EPISODE 528

[INTRODUCTION]

[0:00:00.6] RP: The easiest way to train a computer to recognize a picture of a cat is to show the computer a million labeled images of cats. The easiest way to train a computer to recognize a stop sign is to show the computer a million labeled stop signs. Supervised machine learning systems require labeled data. Today, most of that labeling needs to be done by humans. When a large tech company decides to “build a machine learning model”, that often requires a massive amount of effort to get labeled data. Hundreds of thousands of knowledge workers around the world earn their income from labeling tasks. An example task might be label all of the pedestrians in this intersection. You receive a picture of a crowded intersection and your task is to circle all of the pedestrians. You have now created a piece of labeled data that can be fed into a machine learning model.

Scale API is a company that turns API requests into human tasks. The most recent release of a scale feature is an API for labeling data that has been generated from sensors. As self-driving cars and merge on to our streets, the sensors on these cars generate lidar, radar and camera data. The cars will interpret that data in real-time using their machine learning models and then they will also send that data to the cloud that the data can be processed offline to improve the machine learning models of every car on the road. The first step in that processing pipeline is the labeling, which is the focus of today's conversation.

Alexandr Wang is the CEO of scale and he joins the show to discuss self-driving cars, labeling and the company that he cofounded. A few notes before we get started, we just launched the Software Daily Job Board. To check it out go to softwaredaily.com/jobs. You can post jobs. You can apply for jobs. It's all free. If you’re looking to hire or looking for a job, I recommend checking it out. If you’re looking for an internship, you can use the job board to apply for an internship at Software Engineering Daily. Also, meet ups for Software Engineering Daily are being planned. You can go to softwareengineeringdaily.com/meetup if you want to register for one of those meet ups. In March I'll be visiting Datadog in New York and HubSpot in Boston, and in April I’ll be TeleSign in L.A. Register now, because I expect these might fill up.

So with that, let's get on with this episode with Alexandr Wang.

[SPONSOR MESSAGE]

[0:02:44.3] RP: QCon.ai is a software conference for full stack developers looking to uncover the real world patterns, practices and use cases for applying artificial intelligence and machine learning in engineering. Come to QCon.ai in San Francisco from April 9th-11th, 2018 and see talks from companies like Instacart, Uber, CoinBase and Stripe. These companies have built and deployed state-of-the-art machine learning models and they've come take QCon to share their developments. The keynote of QCon.ai is Matt Ranney, a senior staff engineer at Uber ATG, which is the autonomous driving unit at Uber, and he's an amazing speaker. He was on SE Daily in the past. If you want a preview for what he is like, then you can check out that episode that I did in conversation with him.

I've been to QCon three times myself, and it's a fantastic conference. What I love about QCon is the high bar for quality, quality in terms of speakers, content and peer sharing as well as the food and the general atmosphere. QCon is one of my favorite conferences, and if you haven't been to a QCon before, make QCon.ai your first. Register at qcon.ai and use promo code SEDAILY for $100 off your ticket. That's qcon.ai and you can use promo code SEDAILY for $100 off. Thanks to QCon for being a sponsor of SE Daily, and check out QCon.AI to see a fantastic, cutting-edge conference.

[INTERVIEW]

[0:04:30.3] RP: Alexandr, welcome back to Software Engineering Daily.

[0:04:33.3] RP: Hey, yeah. Thanks for having me.

[0:04:35.0] RP: You're the CEO of Scale API, which is an API for human intelligence, and we'll discuss what that means a little bit later, but people can listen back to the previous episode we did if they want an overview. I thought we'd start with the discussion of self-driving cars. It turns out that self-driving cars have a huge set of data management problems. What kinds of data do self-driving cars need to work with?

[0:05:02.8] RP: . So self-serving cars need to solve a variety of different problems to be able to actually be deployed in the real-world. So the first set of problems that they need to solve are vision and perception problems. So that includes things like getting imagery of streets and being able to recognize various types of objects in them. Self-driving cars also have multiple types of sensors. So more than just images, they need lidar data and radar data annotated with all the different objects that appear in those sensors as well.

Beyond that, self-driving cars also need to be able to sort of capture how everything is moving in the world around them. So they need data about how every object in a 30-second period, or a minute long, how every object behaves. So, for example, a car will stop at a red light and then go on green, and he needs to capture tons of data about how other parties might perform on the road.

