EPISODE 309

[INTRODUCTION]

[0:00:00.3] JM: If you wanted to build a machine learning model to understand human health, where would you get the data? A hospital database would be useful, but privacy laws make it difficult to disclose that patient data to the public. In order to publicize the data safely, you would have to anonymize it so that a patient's identity could not be derived from the data about that patient. True anonymization is notoriously difficult.

In every industry where privacy is a concern, there is a similar challenge. If there's no place with public datasets, there's no place where the machines can go to learn. The possible machine learning algorithms that we can build are limited by the datasets that are available. Auren Hoffman started his company, SafeGraph, to unlock datasets so that machine learning algorithms can learn from those datasets.

In this episode, we talk about the machine learning landscape in both the short and the long-term time horizons. We also discussed some of Auren's strategies for building companies which have been crucial for me in thinking about how to build Software Engineering Daily. I talked to Auren a couple of times prior to starting this podcast and those conversations were very fruitful in terms of deciding how to go about building this podcast.

So, thanks to Auren, and I really think you're going to enjoy this episode.

[SPONSOR MESSAGE]

[0:01:31.3] 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. The Exaptive Studio provides a visual environment for using back end algorithmic and front-end 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]

[0:03:01.3] JM: Auren Hoffman is the CEO of SafeGraph. Auren, welcome to Software Engineering Daily.

[0:03:06.7] AH: Thank you Jeff.

[0:03:07.0] JM: We've had several shows recently about machine learning where we explored the idea that the quality of machine learning models is often proportional to the amount of data that are used to build that model. Explain why that is.

[0:03:25.2] AH: There's been a lot of studies where depending on the amount of data that you have, different algorithms are going to win. You can have a scenario where you've got, let's say, six different algorithms and you're going to have — Let's say you've got a thousand datasets labeled, a million, 10 million, billion, et cetera. Literally, as those datasets grow, you'll see different algorithms winning out.

I've seen many scenarios where like the worst algorithm in the thousand set becomes the best one in the trillion set, or something. It's really hard to know which is the best algorithm unless you have a really big set.

[0:04:06.9] JM: Moore's Law has given us consistent advancements in technology for a long time, obviously suggests that transistor density is what defines the power of our computers. At this point, we know how to distribute machine learning jobs across multiple machines really efficiently. In some sense, Moore's Law doesn't inhibit machine learning. Machine learning is so flexible that it is increasingly defining how our society organizes, how productive we are. Do you think that we've moved beyond a place where the hardware developments are the bottleneck, and now it's the quality of our machine learning models that are the limiting reagent?

[0:04:50.7] AH: In think in every year, it could be different. That's probably true today, like in 2017. Right now, it's probably the quality of the data. We can dive in a little bit more about where we've seen — If we kinda look at the history of machine learning, like if you think of the 90s, chess was this great place where machine learning just did an incredible job in the 90s. Personally, that's 'cause the data was available.

The average chess game — Most of the major chess games, the data is available for the last 150 years. You just have all those games, thousands and thousands — Millions of games. You have this data available, and then the data is not very big. The average chess game has maybe 40 moves or something, so it's not pretty big, and then the notation is standardized across geographies, across everything. That was really, really great.

That moved to the late 90s to the stock market. In the stock market, we have this pricing data just tied to a ticker which, by seconds, we have price by second, or price by time. Maybe a hundred years ago, it wasn't by second, but it was priced by day, but we still have this price by time for a very long period of time, for hundred years or so.

That data is available, it's labeled, it's easy to use. It wasn't free like the chest data, but it's relatively low cost, easy to get. Lots of people were able to get that data and start doing it. I think we could look at where the data is available as to where the innovations are going to come in.

An area I think we won't have an innovation in for a long time is nutrition. We have very little data about what the average person eats by time and then how that relates to outcomes in their life. Also, this is very hard to do, because it's a longitudinal, and so it's a very difficult data.

This is why every three years, there's a complete new fad of nutrition, which is backed up by "science". Margarine was great when I was growing up, now it's considered really terrible. Now, everyone wants to have a high-fat type of thing. Who knows what's going to be in the future?

Pretty much everyone who believes something in nutrition probably — I think as we see this data go in, it could change. Now, I think chips are valuable, and we shouldn't discount chips. If we've seen this like Tensor processing unit that Google uses for their AlphaGo stuff that they're doing, their TPU chip is very interesting.

I think we have these basic chips that are going to be very interesting in machine learning. They're going to have specific jobs that they're going to be focused on, and there probably will have quite a bit of advantages to using these chips in the future. We still need the underlying data, like having these chips without the underlying data, I don't think these chips are going to help us solve nutrition.

[0:07:42.3] JM: Right. At SafeGraph, you are working on democratizing this data for machine learning. Explain what that means.

[0:07:51.7] AH: The core goal is to be able to, if you're actually doing machine learning, is to be able to get access to really great data. That data, first of all, has to be data — It's hard to get access to in the first place. Usually, it's either very expensive, or you have to do all these BD deals to do it, and so data should be relatively low cost, easy to get. It also should be very clean, and labeled, and easy to use. You shouldn't have to spend — Once you get the data, then you shouldn't have to spend all these time then having to deal with it. It should just be like — You should be able to deal it immediately, and then it should be correct.

