EPISODE 623

[INTRODUCTION]

[0:00:00.3] JM: When a patient comes into the hospital with stroke symptoms, the hospital will give that patient a CAT scan, which is a three-dimensional imaging of the patient's brain. The CAT scan needs to be examined by a radiologist and the radiologist will decide whether to refer the patient to an interventionist, which is a surgeon who can perform an operation to lower the risk of long-term damage to the patient's brain function. After getting the CAT scan, the patient might wait four hours before a radiologist has a chance to look at the scan. In that period of time, the patient's brain function might be rapidly degrading.

To speed up this workflow, a company called Viz.ai built a machine learning model that can recognize whether a patient is at risk of stroke consequences or not. Many people have predicted that radiologist will be automated by machine learning in the coming years. This episode presents a much more realistic perspective. First of all, we don't have nearly enough radiologists, so if we can create automated radiologists, that would be a very good thing. Second of all, even if this workflow has a cutting-edge machine learning radiologist in the loop, you still need the human radiologist in the loop.

David Golan is the CTO at Viz.ai and in today's show, he explains why he is working on a system for automated stroke identification and the engineering challenges in building that system. It's a great show about machine learning and healthcare.

Before we get to the show, I want to announce that we're hiring. We are hiring writers and researchers and a videographer. You can find these positions along with other jobs at softwareengineeringdaily.com/jobs. Some of these are part-time jobs, some are full-time. If you are hiring, you can also post your jobs on our job board. It's easy and it's free. Just go to softwareengineeringdaily.com//jobs and see how you can post a job.

[SPONSOR MESSAGE]

[0:02:03.9] JM: Over the last two years, I spent much of my time building a video product. We had issues with latency, unreliable video playback, codecs. I was amazed that it was so difficult to work with videos. As it turns out, video is complex. Figuring out how to optimize the delivery of video is not easy, especially since there is both mobile and desktop and mobile users might not have as much bandwidth as desktop users. If you're an engineer working on a product that involves video, you just don't want to think about any of this. I can tell you that from firsthand experience, and that's why Mux exists.

Check out mux.com and find out how Mux makes it easy to upload and playback video. Mux makes video hosting and streaming simple. Today, you can get $50 in free credit by mentioning SE Daily in the sign-up process. Even if you aren't working on video right now, if you think you might work with video in the future, Mux is a really useful tool to be aware of. Check out mux.com. If you're an engineer who's looking for work, you can also apply for a job at mux.com.

On Software Engineering Daily, we've done two shows with Mux, and I know that Mux is solving some big difficult problems involving lots of data and large video files. To find out more, you can go to mux.com. You can get $50 in free credit by mentioning SE Daily, and you can apply for a job if you're interested in working on some of these large challenges. Thanks to Mux for being a sponsor of Software Engineering Daily.

[INTERVIEW]

[0:03:58.5] JM: David Golan, you are the CTO at Viz.ai. Thanks for coming on Software Engineering Daily.

[0:04:03.7] DG: Thanks for having me.

[0:04:04.9] JM: Your company identifies stroke victims. It does that by looking at CT scans. What is a stroke?

[0:04:16.3] DG: There are two types of stroke; there's hemorrhagic stroke and ischemic stroke. Hemorrhagic stroke is when a blood vessel in the brain erupts, essentially the patient is bleeding into their brain. An ischemic stroke is when a clot is dislodged from somewhere in the

body, can be from the heart or from the carotid. It goes upstream with the blood, until it gets stuck in a vessel that's too small for the clots to – in both cases, the result is that there's a part of the brain that doesn't get the supply of blood and oxygen that it needs. As a result, begins dying quite rapidly.
Ischemic strokes are more common. They make up to 90% of stroke patients, and that's the main focus of us at Viz.ai.

[0:04:58.2] JM: When a patient has a stroke, they're probably not at a hospital when the stroke happens. The response can be not great. How does the patient typically respond to a stroke?

[0:05:14.4] DG: It really depends on the severity of the stroke. A stroke is quite an alarming experience. A severe stroke could manifest as in a one-sided weakness, loss of speech sometimes, some confusion, loss of words, patients can lose consciousness, but it can also be dizziness or headache or something relatively mild. In some cases, patients don't realize they're undergoing stroke. In some cases, the brain is sustaining so much damage that even though the damage is quite severe, they're not realizing it and it's up to the people around them to spot it and call 911.

Sometimes it happens during sleep, and of course then, this is only when you wake up. It can take some time, but once spotted, people should get to the nearest hospital as soon as possible.

[0:06:00.9] JM: When they get into the hospital, the patient will get a CT scan. That's a scan that well, it does an imaging of the brain. Explain what a CT scan does.

