yeah. I think most of the people who worry about safety actually don't know very much about AI. Elon doesn't. I think most of the people who do, ho are actually very knowledgeable about AI, do not worry about safety right now, because it's really just very fancy logistic regressions, and nobody was worried about AI when all we had were logistic regressions.

One day, a really, really advanced logistic regression will be dangerous, but we're very, very far from that. I think Andrew Ng was actually very accurate when he said, "Worrying about AI safety right now is like worrying about overpopulation on Mars. One day it will be a problem, but we've got to get somewhere close to it before we worry about it." We don't have enough information to worry about it right now.

[0:50:21.8] JM: We don't have enough information to worry about it right now, but we're going to have to worry about it eventually, and we already know that. The people at OpenAI, you think they just should close up shop and go work on something that is more proactive, rather than preventative?

[0:50:42.5] XD: No. Actually, I think most things that are happening at OpenAI are proactive and not preventative. All of the papers they've published, pretty much all of them, have not been about safety. All of the tools that they have released, they're not about safety.

They just released a platform for testing AIs on many different video games. That's not about safety, that's about trying to improve AIs on video games, and that's a great direction. I love the research that OpenAI is doing.

Yes, their mission statement is safety, because the billionaires who gave them money are worried about that.

[0:51:20.7] JM: Do you think they're worried about it, or do you think it's a red herring to get some research done in the AI space?

[0:51:26.8] XD: That's interesting. I doubt that they're lying. Perhaps they're actually worried about it, but Elon is worried about a lot of things. He's worried that we're in a simulation. He's worried about Gregu. He's worried about —

[0:51:40.4] JM: I don't think the simulation think is a worry. That's just an observation.

[0:51:43.9] XD: Yeah, potentially. I suppose you could be right. He worries about it in so far as he thinks about it.

[0:51:50.7] JM: Yeah, what's wrong with that? Do you think it's a waste of time?

[0:51:56.1] XD: I do think it's a waste of time. I think it's a waste of time for similar reasons that religion is a waste of time. You can't do anything with it.

[0:52:05.9] JM: It's a more — I agree with you. It's on the same plane as religion, but if you take a probabilistic vision of the world and they say, "Okay. What's the probability that we're living in a simulation versus not living in a simulation." Let's throw religion out of the window, because you and I are obviously both absurd. Sorry, listeners who are religious.

You can take a probabilistic notion if you say, "Okay. We're living in a simulation." To some people, myself included, that's reassuring. I find that a more comforting notion than the ambiguities of reality. Anyway, I don't know — We're getting —

[0:52:49.0] XD: God is comforting too, and we don't have any evidence for God.

[0:52:54.2] JM: Okay, sure. What's the null hypothesis then? What's your null hypothesis for reality?

[0:52:59.0] XD: Does that mean —

[0:53:01.4] JM: Fine. Okay, God is comforting, but it's probably not true. We have no evidence for it. Let's ignore that. The simulation seems probabilistically likely, but we have — Well, maybe not. We have no evidence for it. What's the null hypothesis? Do you say, "The null hypothesis is reality as it seems," or do you say that the null hypothesis is whatever you makes you feel good? It seems like you just don't even care about it. You don't even think about it. You have no evidence. Therefore, you don't even think about it. It's an existential coin flip. Who cares?

[0:53:38.3] XD: I suppose I'm in favor of pursuing any endeavor, where we can collect data. There are some interesting ideas for potentially doing this in the simulation direction. I don't believe any of it has been done, but it was just kind of — I knew very little about this degree of physics. This is just me reading headlines in passing. I'm in favor of pursuing the collection of any kind of data that can be collected.

So far as what is my null hypothesis, or do I think about this? The answer is mostly no, for the reason that I can't do anything with it. If we are in a simulation, does that actually answer the null hypothesis? What happens when you get to the last hurdle? Where's the null hypothesis then?

[0:54:32.3] JM: Okay. Very interesting. Do you have any interesting views on cryptocurrencies that I might not hear anywhere else?

[0:54:41.0] XD: Sure. I think I differ from a lot of the cryptocurrency community in that Bitcoin is not the alpha and the omega. It's kind of become a reserved currency. It's something that there's a lot of support for. It's very easy to turn USD into it and vice versa. It's not a platform. It's very hard to make Bitcoin side chains, and I'm very much not a Bitcoin maximalist.

There's also the fact that the Bitcoin community has been terribly gridlocked and have made no progress on most of Bitcoin's problems for the past couple of years. I'm very much in favor of what Bitcoin maximalists call Altcoins, Ethereum, even Tezos. I'm very much in favor of cryptocurrency platforms. Distributed ledgers, distributed blockchains that make it easy to do more things with them. Even new cryptocurrencies that don't potentially have some of the problems that Bitcoin has, like transaction speeds, et cetera.

[0:55:52.8] JM: When you say that side chains are hard to build on top of it — I know Blockstream is working on those, and you could envision a world where if they get side chains working well and you can get transaction volume up, then, first to market, being Bitcion, is a great utility. You're saying that the bureaucratic morass, the technical issues are perhaps enough to hold Bitcoin back from being that utility currency, that flexible utility platform.

[0:56:30.4] XD: If Bitcoin is able to solve all of its problems and become a platform that's as easy to use as Ethereum, then I'm all in favor of it. I have no qualms with that at all. As it stands, it very difficult to use it as a platform, and Ethereum has a lot of advantages compared to it.

There are even problems with Bitcoin in terms of a massive amount of blockchain power is held by a very small number of miners in the world. That's another problem with Bitcoin that some other blockchains don't have.

[0:57:04.6] JM: Okay. Xander Dunn, I want to thank you for coming on Software Engineering Daily.

[0:57:07.9] XD: Thanks very much, Jeff. It was really awesome.

[END OF INTERVIEW]

[0:57:14.6] JM: Thanks to Symphono for sponsoring Software Engineering Daily. Symphono is a custom engineering shop where senior engineers tackle big tech challenges while learning from each other. Check it out at symphono.com/sedaily. Thanks again Symphono.

[END]

