EPISODE 888

[INTRODUCTION]

[00:00:00] JM: Data as a service businesses offer paid access to datasets. These datasets can be useful for building products or training machine learning models. There has been steady growth in the tools and practices around processing and storing data. For example, we have Apache Spark. We have tools like Snowflake data warehouse. We have lots of tools for manipulating data at scale, but access to datasets remains a bottleneck for widespread development of machine learning applications in a large set of domains.

SafeGraph is a company focused on the problem of data as a service. Today SafeGraph's primary product is reliable, up-to-date location information about places. The data for these points of interest needs to be acquired, verified, cleaned, made accessible through an API and intelligently priced.

In previous episodes with SafeGraph, we've explored the basic premise of data businesses and why they're important platforms for building futuristic data products that are impossible for entrepreneurs to build today. SafeGraph's CEO, Auren Hoffman, returns to the show to discuss the data as a service business model.

Software as a service has existed as a category for more than a decade. Infrastructure as a service has existed for about just as long. Data as a service is much more undeveloped. Auren recently published the *Data as a Service Bible: Everything You Wanted to Know About Running a Data as a Service Business*, and this was a very useful article. It breaks down in considerable detail a category of software that is almost entirely unexplored today, which is dated as a service, and this is something that is in a long line of content from Auren Hoffman that I found pretty useful. He has always been a great writer that I've drawn a lot of inspiration from around business, and strategy, and philosophy. So I encourage checking out that article as well as the previous interviews that we've done with Auren on the podcast. I hope you enjoy this episode.

[SPONSOR MESSAGE]

[00:02:22] JM: Today's shows sponsored by Datadog, a monitoring and analytics platform that integrates with more than 250 technologies including AWS, Kubernetes and Lambda. Datadog unites metrics, traces and logs in one platform so that you can get full visibility into your infrastructure in your applications.

Check out new features like trace search and analytics for rapid insights into high cardinality data, and Watchdog, an auto detection engine that alerts you to performance anomalies across your applications. Datadog makes it easy for teams to monitor every layer of their stack in one place, but don't take our word for it, you can start a free trial today and Datadog will send you a t-shirt for free at softwareengineeringdaily.com/datadog. To get that T-shirt and your free Datadog trial, go to softwareengineeringdaily.com/datadog.

[INTERVIEW]

[00:03:28] JM: Auren Hoffman, welcome back to Software Engineering Daily.

[00:03:31] AH: Thank you, Jeff. Happy to be here.

[00:03:33] JM: Data as a service is a software category that is fairly undeveloped. Why is that?

[00:03:40] AH: Historically, it has been a pretty bad business, and that's because there's just not that many companies in the past that have been able to buy external data and then actually use that data to actually do something with it, and that's because in the past it's been actually pretty hard to like actually make use of some sort of like external data. If you think of like where companies really focused on in the last 10 years, it's really about using my own data and making my own data better and making some sort of like core decisions using my own data.

Almost all the companies that you think about using data from like Tableau, and Looker, to Spark and Hadoop. Just go down the list. It was really about how to use my own data better. Now we're just getting to a point where companies or a lot of organizations are just getting really good about gathering their own data, finding really interesting intelligence from their own data, and then these companies realizable, "Well, I only know about .1% of the world, or .01% of the world," even huge companies like Walmart know about a very, very tiny fraction about the world.

Unless you're like Google, Facebook, or Amazon, you really just don't know that much. So you want to bring in some external data to help you get a broader opinion. So we're seeing just now the ability for data companies to actually make money, because they have more buyers that they can sell to.

[00:05:06] JM: And you're building one of these data companies. What is hard about that?

[00:05:10] AH: Well, one hard thing about a data company is what is data company really? So data company is really just a collection of facts, right? So if you just think of a database with rows and columns, one thing is really important is that you have enough rows to cover the data that a client might care about. You have columns which are kind of a description of that entity that clients care about, and then hopefully as many of those tuples are filled, and then hopefully the data in those tuples are accurate. So all those things tend to be quite hard to do.

Just think about it, about anything that you'd want to have data on. So let's say you want to have data on like the price of Volkswagens in the United States. Okay. Well, you need to make sure that we know all the different Volkswagen models and stuff, and then somehow we have to figure out how do we find out when these things sold and are we just doing the price of new Volkswagens? Are doing the price also of secondary and used transactions? How do we gather the data? What is the – Oh, the price is in dollars. All these different things that we really have to do.

So that would be a fairly valuable narrow dataset, and then maybe, "Okay. Actually, I have the price of Volkswagens. Maybe I'd expand to the price of Audis, or something like that. Then eventually I'm going beyond cars, to airplanes, or trains, or something."

[00:06:31] JM: What kinds of successful data businesses have been built in the past?

[00:06:35] AH: There's been view massively successful data businesses that have been built in the last 20 years. By very few, it may be zero. So there're been really very few essential unicorns, companies worth over a billion dollars that just sell data, and that's because it's just been incredibly difficult to sell just raw data.

So almost all companies that had proprietary data in the last 20 years, instead of selling the data, they built applications around it and they actually sold the application and the data was a core feature of the application, but they didn't sell the data, because it's very difficult for clients to use that. I think just now we're getting to a point where we're starting to see more companies be able to sell data. If you think about – If we take a step back, like one of the ways I think to track innovation is to track where the data is available.

So if we're going to think of like where the innovation will be in the future, we could probably figure out where the innovation will be in the next 10 years by where the data is available today. So if we want to go track innovation, so if we just think about, "Okay, where was the big innovation in like the 80s and early 90s? Okay, it was in chess. Okay, why was it in chess? Because the data was available." The data from every major and most minor chess games is available. All the notation is standardized. The data is actually very small. The average chess game has 40 something moves. So at the time, that was very important not to have huge datasets as well, and it was really easy to categorize, and essentially the data was cheaper, close to free.