Another huge source of data that is often talked about is all the software in car companies today require tons and tons of mapping in the areas where they want to drive in. It's really important for these self-driving cars to know within centimeters of accuracy exactly where they are in the world, and that requires tons and tons of m data, sort of like Google Maps, but with a ton more accuracy and precision, a map of all of these locations.

[0:06:29.3] RP: What are the problems that a self-driving car development companies have with creating and managing that data?

[0:06:37.6] RP: Yeah. So I think the biggest thing is these self-driving car companies have tons and tons of problems to be solving. They are responsible for deploying an actual self-driving car on the road. That being said, a part of — Or sort of a requirement for building these self-driving cars is getting tons and tons of high quality data labeled and annotated so that they can train machine learning algorithms on.

That being said, that's not really a core competency for these companies. These companies are more robotics experts, or computer vision experts, or deep learning experts and it's a huge distraction for any of these companies to have to invest significant resources to be able to figure out this like data labeling and data annotation piece.

[0:07:22.6] RP: This large data set that needs to be labeled, what kinds of labels need to be applied to it?

[0:07:29.7] RP: Great question. So a lot of times in imagery, or lidar data, or radar data, most of it is just understanding what parts of those sensors or what parts of the sensory data correspond to various objects that a human would be able to recognize? For example, an image being able to draw masks around all the cars in the image and having a strong understanding and being able to train a neural net or a machine learning model to be able to recognize all of these different vehicles over time automatically.

[0:08:04.3] RP: So I read this, I think it was in the Atlantic, there was a really long piece about Waymo and how — It was kind of rare. I didn't expect to see it, but they did this really long piece that discussed how Waymo does a lot of its training and their simulations and real-world simulations. They have real-world simulations. They have virtual simulations. They have exercises where they take the cars out into normal circumstances, like driving through a city. They have circumstances where they take them to look weird locations, like a strange roundabout, and then they'll do things like re-creating the roundabout in the future, because they have this huge test bed that's somewhere outside of San Francisco where they can — They have all these land and they could just re-create strange circumstances. Like there was one circumstance where there was a strange roundabout in Texas and the car didn't know how to evaluate it, so they re-created it right outside of San Francisco so that they could do some practices on it. Because you're not a self-driving expert yourself as far as I know, how have you gone about learning or what have you needed to learn about the process by which these companies are teaching their cars?

[0:09:22.8] RP: That's really a good question. So one of the really interesting things about how self-driving as an industry has developed, it's actually a pretty shallow field, and by that I mean there's all these efforts primarily because all these companies need to really need to hire a lot of people who are knowledgeable about how self-driving works and can actually help them with their development efforts. A lot of the information about how the vehicles and these companies are being built is somewhat public knowledge. By that, I mean, a lot of the interesting work has come out of academia actually, and then from academia, a lot of the best [inaudible 0:09:58.2] or the best people working on it go and join a lot of these private companies.

So you actually can learn quite a bit despite keeping your nose to the ground on all the developments and sort of in the outside world. That obviously only gives you a first — An impression of how a lot of these companies are functioning. Beyond that, I mean, think you sort of have to — You have to talk to a lot of people and really understand how these companies are functioning, what their goals are, what their metrics are, etc.

[0:10:27.4] RP: And how does the labeling fit into their development process? They've got these cars driving around in all of these different circumstances. What is their pipeline look like in terms of when they're going to send off some of that recorded data to be labeled?

[0:10:46.2] RP: Yeah, that's a really great question, and it's one of the things that we provide massive flexibility for customers with, because we are an API company. So they have the ability to send off data basically whenever they want to automatically or manually, however they want. So a lot of time we see a couple of different workloads. Obviously, there's, "Hey, I'm just getting started with a machine learning problem and I need labeled data to actually make any progress on the issue." So they'll go through, they'll collect a bunch of data and then they'll send a bunch of data through.

Once you get past that initial hump, then it's all about understanding what are all of the situations in which my robot or my self-driving car is performing poorly and being able to send those situations to us to be labeled and then continuing that iteration cycle over and over and over and over again to just boost the efficiency of these vehicles.

So as an example, one of our customers actually, whenever there's a disengagement of one of their vehicles, like whenever the human driver takes over, they will automatically cap the time around that disengagement and send it to us to be labeled. Similarly, a lot of times, what these companies do is they will have some model performing the real-world. They'll go and drive it a ton and they'll honestly just see what are all the situations in which we're doing poorly and being able to mark those situations and then send us more data of those similar situations.

