EPISODE 1026

[INTRODUCTION]

[00:00:00] JM: Physical places have a large amount of latent data. Pick any location on a map and think about all the questions that you could ask about that location. What businesses are at that location? How many cars pass through it? What's the soil composition? How much is the land on that location worth? The world of web-based information has become easy to query. We can use search engines like Google as well as APIs like Diffbot and Clearbit, but the physical world is not so easy to query. We can't just go to a Wikipedia of the physical world that does not exist. The physical world is not easy to query but it's becoming easier. Location data as a service is a burgeoning field with some vendors offering products for satellite data, foot traffic and other specific location-based domains.

SafeGraph is a company that provides location data as a service. SafeGraph datasets include data about businesses, patterns describing human movement and geometric representations describing the shape and size of buildings.

Ryan Fox Squire was one of the earliest engineers to join SafeGraph and he works on data products for SafeGraph. He joins the show to talk about the engineering and strategy that goes into building a data as a service company and he talks a little bit about how he made his way into the world of software. He does not exactly call himself an engineer. I would call him an engineer, but he's more of a product person, I suppose. He came from academia and learned data science tooling and now he works on a variety of things at SafeGraph.

[SPONSOR MESSAGE]

[00:01:54] JM: I've recently started working with X-Team. X-Team is a company that can help you scale your team with new engineers. X-Team has been helping me out with softwaredaily.com and they have thousands of proven developers in over 50 countries ready to join your team and they can provide an immediate positive impact and lets you get back to focusing on what's most important, which is moving your team forward.

[INTERVIEW]

[00:03:39] JM: Ryan Fox Squire, welcome to Software Engineering Daily.

[00:03:41] RFS: Thank you. Pleasure to be here.

[00:03:43] JM: I've asked this question in some other episodes, but it's worth exploring again. We have computation as a service from AWS. Why don't we have data as a service?

[00:03:55] RFS: Yeah. I think that we do more and more. I think it's a growing area, but I think that, historically, it's been very difficult for companies to work with data in general and especially work with outside data. I think one of the trends that I think is changing is that companies are getting better and better at working with data. They're getting better and better at collecting their own data and using their own data to guide their own decisions, but even just in the last 5 years, the tools around that have grown immensely and I think it's becoming easier and easier. You no longer need to have a giant team of data engineers and data scientists to extract sort of basic insights from data. Those people still have valuable skills and are still helpful in many cases, but there are just so many more tools and platforms now that make it easier to work with data that I think that is now creating more opportunity for there to companies that exclusively all that they do is provide datasets for those users.

[00:04:56] JM: There are some paid data providers. There're also publicly available data providers. Can you break down the different types of data providers that are out there?

[00:05:09] RFS: Sure. Yeah. I mean, the way I would probably frame that is less about data providers and more like what are the data sources. Some of those sources are companies. Some of those sources are organizations like governments and some of those sources are not even any particular entity, right? They're just things that are created because of the internet and exists as byproducts on the internet, things that you cans scrape from the internet or things that are created through internet activity, things like that. I think that there's a wide-range of ways that one, in theory, could obtain data. I'm not sure if that's sort of getting at your question.

[00:05:42] JM: Yeah, and I'd like to get further into that, but just serving the different ways that you can get something like data as a service from the current world of providers. I did see a team at Google posted a list of datasets that are publicly available recently. I think Google actually has maybe an entire team that is working on dataset aggregation. Do you know anything about that? What? Google's data as a service stuff.

[00:06:13] RFS: Yeah. There are cases. One thing that Google has been doing more of recently is essentially publishing datasets that they've used directly that other people can use. This is I think like a really big shift and a very important thing and a very exciting thing that Google can do, because Google does have access to such a rich amount of data that is a sort of a unique dataset. They have been doing more efforts to sort of package and publish some of those datasets and just make them openly available. That's one effort, and I think there's been some cool progress there.

Another thing that's recently happened at Google is that they've released this new product. I forget what it's called. It's called something like datasets of data search, but it's essentially like a Google search function for datasets. You have traditional search for websites. You have Google image search, which today it's like hard to imagine a world without Google image search, but it didn't always exist, and now they're trying to do a similar thing for datasets, which is more and more there's all these datasets being put on to the web, many of them by governments, many of them by academics. It is actually how do you find those datasets. If you have a question and you want to get some type of data, it is hard to search for those things. If you're doing Google searches for those things, those aren't well-indexed. Google is trying to sort of tackle that problem and say, "Let's make a search that's designed specifically to find types of data," which I think is a hard problem, but ultimately just makes that more accessible for everyone. I think that's super exciting, and I think there are just tons of free and publicly available datasets out there that people just can't find. That's a view into sort of the ways that Google is contributing to more and more access and availability of sort of open free datasets.

[00:07:56] JM: When you consume those kinds of datasets, the Google searchable ones, do you have guarantees about the quality of the data?

[00:08:04] RFS: No. I mean, I think that's one of the challenges with the open data world is that, in general, there aren't a lot of standards for how data should be organized or how data should be documented, as is always the case, which one of the challenges with open datasets is that it does take a lot of effort to make those datasets reliable and consistent and updated overtime. That's a lot of time that someone has to spent on doing that. There's always going to be the people on the governments that are managing these datasets and doing a good job with those, but there's also going to be the random dataset that gets published but then is never updated or is published but has various errors in it.

