EPISODE 1066

[INTRODUCTION]

[00:00:00] JM: Social distancing has been imposed across the United States we are running an experiment unlike anything before in history and it's likely to have a lasting impact on human behavior. By looking at location data of how people are moving around today, we can examine the real-world impacts of social distancing.

SafeGraph is a company that provides geospatial location data to be used by developers and researchers. Much of the SafeGraph data is aggregated from cellphone GPS pings which identify where anonymized users are in the world. This dataset provides the basis for SafeGraph's of social distancing metrics which measure how frequently people are coming into contact with one another.

Ryan Fox Squire works at SafeGraph and he returns to the show to discuss social distancing metrics and the real-world research that has come out of studying those metrics.

If you enjoy this show and find it useful, you can help us out by subscribing. You can become a paid subscriber at softwaredaily.com/subscribe. It's $10 a month or $100 a year and you get access to all of our old episodes without ads. That's over 1,300 episodes. There's lots of content in there on anything that you're learning right now.

Also, at softwaredaily.com, you can find question-and-answer and all kinds of other content relating to this episode, which can help you augment the knowledge that you're going to learn from today's show.

[SPONSOR MESSAGE]

[00:01:33] 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.

X-Team is able to support a wide range of needs. If you need DevOps, or mobile engineers, or backend architecture, or ecommerce, or frontend development, X-Team can help you with what you need. They've got a full-range of technologists who can help with AWS, and Go lang, and Shopify, and JavaScript, and Java. Whatever your engineering team needs to get to the points of scale that you want to get to, X-Team can help you grow your team. They offer flexible options if you're looking to grow your team efficiently, and their model allows for seamless integration with companies and teams of all sizes. Whether you're a gigantic company like Riot Games, or Coinbase, or Google, or if you're a tiny company like Software Daily. You can get help with the technologies that you need. If you're interested, you can go to x-team.com/sedaily. That's x-team.com/sedaily to learn about getting some help with your engineering projects from X-Team.

Thank you to X-Team for being a sponsor of Software Engineering Daily.

[INTERVIEW]

[00:03:18] JM: Ryan Fox Squire, welcome back to the show.

[00:03:20] RFS: Thank you for having me.

[00:03:21] JM: You work at SafeGraph, and SafeGraph is a location data company. In addition to that, you have gradually become a COVID-19 data company. Explain why you started studying COVID-19 data.

[00:03:35] RFS: Yeah. Before the pandemic, SafeGraph is a data company. We call ourselves the geospatial data company and we're focused on trying to understand the physical world. In particular, we're focused on understanding all of the consumer points of interest out in the United States, so all the coffee shops, malls, restaurants, airports, all these places that people go and spend time. We've really built some very comprehensive datasets around these places.

What's interesting is that, with the pandemic, it's obviously affected so many aspects of our lives, but in particular it's affected the way that we interact with the world around us so much. As it started happening so quickly, we realized that the data that SafeGraph had was providing a very powerful window into how the world was changing. Basically, what happened is that we have a long history of collaborating with researchers and academics who have interesting questions they're trying to understand that our data helps them when we make those donations to those academics. Then as soon as COVID started happening, these researcher started reaching out to us asking, "Hey, can I get your data to answer these other types of questions?"

We also found that a lot of these academics were collaborating with local and state governments and the federal government. So they were asking, "Hey, I'm trying to help the CDC work on this project, or I'm trying to help the mayor's office, the City of LA. Can I get the data to help with this?" We just realize like, "Wow!" There's a lot of demand and that the data we had really provided a lot of value to these different organizations. We kind of just saw that opportunity and wanted to try to help. We're doing all of these at no cost. This is not like a business that SafeGraph is building. This is just something that we think we can help with right now and we're happy to be able to try to help in any way we can.

[00:05:03] JM: The data that is being generated for measuring social distancing specifically is refined from raw data that you have that you always had at SafeGraph. That's your core business. Can you explain how you have refined the core datasets for SafeGraph and how you've turned those into things that are useful for measuring social distancing?

[00:05:31] RFS: Yeah, definitely. One of the key ingredients that SafeGraph has always used to build our datasets is this anonymized population movement GPS data that comes from smartphones, and this is not data that SafeGraph is collecting directly. We aren't out there measuring GPS data, but it's data that we're able to get in this anonymized fashion. Historically, we've use the data to build these anonymized aggregate measures of foot traffic to businesses.

To do that required a whole lot of work to take that raw GPS data and build these acrid visits involved building a whole concept of how are we going to detect these geo-temporal spatial clusters of GPS points in time. How do we know if there's been actually a visit to a business versus just walking past that business? I had to understand all the quirks around GPS data that

comes with the GPS data. There’re all sorts of weird artifacts that you get and you get data that's coming through, label the GPS, but it's actually simply some sort of IP address lookups. You get these things that we call syncs, which are unique points that have way, way, way too many GPS pings than you'd expect for normal because they're not really a GPS point. They’re just some IP address lookup point or something like that. There's all these quirks that we've gotten really good at understanding how to deal with. To adapt that to understanding social distancing, physical distancing, was pretty interesting.