That's where we saw the innovation in kind of the 80s through the early 90s. Then where the innovation go to after that? Well, the innovation went to the stock market. Why did the innovation go to the stock market? Well, price particular per time data is available. You can go back over a hundred years and get a tick for AT&T. Now maybe the tick 100 years ago was by day, and the tick today is the 10th of a second, but the data is available going back decades. It's really good data. There're probably some typos in the data, but it's probably 99% accurate, that data. So you can back test that data. Then of course it's temporal, so you can get the data going forward as well. So look at where the data is and we can understand innovation.

So, one thing to figure out, where the innovation won't be in the next 10 years? For instance, it's really hard to get data about nutrition. So if you wanted to figure out like what is good for you to eat. Should I have a paleo diet? Should I be vegan? It's really hard, because getting the data about all these inputs, about what people take in, is really hard. How does that interact with your DNA? Your eventual health outcome could be like 50 years away. Then of course this is tied to your wealth and tied to how often you workout and all these other types of datasets.

So I can, with certainty, predict that in 10 years from now we'll still have fad diets, and we still will not know anything. I mean, we don't know today, honestly, if chocolate cake is good for you or not, or if broccoli is good for you or not. We have like some high-level ideas, but we really don't know anything. Now, in other areas, I think in oncology. I think we are making huge strides. The datasets are starting to be available in oncology. They're labeled pretty well in oncology. They could be better, but they're labeled pretty well. The outcomes for oncology are usually 2 to 4 years later. So you get to get a rapid improvement and kind of what's going on. So I think we're going to see huge strides over the next 10, 15 years in personalized medicine and dealing with cancer treatments and things like that.

[00:10:27] JM: SafeGraph has started by focusing on place data, which is location data. How would the engineering problems of handling oncology data compared to location data?

[00:10:40] AH: I think the engineering challenges – And I don't know enough about oncology data. So I'm sure the engineering challenges there are vast and very, very difficult. In every single dataset, they have their own weird nuances. So on oncology data, I'm sure every single type of procedure is coded in some sort of way, and maybe the code – I don't know enough about it. Maybe the codes are standardized in the United States, and then you have to map that to different standardization in the UK or something. But maybe they're not standardized. I don't know enough to know it. Maybe even when they are standardized, they're done by hand. So there's all these different like labeling errors and stuff that could happen in the data.

So in each dataset, you have all of your own unique issues with it. So just figuring out, like just think of place data, which is what SafeGraph does. Just figuring out, "Does that Starbucks on 555 Main Street open on 8 AM on a Tuesday or open at 9 AM on a Tuesday?" Just figuring out that is actually really hard. It sounds pretty easy, but it's actually hard. Then you have all these weird corner cases. Well, that Tuesday turns out to be Christmas. For Christmas, it actually only opens three hours for the whole Christmas Day, but another day, there's like some special and it's open for like 24 hours that day. So you have all these weird kind of corner cases that start to come in, and it's difficult and it's really hard. If the truth is really important, then you are constantly struggling about that. That Christmas case is not like a random use case. Like a lot of local search use the SafeGraph data.

Last Christmas, we sent a whole bunch of people to places that were closed, and those people got extremely angry at all the different local search providers that we sell data to. Of course, those guys got angry at us as they legitimately should, because we screwed up. So it's really hard – When you're a data company, you're really just – You should be judged as a collection of facts. That's what you are. That motto of SafeGraph is that we predict the past is our motto. So you're a collection, you're an archivist and your goal is to get things right, and you'll never be 100% right. It is impossible to be 100% right, but you need to be extremely upset at yourself when you're not right, and you should be doing everything you can to correct it overtime.

[00:13:03] JM: This was one perhaps fault of some of the older data companies, was that they had a bias towards religion rather than truth. You've articulated this spectrum between religion and truth in data companies. Could you illuminate that more?

[00:13:21] AH: Yeah. I wouldn't say that's a good or a bad thing. So I think there're companies in the data world that are backward looking. That's not a bad thing. So that's what SafeGraph is. We are trying to figure out what happened in the past. Maybe the past is like a second ago, but we're trying to figure out what happened in the past. Those I would say are truth companies. They're trying to figure out the truth.

There are separate companies, I wouldn't say necessarily they're data companies. They usually take data as an input, and they're trying to figure out what will happen in the future. So they're making some sort of prediction with different data to figure out. So your credit score is a prediction company, right? They're taking in, "Did you pay off your loans in the past?" All these different types of data. So they're taking data, let's say, from Experian, which would be a truth company in the past. Did you pay off your loan? Did you pay off your phone bill? Etc. Then they tried to then imply a credit score for you for the future as to whether you will pay back your – It's a religion. The reason why it's a religion is because people argue about it. The credit score doesn't really accurately define me, or this doesn't accurately define this, etc.

Some companies are trying to do a little bit of both. So even your SAT test is trying to be a predictive score of how you will achieve in the future, but it's really hard to know how you will achieve in the future, because how one you achieved in the past may not be a predictor. You might be a late bloomer. The whole world might change, and SAT tests measure how good you

are on certain types of things, but maybe those things are not as important in the future, and actually being a great artist is really important in the future, and SAT doesn't. So the SAT test is trying to be a predictor of people's aptitude in the future, and people argue about how good of a job it does on that.

[00:15:07] JM: The five areas of data that you've deemed as viable data business opportunities; people, products, places, companies and procedures. Why do you divide up the world into these five types of data?

