Season Four Episode 4 | The One With the Future of SRE and Matt Zelesko

[JAVI BELTRAN, "TELEBOT"]

STEVE MCGHEE: Hi, everyone. Welcome to season four of The Prodcast, Google's podcast about site reliability engineering and production software. I'm your host, Steve McGhee. This season, our theme is Friends and Trends. It's all about what's coming up in the SRE space, from new technology to modernizing processes. And of course, the most important part is the friends we made along the way. So happy listening, and remember, hope is not a strategy.

—

JORDAN GREENBERG: Hello, and welcome to the Prodcast, the podcast on site reliability engineering. My name is Jordan. I'm joined with my co-host.

MATT SIEGLER: This is Matt Siegler.

JORDAN GREENBERG: Awesome. Today we have somebody who's pretty cool, in my opinion. And we actually had overlap during my time as an SRE in the production engineering team. For our listeners at home, can you tell us who you are, about a bit of your background, and how you landed here?

MATT ZELESKO: Yeah. Thank you, Jordan and Matt. Thanks so much for having me on the show. My name is Matt Zelesko, and I lead SRE. I've been here at Google for about three years, a little bit over. And it's just been a really wild journey, an incredibly rewarding journey as well to see both SRE at Google, the place it was invented, but also to see all of the things that are going to change about SRE and will evolve as we see the industry and the technology landscape evolving around us.

JORDAN GREENBERG: Amazing! Can you tell us what made you want to be in engineering in general in the first place?

MATT ZELESKO: Oh, we're going to have to go way back for that one. This one goes-- unlike a lot of people, I almost always knew what I wanted to do. So the moment that I got in front of a computer-- I think it was middle school. We actually had one computer in middle school, a TRS 80.

I'm trying to remember whether it was a model 2 or a model 3. I think we started with a model 2 and went to a model 3 eventually.

But from the moment that I started interacting with that computer and seeing what I could make it do and learning about it, I pretty much knew that computer science was my path. So this was not a, "I had to do a whole bunch of soul searching and understand myself." This spoke to me almost immediately. And then I pursued that through undergrad. I got a computer science degree. I did a bunch of graduate work in computer science, and particularly large scale distributed systems. So I was always sort of in the computer science, large scale distributed system world. And it's hooked me ever since.

I would say I've been in a bunch of different industries from networking to the intersection of media and technology, and now, of course, Google, which spans a broad set of products in its portfolio. But in each case, there's been a constant through line, which is large scale, complicated systems, challenging problems to solve. And in many of these cases also, a real, impactful delivery to customers and users. In other words, building products that users rely on and often rely on every single day. And I've found that to be really rewarding of how you can use technology to change people's lives. And how do you build products that are reliable enough that people can depend on them?

MATT SIEGLER: Matt, take a step back. Just before you came into Google, I'm curious about your mindset-- what was going on just before SRE and then just after? I'd like to hear a little bit like at the edge of the inception of this concept, what were you not doing? And then what were we doing? And then what was the turnover of that experience for you and how did it change how we approached the problem?

I think for some of us out there, we didn't know what it was like before this existed. And I think that would be a pretty interesting experience to, "what was the approach to this problem before we approached it the way we do now?" And then how did this change how the industry has approached it now? And what was going on? How did that feel? What was going on in people's minds and how has it changed us forever?

MATT ZELESKO: Wow, that is a lot of questions. Let me see if I can tackle them one by one. So before I came to Google, I was the CTO at Comcast. Also a company building products that millions of people depend on every day. If I want to have a family meeting in my house, I turn off the Wi-Fi,

and somehow everybody finds me. So these are crucial products in people's lives.

And both at Comcast and other companies that I've been at, I've seen a variety of models to how you do production reliability. Some of them were models where you had a dedicated operations team separate from the development team. In many of the places, including parts of Comcast, there was also DevOps where the development team was also responsible for production reliability.

But I was seeing gaps in both of those models and was really interested in SRE and started, in fact, talking to Google about that, because I wanted to build SRE teams at Comcast. I was really intrigued by the idea. I thought that it had a lot of things to offer Comcast, and so I started talking to Google just about, how does SRE really work internally? How has it changed since the book? Those sorts of things. That got me so excited about SRE, I decided I wanted to be at the place where SRE was created, and also be at a place where the scale is so remarkably large that you need large SRE teams managing this planet-spanning production infrastructure.