I think that the Google search approach right now is just to try to, "Let's just index all these things and make them searchable." They're not going to solve the problem of do I know what I'm getting and do I have insights into how this data was collected or what are the assumptions going on behind this data? Those are I think going to be – Continue to be challenges always for open datasets.

[00:09:03] JM: At SafeGraph you sell a location dataset. Give me some examples of why location data is useful, some example applications.

[00:09:13] RFS: Sure. Yeah. SafeGraph is a data company. Were focused on trying to produce these high-quality datasets about the physical world. In particular, we're focused on understanding points of interests or places, particularly commercial places of interest or places that consumers would go to spend time or money, so all the hotels, all the airports, all of the restaurants, all of the malls, all these things.

I think there're a couple reasons why places and focusing on those types of data is very interesting. One is that lots of stuff happens in the physical world. Lots of things are happening and place, or more generally, space, is a very powerful indexing key or join key for lots of other datasets. Space and time, in a physical world, those are some of the most important dimensions to any type of data. Many types of data that are collected also could be associated with space or time, and we think that place is a particularly important dimension to be able to index and catalog on. That’s sort of general. I can give you some specific examples of how our place as data is being used.

[00:10:14] JM: Sure. Yeah.

[00:10:15] RFS: Yeah. When we think about this the typical SafeGraph customer, there're a couple different dimensions, I think, to consider. One is sort of what is the different type of industries that are using the data, and then the other is one of the types of users or types of people that are actually using data in those industries.

In industries, SafeGraph has a fairly broad approach. We work in a lot of different types of industries. For example, we do lots of work in like the retail real estate world. So companies that if you’re Dairy Queen or Starbucks and you’re thinking about opening a new location, that's a very important decision for you to make. Starbucks, for example, has a team of people that all that they're doing is trying to figure out where should we open new Starbuckses. Understanding the landscape of what the other places in the world and lots of data about those places is very important for those types of decisions.

Another area that we work in is sort of like digital marketing and advertising. So location-based advertising is sort of an increasingly interesting and popular and powerful dimension in the advertising world. A lot of location-based advertising is based on giving you triggers or giving you signals around that are location context dependent, right? If your target, like the company, like the store Target, and you have the Target app and you have users that are in your Target app, maybe you want to send them a particular coupon or trigger. If that user goes close to a Target or goes close to Walmart, something like that. There's a lot of powerful use cases in that space as well.

Then the third space that I would mention is what I would call like geospatial analytic tools. We work with a lot of product company, a lot of software companies that what they do is they build geospatial analytic SaaS products, and those can vary across industries as well, but often the uniting theme is that, for whatever reason, their users want to be able to visualize points of interest on a map or do analysis that is related points of interest. SafeGraph data ends up powering a lot of those functions in those software companies. In that case, we're not necessarily working directly with the end user of that data. We're working with the company that's producing a product that is then used by the end-user.

[00:12:27] JM: The data that you sell, the way that it's consumed is via a dataset. Somebody downloaded a dataset rather than having an API for requesting the location data. Why that decision? Why make the data available via a dataset that people download rather than what I think will be a more conventional approach of the API?

[00:12:50] RFS: Yeah. That's definitely an ongoing sort of product discussion at SafeGraph, and I think that in the future, there will be more and more ways to access SafeGraph data through APIs, because there are just use cases that depend on that. In the beginning, we sort of made a decision to – I think is often is the case at a startup company, let's do the simplest thing we can to add value to our customers. In the beginning, we realized that a lot of our customers were quite happy to get the data as a bulk download. Part of that is because many of our customers are, for example, technology companies or they're building their own SaaS products and they're ultimately going to stand up the entire dataset in their own infrastructure to interact with their products or interact with their production systems however they want. They're quite happy to essentially just be delivered a bunch of data through the cloud that they will download or stand up however they want.

I think we do often discuss with customers their needs around like more sort of real-time or on-demand querying through APIs, and I think just the way it's played out so far is our customers don't need that to get the value. But there's other type of customers that certainly will want that and that's stuff something that SafeGraph will do eventually, but really, it’s just been sort of a guided by what our customers are doing and what they care about and it’s turned out that most of them are quite happy to get the full download.

[00:14:13] JM: The location data the you sell, there is a set of information about places, geometries and patterns of foot traffic. Describe the types of data.

[00:14:25] RFS: Absolutely. As I said that, all of SafeGraph data products are focused around this idea of points of interest, physical places in the world. It’s sort of the original dataset to think about is what we call core places. That’s all of the sort of essential metadata about a place. It's name, it's address, category information, open hours, things like that. It’s basically business listing information.

The second dataset we have is called geometry, and this is more focused on geospatial data about those points of interests. Many customers don't just want to know where is a POI in – Address space. US address space. They want to know where is that POI in geospatial coordinates terms. A latitude, longitude for that POI. We also have additional geospatial information besides that sort of building centroid. For example, we have building footprints, polygons, sort of two-dimensional projections of these businesses in lat-long coordinate space. Sort of precisely, what is the actual footprint of this business occupying?

We also have another big dimension of the geometry product is what we call spatial hierarchy information. Is this POI inside another POI? Is this a Starbucks that’s inside a Target, or is this a Subway that's inside a Costco, or is this a business that's inside a mall? Sort of trying to understand the spatial relationships of these places is another big dimension that we do in geometry.