[0:06:14.1] DG: A CT scanner is a futuristic x-ray. X-ray takes – you put the patient in front of essentially a camera that takes a picture of them, and the x-rays they go through the body so you can see what's inside. An x-ray provides you with a 2D image. You can imagine you're getting a flat image of what's inside of you. A CT scanner is a sophisticated version of that, where you take a series of x-rays. It get cuts through the body and you can reconstruct those into a 3D image. You get a full picture of what's going on inside. Specifically in the case of stroke, you get a full picture of what's going on inside the brain.

[0:06:55.6] JM: Is the CT scan a single image, or is it a video?

[0:06:59.7] DG: It's neither. It's a sequence of images, but they're not taken over time, so it's not like a video in the sense of a 2D image and the times axis. It's actually a series of 2D images that compose a 3D image of the body.

[0:07:15.2] JM: Okay. It all represents the same timestamp?

[0:07:18.5] DG: Typically yes. It depends on the technology. The advanced scanners today, they acquire most of the slices, those images are called slices, almost instantaneously. Older scanners would have taken longer and would have taken one by one, or four out of four, but nowadays [inaudible 0:07:35.0] – the scans are done very, very fast in a few seconds.

[0:07:39.6] JM: Okay. The CT scan gets created. Then in traditional medicine, the radiologist needs to look at the CT scan. There can be a long waiting time in that timeframe, because maybe the stroke happens in the middle of the night, person comes into the hospital at 3 a.m. and there's only one radiologist on staff. They're already too busy, they were looking at all these different patients and this new stroke victim has to wait an extra hour, or 2 hours, or 5 hours get a radiologist to look at what's going on. Why is that problematic?

[0:08:16.3] DG: Yeah. When at first patient is undergoing stroke, the brain cells die at a very, very rapid pace. We're talking about the number is two million neurons die every minute and that number doesn't really mean anything if you don't know how many neurons there are in the brain, but some other statistics are quite astounding. One sentence that people like to use is save a minute, save a week. Every minute saved in time to treatment actually gives the patient another week of healthy life.

People talk about 15 minutes translating into 4% increase in the probability of a lifelong disability. It's really, really about those minutes and counting those minutes that can translate into very meaningful changes in patient outcomes. Then that patient walks into the hospital, and like you said, it can be in the middle of the night and it can be a smaller hospital, it can be a very busy time where they come in at the exact same time that a car accident came in. There's one

ED doctor trying to juggle all those patients and there's one radiologist that’s trying to juggle all those patients.

Actually in the case of stroke, there's something considerably more dramatic. Starting in 2015, there's an approved treatment for stroke, for scanning stroke that's really making a huge difference for patients. That treatment is mechanical thrombectomy, which is for my perspective nothing short of magic. Mechanical thrombectomy is the act of through a small puncture in the groin and interventionalist was a specially trained physician, can essentially take a wire and extend it all the way to the patient's brain and retrieve the clot, such and removing it from the patient's brain, thus, renewing the flow of blood.

This treatment is so effective that in 2015, five clinical trials were stopped halfway through, because the effect size was so huge that they determined it was immoral to deny the treatment from the control group.

[0:10:21.3] JM: Wow, wow. That's incredible.

[0:10:23.4] DG: Yeah, that was a big moment in stroke care in 2015.

[0:10:26.5] JM: It's called mechanical thrombectomy, is that what you said?

[0:10:28.8] DG: Yes.

[0:10:29.6] JM: Mechanical thrombectomy, this is a minimally – well, maybe not minimally invasive, but it sounds a fairly low-risk. Is it low-risk, or medium-risk?

[0:10:38.3] DG: It’s considered very low risk. You talk to some doctors, they would say they operate on a 96-year-old patient without thinking twice. There's always risk in –

[0:10:48.2] JM: Always a risk.

[0:10:49.9] DG: Relatively minimally invasive. People undergo an intervention in the heart. The cardiologist do a – many people undergo them, then consider a fairly standout nowadays.

[0:11:02.6] JM: Always risk. Obviously, if you walk in to a hospital and you get your CT scan, you don't want to just have a mechanical thrombectomy while you're waiting for the radiologist to come look at your stroke. You do want to at least have the radiologist look at your stroke, or look at your CT scan and see how severe your symptoms are and whether you're having a stroke, or whether it's something else along the differential diagnostic. The point is that if you can get the stroke identified, the severity of the stroke identified and it warrants a mechanical thrombectomy, the point is that we have a very high expected value treatment that we can apply to you as a patient.

[0:11:49.5] DG: Exactly. There’s really a golden opportunity for the stroke patient if they are identified on time and identified as a patient who is eligible for a mechanical thrombectomy. That’s literally a magic wand that the interventionists can waive and remove that clot from their brain and generate huge benefits for the patient.