One of the key concepts that we also have used historically is this concept of a home base or a common nighttime location. One of the things that we ran into early on was we would find that we were often getting lots of artifact visits that when people's home base and near retail POI. If you live near Starbucks or a coffee shop, we run into these problems where we were labeling someone as visiting that Starbucks instead of being at home. So we had to think really hard about how to build this concept of a home base so that we could sort of identify when that's happening.

In the context of physical distancing, we’ve now sort of turned that around and said, “Okay. Rather than filtering out home visits, it’s interesting to understanding how was the population in general staying home or not?” We’ve built these datasets that are all based around on average we know where people's home bases are and we know on average how much they're leaving those home bases. So we’re able to produce these metrics like what percent of the population is leaving home each day on average by different census block groups.

[00:07:40] JM: The data that has been refined, it’s GPS data. GPS pings from, for example, a smart phone. You have lots of GPS pings that get refined into locations and then you can figure out has somebody been staying at a location for 14 out of 24 hours. Well, that's probably their home, and you can look at that data point before and after the social isolation started and see if a given individual has been staying in their home.

[00:08:14] RFS: Yeah, that's basically correct. I think the only thing I would sort of adjust there is at the end of the day these are all aggregate measures. We’re not identifying individual devices and individual people per se. The data that we've been sharing with researchers and academics and governments, it's all aggregated at the level of different census block groups.

What that lets you say is for an entire census block group, what is like the average aggregate behavior of people who live in that census block group. For every purpose, that's all you want to know. You don't want to know data about individual people anyway, and that's sort of the power of this aggregation, is that you can use the data that is in its raw form quite sensitive and has all these privacy complications, but if you aggregate it and treat it with the sort of correct privacy algorithms, then you have a very privacy-safe view in summary that is very actionable and helpful for policymakers.

[00:09:03] JM: You refined the data to create social distancing metrics. Can you describe the specific social distancing metrics that you defined?

[00:09:12] RFS: Definitely. This was like a case where we had these users coming to us asking for data to answer certain types of questions. People were asking, "How can we measure the effectiveness of these stay-at-home policies?"

At SafeGraph, we like got into a room virtually and brainstormed, "Okay. What are all the things we could do that would help people answer these questions?" and we came up with a big list of ideas and we started trying to build them and test them out and we built this product in a week. It was this very exciting, like, "Let's put this together quickly as possible so we can get in people's hands."

One of the things that we came up with out of this product, the first thing was we decided to aggregate everything by census block groups like I mentioned and summarize different statistics for every census block group. You can imagine the dataset as like a big spreadsheet or each row is a different census block group and there's like 220,000 census block groups in the US or so. Each row is a census block group and then we have these different columns that describe different attributes of the people who live in that census block group, and we tried to make this sort of as agnostic as possible. We don't want to give people highly engineered, highly manipulated summaries, because there're all these assumptions that go into that. We want to try to give people as close to a unbiased perspective of what's in the raw data as possible. The columns are very simple. They are things like, in our sample, how many devices do we have living in this census block group? What's the sample size? That's one column.

Next column is of those devices that we have in the sample, what number of them did we see never leave their home this day? Never leave their home base? Then we have another column that says what number of devices did we see leave the home for a certain number of hours, and we called that a part-time work count. We don't actually know if those devices are working or not. We don't know what they're doing, but we do know that they went out and they stayed at some other non-home location for between 3 to 6 hours.

we also have another column we call full-time work, which is they went out, left their home and they stayed at some other location for more than six hours. We have these different types of counts. Then we also have some other columns that we're trying to summarize things like when they did leave, how far from the home did they go? We have this summarized in different buckets of did they travel less than 1000 meters from home? Between 1000 to 5000 meters from home, et cetera? So the people can try to understand what people are leaving, how far are they spreading away from that home base. Those are some examples. I think there's maybe even a couple other columns that I'm forgetting off the top of my head right now, but the core idea was trying to give summary data about whether people are leaving home or not, and when they do leave home, how long and how far did they leave home for?

[00:11:37] JM: Now, of course we could say that people leaving home is going to have some correlation with whether or not those people are engaging in social distancing, but the more critical question is, is this person interacting with other people? Do you have any understanding of how closely these people are coming into contact with other people?

[00:11:59] RFS: Yeah, that's a great question and a super important point. One of the ways that we've seen some of the researchers and governments that we're collaborating with try to address this question is actually by looking at not our social distancing metrics data, but actually what we call patterns, which is the business foot traffic data. What they're looking at is they're saying, "Okay, let's look at all the places that we see people are actually going to visit during this shutdown." People are still living in their homes and go to the grocery store, go to pick up food delivery or whatever. Let's look at all those businesses. Then for those different types of businesses, they can classify different danger levels based on both the density of visits in time and the density of visits in space, because SafeGraph data also has data about the shape of

that building and how big that building is. It's super interesting, because there are some researchers that just count for study a couple days ago where they tried to classify different categories of businesses in terms of the relative benefits to the economy for being open and the relative risks to public health for being open. You can sort of imagine like a 2D space where different categories of businesses are in this 2D space, and some businesses are higher risk than others because they cause lots more crowding. Things like grocery stores, or sit-down restaurants are higher risk, whereas is things like a bank also has a big impact on the economy, but it's a much lower density of visits. Their argument is something like, “Maybe we should think about opening banks first, because it's going to help the economy, but it's not as high risk to public health.”