Then the final product, called patterns, as you said, it's summarizing human movement or foot traffic around these places. That sort of provides a whole another rich dimension to understand what's going on with these places. That data set includes things like how many people are coming to visit this place every month. What are the popular times of day that people coming to visit this place? When people traveled to visit this place, on average, how far are they traveling to get there? On average, what are the census areas of visitors to this place? Things like that.

That dataset obviously connects interestingly to the geometry in the core POI, the core places data, but sort of also has orthogonal information to it. All those datasets all has sort of share the same primary key, which is this unique place ID for places that SafeGraph maintains. Some of our customers buy all three of those datasets and join them altogether based on this unique place ID. Some customers only buy some of those datasets. It really depends on what they're doing.

[SPONSOR MESSAGE]

[INTERVIEW CONTINUED]

[00:17:54] JM: As you said, this patterns data has a population of people that are anonymized. I think the numbers I saw was46 million anonymized mobile devices and the places that they visit. Then the data about individual user traffic is not revealed, but you're able to provide some detail about the users who are visiting these different places by providing the counts of different users from different census block groups. Can you explain what a census block group is and explain how that could be used to infer something from this anonymized dataset?

[00:18:35] RFS: Yeah, totally, and these are the details that I like get super excited about. Happy to talk about this. First of all, what is a census block group? A census block group is the finest grain, a division of the United States that the US Census Bureau maintains. The US Census Bureau divides the whole United States up into like sort of Russian doll hierarchical, increasingly granular divisions territory. We have states that the high-level. We have counties.

within counties, you have different types of census areas called things like tracks, or blocks, or block groups. The block group is the sort of highest resolution area that the census reports all of its data for. Those areas are on average going to hold something like a thousand households or something like that.

They're relatively high-resolution and the census – We’re doing another census in 2020 here, but the census tracts, a huge amount of data in all those areas. It’s this monumental service that the federal government does. It's a very rich dataset keyed on these census block group areas. This census block group.

The way that SafeGraph works with the data is that I think one of the one of the things that is part of the value proposition of a company like SafeGraph is that things like individual location data is a very sensitive type of data and there's a lot of questions that you might want to ask about population dynamics are human movement that don't actually require having visibility or access to individual devices or individual people data. If you want to know what's the most popular time of day that people visit a McDonalds, you just don't care about like individual visiting times. You only care about the aggregate visiting times.

One of the things that SafeGraph does as it tries to help manage all those privacy concerns on behalf of our customers so that some of the data that we work with is more sensitive than the data that we’re sending to our customers, and that's part of sort of a value prop of what we’re offering. For example, one of the things that you mentioned is that for any given point of interest, one thing that’s very interesting is to think about like, on average, what is the sort of type of person visiting this place? That's a question that you can answer in generic terms without having to discuss sort of individual privacy details. So we can say on average, people that visit this McDonalds come from these types of census block areas, and therefore on average they're going to have these types of demographics, these types of household incomes, these types of jobs, things like that. That data is obviously is super rich and valuable if you're trying to decide where to open a McDonald's. You care a lot about what are the types of people that are doing commerce in this area or living near this area, and that's a big part of what I think is valuable about the patterns, the SafeGraph patterns product.

[00:21:20] JM: Yeah. One example I saw that resonated with me was I think you did a study that Starbucks visits declined 6.8% in San Francisco after the open bathroom policy was implemented.

[00:21:35] RFS: Right. Yeah. The first thing to clarify about that is that – One of the things that SafeGraph does is we actually donate a lot of data to academics who are doing economic research or economic development research or sociology research. We just think that like those people are super smart and part of the big SafeGraph vision and goal is we want to make data more accessible to everyone. We select certain types of academic partners that we think are very good and then we'll donate data to them and then they do their own independent research and we're not actually involved in it at all. This is a case of an academic doing an independent research study that we gave them data. They came back many months later in they were like, "Hey, we found this really interesting thing. We're publishing it." We're like, "Okay, cool. They publish it. It gets picked up by the press. They talked about it on like Bloomberg news and these things. Yeah, the story there was that, briefly, Starbucks had this sort of unfortunate PR incident regarding potential racial discrimination of some of its bathroom policies. It was a big news event.

In response to that, Starbucks sort of clarified their bathroom policy, which they call sort of a public place policy in which being super clear that anyone is welcome to use the bathroom, whether you're a customer or not. These academics, one of the things that they're interested in is this concept of what is the cost or value of these sort of public goods services that sometimes corporations are offering. Things like public restrooms, you could think of as a public good that a local government may or may not be meeting. If a corporation is doing that, and sometimes the government sort of outsourced that public good to that company. Then the question is how do you quantify the value of that public good that this company is now handling?

It's a super interesting idea and perspective, and essentially they were able to show ways in which this policy, which certainly has a high-benefit for the public good in many ways also had some cost to their business, and it's a very nuanced picture too, right? Because to do that study correctly, you don't want to just know what is the traffic, the foot traffic to Starbucks. You want to understand what are the right comparables, because it's not just whether traffic at Starbucks went up or down. It's did it go up or down compared to its sort of relevant competitors in adjacent areas? Those are the types of nuanced things that these academics were able to pull out using SafeGraph's data.