[00:15:24] AH: I think those are the five most common. You could probably come up with 500 different types of things, and then even with those five, you can layer on time. So you can have some sort of temporal element to things. You can layer on other types of things like price, is usually common thing that people may layer on as well. But if you're having a data company, you don't want to start a data company about everything. It's like, "I have data about everything in the world." That would be really hard. You need some sort of narrow focus so that you can be correct about what you're doing, and then it's helpful to have some sort of key, some sort of common key, and that could be either a primary key or a foreign key, but some sort of common key where you can start to link all your different datasets together.

So if you think of a person. Okay, well, there's a key to a person. So maybe Jeff has an email address that's a key. Maybe name plus address is a key. Maybe your telephone number is a decent key. You've got a Social Security number. You've got a fingerprint. I don't know. You've got all these different types of things, your DNA strand. There're lots of different things that define you as a key. Some of them don't perfectly define you. You might share that email address with somebody else. You might share an address with another person. There might be other things that don't perfectly define you, but there may be some sort of semblance of a key or a series of things together will equal a key about you.

A place is similar and organization is similar, a product. So if we go back to Volkswagen, okay, there're these specific type of Volkswagen, a Volkswagen Passat or something like that. But then there's also the VIN, right? The VIN is this exact Volkswagen. There's only one car in the world has this exact VIN. So it's not even just a type of car. It's an individualized car. It's like a

Social Security number, and that VIN defines a lot of this, and then we can track this VIN overtime, the price of this VIN overtime, etc.

[SPONSOR MESSAGE]

[00:17:28] JM: At the beginning of 2019, we had problems with Software Daily, which is our custom-built website and mobile app set. The website was not engineered properly, and our iOS app was buggy. Everything needed a redesign. To help us refactor our cross-platform application, we brought in Altology. Altology a full stack software engineering firm that helps innovators build worthwhile products.

Altology will help you get your project or your company to where you want them to be. They can rescue your project. They can augment your team. They can help you get a new version of your product out the door. If you're building a brand-new product from scratch, Altology can also design and develop web and mobile products that are brand-new.

The Altology team is entrepreneurial. They're design-focused, and they're able to work across the stack. To get help with your engineering projects, check out Altology today by going to all altology.com. That's altology.com.

Thank you to the Altology team for helping us get Software Daily to where it is today, and for being continued friends of the show. If you need help with your application, check out altology.com.

[INTERVIEW CONTINUED]

[00:18:59] JM: The veracity of data is an important facet that you need to ensure as a data company. This is very hard, because there're a lot of rows. How do you ensure the correctness or the truth of data?

[00:19:15] AH: Well, first of all, starting with just the fact that having some humility is really important. It's really hard to know if you're right. So you should assume that you've got a long way to go to being correct, and people will label their daily life, "We're 99% accurate," and you

just should be skeptical of those types of claims. Then it's a constant struggle to continue to be correct.

So even if you are right about things like things change, any good dataset is a dataset that changes. Otherwise, it has very little value. So a dataset that has less value would be – I don't know, pictures labeled as cats or no cats. That really has very little value overtime. Once you use that data, you may never need it again. So a good dataset would be something that is changing overtime. For instance, the attributes about Jeff change over time. Maybe at some point you're single, and then you become married, or at some point you live in San Francisco, and then you live in Seattle, or at some point – I mean, people's genders even change. So lots of things about people change overtime. So you really want to understand. Then there might be a core thing about you that doesn't change, like your birthdate might not change. But it still might not be right.

So we might need to go find the birthdate about people. By the way, people not only have birthdates. They have end dates. People eventually die, and you have all these different things about people that are really important. So just having the ability to understand that you're going to be wrong a lot is important.

Okay. How do I figure out when I'm wrong? So how am I testing if I'm right or not? What am I doing? How am I getting feedback? How am I getting feedback when I'm wrong about something, but also how am I getting feedback when I'm right? How am I measuring my precision? How am I measuring my recall? Even doing that becomes really, really difficult. Even the measurement of your precision and recall tends to be really, really hard. Then how do I make those better overtime? Then how do I get the feedback in? What's my rate of trying to improve? All these other things tend to be a struggle at a data company.

[00:21:21] JM: How do you organize engineering within SafeGraph?

[00:21:24] AH: What do you mean?

[00:21:25] JM: Like in terms of teams, or departments, or role types.

[00:21:31] AH: Well, at SafeGraph, we're dealing with a lot of data. So the typical engineer at SafeGraph is somebody that has a history of dealing with a lot of data and being able to – So we use Spark, which is like any company. I guess a lot of companies started in 2016 would probably use the traditional Spark stack that we use. So someone who has some sort of ability to feel comfortable with Spark, or maybe they use like a predecessor environment like Hadoop or something like that. So they feel comfortable with a lot of data.

Then we have to like model that data. So there's a decent amount of machine learning that has to happen on that data. Then there's a matching problem. Usually, when you're doing like merging of data, you're going to have like some matching problem, because you may – If you think of like, "Okay. I learned about – If have data about people, going back to the people dataset, I've got some fragment of Jeff on one dataset. I've got a fragment of Jeff on another, and then I've got to figure out, "Is this the same Jeff? Should I merge it?"

Obviously, the merge can be really disastrous if I'm doing a bad merge. But if I don't do the merge, it also could be bad, because now I'm representing that there's two Jeffs, when actually there's one Jeff. We have data about a place. So let's say we've got a data about that Starbucks on 555 Main Street. We get data today from 13,000 data sources. That Starbucks on 555 Main Street might be in 20 of the 13,000 data sources.

So making sure we do that merge correctly is really important, and we don't want to merge 15 out of 20. But we also don't want to merge 25 out of 20, right? So how do we make sure we're getting as many of those 20 out of 20 as possible? Getting those in there? Then of course once you do the merge, now you're getting all the conflicting data that happens.

So one data source says the Starbucks opens at 8AM. The other source says it's open at 9 AM. Now you have to figure out, "Okay. What's correct, or what's the most likely to be true so we can make some sort of assessment of that."