[0:12:13.0] RP: I guess it varies from company to company what they want labeled. Are we talking about pedestrians, or stop signs, or streetlights What exactly are the things that people are — So, basically, I guess we should give a little bit more context. What we're talking about here is that your Scale API is essentially an API for sending tasks to people, sort of like Mechanical Turk, what most people would know as Mechanical Turk, but it's a little more of a clean API and you have very specific tasks. So one of these new tasks that you've developed is a way for sending labeling tasks to people. So you have a self- driving car that's driving around and it's recording all these information, and the company can make an API call with that recorded information to a human that is waiting to receive that data and the labeling requirements, for example, label all the cars or label all the pedestrians in this video. So the company has to provide, I believe, both the labels that they're asking for and the video. Is that correct? What are they asking for?

[0:13:29.6] RP: Right. These companies, they care about recognizing basically anything that as a driver of the vehicle you would obviously — You would see and it would affect how you would drive. They're interesting in getting all that stuff labeled. So it's stuff like cars, pedestrians, stop sign, like any sort of street sign, any sort of streetlight, any sort of obstacle in the road, like construction in the road or anything that might block off certain areas. They care about cyclists. They care about motorcyclists. They care about debris in the road. They care about the lane markings to be able to learn what the lane structures are. It's a pretty wide birth of different things that they want to be labeled. When they want them labeled, they really just want them recognizing all of these images and annotate with really high quality.

[0:14:17.0] RP: And that's important for training, but also for benchmarking, because they can also run their vision recognition systems on that video and get their computer-generated labels, which are much cheaper to generate, which eventually all these cars will be, hopefully, someday generating their own labelings, but the they can look at those labelings and do a diff between those labels and what the humans have labeled. Is that what they do?

[0:14:50.8] RP: Yeah, exactly. You've hit the nail on the head. Another huge use case for a lot of these companies is sort of getting this massive regression testing dataset where they can use the ground truths and compare how exactly their systems are performing relative to it.

[0:15:08.2] RP: Where are their systems failing? Is it mostly edge cases or are there things that would seem totally obvious that are not edge cases at all where the cars are still failing to recognize like a pedestrian or a stop sign?

[0:15:21.0] RP: Yeah, great question. So with any type of deep learning approach, you're going to get weird — You're going to get performance that you don't expect necessarily. So with any type of system, you're going to get some small failure rate on cases that you normally would expect the neural net to perform well in. For example, you'll look at an image and it won't look like there's any strange artifacts or anything strange about the situation, but you'll notice that the model performs — Or has made a mistakes. Some of those types of errors will just surprise you.

That being said, the other shying class of types of situations where their models would perform poorly are situations that they haven't encountered before or haven't labeled before. So any type of deep learning system or supervised learning system today can't extrapolate beyond what the training data has seen. So depending on what this giant corpus of training data is, a lot of the situations that they've seen, the models will only be able perform within those situation. So, suppose, like I have this massive dataset trained on the streets of San Francisco and I try to take that to the streets of New York, all of a sudden the whole city looks entirely different and you're getting a really poor performance as a result.

[SPONSOR MESSAGE]

instances have gone down too. You can check out all their new deals by going to do.co/sedaily, and as a bonus to our listeners, you will get $100 in credit to use over 60 days. That's a lot of money to experiment with. You can make $100 go pretty far on DigitalOcean. You can use the credit for hosting or infrastructure and that includes load balancers, object storage. DigitalOcean Spaces is a great new product that provides object storage and, of course, computation.

Get your free $100 credit at do.co/sedaily, and thanks to DigitalOcean for being a sponsor. The cofounder of DigitalOcean, Moisey Uretsky, was one of the first people I interviewed and his interview is really inspirational for me, so I've always thought of DigitalOcean is a pretty inspirational company. So, thank you, DigitalOcean.

[INTERVIEW CONTINUED]

[0:18:49.6] RP: During our last conversation we explored a lot of different topics of your platform. We explored the business. We explored the interaction between you and the companies that make API requests and also the interaction between you and the scale workers on the platform. So we'll take that a little bit for granted here. Why did you decide to go after this selfdriving car labeling task? To give people a little context, your platform, you focus on very specific tasks, whereas on something like Mechanical Turk, it's totally wide open, where people can specify basically any kind of task. You have schematized well-defined tasks that people can do, like audio transcription or categorization, and then this new one, sensor fusion annotation, where people are labeling data that's coming into cars. So why did you choose this task to work on?