The biggest issue with this is first of all, most hospitals don't have newer interventionalist units. They don't have the equipment and they don't have the personnel. Those are typically concentrated in the main hospitals, also known as hubs, or comprehensive stroke centers. You're in one hospital, you're being treated by a team there, but actually the person who needs to know about this case and really make a clinical decision is elsewhere, is in different hospital altogether.

Then, I mean, we've open to hospitals. The doctors are doing their best, but sometimes there are inefficiencies in the system. Those doctors trying to figure out what's going on with the patient and whether they're eligible for thrombectomy, they would at some point try to contact the other hospital, consult with them, somehow transfer the images so they can also take a look on the other side, interventionist then needs to decide whether they want to go in and try to retrieve the cut. Then the patient will be transferred to that hospital and operated on.

It's a very, very complex logistical effort. Well, a lot of communication needs to go on and all decisions need to be made by many, many people, and not all these people are in the same place and of all of these people are notified when this patient comes into the hospital.

[0:13:30.6] JM: In the meantime, every minute is a week.

[0:13:34.1] DG: Exactly.

[0:13:35.0] JM: If you can save any piece of time along this series of logistical time transactions, then if this can be a potential boon, this can be that the difference between severely negative outcomes and severely positive outcomes. If there's nothing wrong with the patient or if it's something that is not a stroke and it's just maybe the patient's dehydrated or something like that, then if we couldn't just get the patient out of the hospital as quickly as possible, that would be an added benefit as well. Not to bury the lead, this brings us to Viz.ai, what you're doing.

What you are targeting is when the CT scan gets created, instead of handing it off to a radiologist, which may take three or four hours, you just send it to the cloud and your software analyzes the CT scan. Explain what you can detect by running the CT scan through the algorithms that you've developed at Viz.ai?

[0:14:32.2] DG: Sure. At Viz.ai, what we do is identify – we get all the scans from the city scans to the cloud like you said, and the scans are analyzed by our deep learning engine. We are identifying those subset of patients that have an ischemic stroke of a subtype called the large vessel occlusion. Large vessel, meaning you can actually do a mechanical thrombectomy on them. We identify those patients that can benefit from this magic treatment. What we do is alert the interventionalists, the decision-makers who really need to know about those specific patients and can help them and can provide them this life-saving treatment, we notify them on those cases and they have an app. The app rings very loudly to attract their attention, and they can view the scans and patient information on the app to verify our finding and act on it.

This is done within minutes of the scan while typically, they would be notified way, way, way later. One thing that's really important to stress out is that we're not doing this instead of the radiologist reading the scan. Radiologists, they're not just looking at one thing, or doing one thing, identifying those patients. They're looking at a bunch of other things. If for example, there's evidence of cancer in the scan, the radiologist would pick it up.

We're not replacing them. The patient would still be – the scan will still be read by the radiologist in the hospital, treated according to the standard of care. We've identified that there's a subset of patients that could really benefit from expedited care. If you look at the ER room, that's packed with so many people. One of them can really benefit if we grab the attention of the right doctor at right time. That's our mission with Viz; getting the right doctor, the right patient at the right time, out of all that crowd that's there in the hospital, and get them the treatment that they need. We've been able to show really high accuracies and reduction time.

[SPONSOR MESSAGE]

[INTERVIEW CONTINUED]

[0:18:22.4] JM: The workflow that we discussed before where patient comes in, patient gets CT scan, waits for the radiologist to look at the CT scan. The difference between that workflow and the workflow we're describing with Viz is patient gets the CT scan, Viz processes the CT scan. If Viz notices something that is quite alarming, you can basically say, “Okay, we need radiologist right now. There's something severely wrong and we really need to alert you and get your attention,” is that right?

[0:18:57.5] DG: Yes. Just a minor comment is that we wouldn't alert the radiologist. We would actually alert the interventionalist, the doctor, specialist doctor is going to perform the –

[0:19:06.6] JM: You’d skip the radiologist in this scenario?

[0:19:09.9] DG: Yeah. In that case, because for those specific cases for those patients, at the end of the day the intervention list is going to look at the scans, look at the patients and make a decision. Why wait? Why pass through first the radiologist, a neurologist and then they need to get a doctor and at the hospital call them, send them the images. Why let this chain of event go and lose time for the patient, while we can identify those specific patients that have a lot of benefit, and send the information directly to the interventionalist’s pocket?

[0:19:43.7] JM: Got it. Okay, so think we understand at this point the value of this solution and how it improves the stroke triage system. This sets us up to actually get into the engineering around this problem. You need to be able to train an algorithm to look at three-dimensional brain scans and identify is this a high-risk stroke victim or not? Where did you get the initial training data for this problem?