[00:23:28] JM: With the importance of those merging operations that can be risky tear datasets, I can imagine that versioning might be important. Have you had to invent any strategies around versioning your data so that you could roll back in case a dataset gets corrupted?

[00:23:45] AH: Yeah. Well, one good thing is if you're getting data from many sources, is you have the data from like the individual place and you're keeping that data from the individual place. Hopefully you're keeping it well. So your merge algorithms may change overtime right.

You can imagine like an aggressive merge that all the sudden merges all your entities into one entity. So if you think of like let's say you're tracking all the people in the world and all of a sudden instead of 7 billion, you have 1 or something, right? It's totally possible that you could do that very, very quickly.

Of course, instead of 7 billion people, you could also have like 7 trillion people too. You can make a merge the doesn't work as well. A lot of people will say like, "Okay. The visitors to a website –" A lot of people have like their visitors to their website are like 100 billion people a month or something like that. So, those stuff. I mean, yes, there are a lot of bots out there. So maybe there legitimately are a hundred billion different entities that are coming to their websites. But they're also just not doing the merges in a way that makes sense for them.

[00:24:47] JM: You've mentioned Spark. You've also mentioned in a written post, I believe, that Snowflake is a very powerful tool in terms of your usage at SafeGraph.

[00:24:56] AH: We don't actually use them at SafeGraph itself.

[00:24:58] JM: You don't use Snowflake.

[00:24:59] AH: We don't use Snowflake. Maybe we should. I mean, they're a great company. A lot of other companies I know use snowflake. I'm a huge fan of their product, but we don't use them at SafeGraph.

[00:25:10] JM: How has your criteria for choosing software vendors changed in the last year?

[00:25:17] AH: Well, I don't know in the last year or so, but we've been much more aggressive about using software vendors for everything. So not just for engineering stack, but for marketing, for sales, for products, for just general productivity, for recruiting. We've been much more aggressive by using different software vendors.

Our belief is, the more you can leverage by using all these different APIs and software tools, that means you can focus your employees on just working on like the really hard problems and you don't have to rebuild something that's already done. That means you can also recruit fewer people. So you can keep your team smaller. You can iterate faster.

One of the things we track at SafeGraph is revenue per employee, and we think that's an important metric to understand. The leverage of the company, of how we're getting leveraged overtime. You have a lot of these B2B companies, like they'll do like hundred million – They're doing 100 million ARR and they have a thousand people. So let's say that's a 100,000 per person. Then they do 200 million ARR and they have 2,000 people. They're just not getting any more leverage as they grow.

Whereas if you look at like a great company, like Google, or a Facebook or something like that, every – I mean, there may be quarters where they're just investing heavily in people for whatever reason, but most – If you go year-over-year, almost always they're getting more leverage. They're growing that revenue, or net revenue per person, and that to me is a sign of a great business. If your revenue per person is flat or it's declining, you still could have a good business, but you don't have a great business.

[00:27:03] JM: Do you set any expectations for how you want that revenue per employee to grow?

[00:27:09] AH: We talk about it internally. Yes.

[00:27:11] JM: There are not many buyers of data today.

[00:27:16] AH: Correct.

[00:27:16] JM: What kinds of companies do want to buy data today?

[00:27:21] AH: The companies that want to buy data are the companies that either have gotten really good at using their own data extremely well, and now they're at a point where they want to

go look for other datasets externally. So those were probably like the early adopters of all these different tools in the past, or they're like software companies that help these other companies.

So if you just think of like the financial industry, like hedge funds. So there're 11,000 hedge funds in the U.S. There are probably just 100 of them today by alternative data. So very small number of them by alternative data. So if you're selling data to hedge funds, then you have maximum of probably 100 customers.

However, there're also a lot of software and research companies that sell to hedge funds, and there's probably a thousand companies that might sell to these hedge funds. Many of those could potentially buy data. You cannot only sell to – It's like if you're selling data about retail, maybe you could sell directly to Walmart. Walmart is an incredibly sophisticated company and are really good about buying data. But maybe most retailers can't buy.

So you could sell to Walmart and you could sell to a few other retailers who are extremely sophisticated, but then you can also just sell to a lot of software companies that are selling to them, or you could sell to Shopify that's selling to them, or all these other different types of tools that are selling to different retailers.

[00:28:53] JM: Since you're buying and selling data, there many negotiations that you need to make. Because there is not a data as a service playbook, because there just haven't been very many data as a service companies, except for the data service playbook that you wrote. You don't have a whole lot of guidance in how to go through these negotiations whether you're buying data or you're selling data. Have you learned any general lessons on data vendor negotiations that you can share?

[00:29:30] AH: Well, personally, when I'm selling data as a seller, when I was a buyer. But when I'm selling data, my goal is to do everything possible to make sure that the customer is going to have an extremely high return on investment. So I do not want to optimize for price.

My goal is to optimize to make sure that they're using the data, they're getting value out of the data, that I'm getting the data in their hands as quickly as possible. So I'm not trying to try to

figure out how do I – The classic Orcale type of thing, of how do I extract every last dollar out of the customer. I want to make sure that the customer is happy.

That's another thing about a data companies, is data companies tend to often have some sort of like winner take most ability in their market. So the goal isn't to extract super-high price. The goal is to have very high market share. Sometimes actually having high prices, actually, oftentimes having high prices are in direct conflict with having high market share.

So you should be pricing – In my belief, is if you're a data company, you should be pricing your data to move. You should be pricing data so lots of people who are using it and benefiting from it and giving you feedback on it.

So at SafeGraph, we do not optimize for price at all. We are optimizing for people to use that data effectively, and we're trying to deliberately underprice it so that people use it. Then the other thing we're trying to figure out is how do we keep lowering prices. So maybe the customer is still – If the customer spends X-dollars on us per year, maybe next year, they're still spending X, but instead of getting X, they're getting way more data in return, or away more value from that. So what can we do every year to make sure they're getting more value for their dollar?