[0:19:55.5] RP: Yeah, that's a great question. A lot of what we focus on as a company is trying to figure out — Or a startup in general, is figure out what are the types of task that we can do where we'll make a huge difference to our customers and ideally work in growth industries, right? One of the reasons why we focus on self-driving in particular is we looked at sort of the state of the world and we noticed, "Hey, artificial intelligence is this massive trend. The technology is sort of there, but largest bottleneck is access to labeled data." There's two large bottlenecks; access to labeled data and access to machine learning talent.

We've recognize that, and so we have this understanding that we want to help this massive trend of AI actually happen and figure out what are the highest value ways that we can help produce labeled data for various applications of AI?

Now if you look at where the world is at right now, the biggest capitalize sort of flavor of AI or industry where AI is absolutely crucial is self-driving cars, partially because it's this big existential threat for all of these gigantic companies. But what that means is, is if you want to have an impact by producing AI data labeling and you look at where most of that work is being done in the world, most of that needs to be done for self-driving today.

That being said, as a company, our goal is to help artificial intelligence or help supervised learning actually realize its potential by sort of solving this data labeling bottleneck. So as self-driving cars goes to production, we'll want to look at the new types of industries that are really important and the new applications of AI that are making a big difference.

[0:21:40.9] RP: And the term here is sensor fusion annotation. Explain what that term means; sensor fusion annotation?

[0:21:48.9] RP: Right. So all of these companies have a variety of different sensors that are on their vehicles, so they have cameras obviously, they have lidars, they have radars, they have GPS, they have IMUs. They have a ton of different sensors so that they have a couple of things. Some of the sensors give them different things than the other sensors. For example, lidar gives you a really good 3D and depth perception, whereas as images [inaudible 0:22:17.0] can let you recognize things both close and far very well. In addition, they provide redundancy on top of each other.

So for an object that you see in lidar, you can verify and cross reference versus the images or the radar that there actually is an object there, and that helps even in cases where your machine learning models might fail. It helps you be redundant and still have overall very high accuracy.

So with our sensor fusion product, all of these companies, we noticed that they have [inaudible 0:22:47.5] sensor data and they really want to be able to combine it all and label it altogether. That's what we've been able to provide with our sensor fusion product.

[0:22:56.5] RP: So when you take a car, for example, that's got lidar and — What? It also has radar? Like it will have both of those things going at the same time?

[0:23:05.8] RP: Right. Exactly. So our labelers will label lidar, radar, camera images, basically whatever sensor that these companies want to throw at us altogether.

[0:23:16.2] RP: So you send all of those sensor data time series to the same scale or do you send them to different workers?

[0:23:25.0] RP: That's a great question. So in general, with our task pipeline, we do a lot of work to optimize everything about our process to make it as robust as possible. Similarly, we want to make it as redundant as possible, as high-quality as possible. Actually, we do a combination of both where we'll split up the sensors, give them to different scalers, but we'll also sometimes give the scaler more than one at a time so that they can do a higher-quality job while they're doing the task, and then we [inaudible 0:23:53.4] where we combine all of these results and then send them back.

[0:23:56.9] RP: So why don't we take a step back? Why don't you explain the workflow, the end-to-end workflow for a company? So I think you've got like Uber, and Alphabet and voyage. Explain the workflow for one of these companies who is using the labeling platform? How do they onboard? How do they start sending requests and what do they need to give you?

[0:24:21.7] RP: Yeah, great question. So the onboarding is easy. Obviously we're an API first company, and so we try to just — One of the developers at one of these companies will interact with us and read the API docs and let us know if they have any questions and be able to play around with our test API keys, etc.

How they get started is they will decide what all the data they want labeled is, and usually that's a pretty large corpus of data that they want labeled, and then they will send us through the API all of the image, or lidar, or radar data that they want us to label. They'll send it all through the API and then pass that. They don't really need to worry about it. We'll do all the work behind the scenes of getting all these data labeled with the combination of machine learning in humans and then just send the results back to them via a callback.

[0:25:11.0] RP: So you have some machine learning on your own platform for labeling.

[0:25:14.3] RP: Exactly. Yup.

[0:25:15.6] RP: Okay. I didn't realize that and I guess you get economies of scale, no pun intended, since you have all of these different companies that are putting data through your system. So you get entire set of all of those requests that you can feed back into your algorithms. That's pretty interesting.

[0:25:36.5] RP: Yeah, definitely. I think the main thing that we're excited about is it's really clear that for all of the data that needs to get labeled, to be labeled for self-driving cars to get on the road, you can't really afford to have it be this totally manual process. We've actually seen this. This has happened in the past through sort of history of tech, where they are sort of two famous examples. The first is Google maps with all their mapping operations. There's obviously tons of human labeling that went into producing the high quality maps that you can enjoy on your phone or your computer, and that was done as this massively manual operation initially, and then over time they were able to build in automation to be able to make it more efficient.