Again, it's sort of a view of like what's the crowding and density of people on average, but it's not going to tell you things like did this device get close to another device? Sort of that contact tracing level stuff? That's just not what you can do with the data, because again we don't have individual devices sort of identified. It's all aggregate and summarized.

But still, with that aggregate summarized view, you can make these very powerful inferences in terms of what are the hotspots in a community where governments should be concerned about people crowding? Where do we need interventions? What's happening on average in a city or in a county? Things like that.

[00:13:58] JM: Are there other datasets that are useful to cross the SafeGraph social distancing dataset with?

[00:14:05] RFS: Yeah, definitely, and in fact that's been one of the most exciting things about the work we've been doing I think during this pandemic, is that to try to enable all these governments and academics and researchers and scientists, we've sort of created what we’re calling the SafeGraph COVID-19 data consortium, and all that really is, is it's a shared workspace on Slack that we've created that we invite people into. It's a way for us to deliver data to these people and it's also a way for them to collaborate with each other and share ideas with each other.

Then the other thing we've done is we've tried to reach out to other data partners that we think have data that's valuable and say, "Hey, we have this community of over 1,500 people in this community across the world. We're making our data available to them at no cost, and do you want to also contribute your data to this group?"

We've had a number of data partners come into share data. So let me give you some examples of some high-value things. SafeGraph is really good at understanding what we call sort of consumer points of interest, these like commercial places that consumers go and spend money, but there's lots of places in the world that we don't really consider our bread-and-butter, things like residential areas, residential apartment buildings, other types of places that just aren't historically where SafeGraph is focused.

We've brought in some data partners that have super accurate address information and geospatial information for residential places in the United States so people can understand where are all the residential areas relative to the consumer points of interest. We've also had partners come in bringing transaction data. SafeGraph data tells you about did people visit the business or not? What was the foot traffic to that business? But it doesn't tell you when they're in that business what did they buy? What was the size of the transaction? How many transactions are happening? Those are data that you can get from payment providers, and there's super valuable data that you can again aggravate at the level of businesses or ZIP codes and things like that to understand it's sort of another angle on commerce that's happening.

We have researchers in the group that are publishing studies using only data that other providers have provided. There was a study that came out just last week in the group that use this address information and transaction data but actually didn't use SafeGraph data at all. We're just happy to be able to provide a place for those different data to come together and we're hopeful that it's a way to, like you said, sort of the chasms of the different datasets and help us figure out the best ways to combine them because they are all sort of originally coming from different companies.

[SPONSOR MESSAGE]

[00:16:23] JM: There are two ways to add analytics to your application, you can build them yourself with basic charts and dashboards using free open source charting libraries, or you can use a comprehensive analytics platform from a partner that you trust. If you've tried to build it yourself, you know that free actually is not so free. There are hidden costs like time, and maintenance, and technical debt, and those hidden costs can really add up.

Check out Logi Analytics. Logi Analytics is developer grade embedded analytics solutions and they make it easy to create branded dashboards and report the scale within your own application. You can stop wasting time piecing together analytics and allow yourself to focus on your core application. You can go to logianalytics.com/sedaily and you can get a demo to see what is possible with Logi today. Go to logianalytics.com/sedaily. That L-O-G-I-analytics.com/sedaily.

[INTERVIEW CONTINUED]

[00:17:37] JM: One thing I found inspiring about this was when I was looking at the research reports, you have a lot of public institutions and public sector organizations that are doing data science work at least as I understood it. So there's some public sector organization from Washington, for examples. I'm just imagining this person who's employed in a fairly bureaucratic state government organization and they just wake up one day and say, “You know what? It's time for data science. I've got these datasets, these publicly available datasets that I can use and I'm going to do data science for my organization.” That's actually an important milestone. That's an important idea, because I don't know about you, but me personally, whenever I see like a study that comes out of NIH or really any scientific organization, it doesn't matter how esteemed the organization is. I just look at these organizations. I see thousands and thousands of people. I see government grants and I just do not trust the results and I don't trust the data. I mean, it's an aggregate of distrust.

Obviously, directionally, I trust where science is headed, but any individual study I'm very skeptical of. I think in terms of going towards a place where we have better accountability in science, having third-party datasets is a very good milestone.

[00:19:01] RFS: Yeah. You're right on here. There's something like very exciting happening, and in some ways I think that the pandemic is accelerating some of this, because the government woke up and there was this emergency happening. In many ways, it was new and we weren't prepared in lots of different ways to think about these new challenges. I think it's been really interesting to see all the different ways that these new types of datasets are coming out and being used by people.