So data companies, I think long-term look a lot more like compute companies. You might spend more money on AWS this year than you spent last year, but you're getting way more value. Your dollar per gigabyte has gone down. Your dollar per CPU instance has gone down. Again, you might be spending more, but you're getting more value for your dollar.

[00:31:46] JM: There're some great general strategies around how to do sales in a data business. One thing I have learned about sales and selling podcast ads is the podcast ad sales market is very strange. It doesn't really map to any other things that I can read about. I mean, there are ways in which it does, but there are a lot of ways in which it does not. I'm wondering if you can go a little bit deeper. Like you've outlined some strategies around sales in a data company, but you've got to hire a sales team to scale it up eventually. What kinds of tactics can you confer to those sales people that you're hiring about how to effectively do sales and customer retention in a data business?

[00:32:32] AH: Well, if you think of like – I don't know that much about the podcast business. But if you think about a data business, in some ways it's almost the opposite of podcast business. So, if you think of like a podcast, there's no one podcast that has even probably 1% market share. There's just so many podcasts that are out there.

So even like the Joe Rogan type of – Like maybe that has .8% market share or something. Even the biggest podcast, it's still a relatively small percentage of market share of all the hours listen to podcasts. So there's no dominant company and even within a niche. So even if like the tech podcasts, or the politics podcasts or whatever, even within a niche, there're no dominant ones that have that. Then if you think of the advertiser, well, there're so many different places to put their ads. So podcast is just one of many, many places they could put their ads.

So even just podcasting at a category, you may or may not choose to put even ads on a podcast. You may choose to put ads on Facebook, or you may choose to put ads on TV, or you may choose to put ads on outdoor billboard. You've all these different options of where you can – And maybe even if you don't even do ads, you can do other types of marketing type of things to get people to drive people your way.

So it's not like a must-have type of thing. So then there's really just like a core – Kind of like core ROI analysis about like, "Okay. Does my ad on this particular podcast yield this particular result that I'm trying to do?" and having them kind of understand that, or do I want my brand associated with this particular type of brand for whatever reason that's out there.

In data, because often data can be like a winner-take-most market, in some ways like is this dataset going to be like helpful for my business? It's kind of like almost a yes or no. If yes, I need to buy this dataset. If no, then I don't need to buy – At any price, I don't even want the dataset.

Whereas as a podcast, if someone was saying like, "Well, I'll give you the ad for free, or I'll give you the ad for a few pennies or some like that." Maybe almost like everyone would buy an ad. Whereas with data, if I was like, "Hey, I'll give you this data for free." Actually, there's a cost of even like having the data and like being able to use it and stuff like that. So you may not even – Unless it's adding like real value immediately, like there's not necessarily a reason to go do it.

So I think there're a lot of different types of – Even within data and SaaS, I think there're different core use cases. In SaaS, it tends to be extremely competitive market. That's when SaaS, the whole game in SaaS is raising your prices. If you talk to any like core SaaS person, any SaaS investor, like Jason Lemkin, like the number one advice they give to companies is raise your price. They'll always say raise your price. Double your price. Do everything possible to raise your price.

In data, I don't think that's a good strategy. In data, it's really about market share. Just like compute is kind of about market share and stuff like that. So you don't want to be raising your price. You want to be growing your market. The other thing you might want to think about is like, "Okay. What's your cost of acquiring a customer?" Well, in a SaaS business, your cost of acquiring customers tends to be really, really high, because it's so competitive. In a business, like a data business, or maybe like a middleware business where it tends to be more of a winner-take-most market, or duopoly, triopoly type of market, your cost of acquiring a customer might go down quite a bit.

You can also afford to charge less, because your cost of acquiring a customer is lower, and you get more customers quickly. So all these different things or all these different levers happen as you kind of think about your pricing.

[00:36:13] JM: Can you explain in more detail the winner-take-most phenomenon of data companies?

[00:36:18] AH: So, it doesn't happen in every data market, but in a lot of data markets, there's a lot of investment that needs to be made for data. Investing in data, if you're really going to have high-quality data, it's extremely high-fixed cost. The variable cost is basically very low, often can be zero. Sometimes the markets aren't that big.

So often, it doesn't make sense to support. If you're really going to be investing in like really high quality data, it doesn't make sense to support like tons of different players that are out there. Once you get to software, in software, there's a whole UI component. Then there's like things that are extremely customized to different organizations. So even if you take similar

organizations like Ford and General Motors, just the DNA of those organizations, the DNA of the other tools that they use and how they're used to using tools might mean that they're just like way more comfortable with one type of UI versus another UI.

So it's very hard once you get to applications for someone to dominate, and there's no necessary reason for someone to dominate. Often, the markets tend to be bigger as well. So there's more reason to build investments into those particular things. You do see it in certain cases, like in middleware, where there are some sort of marketplace. So you'll see some sort of like middleware flywheel where companies can get to over 50% market share. But most SaaS businesses, like the leader has 25%. The number two player has 20%. The number three player has 15%, and all three turn out to be really good businesses. In and data businesses, the leader can often have like 50% market share, and that's the only one you'd want to invest in. The number two often makes it hard to do well.

[00:38:07] JM: Again, you started with the place location data business. Have you thought more about what the next adjacency will be?

[00:38:17] AH: Well, for SafeGraph, we're going to do places for a long time. So there are more places in the world than people. So there's just a lot of places in the world, and they're really hard to define and there's not a lot of good data about most places. So we started with places in the U.S. and Canada where people can spend money, like swipe a credit card or pay in cash for something, and there's really only like 5 million of those places in the U.S. and Canada combined. Then we started to add things like parks, and schools, and churches and a few other types of places. So we're really at the very – We have 1/1000th of all the places in the world today.