The problem with most datasets out there is that a large portion of the data is incorrect, and it has errors, it has all these other problems. You should believe that the data is true. Of course, data was never 100% true, but you should believe that it's very close to 100% true. Certainly, if data is 50%, or less, this is really bad.

Depending what you're doing, if it's below 90% true, it could be very bad; or below 99% true, it could be very bad. You need to understand what the false positive and false negatives are on that data and have confidence that those false positive and false negatives that are advertised are actually correct. Ultimately, you don't want to actually have to stand the data up. Ultimately, it'd be great if there was someone else who stood the data up and you could just run your algorithms on it.

Then, if the data is dealing with people, there are lots of privacy issues. If you're a machine learning company, you might not want to become a privacy expert as well. Hopefully, they can take that burden on as well. There could be a lot of different things that you'd want some sort of data entity to help you with if you're doing machine learning.

[0:09:35.5] JM: Once you have these datasets acquired, what's the best way to expose them to machine learning programmers? Is there an API that you have in mind for what is the best way to share a dataset?

[0:09:47.0] AH: It's interesting, and it's still something we're personally trying to figure out. If you have an algorithm, the long-term goal would be if you could put your algorithm in a box so that no one could see that algorithm, and then run it on a dataset that's already stood up somewhere so you don't have to go stand it up. Maybe you don't necessarily see the underlying data, but your algorithm can get access to the underlying data.

This would be like the ultimate point, because now you can have really sensitive data that you could use your algorithm on that maybe they would never want anyone looking at the underlying data. Then you could have your algorithm in a box and no one could see your underlying algorithm.

Imagine if you could run your underlying algorithm on Medicare patient data or something like that. Obviously, this is a super sensitive data, so you never want a scenario where someone could actually see the underlying data. Maybe if you could run an algorithm where you couldn't see the underlying data, but you could actually just run it on the data, you could learn a lot of things and maybe help a lot of people out, maybe help people understand what treatments to get, et cetera. This could be something that could be really, really cool.

I don't think we're going to be there anytime soon, so certainly not 2017 is when we're going to be able to do this type of thing. I could imagine in the next five years where this would exist, and then I could see a lot of really interesting things happening when this exist.

[0:11:10.9] JM: It's a regulatory bottleneck, it's not a technological bottleneck, because you can very easily imagine throwing your algorithm in a Docker container, sending it to SafeGraph, having it run on Medicare data, but the regulatory burden to getting that Medicare data is pretty onerous at this point.

[0:11:31.0] AH: There's a technology burden as well, because to go through the regulatory burden, you have to prove that someone couldn't see the underlying data. There is a technology burden. That definitely has to be solved first before you can go through your regulatory burden. Certainly, both exist.

[0:11:48.9] JM: Safety refers to privacy in this instance in SafeGraph.

[0:11:52.1] AH: In SafeGraph, our goal is to graph lots of datasets together and to make it really interesting. Then, it's really important that these datasets are kept safe, so there's a privacy component to that if you're dealing with data about people. Also, if the data has an underlying owner, they're worried about data leakage. Sometimes the data — The security is really important, et cetera. Making sure data is kept in a very safe way is really important.

At my last company, we managed the internal databases — consumer databases for most of the biggest companies in the world. We both had a privacy burden of making sure that we handled the data in an appropriate way, and then we also had a security burden, because you wouldn't want a scenario where someone's data — We handled large banks, and we handled large retailers, and telcos, et cetera, where if their data got out in the wild, this could be a huge problem to the underlying company that own that data.

[0:12:50.2] JM: From what I see, the way that you think about company building is that you architect a pretty big vision and then you define a sequence of short term and longer term profit centers that can definitely reached along the path towards getting to that vision. You're talking

about some problems that are going to be difficult to overcome in a visible timeframe. Have you thought about what are the shorter term and what are the midterm profit centers that you might get to with SafeGraph?

[0:13:30.5] AH: Yeah. By the way, my advice to entrepreneurs isn't necessarily to define a massive problem. I don't think that is necessary. I think that's helpful. I think it's super helpful for recruiting talented people to join your company. It may be helpful for raising money, et cetera, but I don't think that's a necessity. Sometimes it can be a huge distraction to have a really big vision. I think the most important thing is to define the small vision, the area that you're going to do really well and that you're going to excel in. I think that's really important, is to figure out what's your niche and that you're going to dominate.

I think, too often, people have just a very large vision, then they have tons of competitors who are competing with them in that really large vision. This becomes a very hard game to win when you have all these really, really smart competitors doing exactly what you're doing. What you want to do is go after relatively small market where there is a too high of a burden for lots of other people to go after that same market. Let's say that market is $10 million, or something that, it wouldn't make sense for lots of really smart people to go after this small market. Then, you could do really well in that market.

Then the question is; "As an entrepreneur, are you smart enough to move into adjacent markets when that market is tapped out? When you're no longer growing at 100% year over year? Are you smart enough to move to other markets?" Then this is also true for funder. I think for a venture capitalist, or other types of funders, investors, I think they care about total addressable market too much, and they spent too much time thinking about it. What they should be thinking about more is when this company hits a wall where they can't grow at 100% year over year in this particular market, is this team smart enough, and they adept enough to move into adjacent markets where they can keep growing? Evaluating that team is really, really important, rather than evaluating the market.