I agree. It's been really cool also because we've seen at various levels of government sort of these data taskforces assembled consisting both of government employees as well as people from tech who are volunteering or trying to help out, and there's tons of different ways that's helping, right? There''re been all sorts of companies where I've read about an app that helps track flu symptoms and suddenly they've realized that their data is valuable for tracking fevers across the United States. Maybe that's an input variable to seeing where the pandemic is spreading. You have SafeGraph data that's all about where people are staying home or not staying home.

I think one of the things that we're most keen on right now is as the government starts thinking trying to reopen the economy, SafeGraph data is going to be very viable to help inform what's the right way to do that, and it's not an all or none thing, right? What's sort of the right amount of shelter in place that we should implement and how to ease that up? Again, yeah, these are like data problems that I think we're in a unique time where the data is available and can it really help answer these questions in a unique way? That's sort of a forcing function to drive all these organizations to make sure that they're ready to do it and they're up to do it and they have the tools to do it and they have the skills and the people to do it. It's only time for data scientists and technologists to shine who are public servants already trying to help from this perspective for a long time. Now is definitely a moment for them to show how valuable that kind of work can be.

[00:20:36] JM: Who are these people that have developed a background in data science and wound up in a public sector organization? Like I think of data science as something that is still very much in the technology industry and somewhat making its way into other businesses. I haven't done much coverage on government organizations and to what degree they have data science or they have tech teams but it seems like some of these places have completely qualified data scientists. They have a real public servants with engineering backgrounds.

[00:21:10] RFS: Yeah. I'm not sure I have a great window to sort of characterize that movement in general, but I have a couple clues about that that I know have sort of been happening and it is more recent, and I think a lot of it has been driven from starting with campaigns. There was a big story in the 2008 election around how Obama used social media and technology and sort of adtech strategies in a way that had never been done before, and I think anyone who runs for office is starting to realize how powerful ad tech is, because running a campaign is marketing and there are all these tools that have been built in ad tech that now can be applied to campaigning that are very powerful. I think that's been one of the windows that sort of drive technology into government is by people realizing that they can get elected better if they use this technology. Then I think it bleeds over into governing.

In the Obama Administration, we had the first ever chief data science officer of the United States, this guy DJ Patel. I think you use the technology to get elected and then you realize you can use these things to work when you're governing, and I think that has driven a lot of the talent into that space. I think there a lots of people who care about politics and care about government who can get into it that way, and I think that's at least one part of the change that we've seen in the last 10 years around this.

[00:22:21] JM: Okay. Let's talk more about what you've discovered in terms of how people are carrying out social distancing and how it's affected other aspects of social life. What's the correlation between people who are staying home and the growth rate of the virus?

[00:22:39] RFS: That is a great question, and I don't have a good concise answer to that. That's definitely been one of the hot topics that a lot of people are working on. It turns out to be maybe it's not surprising. It's a complicated question. One of the reasons it's hard to answer in a simple way is that if you just look at the data sort of first pass, what you see is that the places where people are sheltering in place the most, where people are staying home the most, are the places where the virus has spread the most. I think that makes sense because those are the places where there's been the early outbreaks, there's been the most urgency, there's been the strongest sort of government reaction, but it does create this weird correlation where you're seeing if you don't do anything smart and just look at the data sort of in a raw form, you see

there's a positive correlation between staying home and virus growth because there's a different causal relationship there.

What you need to do to answer this question correctly is control for all these factors. You have to control for how much were people staying home given a certain timeline, given the start of a state policy, given the state of a virus when the policy was enacted? You have to model the virus growth controlling for all those variables.

We only sort of have at this point still anecdotal data about this. There was a big report that came out the other week from the CDC using SafeGraph data. This is their weekly morbidity report, and it was sort of a case study. It was, “Let's look at these four areas in the US where there's been early COVID outbreaks.” They’ve looked at Seattle, San Francisco, Bay Area, the New York, New Jersey area, and the first-order question was just our people staying home at all? Do we have any hope of these state policies working? They had some very interesting charts showing how people started staying home more relative to the enacted of these local policies. But it doesn't actually answer your question, which is how do we know how much shelter in place actually slows down the growth of the virus, and I think honestly that work is still trying to be put together, because there are all these confounding variables and factors around how much has the virus already grown by a certain point, and that data unfortunately has been very limited due to all of the limits and testing. It's a picture that’s still been put together. It’s a very important piece of the picture, because ultimately you want to be able to have very concrete statements to say sheltering place this much slow downed the virus this much. But I don't think we have a clear picture yet.

[00:24:44] JM: How has travel changed?

[00:24:46] RFS: Yeah. I can say a couple things about that. I think one of the things that we've looked at and we actually have a dashboard on our website right now that is updated daily looking at what's the foot traffic levels, the different categories of businesses and categories of places? One of those categories that we’re tracking are airports, and airports has fallen off a cliff. I think airports is down something like 95+ percent.

That data suggests that almost no one is going to the airport and we know from other sources of news that airlines are suffering immensely right now and there's the big federal bailout stuff around airlines and things like that because no one's flying. That's sort of like long-distance travels seems to be way down. Similarly, even sort of medium or local distance traveling, and we can talk more about this, but we've seen the rate of people staying home has just skyrocketed across the whole country. It's not the case that literally everyone is staying home 100%. There is significant variance across states and even within states in terms of how much people are staying home.