So if you think of the number of rows in our database, it's actually quite small compared to all of the places in the world. Our goal is eventually to have all the places in the world. That's going to take a really long time to be able to get there. If you think of the number of columns, well, today, maybe SafeGraph has 150 columns, descriptions about a given place. There's probably tens of thousands of actually really interesting columns you could have about a place. Not all of them are applicable to each place. If you think of like a health inspection rating would be applicable to

a restaurant, but maybe not. Maybe don't haven't health inspection readings for like a retail or something.

So the number of columns could also grow. So we have a situation where we should be growing the number of rows. We also need to massively grow the number of columns. We need to have the fill rates, the fill rates of those tuples higher. Then of course building the accuracy, and then of course once you do those four things, how do you deliver that data? Can you deliver that in an easy API? Can you make it easy to download? Can you integrate that with the application? Maybe someone's using Snowflake and you can get the data directly into Snowflakes. They don't have to actually build some integration layer into that. So there's all these different things that one needs to do in a data company to make that data really, really valuable.

[SPONSOR MESSAGE]

[00:40:24] JM: Software Engineering Daily is a media company, and we run on WordPress just like lots of other media companies, although it's not just media companies that run on WordPress. I know of many organization that manage multiple WordPress sites and it can be hard to manage all of these sites efficiently.

Pantheon is a platform for hosting and managing your WordPress and Drupal sites. Pantheon makes it easier to build, manage and optimize your websites. Go to pantheon.io/sedaily to see how you can use Pantheon. Pantheon makes it easier to manage your WordPress and Drupal websites with scalable infrastructure, a fast CDN and security features, such as disaster recovery.

Pantheon gives you automated workflows for managing dev, test and production deployments, and Pantheon provides easy integrations with GitHub, CircleCI, Jira and more. If you have a WordPress or a Drupal website, check out pantheon.io/sedaily.

Thanks to Pantheon for being a sponsor of Software Engineering Daily.

[INTERVIEW CONTINUED]

[00:41:46] JM: Going back to a point on sales, you mentioned in the data as a service guide that it's important to be able to upsell a customer in a data company. Can you give me a short master class in how and why to upsell within a data company?

[00:42:06] AH: Sure. I'll give you an example of SafeGraph. So maybe somebody wants to learn about all the Dairy Queens in the State of Ohio, for instance. The reason you want to learn about is, well, you own a bunch of Dairy Queen franchises in Ohio. So you buy that data from SafeGraph, and maybe that data doesn't cost you very much to buy, because it's not that many Dairy Queens in the State of Ohio that you buy. So maybe that costs you $1,000 to learn about all the Dairy Queens. I'm just making that up. It probably costs less than that, but all the Dairy Queens in the State of Ohio.

Now you know everything you can about the airy Queens in the State of Ohio. Well, maybe actually your franchisor like to buy Dairy Queen. So maybe now I want to know about the Dairy Queens in the State of Pennsylvania. That's not a big of a drive from the State of Ohio. I can get to the State of Pennsylvania. So maybe I want to go buy that. Oh, maybe I actually want to know about the competitors to Dairy Queen. So for all my Dairy Queens in the State of Ohio, what are my competitors? How would my competitors look like? So what are the McDonald's look like that are nearby? What are the other types of core competitors to me?

Maybe it's not just competitors. Maybe it's a compliment. Maybe after people go to the Dairy Queen, they like to go to the gym to work off that Dairy Queen or something like that. So I got to figure out, "Okay, how are the compliments doing? How are the compliments doing well? Are there more compliments coming or less or are they doing? Because that might tell me a lot about the Dairy Queen market. Again, I'm making this up. I don't know that much about Dairy Queen, but it gives you an example of like all the different things that you can upsell.

So if you think of a data company, again, you've got these rows and you've got these columns. So if you got these natural things where you're hopefully building more rows overtime of entities that you could upsell people. Again, that's the Dairy Queen in Pennsylvania. Now, Dairy Queen in Ohio. Now, the Dairy Queen in Pennsylvania. But Now you're also building more columns. So now you know like new things about the Dairy Queen. Maybe – I don't know. The average temperature inside a Dairy Queen. I don't know how you know that, but maybe that's really

valuable, because you want to understand like how much energy usage do people do, or is that ice cream like really making the place cooler, or maybe temperature of the Dairy Queen actually equals sales.

So instead of putting it at 69°, if you moved everything up to 72°, more people would buy the ice cream. I don't know. I'm just making – All these different types of things could have some sort of core value. So all these different things are opportunities to upsell. So you already have these customers. They already love your data hopefully. They're already happy with it. How do you give them more things to upsell them? It's not just upsell, because that's a crass way of saying I want to get more money from customers, but how do I get the more value? It might just be like how do I be able to push this data into – Maybe they're already using like ArcGIS, which is a great tool for understanding your like spatially, things spatially. Maybe they're already using ArcGIS.

Can I give them the data in the ArcGIS format so that it's easy for them to input it so they don't have to like do a transformation of that data before putting in ArcGIS. Oh, maybe even simpler. Maybe I can even like get a button in ArcGIS so they can press a button and the data just flows in automatically and they don't even have to do like the three steps.

Part of selling people any type of solution is trying to save them time. Your goal should be to save people time. This is like my whole hobby horse drives me crazy in general about software things on the Internet, is like a lot of them are doing everything possible to make it really hard for users. I don't know if you spent – If any of your listeners have spent any time in Salesforce. It is the slowest piece of software. It's so slow. It takes me so much time to do simple operations. It's so hard to find something and so long to run a report. Just to do like simple search is really hard.