[0:20:14.5] DG: Data is always a hard problem, and especially in healthcare. Healthcare and this is all very protective about your data for rightfully so. Patients’ health information is sensitive and you want to protect patient privacy. Our original datasets came from academic collaborations from researchers who bought into our vision, who are excited about the opportunity just as we were, and could literally – imagine a world where they get alerted and we're living this problem day-to-day, so that's one venue. The other venue was partnerships outside of the US with radiology companies, hospitals, research organizations that were able to provide us with the right data.

[0:20:59.6] JM: How much data do you need to build an accurate model here?

[0:21:06.5] DG: That's a good question. The amount of images that we ended up using is I think in the hundreds of thousands. I don't remember the exact number, but a lot of effort went into making deep learning work with a relatively modest amount of data. A lot of effort went into developing methods for data augmentation, for example, where you use existing data to artificially generate additional data points and making everything work.

We've designed special deep learning architectures that were designed with the constraints in mind and various ways. Deep learning is often viewed as this black box, where you throw in a lot of data and it provides wonderful solutions, which is quite often accurate. However, when you have not huge amounts of data, you sometimes need to help the network learn. We develop various ways to inject domain knowledge expertise into the learning process.

[0:22:10.6] JM: This data set you needed to build a model to train the algorithm on that data set, describe the training process. What was your sequence of steps to building this machine learning model and training it?

[0:22:28.7] DG: First of all, a CT scan is a big piece of data. It's actually, each slice is 512 by 512 and then you have a few hundred slices you need to work with. That's all of the data. For example, one thing is that you don't throw all that into a network. One, to give you for example is focus on regions. We've identified ahead, a CT scan can start from the chest and go all the way to the top of the head, and the brain is just a small part of that. For example, you can crop out the head out of the scan, then you define the output that you want that typically involves annotation on various levels. There's a high-level annotation saying this scan is normal, this can is abnormal.

There's lower level annotation, so you can annotate things on the pixel level. Once you have all the data ready, you have your inputs, you have your outputs, you start playing with your deep learning architectures trying to get something that works.

[0:23:31.9] JM: How do you get those annotations done? Do you need specialists to label the data?

[0:23:38.2] DG: We have a small duration of annotation definitely. Some things require specialists, some things can be done by more junior doctors, and some things can be done by a layman. We have an annotation platform and we have pagers of various degrees of expertise, and we leverage them for different tests depending on the difficulty and level of expertise required for the test.

Basically for example, we define something that needs to be annotated. Just to give an example, so we'll have something to run with; say we want to identify the brain. We could have someone segment the brain on many, many slices. That's not a very hard task that's doable with minimal training by people with no specific medical training. That would be an easy test that you can send out to minimal trained individuals, but if you want to identify a small hemorrhage, or a small abnormality in the brain, that really requires the eye of a specialist. That task will go out to specialists who can log on to the platform when they have some time and annotate a few scans.

[0:24:49.6] JM: To be clear here, you had to build your own annotation platform. You couldn't outsource this to somebody like, Scale API, or some CrowdFlower or something like that? We didn't start by playing with and trying to work with various existing platforms. They were lacking some features that were very, very specific to our context. Some examples are the fact that we need 3D data. We need the ability to scroll through as if it were video. Another thing is that CT data is not standard images. For example, images are typically either 8-bit if they're grayscale or you have 24-bits if you’re RGB. CT scan, every pixel is a 12-bit grayscale.

This human eye can't really perceive a 12-bit scale, so what radiologists do is they apply a trick called windowing, where you essentially zoom in on a sub-range of this 12-bit scale and focus on the subtleties in that scan. This is a dynamic process. As they read through the scan, they play with this window to identify for example the edges of a lesion, or where something starts and something stops. This interactive nature, it just wasn't available in any existing platform. Another thing that was really important to us is to build in, for example an hierarchy into the annotation tasks.

For example, we were worrying with people from doctors for many parts of the world, and we wanted to have for example a unified lexicon. Our platform support a hierarchy where you can define this scan – you can define the hierarchy of questions and tasks that they need to answer.

For example, first question is is the scan normal or not? If the answer is yes, well maybe there's nothing more to do. If the answer is no, then you can ask, “Well, what type of abnormality?” Then the annotation test can depend on the type of abnormality. You really – we've organized this in very fixed manner. Then when different people annotate different things, they all use the same lexicon and it's very easy to put all the data on the same frame of reference, so to speak. Yes, we've ended up building our own platform, which was – it was a fun experience, I can tell you.

[0:27:20.8] JM: I want you to stay back here, because this domain-specific data labeling sector of the computer science economy seems to be something that is really developing. I did a show with Scale API a while ago, and are you familiar with Scale? You know what that company does?

[0:27:39.5] DG: Actually I’m not.

[0:27:41.1] JM: Okay. It’s a pretty interesting company. It's like upmarket Mechanical Turk thing. It's basically a Mechanical Turk API, so you can make an API call to get an audio file transcribed, or to get a picture labeled by a human that has been vetted. It's a very high-quality API for Mechanical Turk style stuff. You might be able to make use of it somewhere in your pipeline, but they recently built a set of APIs around labeling self-driving car data. They had to build an entire platform around this.