Similarly, with Facebook and their content moderation, they've made major strides in being able to automatically categorize content depending on whether it's like automatically moderate content, using all these data that they have while still kicking back their human when they're unsure.

[0:26:40.0] RP: I'm surprised that Alphabet — You have Alphabet as a customer. You would assume that they have this on lock. Are they using you just as another point of reference, or can you say anything about like why on earth would they use — I mean, no offense to you, but I mean it's Alphabet. Like why do they need scale for this kind of task?

[0:27:02.0] RP: Yeah, that's a great question. Honestly, I think the answer basically lies in the fact that big companies are big companies, and so while Alphabet itself might have core competencies in producing labeled data, and they definitely do, they definitely, definitely do, it's still the case that interacting with those internal teams could be a slow and low NPS. Like, basically, in an unhappy process for the internal customers, and if these internal customers really want to get moving quickly, then they want to be able to work with, basically, the best people around to solve that problem for them. If that means that a startup is going to be able to provide them better response times, better quality, a better API, etc., then they'll make the internal decision to use it.

[0:27:47.1] RP: Yeah. I mean, it's not like that saves them money, because if they were to go through the processes of finding the right internal team and getting into their internal teams a work queue and getting the task going, then all of that time spent, relative to what they would be doing with Scale API, is money spent. So that's pretty cool.

So when someone tees off an API call to you infrastructure, so they've got, let's say, lidar and radar data and the call says, "Identify all the bikers and people walking on a sidewalk in this video. What's happening? Give me the end-to-end thing that happens from your infrastructure to the scale workers who receive the task? Give me the end-to-end story.

[0:28:36.7] RP: Yeah. So it's hard to go into too much detail primarily because of two things. First, we're experimenting a lot and it varies a lot from request to request just because we're trying a lot of things to try to improve the quality and do the best job for our customers. The other thing is there are some things that we do that we feel are proprietor or that are sort of special to being able to produce really high quality.

That being said, roughly what happens is a request will come in to our system and we'll initially evaluate it with our machine learning algorithms, and then depending on whatever that result is, it will kick to human. That human will go through and verify the quality of it and we'll do a bunch of work to basically ensure that the quality of it is very good. Then we have a bunch of preprocessing and post-processing steps on the data just to make sure that we're able to ingest it properly, that our customers are seeing it properly and they're able to send it back properly, and then we'll fire back or call back to the customer.

[0:29:32.6] RP: Can you give me a description of your infrastructure, just kind of the underlying infrastructure?

[0:29:39.3] RP: Yeah. That's an interesting question. There's basically, what I would say, there's two main types of buckets of types of infrastructure that we work with. There's machine learning infrastructure and then there's like core, just like systems infrastructure to handle all the data that's coming in and being able to processes it and being able to handle it. For all the machine learning infrastructure, we happily use TensorFlow for essentially all of it and are able to get quite far using a combination of GCP and AWS using TensorFlow. Then for the rest of it, I think we're getting to the day and age — And this is really interesting, where a lot of systems infrastructure is sort of solved if you use Docker — If you use Docker intelligently with — Or these AWS services, like ECS, or they just launched like a Kubernetes solution. We use a lot of that to basically make the problem easier on ourselves.

[0:30:32.4] RP: I was talking to the guy who manages a lot of the infrastructure thumbtack and he was talking about when they moved to ECS, their sleepless nights went away. It was just so much easier to manage their infrastructure.

[0:30:49.3] RP: Yep.

[0:30:49.5] RP: It sounds like something similar has happened with you.

[0:30:51.8] RP: Yeah. I mean, I think general, it's really interesting for me. So I used to be at Quora running much of infrastructure, and everything was on EC2. There was so much infrastructure work to make it all function properly. But that compared with the world today where you can use Docker and get so much like reliability and niceness out of how everything you do works, it's just kind of crazy.

[0:31:16.9] RP: Speaking of Quora. So for a very long time, I wanted to work at Quora, because I found the company — I found the product incredible. I found the people who work there to be totally brilliant. I also find the company really underrated. What did you learn there? What did you learn that you think was unique from your time at Quora?