[00:23:51] JM: And a more directly profitable example that you sent me recently was the ability to judge risk scores for insurance based off of the datasets that you're generating. You had this detailed blog post about how you could create a risk score based off of this location data and use it to judge. If I'm an insurance company, how much should I charge my customers for fire insurance, for example? Can you give a little bit of color on that just to give people an example of a business use case?

[00:24:33] RFS: Yeah. Right. Commercial insurance is obviously a massive industry and it's a super data-driven procedure where the goal of the insurance company is to figure out what are the high-risk opportunities. What are the lowest risk opportunities and then to price their policies accordingly. The more accuracy their insurance company has about making those predictions, ultimately, the more profitable it can be because it can correctly risk adjust its policies.

There're many, many dimensions to figuring out an insurance policy. One of those dimensions – Many of those dimensions have to do with physical space, and literally what is the physical place that this business is and what is the context of that place? Is this a place that – I think everyone could understand, if a business is located next to a river that floods often and that parcel has had flood many times, that's going to be a much higher risk opportunity for an insurance company to give flood insurance to.

There're many, many examples like that where your physical location has a lot of implications on your risk. This blog post was focused in particular on this one aspect of that, which is fire risk and what's called co-tenancy risk. Co-tenancy risk is risk that's associated by the other types of businesses that are in your same building or in your adjacent vicinity. Sort of the very, very simple proof of concept that we were showing in this blog post was all things equal. If you're ensuring against fire risk, you're going to have a lower fire risk if you're next to a clothing store than if you're next to an industrial kitchen or a bakery or something that has these know very hot fires and ovens going all day long.

You can calculate that using SafeGraph data, because there’re a couple of things you need to be able to do that. The first thing you need is you need to have very accurate category information because just knowing the name of a business or the address of a business is not sufficient to know is this a business that has fires like a kitchen, like a restaurant, or is this a business that just sells clothing? You need to have the accurate category information and then you also need to have the accurate geospatial data about is this business that has a similar address number, like 525 and 523, are those in the same building? Are those sharing a wall with two different buildings? Are those actually – There's a parking lot in between them? That makes a big difference for your fire risk and that's the kind of stuff that SafeGraph data, geospatial data is covering. We should've put together an example of showing how you could do a simple calculation and look at some examples, high and low risk places that we found from that.

[00:27:04] JM: I’d like to talk a little bit about the collection of the datasets and the cleaning of it. My understanding is that the datasets that you’ve collected at SafeGraph, there's no one weird trick to generating them. It's a combination of buying from some providers, getting from some free providers, like the census, and also a matter of scraping the Internet to – For example, if you have – Well, you could imagine all kinds of things that you want to scrape up the Internet. Maybe hours of operation. Things like that. Tell me about the data collection process.

[00:27:44] RFS: Yeah. I mean, I think you're on the right track there, which is that it's multifaceted and there's not just one sort of one solution or one source that we’re using. SafeGraph data is built from thousands and thousands of sources that we’re ingesting. Many of those sources are things that we’re finding publicly available. Some of those sources are things that we’re buying or are managed by governments like census. Census is an incredible resource that’s technically open source. Yeah, we’re using all those types of things.

It's actually not – I mean, there are challenges, but it’s not that hard to scrape the content of like 10,000 websites. I mean, there're some technical challenges there, but I think what's hard about it is taking all that raw data and combine it into a common schema. Even if you imagine, for example, scraping or crawling the store locator of a restaurant, let’s say a chain restaurant that has 10 locations in California. They list all those locations on their website. SafeGraph is going to go grab that information to know the addresses for those locations.

That part is not hard. The hard part is, is that data structured the same way or different than the other 999,000 websites that we're scraping? How do we make sure that we're getting the right information from that page? How do we do that in a programmatic and automatic way? It's not scalable for us to sort of manually configure 10,000 different scrapers. How do we build systems that can find that information programmatically and automatically bring it in, handle the millions of edge cases that we hit where – You mentioned open hours, right? That's a good example. We scrape open hours from things like store locators, and it seems like that should be straightforward. But, of course, every store formats their open hours differently. Then you have these – All of the cases you can imagine come up where suddenly in the field where there's supposed to be open hours, they just right coming soon!

[00:29:31] JM: Right.

[00:29:32] RFS: Or in their open hours they say, "Open till 2 AM," and they don't have an open time. They just have this sort of closing time. There's like a million things like that that happen when you're doing this really at scale, and I think that's the hard part. That's where like the sort of engineering and machine learning comes in to handle those types of cases.

[00:29:51] JM: Have you looked at Diffbot at all?

[00:29:53] RFS: I'm not sure if I'm familiar with Diffbot.

[00:29:55] JM: Diffbot is cool. It's one of my – Gets the Software Engineering Daily award for most underrated API. It's a system that takes unstructured web data and structures it. I don't know if you're on the scraping team or whatever, but if you haven't looked at Diffbot, you definitely should. Basically, if you take just a random webpage and you want query that webpage for all the entities that are on the webpage, it uses a lot of natural language processing and figures out like the entities. Rather than having to scrape the HTML and parse the HTML and figure out what are the entities on this page, you can just use Diffbot, query it for the entities, like, "Oh! There're places on this page. Maybe I'll query that a little bit deeper." That's one of the more interesting – Actually, I would call it data as a service company.