You get a three-dimensional image from a self-driving car and there's not a good algorithm for noticing at least publicly available, maybe somewhere at Waymo, but for doing labeling of the setting within – can you label a pedestrian? Can you label a strange circumstance in maybe it's a road structure that is unique, like a roundabout, or something like that that's confusing? Because there's so many edge cases, they had to build their own platform for how people can manipulate an image, a three-dimensional image and create labelings within that image.

It sounds very similar to what you're doing at Viz, where you had to build – it's not a straightforward thing of, “Oh, we just had to bring in the humans and get them to label stuff.” It's like, we actually had to give them an interface for the labeling platform. Are there are a lot of domains where this custom data labeling platform is going to need to be built within whatever machine learning power domain we're talking about?

[0:29:24.1] DG: Honestly, I think every company I know of in the medical imaging domain is supporting its own annotation platform.

[0:29:31.9] JM: Wow, that’s bizarre.

[0:29:34.7] DG: Yes, because everyone is a snowflake. Here you need some semi-automatic tools, and here you need something specific for CT. Someone is doing – there's a few really cool companies in doing pathology. Then you have 6GB images, like huge, huge images, very high-resolution. That’s just not supported by those platforms. Everybody's running into their own problems. So far, back then when we started talking about this, we were interacting with a few companies about doing things like you mentioned, but back then, we couldn't find any platform that was suitable for us, so we ended up writing our own. Now it’s already integrated in all our data pipelines and everything is working very well, so I don't know if you jump off and switch to something else.

[0:30:27.6] JM: Of course.

[0:30:28.3] DG: If I were to start a company now, there's a good chance that something better is available.

[0:30:34.4] JM: Yeah, interesting. You got this training process, you've got your data labeled, and then you need to put it through an algorithm to basically now have your – since you have your labelled data set and you have, I guess at some point this data is getting labeled as stroke, or not stroke, right? You have the lower-level pieces of data that you're labeling about the image, and then you have the yes-or-no case, right? The yes or no, we should escalate this to the interventionist versus not escalate it to the interventionist, is that correct?

[0:31:10.1] DG: Yes.

[0:31:12.0] JM: Okay. Once you have those set of labels, what do you do? I mean, remind us the process of you have to train the data, you have to maintain some test data set. Give us an overview of these processes that applies to the stroke identification problem.

[0:31:29.3] DG: Okay. You lucked out and you got big data set shared with you. Typically what we do, first step like you said is to define this into training test data, and test data is kept aside and keeping it away from the hands of the algorithms team, so they don't get a chance to peek into it. Now just say we finished the process of annotation, we have your inputs and outputs and you start training models. You can start training models, or you can start working on pipelines.

A pipeline would be a sequence of steps, versus a black box algorithm that’s scanned as input, label as output. We mix and match according to the context. One thing I mentioned earlier is this idea of for example, cropping out the head out of the scan, which is very, very simple but, actually very powerful, because then you've reduced the amount of data needed to be processed by all those affected.

Basically you iterate on that. You investigate the successes and failures and try to figure out what you've done well and what you've done wrong and how you can improve. Until you get to a point where you have something that's working and you're happy with the performance and you have a good feeling about it, at that point you would go to test it and get a more objective quantification of your performance. Then once you've hit the target, which is a level of accuracy that we've decided that is reasonable for this product to be clinically useful, then you can go to the next step, which is getting this cleared by the FDA.

[0:33:16.7] JM: Right. Okay, so we'll get there in a second. Just to put a finer point on what you said, you've got these different parameters of the data set, or every piece of data in the data set has different parameters associated with it. Your process of iteration is going to be changing the weights of those parameters, and being able to determine which of those parameters are most highly correlated with the stroke victims that should be sent off to an interventionist. As you are modifying those weights, your model is getting better and better at identifying which are the parameters that matter and you're going to be doing that on the training data for a long period of time, until you have a set of weights on those parameters for the training data that works for the training data. It seems to be like, okay every time we put – now every time we put the training data through this algorithm based on the weights that we've adapted for these parameters, we tend to get a really high accuracy of – or really high precision of how the training set is being labeled as accurately or inaccurately.

Now let's put it on the test data set. If the test data set also is being accurately measured, that means that you have not over fitted your model to the training data. That means that it's probably good enough to move on to the next step, which is actually putting this training, this model into some bigger application. Do I understand that correctly?