In general, the biggest headline from the data is that, yes, in fact everyone in the United States is staying home more than usual and it's very clear. I think there's a lot of signals pointing to the fact that people are traveling less and it sort of depends on what view you want to take to try to dig into that more.

[00:25:55] JM: Do you have any understanding of how people actually staying at home is correlated with what their local governments are saying versus just what the overall populous is doing? Do you know any correlation between the state advisory, the governor's advisory and what people are actually doing?

[00:26:16] RFS: Totally. This is something that there's been a couple really great studies come out using SafeGraph data to show. The big effect is that, in general, people are all staying home more than usual, but like I said, there's significant variance across different states. One source of that variance is the difference in state-level policies that are happening across the US. As you know, right now we have this nationwide experiment happening where every state has implemented shelter in place, stay-at-home policies on different timelines and to different degrees of severity. What you can do is you can collect all the data for all these different timelines and policies and control for other variables like what is the state of the virus growth and what is the state of the overall calendar timeline to see, "Okay. What is the actuality effect for any given state of having a state policy?" The data is very clear, like having a state policy does make a significant difference and people stay home significantly more after the policy goes into place.

Now, that's not necessarily surprising, but I think it's very important because it does emphasize that state-level policies have a very important role to play in encouraging physical distancing if that's a public health strategy that people want to implement.

Let's talk a little bit about effect sizes. Overall, the population on average is staying home about 20% more than early February, and by 20% more, I mean, sort of in absolute terms. At baseline, maybe we're around 10% to 20% of people would stay home at any given day. Now, that number is up to like 40% of people are staying home on any given day in the US.

That 20% change happening on average across the country is pretty big, and interestingly the effect of a state policy is only about 10% of that. It's not that every state sort of goes from zero to none after the state policy comes in the place. It's more like overall all states are sort of creeping up. Over the course of March, people are staying home more and more. Presumably that's driven by a number of factors and the increased national media attention and all these things.

One of the big effects also is on March 13th when there is the national declaration of an emergency from President Trump, that seemed to have a really big effect across all the states. There's this macro trend across all states of people staying home more and more. But then on any given day when the state institutes its own policy, we see a significant bump in that state on the order of about 10% of the overall amount of staying home.

It's not like it's doing everything, but it certainly is a very significant amount of the overall staying home, and there are still states in the country today that don't have stay-at-home policies at the state level and our data shows that those states are all among the bottom of the list in terms of states that are staying home. There's a very strong correlation between these state-level policies and what actually is happening with the population in that state.

[00:28:51] JM: This whole social distancing experiment has been quite an experiment in our political partisanship and I know you have some data on the correlation between political leanings and adherence to social distancing. Tell me about that.

[00:29:11] RFS: I'm glad you brought this up, because that was sort of the next thing to talk about. Another thing that people have been looking at is, “Okay. So not everyone in the country is staying home with the same amount. What are the different sources of variance that are causing people to stay home different amounts?”

A couple researchers that we've been collaborating with, a couple different groups using SafeGraph data has tried to say, “Okay, how does political affiliation correlate with staying home or not?” Again, it's a complicated nuanced picture because you have always different timelines, you different state policies. We have to control for all those variables, but what these researchers did was they said, “Okay, after controlling for all these are the standard variables and timelines, let's characterize a county based on what percent of that county voted for Trump in the 2016 election, and that will be an indicator of the Republicaness or Democraticness of that county and then let's use that as a variable in the model to explain how much people are staying home or not.”

There's a couple interesting affects. The first effect is, in general, Democratic counties are staying home much more than Republican counties. It's a pretty big difference. I think the difference between the most Democratic counties and the most Republican counties is something on the order of like a 20% difference in how much they’re staying home. There is this interesting effect where more Republican counties seem to be staying home less overall. In addition, we see that the effect of that state policy where I was just talking about how the state policy has its boost at the state level to get people to stay home more. The effect of that policy is larger on Democratic counties than it is on Republican counties. Not only are Republican counties staying home less in general, but they also are responding less to these state-level policies.

The interpretation of this is that it has to do with the credibility that people have about their authority figures. One of the result that was interesting is that if you're in a Democratic county in a state that's held by a Republican governor, you will actually respond less to that state order policy than if you're a Democratic county led by a Democratic governor. It's not just a Republican versus Democrat staying. It seems to be also what is the relationship between your political affiliation and the political affiliation of the governor or the authority figure that's giving that recommendation.

Yeah, a lot of interesting complicated things happening there, and I think one more thing to say about that is that this is not SafeGraph data, but other survey data has shown that there is a very wide divide right now between if you ask people the survey question, how well do you think President Trump is handling the coronavirus pandemic? There's an immense difference between Democrats and Republicans in response to that question. There's like a 60% difference in response. It's a huge difference.

I think part of the interpretation also is that, Republicans, they give a lot of credibility to the Republican president, and so the Republican president's messaging around the urgency of the pandemic seems to be an important factor. If you're a Republican, you're more influenced by that. But if you're a Democrat, you're not really paying attention to that. It's sort of different selection in terms of which authority figures are driving your behavior.