[00:30:41] RFS: Yeah, that's interesting. I know we have looked at a bunch of things like that. We have looked at a bunch of things like that, and I think one of the sort of surprising things to me that I didn't expect when I was looking a lot at these vendors is that I realize there’s sort of two different types of like large-scale scraping that happens. One is the type that we’re doing, which is we want to scrape a small amount of data from 10,000 webpages. The other type is essentially the use case of I want to scrape all of Amazon’s inventory or all of united.com flights inventory, which is I want to scrape a huge amount of data from a single webpage. I haven't examined all these solutions in detail, but one thing I found is that a lot of those solutions are better at those latter types of use cases, where we’re going to crawl 10,000 pages from amazon.com and pull out all the product listings, but it was much harder to generalize those solutions to I want to find the 10 restaurants on 10,000 webpages that all have sort of different formatted entities.

That’s all to say that I think that there are definitely ways to automate this. That’s some of the stuff we’re doing in-house. When we looked at vendors for this, we found that we thought would be able to outsource a lot of this, but in the end, we just ended up needing more control. Our air tolerance is also quite low. That's another difference, I think, is that many use cases that’s scraping, you don't care if you’re getting some things wrong. It depends what you're doing, but at SafeGraph we like really care a lot about getting every single thing right. That just makes it like a different type of problem in some ways.

[00:32:03] JM: Definitely. Did you work on the scraping infrastructure?

[00:32:07] RFS: Yeah. I was working on it, especially in the beginning when we’re first building it. We sort of went through a series of phases of our first – Often at SafeGraph, our first instinct is how can we outsource something? Because one of our core values is this idea of like how do you get leverage, and a big a big part of that is outsourcing. Vendors are great, and Auren is big on vendors.

[00:32:26] JM: Oh, yes.

[00:32:27] RFS: We’ve talked a lot about vendors at SafeGraph. IN the beginning we looked to outsource and we started – We actually did a lot of outsourcing and we engaged in some contracts with some consultants and some outsource technologies to do a lot of scraping. That was sort of phase 1, and that was okay, but as we got deep into it, we just realized that there was a level of quality that was extremely hard for us to maintain with these outsourced vendor solutions.

So then we moved into phase 2, which is, “Okay, how can we build a platform to enable this type of scraping internally while outsourcing some components of it?” That’s sort of where we are now, was trying to figure out, “We don't want to have to write 10,000 scrapers. That's not scalable, but there are a lot of people out there that can write web scrapers. Can we create a platform or a framework that like makes it easier for us to work with them while also doing all of the quality checks that we know we need?” I was involved sort of on the product management side of working on that in the beginning in the first year or so of that project.

[00:33:26] JM: Is it costly your bandwidth-intensive or anything to build a scraper or is it cheap enough?

[00:33:32] RFS: In general, I think a single scraper is pretty cheap. What we didn’t – Or what we learned. What we didn't anticipate and what we learned was that the complexity really was just figuring out how to handle all of the edge cases that come up when you're doing this at a really large scale. I think what was hard for us to anticipate in the beginning was, “Okay. Well, we did 10 of these ourselves. It seemed pretty straightforward. Let's see if we can outsource this.” Then when we’re doing thousands and thousands of them, there's just all these things come up. A big part of what we've been building is how do we automate the QA and how do we automate the evaluations? How do we provide feedback automatically to whoever actually is the person working on that scraper or configuring at scraper when we can't configure it automatically.

I think doing a single one is cheap, but there're like an exponential difficulty that happens when you start doing many of them. That's just been the sort of the surprising thing to me, was that it wasn't like doing a thousand of them was a thousand times harder. It was like doing thousand times squared harder.

[00:34:35] JM: I don’t know how much we want to get into it. I’m just very intrigued by the scraping, engineering scraper. Do you like scrape a bunch of information and then you have it all stored in a bunch of blob storage and then you clean it and then you actually turn it into a canonical data, or can you tell me anything about the cleaning process?

[00:34:53] RFS: Yeah, I can tell you a little bit. I mean, I think we’ve had a lot of different approaches on this, and I think in general the perks that we’re most invested in right now is less of the let's sort of scrape the raw data and then push it into some sort of centralized information extraction and more like let's do the information extraction pretty close to the page to that individual scraper. You could imagine the two spectrums of that. One spectrum as I'm literally just going to scrape the entire webpage code and then throw it into some sort of thing that extracts information.

We've just found that we've had more success doing sort of the opposite end of the spectrum, which is let's try to get as targeted as possible collecting from the page directly so that by the time the data is entering our system, there’s already some level of structure to it. Then we have a variety of systems that then handle all of the types of problems that come in from that stage. Open hours is a good example that we’re talking about where we've at this point just built a lot of logic around the types of things that come up with open hours and the ways to correct it and what to do when there's no information and things like that. Yeah, I think that kind of answers – Gets to the question a little bit.

I mean, there’s a lot of stuff we do post-ingest to the data. That's where most of the work is actually happening.

[00:36:08] JM: Like what kinds of stuff happens there?

[00:36:10] RFS: Yeah. Another one of the big challenges of ingesting 10,000 data sources or whatever is that you have all these different systems giving you information and you have to figure how do you join those and merge those together. The sort of merging problem is I think sort of the big important but unsexy problem in data companies in general, which is that if I have three different data sources and data source A tells me there's a Starbucks at 545 Main St., and data source B tells me there's a Starbucks at 543 Main St., we have to decide like are those to Starbuckses different or is not the same Starbucks that just had a typo somewhere?