[SPONSOR BREAK]

[0:15:38.2] JM: When you are continuously deploying software, you need to know how your code changes affect user traffic around the world. Apica System helps companies with their end-user experience, focusing on availability and performance. Test, monitor, and optimize your applications with Apica System. With Apica Zebra Tester, Apica Load Test, and Apica Synthetic, you can ensure that your apps an APIs work for all your users at any time around the world.

Apica Zebra Tester provides local load testing for individuals, small teams, and enterprise DevOps teams to get started quickly and scale load testing as your needs evolve. Apica Laod Test ensures that your app can serve traffic even under high load. Apica Synthetic sends traffic to your website and your API endpoints from more than 80 different countries, ensuring wide coverage.

Right now, you can go to softwareengineeringdaily.com/apica for a webinar about the real ROI of API testing. You can also find past webinars, just how to optimize websites for fast load time. Go to softwareengineeringdaily.com/apica to find the latest webinars on load testing and lots of other topics, and check out Apica System for testing, monitoring, and optimization.

Thanks again to Apica for being a sponsor of Software Engineering Daily.

[INTERVIEW CONTINUED]

[0:17:08.2] JM: Talking about SafeGraph specifically, the steps to building the SafeGraph platform are; one, acquire the data. Two, host the data. Three, prepare the data. Four, understand data privacy. I know that you are a fan of doing things serially rather than in parallel, and this seems like a serialized strategy. From a management perspective, are you thinking about all four of these things right now and delegating specific people, or specific teams to each of these serialized tasks, or are you having the whole company think about each of these problems? Does it still kind of in the ideation step where everybody is working on all four steps?

[0:17:52.2] AH: SafeGraph has 13 people as of today, and mostly software engineers. I think when you're a relatively small company, you really have to — I think it's true when you're at any stage of the company. Whether you're a one person company, or a million person company, you need to really prioritize and really, really figure out what are you good at, what are you going to

focus on, and get everyone focused on a very small number of things and really make sure you excel at those things.

It's important to figure out what those things are. I don't think you have to optimize always too much to figure out what those are. It's just important that you do have things that you are focused on. What you don't want to be doing is lots of things. You want to figure out, let's say there's a hundred important things to do. You don't have to spend too much time optimizing on the two of the hundred things are the most important thing, 'cause you could probably spend just a little bit of time winding down the hundred to 20, and then like picking up randomly out of the 20 is probably okay, or you go with your gut, or whatever you're going to do.

What's most important is that you're relatively focused, and this is like really hard for especially entrepreneurial people to do, because there's just so many opportunities that they see and are constantly seeing all these really interesting things. You can could always like rationalize stuff. So everybody I know, including myself, struggles with this. It's a constant struggle to deal with this. Usually you have way too many ideas, not too few ideas, and you could do too many things.

I think this is true in life. I think people in life should literally take things off the table. I think they should actually declare to them self and their friends that they will never do something. Not that they're going to do it later. The should declare they will never do it. I'll give you my life example, which is I've always wanted to write a novel. This is always something in the back of my mind. Since I could remember, since I was really young, I've always really wanted to write a novel. A lot of people have this thing.

Not that long ago, a few years ago, I just declared to myself, "I will never write a novel." It's just never going to happen. I just think it's never going to be in my top 10 things to do. It's always been in my top 50 things. It's never going to be in my top 10. Therefore, it's never going to get done. It doesn't mean I can't change my mind 20 years from now, and maybe I may change my mind, and maybe I end up writing that novel 20 years from now. I've taken it out of my mind as a thing that I like hope I'm going to do one day as some sort of like bucket listing. I've removed it from my mind, and that freeze my mind to now focus on the most important things.

[0:20:31.2] JM: I have a segue way to my next question that plays off of that novel. There's a platform now called Book in a Box where you basically pay them $15k. They send somebody to interview you about what book you want to write, and then you basically outsource a lot of the work that goes into writing a novel. We sometimes see these —

[0:20:53.3] AH: I think that's probably for nonfiction, rather than fiction.

[0:20:56.4] JM: I think they have both.

[0:20:57.2] AH: They do? Okay. That's cool.

[0:20:58.9] JM: Yeah, so it's kind of interesting.

[0:21:00.1] AH: That's cool.

[0:21:01.2] JM: The reason I say that as a segue way is you have written about how the necessary size of employees in a company is shrinking, because we have these products like Twilio, and Heroku, and Asana and these amazing SAS tools. Not to mention, more open-ended outsourcing tools, like Upwork, and Fiverr, that are actually pretty good if you want to do outsourcing, because they have an —

[0:21:23.8] AH: Yeah, or even Mechanical Turk, or CrowdFlower, or whatever.

[0:21:26.2] JM: Exactly. These things are getting better, and this is a pressure to have a smaller workforce, at least, early on, because if you can just outsource these tasks to services that do it, or if you can outsource it to contractors that you don't have to give equity to, for example, and they are incentivized to do a really good job, because they have some star rating system. It just reduces the amount of workforce that you absolutely need.

How has your management strategy changed in the era of the smaller workforce?