[0:35:01.1] DG: That's exactly right. You go through the process of training, which is one part, like you mentioned, updating weights, using deep learning infrastructures like Tensorflow to optimize a deep learning model, so that's one part. The other part is more adding, throwing in some domain expertise. From simple things like here is a head, some more complex things like, sometimes some people have metal in their brain, metal implants, and that's very, very disruptive to the CT scan. We found out for example that removing those cases, or including those cases really makes the life of the network hard, because it's an event that's very rare and expected, and just metal on a CT scan is this shiny white star. It's a very dramatic thing.

For example, we came up with a simple heuristic that identifies those scans and sets them aside. It's a combination of knowledge-driven heuristics with really hardcore deep learning, which is like you described and well, iteratively improving those weights until you get the results accurate enough.

[0:36:11.9] JM: How do you fit in those outlier events to building a good model? Like when somebody has metal in their brain, how do you adapt to that?

[0:36:23.4] DG: That's a terrific question, and it's a very interesting product question. What does this mean? Can we have a product that doesn't work, for example of when there's metal in the brain? This specific example was a nice decision, because metal in the brain often occurs when a person had an aneurysm and the aneurysm was clipped, so there’s little metal clip inside that person's brain. That's just so disruptive to the image that even experts can often can't make sense of it. In that case, we would not analyze that scan and we would alert the patient and notify them that there is a scan that we couldn't process.

Another example is that if someone had a major surgery recently. One thing that's common in the context of stroke is removal of a piece of the skull. If a major piece of the skull was removed and there's a lot of deformation, sometimes death can throw you off as well. Again, because this is actually a very special case, it's not so a case where you're expected to apply it well. Maybe it's okay to say to the doctor, “Hey, there is this scan here, I'm not sure I know how to deal with it.”

Other examples are less technical. For example, scans of children. The brain of a child, the head of the child look very, very different. Just at that point, we just didn't have a lot of that data in our data set and we just felt uncomfortable that we would – we just said, we can't train accurately, we’re not comfortable with running us this device, this algorithm on scans that we didn't have enough data from. Our labeling, this is regulatory part, but the label of the device, it's a medical device that has a label and instructions, clearly states they shouldn't be used on patients that are children. If the metadata of the scan indicates that it's a person under the age of, I think 18, I’m not sure the exact date, we would refuse to analyze it.

[SPONSOR MESSAGE]

[0:38:38.9] JM: Citus Data can scale your PostgresSQL database horizontally. For many of you, your PostgresSQL database is the heart of your application. You chose PostgresSQL because you trust it. After all, PostgresSQL is battle-tested, trustworthy database software.

Are you spending more and more time dealing with scalability issues? Citus distributes your data and your queries across multiple nodes. Are your queries getting slow? Citus can parallelize your SQL queries across multiple nodes, dramatically speeding them up and giving you much lower latency. Are you worried about hitting the limits of single node PostgresSQL and not being able to grow your app, or having to spend your time on database infrastructure instead of creating new features for your application? Available as open source, as a database, as a service and as enterprise software, Citus makes it simple to shard PostgresSQL.

Go to citusdata.com/sedaily to learn more about how Citus transforms PostgresSQL into a distributed database. That's C-I-T-U-S-D-A-T-A.com/sedaily, citusdata.com/sedaily. Get back the

[INTERVIEW CONTINUED]

[0:40:23.9] JM: I think we have tackled the machine learning part well enough at this point. I'd like to zoom out to the broader engineering setup at Viz. If I'm not mistaken, this is really the hardest engineering problem. What we just discussed is the hardest engineering problem in your organization. The other hard problems are things around FDA approval, some subjective decisions around, for example, should we skip people with metal in their brain? Probably some other subjective product development questions. As far as engineering, is there anything else that's particularly hard, like questions around the platform, the scheduling system for spinning up the machine learning infrastructure? What else is difficult in your engineering organization?

[0:41:14.5] DG: Okay. First of all, I would like to say that I really don't think that the algorithm part is the hardest part.

[0:41:20.2] JM: Okay.

[0:41:20.9] DG: What I like to say is that we have the technology side. We're still resting on three legs. There's the algorithm, there’s the infrastructure and there's the app. Really, I don't think any one of them is harder, or more important than the others. Specifically, so let me walk you through some of the difficulties here. First of all, a CT scan is a lot of data. It can be several 100 megabytes to a few gigabytes. You need to get it from the hospital to the cloud and you need data infrastructure to move that data around in ways that are efficient, because every minute counts. That’s a major engineering challenge.

Another thing is that literally, when you count minutes and seconds, you need everything to run very, very smoothly. On top of that, you need very high redundancy. We're trying to design a system in the cloud that's very fast, very robust, that we can guarantee to our customers that it would work when they need it. That means, high-availability standards, cutting-edge and so on. Another aspect is cyber security. Because we're in the healthcare business ,there's a big emphasis on cyber security and everything needs to be done in a very, very secure manner adapting to standard care practices; very detailed and very elaborate architecture for cloud environment, making sure that access is very well controlled and we comply with the various standards.