[05;41] So I was very excited about the work that was happening at Google and the ongoing thought leadership in SRE and said, rather than try and duplicate this somewhere else, maybe I have an opportunity to go and work in the place where it was born. And that was really, really exciting to me. I think we see companies adopting a bunch of these different models. And I think depending on where you are in terms of your product maturity, in terms of the type of innovation you're trying to do, and frankly, in terms of the scale of your infrastructure, different models may make sense.

For Google, the sheer scale of our infrastructure almost compels us to create models like SRE. And I think one of the superpowers of SRE that I've seen is this ability to look across system boundaries, to look across product boundaries. When you've got a production incident, you often have dependencies between those systems, dependencies between those products. And SRE knowledge spans those things and is really the team that is best suited to go solve that.

If the BigQuery development team has an issue that ultimately is networking underneath it, it's really hard for the BigQuery development team to debug that. But you call in SRE, and SRE has that horizontal knowledge across systems that help you really solve those problems. And to me, that is-- like I said, it's a superpower of SRE, and I don't see that replicated in many other models when you're at the scale that Google is operating.

JORDAN GREENBERG: OK. So thinking about superpowers and how their role has changed, we can

also think about what supports superheroes, and that is the tools that they use every day. So we've got monitoring. We've got alerting. We've got Gemini now. So how does the role change now that we have these tools that have been developed over time and hammered into shape? How do they push into incident response? How do they push into postmortem creation? How do they change how the role has developed over time?

[08:05] MATT ZELESKO: Yeah. It's a great question, Jordan. You mentioned platform reliability infrastructure before, and those are the teams that are building some of those tools that you just mentioned. And so let me talk about those tools, and then I'll move over and talk a little bit about AI, because I think that deserves its own conversation. So as part of our reliability practice, SRE published a set of production principles. And not surprisingly, this set of principles includes things like making sure you've got actionable reliability data, that you do change management in a safe way, that you have thought about failure domains and fault isolation for your system, and that you really have a strong practice around data integrity as well.

We have this set of things that we believe any well-run production system needs to think about and is encoded in those production principles. And we've seen Google products, for example, YouTube, go through a multi-year journey to get everybody onto a common set of production tools for doing things. And as you said, Jordan, these are things like monitoring and observability, rollouts, incident management, et cetera.

And that has had a really strong impact, both in terms of putting our investment into one set of tools instead of multiple sets of tools. But it has also really enabled us to drive much more adoption of these production principles, because we've made it easy to do so. And that's been a really important shift for SRE and for Google.

MATT SIEGLER: And how is it feeling right now that we're moving really fast in some of the AI-related launches that we're doing? Where we're having to make decisions between go fast and go slow on making decisions about reliability and velocity, and the industry at large. Because I know those who are not at Google, who's just about everybody listening to us, are having to make these decisions for themselves, and we're having to be good stewards of our own products.

And people out there are also trying to make these decisions about-- and these pendulum-swings all the way back and forth between “OK, we got to launch” or “no, we have to be cautious”. And we are internally always making those decisions, and we're making them very, very thoughtfully. What

are you seeing right now and what are you seeing especially across companies, across-- you're talking to other leaders out there. They're having to make these decisions too. How are we thinking about this? Especially not just with the tooling, which I think informs people down on the ground, but also big thinkers, business decisions here, what are you seeing?

MATT ZELESKO: Yeah. No, it's a great question. And you're absolutely right. The last two to three years in [12:00] AI has been a really high velocity. We understand that we're at this transitional moment in the industry and innovating as quickly as possible, whether that's at the model level, getting our new and latest frontier models out and available to the public, whether it is incorporating AI into every one of Google's products, whether it is launching the Gemini product itself, there's a lot of different ways in which we are trying to move faster than I would say Google has moved historically or has moved for a long time.

I think it's helpful to ground ourselves in the original philosophy of SRE, which was we want to enable our partners to move as quickly as possible while still meeting their reliability goals. That has always been the core principle or mission, if you will, of SRE. And what we find is there are certain users who want to be on the leading edge of technology, and there are other users that really value reliability more than anything else, even if they don't have the latest and greatest. And so ultimately, the question is, what level of risk is the business willing to take or choosing to take? And SRE has the tools and experience to work with all sorts of different levels of risk versus velocity and to support our customers with whatever balance of velocity and reliability they need for their business. And so we've started offering tiers where you can turn the knob between “how close are you to the frontier of new models” versus “how reliable do you need this to be for the work that you're doing?” And I think expressing that sort of flexibility to our internal customers and our external customers has really changed that conversation as well.