[SPONSOR MESSAGE]

[00:32:05] JM:
Over the last few months, I've started hearing about Retool. Every business needs internal tools, but if we're being honest, I don't know of many engineers who really enjoy building internal tools. It can be hard to get engineering resources to build back-office applications and it's definitely hard to get engineers excited about maintaining those back-office applications. Companies like a Doordash, and Brex, and Amazon use Retool to build custom internal tools faster.

The idea is that internal tools mostly look the same. They're made out of tables, and dropdowns, and buttons, and text inputs. Retool gives you a drag-and-drop interface so engineers can build these internal UIs in hours, not days, and they can spend more time building features that customers will see. Retool connects to any database and API. For example, if you are pulling data from Postgres, you just write a SQL query. You drag a table on to the canvas.

If you want to try out Retool, you can go to retool.com/sedaily. That's R-E-T-O-O-L.com/sedaily, and you can even host Retool on-premise if you want to keep it ultra-secure. I've heard a lot of good things about Retool from engineers who I respect. So check it out at retool.com/sedaily.

[INTERVIEW CONTINUED]

[00:33:41] JM: Just a few things that I wonder affect what we can actually draw conclusively from this. Specifically, the geographic sparseness of a region seems like it's relevant in terms of the social distancing question, because if you live in a really sparse place, it's actually not that bad if you go outside. You could be adhering to the guidelines maybe not exactly in specificity, but in spirit, because you're going outside into a really sparse area versus a more crowded area. Did you do any work to try to normalize or counterbalance against the sparseness of a geographic region?

[00:34:18] RFS: Yeah, and it's a good question, because I think importantly there's already a latent correlation in the United States between political affiliation and sort of the density of where you're living, right? On average, less dense, more rural areas tend to be more Republican and more dense urban areas tend to be more Democrat. So you'd want to make sure that any differences describing to political affiliation are not simply due to these other correlated variables like density of the population.

One of the ways that you can handle that is essentially to control for population or to do things like compare counties with similar densities that are Democrat and Republican and see how the effects of these analyses vary just for the sort of those subsets or when you're controlling for population density. In those cases, you're able to see that there still does seem to be a difference between Republican and Democrat for example. You can try to control for that using things like population density measures, but I do think it's still an important point and it's one of the things that makes it hard to know what exactly is the right best measure of physical distancing. Certainly, a limitation in the way SafeGraph has aggregated the data, because we don't have a sort of adaptive concept of home base depending on whether you're in a rural or in an urban environment. We sort of just have a singular concept of home base. I think there's inevitably going to be some bias because of that. The best you can try to do is control for it sort of in the statistics and in the analysis to make sure that when you're trying to study a particular effect, you've correctly controlled for these other latent variables that are going to be correlated as well.

[00:35:42] JM: Another question I have about this, this is not directly related to the SafeGraph data itself, but I am curious about your perspective, because we've been talking to a lot of people, a lot of smart people around whether any of this makes sense, whether we're just going crazy with this social isolation. I've heard a few podcasts about this. I've read a few articles about it, but have you heard any compelling arguments that all of this is just crazy and we should definitely not be shutting down the economy in favor of social distancing?

[00:36:13] RFS: Yeah, so it's a super important point. One of the challenges of this whole situation is that you have different experts or different areas advocating for different types of things, and I think it's very hard for someone whose background is in epidemiology public health rightly so. They are sort of singularly focused on trying to save as many lives as possible and it's not their expertise to think about things like what is the impact on the economy or what are the costs of telling everyone to stay home?

Similarly on the other side, you have economists and things like that who are sort of singularly experts in understanding the impacts on the economy and how shutting down movement is going to impact the economy. It's not that they don't care about saving lives. They do. It's just that they don't have the expertise to think about sort of the benefits of staying home. I think that's one of the challenges, is that you have different groups of expertise trying to come together to make policy decisions that's why it's hard.

I also think that's why we hope that the SafeGraph data will continue to be so valuable for a lot of these governments, because ultimately what you want is – What I sort of might call like the Goldilocks lockdown. I don't say that to make it lightly, but I say that to make a point which is that you don't want to be too little or too much. There's an amount of social distancing and sort of locking down the economy, which if you do it too little, then you're going to enable viral growth, and that's bad, and people will die who don't need to die. But if you do too much, there's a possibility that you're doing it more than actually is necessary to slow down the virus growth and that extra amount you're doing is unnecessarily harming the economy. We want to find sort of the sweet spot of what is the right amount of shelter in place, stay-at-home, closing of businesses to sort of balance the two things? I don't think that we have – I'm speaking a little bit outside my area of expertise now at the national and state levels. I don't see any sort of super rigorous approach to this happening. I think people are trying though, and I know people are

working with our data actively trying to sort of understand this question. I spoke earlier about that analysis showing different benefits and risks of opening different types of businesses. I think that another key insight around this is that not all businesses are created equal in terms of either the benefit on the economy or the risk to public health.