[0:31:38.1] RP: Yeah, that's a good question. So I think the first thing, and this really was just like this culture of the company, was really aligning around metrics and being very data-driven as a company. This manifested in a bunch of ways. Obviously, there were metrics that we used to track our progress, but there's this gigantic data science — Or not gigantic, but there was this dedicated data science team all focused on building these metrics and being able to understand what was going on behind the scenes, and that was a — Now that I'm learning, like a very unique environment compared to a lot of other companies just because it was so focused on the data and so driven by these changes in the metrics. Quora actually built their own AV testing platform and that's actually really, really good. I mean, it's better than Optimizely, for example, when I tried Optimizely later, and that was really impressive and it just sort of showed this massive dedication to data.

The other thing is Quora had this large authentication of machine learning as well. Machine learning is really important to the product in terms of feed ranking, answer ranking, understanding answer quality, being able to send you the content that you want, and being in an organization that value that so heavily was really great for me being able to understand like what every company would have to work, what every company would have to operate like in a world where machine learning is sort of table stakes.

[SPONSOR MESSAGE]

[0:33:12.0] RP: When you're building an application, you need it to be fast, secure and always evolving. With Kubernetes engine on Google Cloud Platform, developers can deploy fully-managed containerized apps quickly and easily. Google has been running production workloads in containers for over 15 years. Google builds the best of what they learn into Kubernetes, which is the industry-leading open-source container orchestrator.

Kubernetes engine combines automatic scaling, updates and reliable self-healing infrastructure with open source flexibility to cut down development cycles and speed up time to market. To learn more about Kubernetes engine, visit g.co/getgke. That's g.co/getgke. G.co/getgke.

Thanks to Google Cloud for sponsoring Software Engineering Daily.

[INTERVIEW CONTINUED]

[0:34:17.0] RP: It sounds like that has been deeply integrated into your product, because you've got this constant testing that you need to do between the machine learning models that you've built and the people that you can tee off tasks to. Have you had to build your own internal A/B testing system or are there useful solutions that you can build off of that you find to be flexible enough to suit your needs?

[0:34:46.0] RP: Yeah, that's a great question. I mean, Quora, because of sort of when they got started and, honestly, sort of the background of the company, they'd ended building a lot of things in-house, which a lot of it is because when they got started, like there was a lack of ecosystem around all of these different sure products for companies to use to be able to move faster.

Fast-forward to now, we're working on scale. It's actually pretty incredible. We're in this time of product development and technology development where a lot of the building blocks out there are already built for you, to like Stripe, and Twilio and Optimizely, etc., Segment, and so you can —

[0:35:27.5] RP: React.

[0:35:28.2] RP: Exactly. Yeah. We take as approach, and a lot of startups an approach of relying on a lot of the technologies that have already been built just to give you the opportunity to move faster, to build stuff that's core to your company as complete as possible.

[0:35:45.5] RP: But A/B testing, not nothing you can comment on.

[0:35:49.3] RP: Right. A/B testing, we have not built out a totally in-house solution. I don't think we haven't needed to.

[0:35:54.8] RP: Got it. What did you learn from Adam D'Angelo, the Quora CEO?

[0:35:59.7] RP: In general, a lot of things. He's an extremely thoughtful guy, very intelligent, and also makes great decisions. I would say the most important thing that Quora really adopts, and they're all very aligned on this, is having a extremely long-term view for the company and being super focused on how do you build sustainable consistent value over time as a company and not get caught up in sort of any metrics or caught up in things that actually aren't productive for producing long-term value.

Probably one of the best examples of this is Quora is so focused on having high quality content on the website that, honestly, for any type of social product, like if you enforce that level of quality, it obviously limits like how viral it's going to be or how social it's going to be, because you can't have some — You can have low-quality content. So it's not going to sort of blow up like a Twitter or a Snapshot. But what that means is all these content that you're building on this platform over time, it's indexable, it's all really high-quality content, it's all very valuable.

As you build up —

[0:37:07.6] RP: It's durable.

[0:37:08.7] RP: Exactly. So as you build that up over time, you're continuing to build long-term value into the future.

[0:37:14.4] RP: Yeah, definitely. Back to self-driving discussion. So you have, again, most of these major players as customers. You got Cruise, and Alphabet and Uber. Hadn't these companies built their own labeling pipelines? I mean, does is come back to the same question, as the company can be so big that it's just hard to — If you're some random engineer at Cruise and you don't have an ear with the labeling team, you go off to scale. I mean, it seems like labeling is such a core competency at these companies. I may, I know you said earlier, it was not. But yeah, I mean, do the scale manage to fit in comfortably into their labeling pipelines? I mean, what was the process of winning them over or onboarding them?