JORDAN GREENBERG: OK. So it seems like the manual that our superheroes need, the abilities that they need to develop needs to change to adapt to these new technologies we have and the new focuses we have on how this work is. So it seems like we need a new book. So can you tell us, do we need to write a new book? What are some of the ways that we should change what the role is defined as to take into account the way that this technology is coming in today and future proof the way of thinking so that SREs that are new or SREs that are working on the role can adapt, change, and grow?

MATT ZELESKO: Yeah. So a couple of thoughts here. I think one is, we're coming up on-- I think maybe next year is the 10 year anniversary of the SRE book, if you can believe it. So there's a lot that has changed in 10 years. And maybe call out two really big transitions. The first one was the transition to Cloud. The number of companies that have shifted from on-prem to cloud-based systems, resources, relying on the Hyperscaler Cloud infrastructure. And Google is the same. We have shifted a lot of our systems and products to the Cloud infrastructure as well.

And then the other one is, as we've danced around a few times here, AI and ML. And there's one element of that is, how do we enable the company to create great AI and ML products? The other one is, what is the impact of AI and ML on SRE? And I think we are really starting to understand the potential of AI. And it's a pretty exciting thing.

I think there's a huge potential, and we've already seen some of that in the early work that we're doing, to really change and elevate the ways that SRE operates. And that's pretty exciting. I do think it is going-- we've had a lot of things over time that have changed the balance of how much time an SRE spends in operations versus how much they spend on reliability engineering and investing in tools and systems.

And I think this is another shift along that spectrum. I'll make a very clear point here, that SRE will never get out of the operations business. I think that there is an incredibly valuable piece of living with the production infrastructure. That on-the-ground experience of operating the production infrastructure is something that really informs everything that we do as part of SRE, and I never want to get too far away from that.

But I think AI and ML really holds a lot of promise in terms of being that assistant that is going to make our jobs better, and it's going to actually make our customer outcomes better, too, because we'll detect incidents faster, we'll fix them faster, and hopefully, we'll fix them for good.

MATT SIEGLER: This is a pretty-- I find this particular line of inquiry really fascinating. Operations as software, automation, I mean, we've been on this line for decades. It's not sitting around a friend and watching a button go red and then pushing the button. Trying to get a software to watch this for us, waking us up with a pager rather than watching the dial. And that's what we've been doing. That's our business.

Then having a smart alerting, knowing when something is going south and then letting us know.

We've been increasing that automation and our trust of that automation and being living happily aside that as SRE for quite some time. Now we're talking about insights and automating some of those insights and giving us evidence of those in the alerting, in the monitoring. We're about to see, I think, a step function in that, and I want to know if you have specific thoughts about that and how we're going to accommodate those increases, which are going to come, I think, pretty rapidly with the AI insights at a pace that we're not used to.

Because we've been doing that in a very gradual, thoughtful, curated way, and I think we're about to see them in an explosive-change way. How are we going to handle that and what are we going to do to accommodate that rate of change? And I think the upside is a fairly huge increase in accelerated ability to work, but I don't know how to do that. But we probably will have to. Thoughts?

MATT ZELESKO: Yeah. I think this insight is really important, and it's something that I've been spending a lot of time thinking about, because I think you're going to see this step function in SRE, but in a whole lot of other things where AI is being applied. We're seeing things move at a pace that is much faster than we have historically. We often find ourselves trying out a model, and we go, OK, the model gets us this far, and so then we're going to have to have a team build these elements afterwards that'll get us all the way to the outcome that we're trying to achieve.

And so we sent a team off, and they spent six months creating all of those pieces at the end that pick up where the model leaves off. And we get to the end of that six months, the team delivers it. We're like, awesome. This is great. We go back and try the latest frontier model, and it does all the stuff that we just spent six months investing in a team doing. And so we are not-- we need to get to a much better intuition and understanding about how fast this technology is moving, and then get ourselves in a mindset of constantly experimenting, constantly updating our assumptions-set around it.

I agree that right now, we are using this at Google for software development and for software quality, and as I mentioned, also production management, production engineering. And we're seeing a lot of benefits. I would articulate a lot of those benefits as being this buddy next to the human. But you get a sense that we could get to a point where there are less things that require that human in the loop.

I hesitate to predict how fast we get there. But again, a lot of that work tends to be manual work. It tends to be toily work. And so if we are eliminating that from the SRE diet in the favor of them doing

much more interesting, engineering work and innovation, I think that's a great thing. We always talk about SREs automating themselves out of a job, they still have a job. And I still expect them to have a job in 10 years, because there is much harder work.