This sort of entity resolution is a really hard problem when you're working with many, many different data sources, because there's no easy common join key to merge them all together. We care a lot about getting the duplicates, removing the duplicates, being good about duplicates. If you're not – I think the biggest issue you're going to have with most datasets that you get are going to be duplicates, because there's this problem of merging that's really hard.

We've built a lot of systems including machine learning applications to try to help us understand when we have these two sort of noisy datasets that are sort of fuzzy information about places, how do we decide? Is that the same place or two different places? Because it's also not unheard of to have two Starbuckses like across street from each other. So it turned outs to be like a super complicated hard problem to do even for a human looking at it, let alone, a machine. That's like a big part of the post-processing that’s happening during cleaning.

There’s also data imputation. So categories are a good example where it's hard to get good category information. If you do get good category information, it's going to be at some random system that whatever that source has invented to do categories. Categories are hard because there's not really a ground truth for like what a category is supposed to be. If you ingest data from many different sources, they might all use different category systems that they've invented and you're trying to resolve that into some unified category system that's convenient for your users and often category information just missing completely. We do a lot of data imputation around categories where were saying, “Okay, we didn't get any category information for this particular place. Let's use everything else we know about this place to make a good guess about what we think the category is,” which is also like a very fun machine learning problem. Yeah. Those are some examples of the types of transformation that we’re doing to the data once we get it before we send it out.

[00:38:30] JM: So the imputation thing, the categorization, that's like we might take – What kind of information can you get that will allow you to – I mean, imagine, you could do some kind of clustering. The cluster businesses that you don’t – Is this a car's repair store? Is this a bakery?

Is this a pizza shop? It's called like Tom's Slices or something. It’s like, “Okay, Slices. Maybe it's a pizza store.”

[00:38:57] RFS: Maybe it's like a paper company or something.

[00:38:59] JM: Yeah, or a knife company.

[00:39:01] RFS: Yeah, right, exactly. Yeah. You're hitting the nail on the head in terms of why this is hard, and there’s I guess two things I can tell you about that. One is that, often, there is some sort of category information that we’re getting from the source. For example, that might be that let’s say that’s a store locator that we've crawled, like walmart.com. We have gone to the trouble of essentially hand labeling all of these major – Thousands and thousands of what we call brands or corporate chains.

In that case, like that information is sort of already tagged with a category when it comes in. But there are other systems that we get sources – Other sources that we get data from where they have some sort of category information, it’s just a different category system that what we want our customers to have. So you have this like complicated, maybe it's a tagging system or maybe it's just some unique category system whether you’ve given that one category. That could be a many to many mapping or it could be a one to many, or many to one mapping for going into our category system. Some of that is you could build heuristics to solve, but some of that you can also just say, “Okay, I –” The way we ultimately approach it is we’re going to take all these different types of information about a place. Some of it might be category information for a different system. Some of it might be name information. We do a lot of that. We do a lot of name analysis to try to make a guess.

There are other things you could do too. You could imagine like – I mean, you think about this stuff that you do if you were a human Googling this. Where that place is located is also going to inform you something about that place. If you see something that's in the middle of FiDi in San Francisco, it's probably not a car dealership. It's probably something else. There's things like that that – Not always. There are some car dealerships in San Francisco. But things like that are rules that you can learn and make educated guesses about. Yeah. There's a lot of different ways that we're pulling information in, and ultimately we're trying to compile all these like small little signals to make a best guess.

[SPONSOR MESSAGE]

[00:40:55] JM: This episode of Software Engineering Daily is brought to you by Datadog, a full stack monitoring platform that integrates with over 350 technologies like Gremlin, PagerDuty, AWS Lambda, Spinnaker and more. With the rich visualizations and algorithmic alerts, Datadog can help you monitor the effects of chaos experiments. It can also identify weaknesses and improve the reliability of your systems. Visit softwareengineeringdaily.com/datadog to start a free 14-day trial and receive one of Datadog's famously cozy T-shirts. That softwareengineeringdaily.com/datadog.

Thank you to Datadog for being a long-running sponsor of Software Engineering Daily.

[INTERVIEW CONTINUED]

[00:41:49] JM: How do you know what to prioritize? I know you've been with the company for a long time. So I'm sure you've like worn a lot of different hats. You have a pretty good understanding of what's going on throughout the company. I can imagine a number of different states that the company might be in. I can imagine a state in which you've got these new data products and you're really trying to figure out how to just get them to customers so that you can start getting feedback. I think you've made it beyond that phase. Then the customers start to use them. They started to say, "Okay. We're using this location data. We're using it in this way." You interface with them closely. You get feedback that they want this additional field in the dataset maybe. You can kind of imply customer success. You can iterate on the product just by virtue of being customer success for people who are consuming the location data.

Eventually, you're going to reach a point where you'll probably expand into some other data category or some other product altogether. I'm very curious as to how close you are to that additional product, or is it like the more you look at location data, the more depth you see to that problem that you just want to go deeper and deeper and deeper into it.

[00:43:02] RFS: Yeah. I think it's a great question, and I think that I think is often the case, there's always more depth than you realize there ever could be. I think, definitely, we've discovered more and more depth around through the core products that we've been working on as we've been doing it. But also even within the last few years, the products that we've built have been influenced highly by that sort of customer feedback you're talking about.