That's actually a very big challenge as a small startup with a small team operating under various standards, such as the ISO, 27K, cybersecurity efforts, HIPPA compliance, iTrust. Europe now is GDPR, all that are things that you need to support with a small team of single digit number of engineers. That’s very hard. The third part is the mobile app. Again, I think it's a very complex and interesting mobile app, because we actually send all this data. We send the scan to the physician. It's just a huge piece of data. It needs to get to them quickly and does anything from implementation decisions to be made, to provide any compression algorithms, to get. On top of that, what we really want, I'm sure you've been around hospitals, the IT infrastructure in hospitals is sometimes I think old.

We see this as a natural opportunity to provide a novel platform to the doctors that's really delightful. They all have smartphones to use to a delightful experience. Then they walk into the hospital and they’re working on a computer that's running Windows XP. Really want to provide them with tools to make their life easier, and that's another challenge, so it's an engineering/design/UX challenge. It's also a very interesting challenge of how to make this an app that they would want to use for the other patients as well, because it's so much fun for them to use.

[0:44:35.6] JM: I have heard from people who have tried to build products for healthcare, and they just end up in this horrible situation where it's not anybody's particular fault. It's just the collision between Silicon Valley-flavored product development and the process of getting something approved by the FDA, or just getting somebody within a healthcare organization to take a risk on you and your product, those two things do not collide very well. Can you talk about how you managed to get people using this thing? What was the process of – for anybody out there that's listening that is trying to get a product approved by a hospital, or to get a product trial by a hospital, or medical organization, what advice do you have?

[0:45:34.7] DG: Okay, so that's a really, really good question. I think healthcare does present some complexities. At least in that case, maybe some lesson from general B2B can be transferred. Generally, you want to do something, you want a solution that helps three people in the organization; the patient obviously, the doctor, because that doctor is going to be your champion, B2B sales, you always identify the champion that's going to push your product through the system. Then you have the administrator of the hospital, or the CFO, or someone who gets these requests for many, many doctors within the organization and they need to make a decision on the budget.

They need to understand, or see something in it for them in the bottom line, or in the top line, or both. For us at Viz.ai, I think the case for stroke is very, very appealing, first of all for the patient. I think that's obvious on all we discussed; every minute counts, getting there faster, that’s wonderful for the patients. The doctors, at least our feeling was that they have bought in quickly. This is more than once I had a conversation with an interventionist in a conference and said, “This is what I need. This is what I want. This would make my life much, much better.” If you don't have that support, you would have a hard time if you don't have a champion.

The last part I think is the hardest, because often, and I know many stories, I have friends in companies like you mentioned, who've built great products, could save lives, or improve a patient's stay in the hospital, or we just time, do all sorts of good things for the patient, but there's no strong financial incentives for using them. That's really the tricky part. The case for stroke is very appealing financially also. First of all, stroke is in the top five killers in the Western world. I think in the US is number four or five, but it's the number one in healthcare spend, and that's because when someone is having a stroke and is not treated, they can remain handicapped for the rest of the life, they would need assisted living, they're losing income, their family is losing income, insurance is paying for that, that's just a lot of money. There’s just big incentives there. That's actually very important that the system would have something to benefit financially from using the device.

[0:48:02.8] JM: From a product development point of view, what I heard you say there that stood out to me was there's this moment that you can set up for at a conference, for example, where you're at a conference, maybe you have a booth set up and you can demo two doctors. Here is what you do to use our software, here is how it saves you time, any questions?

[0:48:28.0] DG: Yes.

[0:48:29.5] JM: Right? Like you need that demo moment. You need the wow moment, where somebody can say, “This is so concrete to me.” This is not like, you've given me – here's a new CRM platform for your patients, right? Where it's like, here's a tool you can use to change your entire workflow within the hospital. That is not how you convince somebody to start using your software in healthcare.

[0:48:51.8] DG: No, not at all. Not at all. You need to get that aha moment where you talk to them, you talk to them about our institution, you talk to them about the hospitals that they're getting patient referrals from, you ask them how long does it take? Do you have any issues? Do they always call you within the 30 minutes that you're required to by the guidelines? Then they say, “No.” Then like, “Well, how would you feel if you get an alert and there's a positive case?” That's that one moment where they typically realize there's a very interesting value prop here for them and for the patients.

Another thing, if you're talking to administrators and that's actually another big thing in stroke is guidelines, so because probe is such a big burden on the healthcare system financially, hospitals are required to abide by certain guidelines on how fast they treat those patients. This goes back to the offering for the administrators of the hospital, because what you're offering, what we're offering them is we can help to improve your statistics and retain your status as a high performing stroke center, so that's also very appealing.

[0:49:59.9] JM: We also had a show with a company called HeartFlow. You're familiar with HeartFlow?