I think you can be smart about this. I think it's important to save as many lives as possible and no one's arguing against that, but I think there's another side to it, which is how do we it in a smart way and how we not do it more than it's necessary? Because there are also all of these important impacts on the economy from doing it. You have to figure how to balance those things and bring both those expertise together.

[00:38:35] JM: What about any more nuanced plans? Have you seen any nuanced plans or things that we could do once we have mass testing, or is it not even worth thinking about these because we don't have the – At least in the United States, the political will to organize something more nuanced than just everybody stay home all at once?

[00:38:57] RFS: I think that there is the political will to do it at least at certain states and regions. The federal government generally has mostly deferred to state-level authority on this. We could debate whether or not that's the best approach, but given that that's the case, I think that there is a big role for the state governments to play, and we've seen already these sort of consortiums of states banding together. We have the West Coast states sort of trying to work together, and you have this like coalition of the Midwestern states. I think that's important, because the virus does not care about state boundaries. So you want to coordinate efforts across those regions. I do think there is a lot of interest from some of those groups. Some of those groups are considering SafeGraph data and working with SafeGraph data as well. I think they want to get this right and they want to do it in the right way. I do think there's a possibility that we could do it in a smart way. I don't think that it has to be all or none, and I know that there are people who feel the same way and are trying to make that a reality, because the stakes are quite high on both sides. It's super important and people know that it's important to try to figure out as soon as possible.

[00:39:56] JM: Some of the other studies that were produced by various people working with SafeGraph data, or maybe this was produced by SafeGraph itself, I'm not sure, but around

economic implications and consumer activity. I mean, they were negative except for like Costco, but I wonder if you have any perspective just because you're talking to a lot of smart people right now. The long-range economic implications, obviously, the bearish view is, “This is terrible. We have shut down the terrestrial economy,” and the bullish view is we are just accelerating the move towards an entirely digital economy. Maybe that’s the reason the stock market is doing well. Maybe it has nothing to do with the earnings not being announced yet. It's all just because everybody's confident that this is an overreaction and actually we have an entirely digital economy for what matters. Any perspective on that? To what extent has the physical world and the vast majority of the working population become decoupled from the physical economy?

[00:40:53] RFS: Yeah. I can say a couple things about that. I think the first is that one of the big, again, sort of headline insights from SafeGraph data so far is that not all businesses are created equal and not all businesses have been impacted the same way by the shutdowns. There’s some intuition behind that. We know that things like grocery stores are staying open and things like sit-down restaurants are doing delivering only. We have some intuition around this, but I think it is quite striking when you look at the data to see, “Oh wow! Grocery stores on average are maybe only down 15% to 20%.” I say only, but that’s a huge impact. Only down 15% to 20%, whereas things like movie theaters or shopping malls, they’re down 90%, 95%.

I do think that that's a very big difference. That could be the difference between being able to ride-out this issue versus not being able to ride this out. I think the big unknown here is that we still don't really know what the timeline for this is in the future. It’s very hard to predict I think when things ever will get back to normal. So that makes it hard to predict sort of how different industries are going to be able to ride this out or not. I don't know what to degree there is going to be the capacity of things like government-related aid to sustain some of these businesses. I think it is probably the case that there will be categories of businesses that are just going to be forever changed and may not really recover. It’s bordering on wild speculation, because I think we just don't know. Is this going to be three months? Is this going to be six months? Is this going to be a year? It’s actually quite hard to know how that's going to play out, and here's what's crazy, is that a lot of that depends on the choices that we make in the next couple months in terms of how do we try to reopen the economy or not. How do we maintain shelter in place policies or not? That's why it's important to try to get these things right now, because it's going to have ripple effects for many months to come. That's some meandering speculation for you, but I

do think that there is some parts of that that we try to start understanding now and that will make a difference in the future.

[00:42:38] JM: What's the Apple and Google contact tracing stuff? That's something I haven't really looked into. Do you know much about that? I know a little bit about that and really only a little. Contact tracing is a super-powerful technique from an outbreak pandemic control perspective, and the general premise behind it is if you can lock down an area early and you can identify cases early, if you have lots of testing, then you can really go and actually figure out the path of the virus spreading from person to person and it's a tool to control outbreaks really. That's my general understanding. This is, again, sort of outside my area of expertise. That's my general understanding.

In theory, it sounds good, but I think there're a couple problems. One is that virus has spread quite far out the US. It's going to be tough at this point to sort of contain it. I think one of the roles of contact tracing in the future is if we're able to get the virus back under control, then we might expect there to be small outbreaks here and there and we can sort of jump on those and use these tools to try to control those things. I think that's their idea, although you also need to have lots of testing to do that. I think contact tracing is one tool. Testing is another tool, you need to bring these things together. But I think there's also a broader sort of discussion happening around this, which is that the privacy concerns are really complicated and it's one of these things where in the past you have these sort of hypothetical privacy ethical discussions and now it's like, "No, literally. How much are we willing to trade our privacy to save people's lives? To save strangers lives?"

[00:43:55] JM: Right.