[0:21:57.5] AH: I think you should be doing — If you're a company, you should be doing everything possible to think about how to keep your workforce smaller, rather than how to make

your workforce bigger. The best run companies in the world — Let's say the worst run companies in the world have some sort of like N-squared communication challenges, where N is the number of employees. The best run companies in the world maybe have like N over three. That's the best, is that N over three.

Even in the best case scenario, as you increase your number of employees, you have linear scale challenges. It's never like — There's not an asymptote that happens as you scale. You should be doing everything possible to eliminate these communication challenges, and then this world where things are moving really, really, really fast, having communication challenges is a much bigger burden that it used to be, because you're competing with very nimble, very fast moving organization. You want to be able to communicate in your company really, really, really quickly.

There's a couple of different ways you can design your company. One is you can design your company in like pods. I think this is a really smart, innovative way to design your company. This is the Amazon, two pizzas idea where you're designing your companies in pods and you're essentially creating these APIs of pods, and where they're interacting each other in APIs. You could do these with software developers, but you could do these in lots of other types of ways. I think that can work really, really well. Certainly, as you scale of something like you should be thinking about.

The other thing you could do as a company is you could just figure out — You can almost redefine what your company is. Even if your company has 200,000 people that work for it, you could almost redefine it like, "Okay, corporate has these 30 people here, and then we're going to just be like capital allocators," and we're almost going to pretend that these are different. This is like the Bircher Hathaway model.

[0:24:14.7] JM: Alphabet.

[0:24:15.3] AH: Right. Yeah, or Alphabet is kind of starting to get like that as well. They're not there yet, but I think they're moving in that direction. I think that's really — That can be really smart, and they're doing this all for the same reason, which is, "Okay. If we have this small

number of folks, we can communicate really, really quickly, and then we'll just do this capital allocation problem."

Of course, the best case scenario is to remain small so you never have this type of thing. I think this using the services that you mentioned, whatever, it's CrowdFlower, Heroku, AWS, Google Compute, you go down the list — Twilio. All these really great — SendGrid, these great APIs services are out there. You don't have to build something if somebody else has something.

Now, usually, if you build it, it will be better for you. Sometimes it will be actually more cost effective for you to build it. There's this opportunity cost of building it, which means you're not going to do something else, or you're going to have to grow in size, which is quite difficult. By the way, even if you want to grow in size, doesn't mean you can. There's this very difficult burden of recruiting really talented people. It's really hard to find super talented people, and it's unclear once you find them that they'll want to join you.

So it takes a lot of time recruiting. I know lots of companies where the average software engineer spends 25% to 35% of their time actually doing interviews and recruiting. That comes in like chunky places too. It can end up being like 50% of the degradation to their actual output doing all this recruiting. If you can cut the recruit time of the average software engineer to 10%, or maybe even 0%, now you've massively increased their productivity.

While some software engineers love recruiting, quite a few software engineers actually don't like recruiting. It becomes this huge issue, et cetera. You want to be doing everything possible to focus on actually doing what you can do great as a company and not having to deal with all these growth issues that come with companies.

[0:26:08.0] JM: As someone who's building a marketplace for machine learning data, to what degree do you feel obligated to have a technical understanding of how machine learning algorithms work? I host a podcast about software engineering. A lot of the shows about machine learning. I don't have an in-depth understanding of how a lot of these things work. A lot of the people that I interview do and I know the words that I can use to get out the necessary information that I need, but I don't find myself going deep into how deep learning works, for example.

To what degree do you feel that is something that you need to focus? I mean, to your point about focus.

[0:26:45.2] AH: I mean, that's on my list, and I haven't eliminated it like the novel yet, but I also, like you, haven't been able to spend as much time as I would like on it. We have a lot of really talented people at SafeGraph that do have a really good understanding machine learning and do spend a lot of time machine learning. As a former software engineer, but not an active current one, I have not had the ability to spend as much time on it as I would like.

Of course, whenever people say this thing that, "I don't have as much time as I like." It really means, "I don't choose to do it," right? It means I'm choosing to do something else. I'm either choosing to read about when I have reading time, or I choose to read about something else, or I choose to spend time with my kids, or whatever it might be.

[0:27:34.0] JM: Of course. I make the same decisions.

[0:27:37.8] AH: Yeah. These are always difficult things. I'm never sure if I'm making the right decision or not. Whenever you're doing time allocation, you almost certainly aren't doing what's optimal. If you take a meeting with somebody, it might turn out to be good, it might not. Sometimes you actually won't ever know if it was good or not. Sometimes you may not know for 20, 30 years whether something was good. If you don't take it, you'll never know.

These are always hard things to understand from like a time allocation thing. I think it is good to spend time understanding your time allocation. Of course, if you spend too much time understanding your time allocation, then you don't have time to do other things. All these things are difficult decisions to figure out.

[0:28:19.3] JM: I first met you through Quora. You were on my podcast, The Quoracast, and I was intrigued, because your writing tends towards what you call non-obvious ideas. What is a non-obvious belief that you have about machine learning?

[0:28:35.6] AH: Good question. I don't know that I have anything like super non-obvious about it. My core belief is that data is like way more important than most people think is like, now, actually relatively obvious. I think a lot of other people believe that. But I think it is important to understand things like there are often things that you end up believing in that other people don't believe in. I think that those are really important beliefs to have, and then to always check yourself on them.