JORDAN GREENBERG: Right. So in thinking about this, how it's allowing us to do the thinking, whereas some of the more manual tasks can be automated away. It also gives us the opportunity to improve our processes in conjunction with the technology that we have. So for example, better postmortems, better action items gathering from postmortems that we talk through and that we ruminate on as humans. Is there any other sort of improvement that's tangential but still necessary to improve the scope of SRE skills, process improvement, postmortem reviews, et cetera?

MATT ZELESKO: Yeah. Continuous improvement is at the core of the culture of SRE. I continue to remind people that continuous improvement is incredibly important. Our culture of blameless postmortems, our ability to look at every single failure and go back and learn from it, understand that that learning is important, and then instituting things based on that learning is just core to everything we do.

And that continuous improvement takes a lot of forms outside of just incident management, which I think is what you were calling out. There's a lot of work we do ahead of time. And things like production readiness reviews. The healthiest SRE engagements with our partners means we are they're really at the start of system design. And we are talking about reliability incredibly early in the development process. And I think there's a lot of stages there that could really benefit from AI and ML as well.

And for example, we are looking at ways that we could take design docs and have AI and ML start to opine on whether these designs would adhere to our production principles or not. And just get a jump start on some of the things that we should be paying attention to in these design docs. I think another-- we've talked a bit about the evolution of SRE. Another direction that we're really pushing that is in terms of risk management. And both identification and articulation of risk, and then, of course, mitigation of risk. And a lot of that happens much earlier in the cycle. You step back and go, OK. SRE has traditionally been focused on SLOs, availability and performance.

JORDAN GREENBERG: The nines, yes.

MATT ZELESKO: Yeah. And you could argue, those are all trailing indicators of risk. If you have an

outage that impacts your availability, that is because you had a risk upstream somewhere that actually manifested and happened. And so being able to, both in paper and tabletop exercises, but also with tools and automation, start to assess the risks in the system. And if we can understand those risks earlier, then we get ahead of those risks as opposed to having them manifest and show up in terms of availability or performance outages.

MATT SIEGLER: So, Matt, we have the rare opportunity to talk to someone here who has talked to other people who talk about us.

JORDAN GREENBERG: That's a lot of talk.

MATT SIEGLER: That's a circuitous way of saying, what are people saying about Google, things that they think that are important happening here, things that they want from us? Talk about that. I mean, this is an opportunity here, big picture about big customers, other competitors, collaborators. Speak to that. That's an interesting perspective that most of us never hear about.

MATT ZELESKO: So recently, I went to Cloud Next, which is our annual customer conference for Cloud users, and got to talk to a lot of customers there. I would say first and foremost, the question we hear the most is-- and it's the same question I had when I was at Comcast, which was, can we build an SRE team? How do we do it? How can we most effectively create an SRE team, because we're not happy with the production management models we have now?

The second question we get a lot of is, can we have access to your tools? You've built all these great tools to manage production. Can we do that as well? And I think if we get back to the first question, which is, how do I build an SRE team? I think the real piece that you have to internalize there is, this is a change that is as much about culture and process and the way you work as it is about what talent you hire or what tools you use, right?

JORDAN GREENBERG: Yes.

MATT ZELESKO: And it is that that is one of the hardest things to change inside of other companies. And so we start talking about the culture of SRE. What are the important elements of it, and how can you start to move your company in that direction? I think there are a lot of companies interested in doing it, but as I mentioned, it's a lot harder than it sounds, particularly when you're starting with a different model.

I think there's also-- as we've been talking about today, there's a lot of excitement around AI and ML and what that means for really all parts of our business. But SRE is not exempt from that, and I there's a lot of interest about how AI and ML can maybe help elevate how people manage production, whether they are SREs or not.

And this may be-- as a side commentary, this may be a real change in our thinking also, because SRE does not support every single Google product. And maybe this is something that's really important to emphasize here, we choose engagements and we engage where SRE is going to have the most value and the most impact, which means that there are a lot of internal systems, external systems, products at Google that are managed by their development team.

And I think we've started shifting a lot of our thinking to be around not just, how do we help SRE be the best SREs, but how do we help anybody at Google be the best at managing the production infrastructure for their systems or products? And that mindset means syndicating a lot more of our tools and culture and process to parts of the organization that historically are not part of SRE. And it's a real mindset shift that I think leverages the expertise of SRE, but has a much broader impact across Google and our users.

MATT SIEGLER: All right. I'd like to thank our guest, Matt Zelesko, and my co-host, Jordan Greenberg. I'm Matt Siegler. This has been The Prodcast, Google's podcast on SRE and production software. Goodbye.

MATT ZELESKO: Bye.