[0:50:05.3] DG: Yes. Yeah, they’re awesome.

[0:50:06.6] JM: Okay, yeah. It sounds remarkably similar to what you do in the sense that they look at CT scans of hearts, and they make decisions around who should get a stint in their heart or not. How does that problem domain compare to the stroke identification problem set?

[0:50:30.3] DG: I'm familiar with them, with their product. I don't know their technology very intimately. I think these are slightly different problems. With HeartFlow, to the best of my understanding, and again, I apologize if I'm misrepresenting, you're essentially acquiring a dynamic scan where dynamic scan is acquired over time, so would essentially scan the heart again and again and again, while injecting a contrast. A contrast shows up on the scan and you can see the veins and the arteries, and you can see the progression and you can use that to make a 4D model, a hydrodynamic model of vessels around the heart and address those important clinical questions that have for dressing.

For example, pressures, what are they? Do they justify this tent or not and all that? That's very interesting. In our case, there's – their decision, it's not about time. It's about they're there and they need to make a clinical decision, but they want to diagnose the patient accurately and treat them. Our case, the focus is on – it's essentially a yes/no question. Should we take this person for a mechanical thrombectomy or not? Then that question needs to be answered as fast as possible. That's actually something that really separates, I think Viz.ai from many other companies in the medical imaging domain. Everybody are doing amazing things really, I think there's so many interesting and great companies.

We're operating at a very, very high pace, right? We need to give the answer now. There's no waiting. Right now for example, a woman does a mammography scan, she gets an answer within two or three weeks, and it's fine, right? If the startup is working on better breast cancer detection, that's really wonderful cause that's going to save many, many women, but there's no time factor there. It's perfectly fine for an engineering organization to be, for example skewed towards a very strong algorithms team, and some engineering wrapping around it that handles moving the data around.

In our case, that's why I said we have those three legs that we stand on. It's getting everything to work and getting everything to work fast is just as important as getting the answer right. That's all the information you need from many of the other players.

[0:52:49.5] JM: Understood. I know we're up against time here, so I want to ask you in the same vein of the previous question around HeartFlow, this abstract market of we have a problem within the healthcare world that we can solve with machine learning. None of the individual problems within this, whatever problems that we're talking about is particularly hard, we just need the data, we need somebody to build the machine learning model, we need somebody to build the backend and we need somebody to write the mobile apps and we need somebody to do the sale at the conference. We need these different things. None of them are that risky. We just need the right team, we need the right sequencing, we need the right project management. How many of these opportunities are there in healthcare? Is this something where people could just get involved in healthcare, spend five months, find a problem and build a product?

[0:53:51.2] DG: I think that's maybe slightly exaggerated.

[0:53:55.1] JM: Okay, exaggerated.

[0:53:56.2] DG: I think some of the issues is if you define a problem and you say I'm going after this and do all the processing; get the data, build the algorithm, build a new structure, build out a – you will get there. Honestly, I think that one of the biggest challenges is this generating this value offering for the hospitals for the patients, doctors and administrators, it's very, very tricky and you need to iterate quite a bit on that.

Our opportunities, I think we're seeing a wave of companies emerging with interesting volume probes for healthcare, but I think we've also seen some companies who tried with great talent and backed up by VCs and with funding and all that, and they didn't crack. It's because it's not just about building a technology that works, it's building a product that works and is used and is loved and effective. The best algorithm in the world is useless if it's not part of one Lego brick in a product that's delivering value across the entire chain.

[0:55:04.3] JM: Understood. David, I'm sure we could talk for a lot longer. If you got another product in the future or some further announcement as I'm sure you will have eventually, I'd love to have another conversation, something. There's so much stuff that I didn't get to, but I really want to thank you for coming on Software Engineering Daily. It's been great talking to you and I'm inspired about what you're working on.

[0:55:25.2] DG: Thank you very much. I really appreciate the opportunity to speak.

[END OF INTERVIEW]

[0:55:32.5] JM: At Software Engineering Daily, we have a web app, we have an iOS app, an Android app and a back-end that serves all of these frontends. Our code has a lot of surface area and we need visibility into problems that occur across all of these different surfaces. When a user's mobile app crashes while playing a podcast, or reading an article, Airbrake alerts us in real-time and gives us the diagnostics that let us identify and fix the problem in minutes, instead of hours.

Check out airbrake.io/sedaily to start monitoring your apps free for 30 days. Setup takes only a few minutes. There's no complicated configuration needed. Airbrake integrates with all of your communication tools, from Slack, to Github, to Jira and it enhances your current workflow rather than disrupting it. You can try out Airbrake today at airbrake.io/sedaily. If you want to monitor and get visibility into the problems that may be occurring across your application, check out Airbrake at airbrake.io/sedaily. Thank you to Airbrake.

[END]