I think if you believe something that everyone else believes, you need to check yourself on that. If you believe in something that very few other people believe, you also need to check yourself on that. Whenever I find myself believing everything that everyone around me believes, I try to really question it. You can't question everything. I don't spend a lot of time questioning whether the world is round. I don't actually really know is the world is round.

I don't like spend time questioning if we faked the moon landing or something like that. I'm not 100% sure, but I'm pretty sure that was landing on the moon, but I can't prove it. I wasn't there. I wasn't even alive in 1969. You can't question every single belief. If you find yourself believing something with high certainty, like yesterday was the Super Bowl, and if you find yourself believing in the second quarter, with a high certainty, that the Falcons were going to win and you were 100% sure, this is probably something you should probably check yourself.

Do you believing it, because everyone else is around? Or do you believe it because these are just certain pattern matching that you've done before, et cetera? Certainly, they're probably more likely to win at that point, but they probably weren't 100% likely to win.

[0:30:16.9] JM: It certainly seems like we are moving towards a place where more people are picking up non-obvious ideas, even non-obvious ideas, both that are productive to pick up and ones that are on the extremes. I don't know if this is Twitter, or other social networks presenting this to me and maybe this is always been the case, that people are picking up things like Pizzagate, or faking the moon landing and running with it.

Does it feel like there is a degree of societal variance that increases dangerously when people pick up non-obvious ideas? Obviously, I sympathize with your, and I find it very productive to pick up non-obvious ideas and run with them to a safe degree, but it seems like there is an

atmosphere, whether it's just the way that the internet is projecting this to me or if it's real, that many people are kind of losing a shared sense of reality. Whether that's a good thing or not, I'm not entirely sure, but it certainly feels like something is in the air. I don't know if you agree with. To what degree do you think that's dangerous?

[0:31:51.3] AH: I'm not sure that I do think there's probably more social conformity today than there was in the past. I think it's probably — I still think we're kind of on the extreme end of social conformity of like certain — Especially if you think of — I don't know that much about society. So I don't know as much about politics, but if you think about business and stuff, I think probably too often, people want to fund things that everybody else wants to fund. Too often, people will say, "Okay, AdTech is really bad. We shouldn't go fund that."

Everyone just repeats it blindly, but they don't actually go in and really dive in as to like, "Okay. What's good about it? What's bad? There's certainly been tons of really great exits in AdTech. Okay, there's a lot of data that it's good. There are some people saying it's bad. Okay. Why are they saying it's bad?" Really diving into it. I think, too often, people are just like very willing to conform to some sort of basic thinking. I also think people are — There's a lot of pushback on people when they say things that are even like somewhat controversial. Certainly, in business, that's the case so it just causes people not to say things that are controversial.

Jessica Livingston who is one of the founders of Y Combinator, I think one of the really, really smart minds, penned a really interesting piece about a month ago about essentially some of the biggest problems she believes is that people not saying things because they're worried about backlash, or et cetera. I have a lot of friends that before they say anything interesting, they'll spend anywhere between three to six months testing it on their closets friends, people who won't judge them, et cetera. They'll be tweaking it, et cetera, before they're willing to say it in public.

Sometimes this is good because certainly don't want to say things that can radically offend people, et cetera. Sometimes it's good. Of course, this could be bad, because this self-censorship could mean a lot of new ideas aren't coming out and it's just wild, and a lot of things aren't being shared, et cetera.

[0:33:32.3] JM: Yeah. That's the tradeoff, because in some sense it's exciting, the idea that maybe we could move to a place where people will be less restrained in how open they are, but it also is — I guess it's disconcerting to me, just because in my lifetime, the widespread conformity is something that I'm used to. Moving quickly beyond that, if the internet can enable that, is somewhat disconcerting.

[0:33:58.5] AH: By the way, I think it is really important — You can't question everything. Again, I don't question whether the world is round or not. There's some believes you just need to have. You do need to be open to —

[0:34:10.2] JM: But if we can't come to a consensus about that as a society — If we have these big riffs, like as open ideas profligate through society.

[0:34:20.9] AH: I think if there was some good data about the world not being round, we should be open to the fact that it isn't. If there are some data that it was actually flat, or something, or it was round in the past, but now it's gotten flatter somehow through gravitational pull or something. I don't know. I think we should open to hearing about that and thinking about that. If some credible — Or someone who we think is smart, has that idea, be open — It's like, "Oh, that's kind of interesting."

After you listen to them, you're like, "I still don't agree. I still think it's round." That's cool. I love hearing about those types of things, but you can't — Again, if you question everything, you become Unabomber.

[0:35:04.2] JM: That's right.

[0:35:05.4] AH: This is a very difficult world to live in. You can't question everything, but I think you should question some core things relatively often. You also can't question everything all the time. If you question like — Let's say you question your core beliefs, once a year is probably good, but if you question them every day, this could be a real problem.

[0:35:25.1] JM: One of the tragedies of the comments I see today, I think particularly about with the machine learning, and I think you'll agree with me here, is that these public datasets are

hard to access. We have this taboo around sharing medical data. For example, people get very caught up in concerns about health insurance rates that might change if their data gets de-anonymized. We also have these concerns about privacy. People seem more concern with those things than the potential upside that they could provide.