[0:38:06.4] RP: Yeah, that's a great questions. I think with any customer, the way that we win them over is just sort of with the quality of our technology, the quality of the data we're able to provide back, the capabilities that we have. Really, like the main cell to any company is, "Hey, if you scale, you're going to be able to accelerate your own efforts significantly. In the world of self-driving, there's so much competitive pressure that everybody is just trying to ship as clearly as possible, but anything that I can give them an edge in terms of shipping faster is massive or is immediately interesting to these companies. I mean, that's a lot of the reason why they would engage with us.

The other thing I would say is, it does make a lot of sense in general for one company to be focused on this labeling problem, sort of know the ins and outs. Like our entire company is focused on how do you get high quality work out of people doing tasks on the internet? That allows us to operate a lot better and get a lot more efficient, produce higher quality. In general, just do a better job than an internal team at these companies who isn't like solely focused on this problem and getting all of these customer demands from all these [inaudible 0:39:16.7] customers and being able sort of understand where the whole market is going, etc., etc.

By that, I mean — And this is sort of a tentative of any B2B software is like because we as a company are fully invested in the problem and are going to be fully invested into making our product better in every single way, every single day that we operate, we're going to — If we do a better job now than some of these internal teams, then we'll continue to do a better job as time goes on.

[0:39:40.6] RP: How do you think about pricing?

[0:39:42.0] RP: That's a question. I think for us, we take a lot of inspiration, honestly, from AWS, and I think a lot of — We view ourselves as an infrastructure company. A lot of infrastructure companies sort of have to take inspiration from AWS based on their massive, massive success, but a lot of what we try to do is we aren't trying to necessarily gouge or customers, which you might wish some companies obviously do, like a lot of very big enterprise companies do. But we're just trying to do is be able to offer sustainable infrastructure, because we really believe that every single company is going to be needing labeled data in some way, shape or form as they try to make this transition to an AI world. So our goal is to become the AWS of that. By that, I mean, we want to provide fair prices to all our customers that they can rely on and build the businesses.

[0:40:38.2] RP: When you go into a new vertical, like the self-driving car data label — I should stop saying self-driving labeled, because it’s useful for drones and other kinds of robotic cameras and stuff. But when you go from market to market — So you’ve also got audio transcription, for example. So can you kind of replicate the same strategy going from audio transcription to the sensor fusion API, because it's essentially the same thing where you've got a complex set of — I guess unstructured data is how you would define it, like audio is unstructured data, sensor fusion data is unstructured data, and you've got a model that you can run the task against, and then you've also got humans that you can run the task against. Is it basically the same when you're looking at audio transcription versus sensor fusion, or did you really have to reinvent the wheel when you stood up sensor fusion? Like how much of the patterns from previous labeling tasks, like audio transcription, were you able to port to this new project?

[0:41:46.4] RP: Yeah, that’s a great question. So I would say, in general, we try to build as much reusable infrastructure as possible, and this goes into the last point of just how we want everything that we do to continue to build for us, long-term value for our platform so that we can be able to produce a better and better service no matter what over time.

So we try to use as much reasonable infrastructure as possible. Similarly, we have a bunch of sort of institutional learnings at our company about how do you actually do the best job with a lot of these work, and all that institutional knowledge saves a lot of time and development process, so we just know what's going to work and what’s not going to work.

Basically, I would say it's closer to the fact that you're right, we can actually reuse a lot of the same playbook reviews, a lot of the same infrastructure reviews, a lot of the same product learnings as we have in the past to be able to stand up these new types of tasks.

[0:42:38.9] RP: There are so many dimensions to the scale of products. So you've got a bunch of different APIs that you need to maintain. You've got sales. You've got customer success. You've got this giant fleet of scale workforce people. What has been the hardest aspect of scaling your hiring process and scaling your org structure?

[0:43:05.5] RP: So I think if you talk to probably almost any startup early on, a lot of the main sort of challenges or the main the main things to consider when hiring are how do you ensure — Let's just say you're continuing to hire really, really amazing people over time and never compromise on that while still being able to grow and while still being able to grow to hit all the milestones that you need to. Those two things are sort of inherently add-ons. Like you could potentially make a hire for the short term that helps you hit a milestone, but it might not have been the right long-term decision, right?