LinkedIn is the worst. LinkedIn is the slowest, like a large company software tool that I know of. It's terrible. Just doing basics, it is so slow. Why is it have to be so slow? What is going on? Why don't they value my time more? I'm a paying user for a LinkedIn, why they're not valuing my time?

Google Apps is super slow. Gmail is superfast. But if you want to search your contacts in Google, it could take you like 10 seconds to run an operation. Has anyone ever tried to search their calendar? If I want to search my calendar for the last time I met up with Jeff Meyerson, just running that search can take many, many, many seconds. Google can search the entire internet in like 1,000 times faster than I can just search my calendar with maybe only 10,000 entries in the whole calendar.

Companies, for whatever reason, don't care about the time of their user. They're not valuing their user's time. I think one of the important things that people have is time, and you should be doing everything possible to think of your customers and try to figure out how do I give them the most bang for their buck in the amount of time that they're going to do? How do I decrease the amount of time they have to spend?

This is why like Google search I think initially was such a great thing. You don't have to spend that much time, and they're actually measuring you on how little time you spent on their site to get value. Not trying to like get you to spend all these time on their site. Time on site is actually a terrible metric. People should be measuring the opposite. How do I get value as fast as possible?

Actually, if I continue on my rant, this is my problem with podcasts as well. Why does the podcast have to be 2-1/2 hours? You have some of these like long – They definitely aren't better. A lot of superfluous type of things, like how do we get the most value to the listener as fast as possible so they can spend – Books are this way. Books are like a thousand pages. A lot of those books are thousand pages could be condensed into 50. So why can't they do a little bit more work and take that thousand page book and condense it into 50 pages? That would be way more valuable.

In fact, you should be willing to pay more for a 50-page book that you are for 1000-page book. Even TV shows today, they're just like they have all these like long, sweeping views and all these stuff and things, because they're just trying to fill up that hour with as dialogue as possible. The dialogue per minute in the Game of Thrones has gone down every single year every single season. So the amount of dialogue per minute in season one was, I guess, two or

three axes high as the amount of dialogue by season seven. So they're just doing everything possible to draw everything out, to not value the user's time.

[00:49:06] JM: How have the economics of starting a data company surprised you?

[00:49:10] AH: Data companies are very hard businesses to do, because the number of customers you have are small, because not that many people can buy from data companies. So you have to kind of have a very good understanding of your market before you go out and just try to sell data. Obviously, if you're selling data about something that people already buy. So SafeGraph sales data about places. There are a lot of companies that already buy place data. Then that's easier to start that. But you can also be selling about data that people don't already buy, and then you really have to dive in to understand that market and try to understand why would people want to buy that overtime.

[00:49:57] JM: You started to work on a new investment fund. What investment thesis do you hold today that your one year ago self would disagree with?

[00:50:10] AH: That's a really a question. I don't know. I think we're in a really weird time right now where every single asset class is doing extremely well across the board. There's basically like no asset classes that are not at their all-time high. Usually, asset classes, there's not always like completely correlated with one another. But every classes correlate with every other class. So there's essentially no hedge.

So it's like, often, "Okay. I don't want to be in the tech sector. I need a hedge against the tech sector, or a hedge against the real estate sector, or a hedge against bonds, or a hedge against – Whatever it might be. Every sector is doing extremely well with maybe be a small number of exceptions. Everything is at all-time high.

So this is made geniuses out of everybody, except the short-sellers. Every single person out there, every single investor in the last 10 years looks like a genius. I think that is an extremely worrisome time, because everyone can overvalue how good they are about something. Whereas, if I'm a free-throw shooter in basketball, like I have a lot of benchmarks to understand, how I'm doing? There's a whole bunch of under – It's hard for me to lie to myself about how

good of a free-throw shooter I am. Whereas as an investor, most people, for instance, who invest in private internet companies the last 10 years have done extremely well. They can't all be good investors, but they've almost all have done extremely well, because it's been the best time in history to invest in private internet companies. Even people who invest in the stock market. Well, investing in the stock market in the last 10 years have been extremely good for everybody. Almost every investor has done extremely well.

Now, not everyone has beaten the S&P 500, because the S&P 500 has done so well. But in private internet companies, there is no S&P 500 thing. So it's hard to know. So everyone has done well. So everyone thinks they're genius. I think this is always – This is a very, very weird time, because there are so many “geniuses out there”.

[00:52:27] JM: Do you take any solace in the amount of leverage that you are seeing increase in your employees, because you can look at that and you can say, “Wow! There really have been some material improvements in productivity. Maybe this is fundamental shift. Maybe I shouldn't be so worried.”

[00:52:50] AH: Well, even if there is a fundamental shift, you can make a case that all the internet companies on aggregate are fairly valued. Obviously, there’ll be some that are overvalued and some that are undervalued, but in aggregate you could say all the internet companies are fairly valued. But that doesn't mean you're a good investor. If you are already predisposed to investing in internet companies, that doesn't make you necessarily a good investor. You may not even be beating the benchmark. So you might be having a 20% IRR, but the benchmark is a 30% IRR, or something like that. Essentially, if you’re throwing darts, you would've invested – You would have done better than how you're doing, or maybe just got lucky. There’s just one particular investment that return five-X your fund or something like that and you got extremely lucky and then everything else just kind of ends up that way. So I think I think it is hard to know.

I do think we are seeing a fundamental shift in productivity, and most economists don't believe this. Most economists think that we’re not – If you look at the productivity data in the United States, the productivity is not growing at a rate that you would expect. It really hasn't been growing for almost 20 years at a rate that one would expect the productivity rate to grow. But I

think that is because there are – Most people are really not adding productivity. So most people aren't taking advantage of the different tools that are out there, but for people who are predisposed to take advantage of the tools, those people are getting massive, massive productivity gains.