[00:43:56] RFS: I haven't been following that discussion super closely, but I think that that's the kind of stuff that's being talked out right now in some of these discussions in terms of we do have the technology potentially to enable things like digital contact tracing, and do we want to do that and what does that look like and how do we do that in a way that does maintain people's privacy for the rest of their lives outside of contact tracing use cases and things like that?

Again, that sort of has nothing to do with SafeGraph data and you couldn't use it for contact tracing, because there's no individual devices. It's all aggregated, anonymized et cetera, but there are all these other companies talking about doing it and how they could help. I think it's super complicated. It's hard to know what is the right way to do that. But in my opinion, either way, it's really a discussion about how do we get sort of to the next phase once we've gotten this sort of first outbreak under control to some degree?

[00:44:40] JM: Has good social distancing data, the social distancing, I guess you could call it a product. Has it given you any other ideas for creative uses of location data?

[00:44:50] RFS: Yeah, it's a good question. It's funny, because we think about these things a lot at SafeGraph, and startup, you're limited by time, and time is one of your most precious resources and you want to try to get feedback on things you build as quickly as possible. I guess one of the sort of hard things about building new products is that you're trying to walk this line between how much I have to build to get feedback and how much feedback do I need to build?

I think one of the things that from a pure sort of SafeGraph perspective that's been so exciting about this for us is that suddenly we have this huge community of people who are looking in our data every day, and we launch a new feature and we get feedback about that feature the same day. We have 1500+ researchers and government officials working with a data actively. For us, we're humbled to try to help those people and their feedback is also super valuable to us as we think about these different types of products that we could launch.

For example, there's another thing that we just launched today internally to this research consortium is the question of how to we know whether a business is actually operating or not during the pandemic? Historically, before the pandemic, we've been doing a lot of work to try to understand exactly when does a business open and when does a business permanently close? There's obviously lots of turnover in restaurants and things like that. Businesses open, businesses close, and we want to try to track that.

The pandemic is this new weird world where businesses aren't permanently closed, we hope, but many of them are not operating right now. You couldn't go there to do business. They sort of

have this indefinite close status, and a lot of economists are interested in trying to understand how much that's happening and to what degree that's happening. We're actually trying to use the foot traffic data, the patterns to sort of reverse engineer, "Are we seeing foot traffic to this place or not? How much does that compare to recent past and more historical past?" Can we use that to infer whether or not a business is actually open and operating rather than having to call all 6 million businesses in the US to see if they're open? Can we use our data to infer whether they're open or not? If we can, we think we have a pretty good solution to this. We just released it to the group today. Again, it's just a good example of sort of this feedback loop that's very tight right now and it's inspiring us to build new things and it's helping us build things better, because we're getting the feedbacks so quickly.

[00:46:49] JM: Okay. Well, Ryan, great talking to you and I look forward to seeing what other developments you guys come out with. I'm sure you'll inspire some more data scientists do interesting work around your social distancing data.

[00:47:01] RFS: Yeah. Thank you so much for having us, Jeff. Again, if anyone who's listening who wants to get involved, we have all these academics and researchers and government officials working with the data, and we talked about earlier many of them are quite data savvy and they have technologists working with them, but many of them aren't data savvy and don't have technologists working for them. Not every city in the US has these resources, and so we're also trying to recruit technologists and data scientists to join the community as well, because there's often times where you could get paired with a researcher or you could help a researcher with something that they're working on.

If you have these technical skills, if you're a data engineer or a data scientist or an analyst, if you have extra time, if you want to try to help, I encourage you to join. Just go to safegraph.com and look for the data consortium. We have a number people in the group who are not themselves academics and researchers, but they are working and helping in answering technical questions, and it can be dumb stuff like, "How do I load this data into AWS S3 instance or something like that?" things like this. Every level of technical help can be helpful, and if you want to help out, we'd love to have you try to help out the effort.

[00:47:57] JM: Okay! Well, thank you, Ryan. Great talking.

[END OF INTERVIEW]

[00:48:09] JM: Apache Cassandra is an open source distributed database that was first created to meet the scalability and availability needs of Facebook, Amazon and Google. In previous episodes of Software Engineering Daily we have covered Cassandra's architecture and its benefits, and we're happy to have DataStax, the largest contributed to the Cassandra project since day one as a sponsor of Software Engineering Daily.

DataStax provides DataStax Enterprise, a powerful distribution of Cassandra created by the team that has contributed the most to Cassandra. DataStax Enterprise enables teams to develop faster, scale further, achieve operational simplicity, ensure enterprise security and run mixed workloads that work with the latest graph, search and analytics technology all running across hybrid and multi-cloud infrastructure.

More than 400 companies including Cisco, Capital One, and eBay run DataStax to modernize their database infrastructure, improve scalability and security, and deliver on projects such as customer analytics, IoT and e-commerce. To learn more about Apache Cassandra and DataStax Enterprise, go to datastax.com/sedaily. That's DataStax with an X, D-A-T-A-S-T-A-X, @datastax.com/sedaily.

Thank you to DataStax for being a sponsor of Software Engineering Daily. It's a great honor to have DataStax as a sponsor, and you can go to datastax.com/sedaily to learn more.

[END]