We adopt a really, really long-term similar to everything else. We adopt the long-term view with respect to hiring, that we want to hire all of the best people out there and we never want to compromise on that. So especially as we scale the team to new types of functions and continue the org with salespeople, with customer success people, with operations people, with more and more engineers, etc., that the number one thing for us is, "Can we hire the best people that we can?" Then as part of that, is really research the interview process and research how you're going to be able to find the best sales people out there. How you're going to be able to find the best customer success people out there, or how you're going to be able to find the best marketing people out there and really do your research to ensure they you're finding the best people.

[0:44:26.8] RP: . One of the things I have learned, and I have a couple people that work with me full- time, but it's nothing compared to what you're doing and what many other startup founders that I talked to are doing. But there seems to be just a difficulty or it's a learning process of learning to deal with your own psychology as you build one of these organizations and as you learn to manage people. Is that been a challenge at all? Like how have you learned to deal with your own psychology and manage the inevitable mood swings, the highs and lows and whatnot?

[0:45:05.5] RP: I think with any startup you're going to get stuff that goes well, stuff that goes poorly. It's all inevitable. I would say the most important thing is to — These is all harkens back to this focus in the long-term, but, truly, just focus on — Like have core principles for why you're building what you're building. Have core principles for why you believe that what you're doing is the most important thing for you to be doing and is the most valuable problem for you to work on. So, for us, like solving these massive bottlenecks for the deployment of AI are these very important problems and stuff that we're very passionate about. Then whenever stuff goes poorly or stuff goes well, compared to how you expect, go back sort of center along these long-term principles or these long-term motives and be able to use that to continue motivating yourself and sort of balance yourself.

[0:46:00.6] RP: Okay. Yeah, it sounds like you're pretty unphased.

[0:46:03.1] RP: Well, I would say, right now things are going pretty well for Scale. Obviously, things will go worse and better over time and sort of just about weathering the wave while still building long-term value.

[0:46:14.9] RP: Yeah. Indeed. Okay, I think that's a good place to close. Alex, really great talking to you. I remain fascinated by Scale API, and anybody out there who's got self-driving car data, or drone data, or — We didn't really talk about much of the other use cases. What are the other things people can use this API for?

[0:46:35.4] RP: Yeah, and it's really interesting. So self-serving car, drone data. In general, any type of robotics data, satellite data, there's e-commerce. There're tons of images on e-commerce that are useful for data labeling as well. There's tons of security footage out there that needs to be labeled. There're tons of — There's like the action footers which are like the Skydio or Gopros of the world that need a ton of data labeled. So there's basically anywhere that you know there's a camera is somewhere where there can be tons and tons of data labeled.

[0:47:06.9] RP: You had a vision API for a while, right? Like you've had a labeling API before, but this was new.

[0:47:14.4] RP: Yeah, exactly. A lot of the new launch has been really focusing around how do we serve the needs of not just a camera from one of these companies, but a lot of times these companies have either video feeds or hey have more and more data over time, like they have lidar data, they have radar data, and how do you go about supporting all of these really sustainably to produce value for them?

[0:47:36.3] RP: Yeah. I think I really failed to go into this, but I think part of the big draw is what the scaler sees, the UI that you give to the scaler and how easy it is for them to see what's going on in the sensor data. So anyway, if you have a use case for this kind of thing, then definitely check out — I mean, you've got a very nice set of resources that is pretty good at conveying what you have done differently. So maybe I could have gone into that in more detail, but I think we've covered it well enough.

[0:48:02.4] RP: Yeah, thank you so much for having me on the show again. Again, big fan of Software Engineering Daily.

[0:48:06.8] RP: Okay.

[0:48:07.1] RP: Yeah, thank you again.

[END OF INTERVIEW]

[0:48:10.9] RP: If you are building a product for software engineers or you are hiring software engineers, Software Engineering Daily is accepting sponsorships for 2018. Send me an email, jeff@softwareengineeringdaily.com if you're interested.

With 23,000 people listening Monday through Friday and the content being fairly selective for a technical listener, Software Engineering Daily is a great way to reach top engineers. I know that the listeners of Software Engineering Daily are great engineers because I talked to them all the time. I hear from CTOs, CEOs, directors of engineering who listen to the show regularly. I also hear about many newer hungry software engineers who are looking to level up quickly and prove themselves, and to find out more about sponsoring the show, you can send me an email or tell your marketing director to send me an email, jeff@softwareengineering.com.

If you're listening to the show, thank you so much for supporting it through your audienceship. That is quite enough, but if you're interested in taking your support of the show to the next level, then look at sponsoring the show through your company. So send me an email at jeff@softwarengineeringdaily.com.

Thank you.

[END]