In some companies you can see a scenario where a large percentage of the company, a large number of employees are actually could even have negative benefit to the company. These companies could actually benefit from having fewer people and having less communication problems and being able to move faster and being able to leverage the other people more.

So there are all these scenarios where like the U.S. has often lower productivity than other countries. That could be just the case that the U.S. has just way better employment, has lower unemployment than other country. So France has sometimes higher productivity than the U.S. But in France, the average person works a lot fewer hours than in the U.S., and productivity is just like economic output divided by number of hours worked. So the average person works a lot fewer hours than in the U.S., and also there're a lot fewer people working in France than in the U.S. So that just could define why the productivity differences are in some of these different countries.

[00:55:25] JM: You've been building companies for a long time. What's a general lesson about company building that you learned in the last year?

[00:55:32] AH: I think one of the things that I've realized is that that the ability to automate things is growing at a rapid pace and you don't have to be a software developer to do it. You just have to be a smart, technically adept person to be able to start automating things. So if you think of like all these companies, like Zapier. If you're familiar with them, these kind of "no code". You do need a decent amount of technical ability to go do that, but you don't need to be a software developer. You just need to be a smart, technically oriented person, someone who feels comfortable in a spreadsheet, or someone who feels comfortable in PowerPoint, or some other type of thing to be able to use these tools, and these tools can really change your life and change your organization's life. There are so many different things that people do and there are so many hours of the day that people spend where they could potentially invest in automating those things out.

Then the other thing that I have found really exciting is the ability to use other people to help me get leverage quickly, and I am a little bit late to the party in this. But just the ability to go hire somebody, let's say on UpWork, or some other – Your favorite place to go help you with something is amazing. If you want to go find – First of all, the cost of finding somebody, the cost of your time of finding someone has gone down really dramatically over the last few years. I personally use UpWork, but there a lot of other tools that people can use to go find people to help them. If I want to go find someone on UpWork to go do something, I can find them often very quickly. Sometimes within five minutes of my time. I can go do that.

UpWork could probably do a lot of things in the UI to make it a lot easier. So it does have its own problems. So it is actually a lot harder to use than it should be. So they could even probably take that 5 to 10 minutes down to one minute of my time to go find something. Then as the cost come down of finding someone, I'm happy to pay someone a fair hourly rate.

The other thing is using people to help you get up to speed on something. So let's say you're trying to learn how to use podcasting software, right? I don't know that much about podcasting software, but I imagine it's a pretty high learning curve to use it. You have to read a lot of stuff, about doing it, and how do you edit things and how do you put the music in and how do you do all these other stuff that's out there. What software do you use?

[00:58:12] JM: Zencastr.

[00:58:12] AH: Zencastr.

[00:58:13] JM: For recording remote podcast, but right now I'm using GarageBand in-person.

[00:58:16] AH: Okay. So, yeah. So all these different things and probably like all these different like GarageBand probably have all these different tools about how do you use it how do you – It probably took you long time to become – Like you're probably an expert today, but maybe it took you a long time to go do that.

One thing I've found is like I can hire someone as a tutor and I can say, "Hey, can we do a live tutoring session on GarageBand? I really want to learn GarageBand." Let's say they cost $65 an hour. Okay. So for $65, I can save myself potentially five hours of time. They could tutor – They can get me up to speed. I can hire a tutor and then I can go off and do my own thing, and I then can do another tutoring session with them like a week later and they can help me and they could show me the best practices and what mistakes I might be making, and all these other types of things. Those two tutoring sessions, they might save me 10 hours' worth of time or some other type of thing.

So if you start to even like – Even the ROI on all of these different things that one can do is so astronomical. So many of my friends who are developers are like setting up an environment. Any type of environment, like whatever you're doing, even like you're sitting – Even like Heroku, which is like the simplest thing, like just setting up your Heroku thing is like so hard to do sometimes, and you like make all these mistakes and sometimes you go down to some rat hole and teach you like 20 hours. You know what I'm talking about, right? I'm sure every single person is nodding their head right now.

[00:59:36] JM: That's why I switched to podcasting.

[00:59:37] AH: Yeah, exactly. It's so frustrating. You could hire someone, like sometimes for like 30 bucks to just help you, like just be there with you. Even if you have a colleague in your company who's really good, like instead of bothering your colleague – For someone, they just do Heroku all day long and they're actually really good at it and they're actually happy to help. They love showing their knowledge. They love the fact that they could teach you this type of thing.

So I am super late to that party. I'm just starting to do that now over the last few months. I wish I was doing this like many years ago, and I imagine over the next year that I'm going to do more and more and more of this.

[01:00:20] JM: Auren Hoffman, thanks for coming back on.

[01:00:21] AH: Thank you, Jeff.

[END OF INTERVIEW]

[01:00:26] JM: When I was in college, I was always looking for people to start side projects with. I couldn't find anybody. So, I ended up working on projects by myself. Then when I started working in the software industry, I started to look for people who I could start a business with. Once again, I couldn't find anyone. So, I started a business myself, and that's the podcast you're listening to. But since then, I've found people to work with, on my hobbies, and in my business, and working with other people is much more rewarding than working alone. That's why I started FindCollabs.

FindCollabs is a place to find collaborators and build projects. On findcollabs.com, you can create new projects or join projects that are already going. There are topic chat rooms where you can find people who are working in areas that you're curious about, like cryptocurrencies, or React, or Kubernetes, or Vue.js, or whatever software topic you're curious about.

We now have GitHub integration. So it's easier than before to create a FindCollabs projects for your existing GitHub projects. If you've always wanted to work on side projects or you want to find collaborators for your side projects, check out FindCollabs. I'm on there every day and I'd love to see what you're building. I'd also love if you check out what I'm building. Maybe you'd be interested in working on it with me.

Thanks for listening, and I hope you check out FindCollabs.

[END]