There is tension between the gains that we can make from open data and then the societal objections to that open data. How do you think these societal values will change with regard to datasets? What are your projections with how long that will take?

[0:36:11.7] AH: I think these concerns are valid.

[0:36:13.5] JM: They are. Sure.

[0:36:14.1] AH: Yeah. I certainly don't want insurance companies making decisions about me, or about my loved ones based on my medical history, or something like that. I certainly wouldn't want to get it out that I went to the doctor about a certain thing and get that out in the public and have people writing about it., or my tax returns, or something like that, making those public. I'm not public official. I don't want my tax returns public.

There are lots of things that I would personally want to keep safe, and I think that most people presumption is that they want to keep safe, and I think that's probably a good thing about society. Even what I have for breakfast this morning was probably something I don't want people — By the way, it was oatmeal. Now I'm telegraphing that, but I don't want to be able to choose to telegraph that to people, rather than someone else telegraphing it without my choice to people.

I think these privacy concerns are incredibly real, and they should be real, and I think people should be very concerned about their privacy, and that I think we're going to have more and more applications to be able to protect our privacy. I think we're going to see more and more consumers thinking about privacy. One of my crazy beliefs about privacy is that I think old people like me, let's say, like Gen X-ers, care about privacy the least. I think Gen X-ers care about privacy the least out of all the generations, and that the younger people, the millennials,

or the Gen Y-ers, or Gen Z-ers, or whatever we call them, they care about privacy way more than the Gen X-ers.

Certainly, they change their privacy settings in Facebook a lot more, maybe it's because they're more sophisticated about how they use Facebook, but they certainly use a lot more. They've adopted other privacy-centric social networks, like WhatsApp, or Snapchat much more than the Gen X-ers. I think they do care about it quite a bit. I talk to my nephews and nieces who are in high school and stuff, they'll have six Instagram accounts, where they're sharing with different types of people, and so they're well-aware of different privacy implications of how they share it, or what they're doing, et cetera.

I think we'll see privacy on an upswing. I think there will be many more applications that people will care about. I think more and more people will use applications like Secret to do their communications, because they'll care about privacy. I think there will be a lot of really pro-privacy things that happen over the next 10 years.

There's this believe that there's this inevitability that privacy is dead, or get over it on the privacy front. I think there'll be this tension that happens for a very long time, and between that — Certainly, if you're getting benefit, you may be willing to share. You'll also willing to share with certain actors, but maybe not other actors, depending on how you believe they're good stewards of your data. I think it's incumbent upon all actors that have data to be really good stewards of that data as well.

[SPONSOR BREAK]

[0:39:23.5] JM: Simplify continuous delivery with GoCD, the on-premise, open source, continuous delivery tool by ThoughtWorks. With GoCD, you can easily model complex deployment workflows using pipelines, and you can visualize them end to end with its value stream map. You get complete visibility into, and control of your company's deployments.

At gocd.io/sedaily, you can find out how to bring continuous delivery to your teams. Say goodbye to deployment panic and hello to consistent, predictable deliveries. Visit gocd.io/sedaily to learn more about GoCD. Commercial support and enterprise add-ons, including

disaster recovery, are available. Thank you to GoCD and thank you to ThoughtWorks. I'm a huge fan of ThoughtWorks and their products including GoCD, and we're fans of continuous delivery. Check out gocd.io.sedaily.

[INTERVIEW CONTINUED]

[0:40:36.5] JM: You don't think that we necessarily need to have some changes in the norms in order to get to this place where we have enough data to do the kinds of productive machine learning that we talked about at the beginning of the conversation.

[0:40:51.4] AH: I think there's a lot of things we could be doing in the technology front to be better stewards on privacy. A lot of what I think about every day is how to maybe a better stewards of privacy. I know that's a lot of what a lot of companies think about every day is how to be better stewards of privacy. I think this is a good thing. Now, there is tension and I think there's a real tension that we should be thinking about, but I don't think there's an either or.

[0:41:15:5] JM: You can have utility with privacy.

[0:41:17.7] AH: Yeah. I think this is also even true — There's this other public policy debate between civil liberty and security. I think this is kind of this false choice. I think you can have both if you designed it in the right way. Often times, you have neither. There's this point where you could have neither too. This is the worst scenario when you have neither civil liberty, nor security. The best scenario is you have both. I think sometimes people design systems so you have neither. This is like always the worst place to be is to have a neither.

[0:41:52.2] JM: You write that computers are not learning fast enough. What are the costs to our society if we don't get computers learning fast enough?

[0:42:03.2] AH: First of all, I think computers are learning quite fast, and I think there's a lot of really interesting things that they are doing. But in a lot of the areas that they've been focused on are in areas where data is available. In a lot of the areas that data is available, it's kind of interesting, like chess or something, it's kind of interesting. Or even the financial market, it's kind of interesting. I don't know it really helps society.

In the areas where the data is less available is often in the areas where we could help society a lot more, and we could things that are really, really beneficial to society. When you mentioned this nutrition thing, this is an area where there is no data available. I don't expect there to be data available.

[0:42:45.0] JM: We know you had oatmeal.

[0:42:46.2] AH: Yeah, we know I've got oatmeal, but we don't know what I had yesterday, or the day before, or etcetera. This is an area where probably we're not going to make a lot of progress. Certainly, we would be able to benefit a lot. We don't actually know, like people talk about things like is wine good for you? We have no idea.