In the beginning, we only really had one product, which was sort of a merge of what we call core places and geometry today. Then as we went on, we realized there're sort of distinct needs around core places versus geometry. We decide to split those into two different products, which are often brought together. The human movement stuff, the pattern stuff was something that we added a little bit later and took us a while to figure out the right way to do that. But I think there is a lot of – Still a lot of opportunity in just sort of points of interest in places that we have on our roadmap.

One obvious way to expand that is right now we've mostly been focused on sort of what we describe as these places of commerce, like these places where consumers can go and spend time and money, but there's a lot of places in the US that aren't in that bucket. Things like office buildings. The SafeGraph office is currently not in a SafeGraph dataset, because it's not a place that consumers go and spend time or money, but there're a lot of reasons why you might want to know where offices are. Is this an office building or residential building or a mixed-use building? There are things that we call sort of like industrial POI, which are like loading docks, factories, warehouses, things like that are currently aren't in the dataset. There are a lot of things like that that we could imagine trying to go out and find sources for and include and some of our customers want some of those things and there's new types of use cases that are opened up by doing that.

Then there's another is another way that I think – There's also internationalization, which is sort of I think a cop out for how to expand, but definitely something that's on our mind. Right now we're focused on North America and United States and Canada, but there's lots of reasons why you want to know data about everywhere in the world, and that's something that SafeGraph ultimately wants to be able to do. We talk about that a lot too in terms of what's the next country to launch in and when should we do that. Just in this last year, we added Canada. So that was new for us. Just in the last 6 or 12 months. There's internationalization stuff.

Then the final thing that I would add, I'm just giving you like a laundry list of our product roadmap here, but the final thing I would add is that there's what I would put on the bucket of “integration”. We talked earlier about mostly SafeGraph data is delivered as sort of a bulk download, and many of our customers are quite happy with that. But there’s a lot of innovation and iteration we can do on sort of this delivery component of the data. In some cases, that means delivering the data into other types of systems that people want to use it in.

For example, one of the largest, most prolific geospatial analytics software is this company called Esri, and a lot of people use Esri today for a variety of things and a lot of people want to use SafeGraph data inside Esri. We've actually done a lot of work to make it easier to use SafeGraph inside Esri. Sometimes that’s by partnering directly with Esri. In fact, now, like some of SafeGraph data is natively in the Esri product. Some of that’s also just helping figure out is there anything we can do to make the data easier and is there a different type of format or a different type of ingest delivery style that we can do that would make it easier to ingest it into these platform? I think there's a lot of stuff thereto that we haven't done that we could do. I think there're a lot of things to do sort of around this core value proposition of physical places.

[00:46:21] JM: No. I think that's pretty fair. People you Spark for ETL because it's more flexible, ETL it from your data lake into your data warehouse. Maybe in some cases that data warehouse is Spark itself. It’s not exactly a data warehouse, but can function as one.

[00:46:37] RFS: Right. That's sort of how we do it SafeGraph today, right? SafeGraph, our “data warehouse” is essentially just AWS S3 that is organized in different ways. Then we’re a small team and we’re a fairly technical team. So we’re quite comfortable to go in and write Spark jobs to query against that data. But I can imagine in the future, maybe we want to have people working what that data at SafeGraph that aren't data engineers or aren't super technical and we want to have a way that like let them write SQL on that more easily, things like that. Yeah, I think that it depends on what the user of the data needs and how you're trying to deliver that. I think it also – Data warehouses, you have to make decisions and assumptions about how you’re structuring that data so that that determines how you can query it. Sometimes that's fine and that's what you want, but other times that's going to limit what the user can do. Those are all the considerations you have to think about when you're building that pipeline to put the data in.

[00:47:30] JM: Do you use Databricks or Amazon?

[00:47:33] RFS: Yeah. We're heavy users of Amazon and Databricks. We're using Databricks every day both for ad hoc things and data exploration, but also like building pipelines and running jobs and things like that. Yeah, we use Databricks a lot.

[00:47:47] JM: The value from Databricks, that's the fully integrated, you get like a really nice frontend to work with.

[00:47:53] RFS: Yeah. I think I would say there're two value propositions that I see in Databricks. One is like it manages your AWS resources for you. So you don't have to spent much time on sort of like the DevOpsy, like setting up resources, setting up servers. It's very easy to like click a button and Databricks spin up a cluster and start working with it.

For me, especially that I don't have this data engineering background, that's super helpful to me, because I don't have to be like mucking around in the AWS like command line interface and like spinning up things and configuring these Spark clusters and things like that. I think that's like the first thing that's super valuable.

The second thing is it has this notebooking system that I think is just a very convenient way to interact with data and write little scripting functions and write little queries and things like that. That's sort of this UI component. It's not a Jupyter Notebook, but it feels a little bit like that, and that's just a very comfortable place to work if you're doing ad hoc analysis or even building pipeline, prototyping pipelines, things like that.

[00:48:46] JM: As we begin to wrap up, before getting into software entrepreneurship, you were in academia. You were working on neuroscience. How was that shift out of academia?

[00:48:59] RFS: Yeah. I was doing a PhD in neuroscience for a little while. One thing that I realized is that what I liked the most about science was I liked working with data. I liked communicating about data. I liked doing analysis. I liked designing experiments, thinking about experiments, but there was this one very important part that I did not enjoy, which was I didn't really liked doing experiments or collecting the data. I found that to be sort of this weird combination of like boring and stressful. But I definitely came up as a scientist and science is a big part of my identity, because I just spend so much time working in that space in college and grad school.