We don't even know if broccoli is good for you. We have no idea that broccoli is good for us. There's just no real good data to suggest that it is. It probably is. I eat my broccoli. It probably is. I wouldn't like to tell your listeners to like just go on a chocolate cake only diet, or something. It's probably not a smart idea. We don't have a lot of data to suggest that it is good for us.

I think as we get more data — Nutrition is not an area that I don't expect to have a lot of advances in the next decade, but there are a lot of other places where I think we can have advances in. I think we can get at a lot of really interesting questions that have been vexing society for a really long time. With more data, I think could go do that.

[0:43:47.3] JM: We've mostly been talking about how — We have been talking about how machine learning affects things, because we are — From a technical perspective, because we are technical people. This is also going to affect the blue collar class. In the near term, I can imagine a lot of jobs that are somewhere between a Mechanical Turk and what an accountant does, because there's these kind of midlevel — Things that are somewhere between the Turk style tax and knowledge work.

These are things like translation between two different languages, or some abstract image identification. There's going to be a lot of these jobs, because the human in the loop computing is obviously necessary for training a lot of the machine learning models, at least in our current

paradigm. How do you think this blue collar knowledge work is going to evolve? Is it going to be enough to sustain a lower middle income class?

[0:44:48.9] AH:I think there's lot of creative distraction that's going to happen. If you're thinking, like if you're someone who cares about society, this is definitely something that you should be thinking about and worrying about. I think there are some good things about it, but there's also a lot of bad things and a lot of displacement that's going to happen.

I think there's a lot of people who've been writing about this displacement of the blue collar worker, which I think there will be a lot, and there already has been. That has happened a lot from things like outsourcing already. Some of it is due to technology, but a lot of it's also due to foreign competition to them or just kind of other types of things that have been happening there, but some with robots, et cetera.

I think we're also going to see a massive trend to displace the white collar worker, that college educated worker. I think we're going to see lots of college educated workers who are going to be underemployed in the future. The problem is it's really hard to predict which ones. It's not just the college educated worker that has the two year college degree, or even a four year college degree, it could be just someone who has like the 10 year college degree. We're talking about like the radiologists, et cetera, who literally studied for 10 years, who is top of their high school class valedictorian in high school.

[0:46:45.4] JM: Or the sociologist, the history professor.

[0:46:47.7] AH: Yeah, it could be a lot of really, really impressive people. The person like when you were in high school, you're like, "Wow! This was the smartest person. They got all the best grades. They worked the hardest. They did everything right. Everything society asked for them. They didn't take drugs. They did all the things that — All the things like society is supposed to do.

Now, all of a sudden, through no fault of their own, they just couldn't predict the future perfectly. They didn't know where computers are going. By the way, none of us know where the future is going, and they didn't predict where the future was going and now all of a sudden, this job that

they were doing, this incredible job that they are banking on, that for decades and decades was incredibly good, stable job. Now, all of a sudden, gets replaced through computers. I think we're going to see lots of these happening.

Now, this person who is making in the top 10% is now making in the top 50%. Maybe they still have a job, they're still relatively smart person, but their entire life has to be changed about how they've been thinking about it. I think we're going to see lots of these happening as well, and not enough people are talking about that. I do believe the blue collar piece is really important, and we should be talking about that, and I think we are.

I think this is now like a common thing, the lexicon, but this white collar college educated thing — and this is why, I think, in society right now, we're trying to push people to go to "college" without actually having a plan for their future. I don't think this going to help things. If you go to college and you spend four years at college is not working. You not have some sort of expectation of a much higher income, et cetera.

I don't know that you're going to have a higher income if you go to college, if you don't pick good future path in the future. Of course, it's very hard to know what computers are going to do. I think this is something that we should be very worried about as a society, because we're going to have lots of displacement that's going to happen.

I think the lawyer is going to get displaced, the accountant will get displaced, a lot of med people in the medical profession will get displaced. These are areas where if you are apparent, you'd be like pushing your kids to go into, and you'd be so proud that your son or daughter became a lawyer and you'd be bragging to all your friends, and then it's possible that now this person is now going to be in the unemployment line, because they picked that path.

[0:48:24.9] JM: This gets at the question of narrow A.I. versus generalized A.I, which we've been discussing on a lot of episodes recently. We wouldn't think that if a computer can play poker and if a computer can take all the tasks of an accountant, or a doctor, it seems like this qualifies as generalized A.I. in some sense, because in order to do all the tasks of a doctor, you perhaps have to have really good natural language understanding. You have to have good

ability to communicate. You have to understand a large corpus of knowledge, and yet we keep repurposing that generalized A.I. term.

Have you thought deeply about what your definition of generalized A.I. is? Do you think it's a false dichotomy between generalized A.I. and narrow A.I.?

[0:49:09.0] AH: I don't. We've spent too much time on the definition. At least the way I think about it, I don't think we're going to be hitting general A.I for quite a long time. I think we're going to have these narrow things. I think there's going to be a bit role for humans to play in it. I do think the human lawyer is going to be working with software, the human radiologist is going to be working with software more.

In some cases, it's going to lead to displacement, because they're going to be out of their job. In other cases, they're going to just be a lot better at their job and they're going to end up being paying more. What's going to lead you — It's not like all the lawyers are going to go away. We're going to have tons of lawyers and it's still going to be a really good field to go into for some people. We're just going to probably have fewer lawyers, and the ones that are there are going to get paid a lot more. Then, there's going to be a lot of other people who are going to get paid a lot less.

We're going to get these barbells that happen way more often in all of these professions that happen out there, because it's some sort of narrow A.I, or machine learning, or computers, or whatever you want to call it. It's been going on for years. It's just now more apparent to people than it was before.

It's not just A.I., it could just be a tool, like Excel, or some sort of — Any tool that you could be using as — You're going to have getting people way, way more efficient. They're going to be able to take on more tasks, do more things, and there may be need for less people. As we lead to a globalized world, you're going to be able to — There's still maybe some people who are not sure about whether or not we're moving more and more to a globalized world.

As you move to a world where you're going to — You can increasingly be competitive in many, many different markets because of technology. It makes sense that these winners are going to win really, really big, and the people who don't win are going to have much more difficult lives.

I think we're even seeing this in software development. In software development, there's a been a tide, which has lifted all boats. We're seeing salaries for all software developers rise, but we're seeing the rise of the best software developers grow at a much faster rate than the rise of just the okay software developers.

I think there's this whole other kind of place in society where they're trying to get all these people to code, and all these things to do. I don't know that this is a smart move. I think this is a short term good thing. You'll probably be able to get a job in the short term. This might be a very good short term thing next five years, next 10 years. It seems like it's probably a good thing to be able to — it's probably a very good skill to have to be able to code.

But long term, you could have a scenario where someone who's making very good living, they're making $150,000 a year, they're making an incredible living, you could easily see that job go from $150 to $50. There may be something else they could have, like repair trucks, or something like that, become a mechanic where they could have been making $100k in between or something like that and much more where technology might have not been able to jump in as quickly.

[0:52:01.5] JM: Okay. To wrap up, there's something that's kind of been pervading the conversation we've been having, and something that I see in my everyday life. There seems to be a growing gulf between technologists and non-technologist, or maybe you want to call it globalists versus non globalists. How do you bridge the gap there? When I'm talking to my friends who I went to high school with who seem anti-technologic, like anti-Uber narrative. Do you sense this in your everyday life? Does it seem like a reality to you, a growing reality?

[0:52:39.9] AH: I have not sensed that as much. I think there's a very pro-technologist strain. I think people yearning for more technology. Most people are using technology all the time, especially in the U.S.

[0:52:51.2] JM: You're saying across the world.

[0:52:52.5] AH: At least in the U.S. People are using it to get their entertainment. They're using it to get their groceries. They're using it to get their transportation. They're using it to get their education. They're using it to get their housings. If you just think of like all of the core things that we do and we spend money on, we spend our time on, technology is pervading that and people are happy with that.

Now, there is lots of problems because of the displacement that we talked about, and so I think we're going to see lots of displacement happen because of it. If you're only benefitting from it and you're not getting displaced, you like it, or if you're getting displaced from one thing, but you're benefitting from the other stuff, or you don't like what you're getting displaced on, et cetera. If you're a truck driver, you might be worried about self-driving cars, but you love the fact that you can listen to your podcast while you're driving the truck.

There's all these interesting tensions. By the way, I think this is going to be true for all of us. All of us have the ability to get displaced. Every one of us, every one of our professions have the ability to get displaced. Some of us are going to be doing things like investing in ourself, so we're less likely to get displaced. Some of us are going to be investing in regulatory things, so we're unlikely to get displaced. Some of us are going to be changing careers. There are all these different things that we're going to be doing.

Some of us are just going to be okay with getting displaced and having a different type of life. This has happened over large periods of time throughout history. In general, I think we're going to have — We're going to spend more and more of our time in technology today. Most of us are already spend close to 100% of our waking hours with technology.

[0:54:32.0] JM: Yeah.

[0:54:34.8] AH: It is a huge piece of time, and it's already augmenting a lot of stuff that we're doing. People talk about augmented reality. We already have that. It's already been a huge augmented thing, and it seems — and this might be where I'm way too conventional in my thinking, so I probably should check my thinking. It seems like this is going to continue to

happen. We're going to continue to have technology augmenting our reality in every single scenario, in every single thing that we're doing in the future.

[0:55:02.6] JM: Auren, thanks for coming on Software Engineering Daily.

[0:55:04.5] AH: Thank you, Jeff.

[END OF INTERVIEW]

[0:55:10.0] JM: Listeners have asked how they can support Software Engineering Daily. Please write us a review on iTunes, or share your favorite episodes on Twitter and Facebook. Follow us on Twitter @software_daily, or on our Facebook group, called Software Engineering Daily.

Please join our Slack channel and subscribe to our newsletter at softwareengineeringdaily.com, and you can always e-mail me, jeff@softwareengineeringdaily.com if you're a listener and you want to give your feedback, or ideas for shows, or your criticism. I'd love to hear from you.

Of course, if you're interested in sponsoring Software Engineering Daily, please reach out. You can e-mail me at jeff@softwareengineeringdaily.com. Thanks again for listening to this show.

[END]