What I realized was like these things that I like about science that are like working with data design and experiments, communicating about data, there were a lot of opportunities to do that outside of academia. Increasingly so, companies care a lot about data and were working with data, and I knew very little about the real world before I left academia, but I was fortunate living in the Bay Area and have friends of friends that worked at tech companies and just got a sense that like there's cool stuff to do there. That was sort of my transition out, was had this vague idea that there'd be interesting thing to do and data. I also had an opportunity to work on a startup when I first left grad school, which was serendipitous. But yeah, it's really just been trying to figure it out as I go, and I think in some ways it's interesting to look back now, because you realize that there are lots of things that are inefficient about how academia works and lots of things that can be improved in academia and it is interesting to understand, to witness sort of the difference in the pacing and the different in the way that you think about things. I think there're pros and cons. I've written in the past about things that I miss of academia, but I'm super happy. No regrets at all. It's been the right choice for me for sure.

[00:50:40] JM: Neuroscience is really cool. I took a few classes that are in that realm, and obviously a lot of it has not been figured out. Do you have any beliefs about neuroscience that run contrary to popular opinion, or maybe certain appreciations of the field that might come in surprising or novel to a listener who is not well-versed in neuroscience?

[00:51:08] RFS: Yeah. Sure. I mean, I think it's funny. At SafeGraph I actually – We do an annual retreat every year at SafeGraph, and the last three years I've given like a little neuroscience talk at ever retreat, talking about like some topic in neuroscience that I think it's interesting, because I think neuroscience is like an inherently interesting thing to people, right? I think people understand that the brain is the super important part of who they are. It's the super complex structure. I think people are inherently interested in psychology and neuroscience for that reason. I think in terms of like beliefs that I have that might be contrarian, or whatever, to neuroscience at large. I think the one thing I can say about that is that I think there's like this very hard question to understand, these things of things like what is consciousness? What is identity? What is free will? These are super-hard problems that in general like I don't think we have any good answers for today, but there's been a lot of philosophies, time spent on those questions and there's been some neurobiology time spent on those questions, and I think that, in general, I think neuroscience and neurobiology still has like a big reckoning to do with how we are going to answer those types of questions. I think some neurobiology in the past have sort of said, “Well, we’re just not going to worry about that stuff.” I think, to be fair, there's a lot of stuff to figure out first. We have to understand how neurons work and how networks work and all these things, but I think there is some sort of philosophical perspectives around what does it mean to be a conscious being. Do I have free will are not? That, ultimately, I do think neurobiology should be able to answer, and I'm not sure that we’re like totally ready to do that yet. So I think it will be interesting to see how that evolves across our lifetimes. Yeah. I mean, I think it's a super interesting area.

[00:52:43] JM: To what extent do you think the human brain is a good model for the computer?

[00:52:47] RFS: I think it's pretty bad model for a computer. In general, there's I think some very important distinctions to understand about the difference between a brain and a computer, and the biggest one is this idea of parallel processing versus serial processing. Because, in general, computers are serial processors, and we've generally thought about computers as serial processors. That’s super powerful, right? You can do – If you can do many, many competitions very fast doing them serially, it’s fine, and that's why computers are powerful. But that is not at all how the brain works. The brain is this massively parallel distributed processing machine, and I think in some ways the computer can actually be a bad model for the brain and get in the way of understanding how the brain works, because there's just so little about the brain that is serial. When you’re having a conversation like this or our in the world, information streaming into your brain through many different senses, that information is being processed in a highly parallel way. I think the ways in which ultimately that guides behavior and makes decisions is very complicated. In our heads, it feels like a serial process. It feels like I’m a singular entity making serial decisions, but in reality there's a lot of evidence that shows that there is not a lot of singularity to your behavior and experience, a lot of stuff happening parallel, a lot of stuff is happening unconsciously, and I think it's really hard to wrap your head around just how different that is from a serial processor.

[00:54:05] JM: Ryan, thanks for coming on the show. Great talking to you.

[00:54:06] RFS: Been a pleasure. Thanks for having me.

[END OF INTERVIEW]

[00:54:17] JM: When I'm building a new product, G2i is the company that I call on to help me find a developer who can build the first version of my product. G2i is a hiring platform run by engineers that matches you with React, React Native, GraphQL and mobile engineers who you can trust. Whether you are a new company building your first product, like me, or an established company that wants additional engineering help, G2i has the talent that you need to accomplish your goals.

Go to softwareengineeringdaily.com/g2i to learn more about what G2i has to offer. We've also done several shows with the people who run G2i, Gabe Greenberg, and the rest of his team. These are engineers who know about the React ecosystem, about the mobile ecosystem, about GraphQL, React Native. They know their stuff and they run a great organization.

In my personal experience, G2i has linked me up with experienced engineers that can fit my budget, and the G2i staff are friendly and easy to work with. They know how product development works. They can help you find the perfect engineer for your stack, and you can go to softwareengineeringdaily.com/g2i to learn more about G2i.

Thank you to G2i for being a great supporter of Software Engineering Daily both as listeners and also as people who have contributed code that have helped me out in my projects. So if you want to get some additional help for your engineering projects, go to softwareengineeringdaily.com/g2i.

[END]