EPISODE 642

[INTRODUCTION]

[0:00:00.3] JM: Medical breakthroughs require medical research. Medical research requires patient testing and data collection. The most common form of capturing patient data is through surveys and most of those surveys today are done on paper. Surveying patients to understand the side effects, or the benefits of trial drugs, or treatments and getting accurate results out of these are critical aspects of medical research.

Traditionally, these surveys are filled and read manually and they're entered into a database by a human operator. In these steps, there's too much room for human error, from unreadable handwritings, to typos being entered into the database. Electronic data capture platforms were created out of this need for easy and accurate data collection for researchers, by enabling online survey creation and result collection, EDC or electronic data capture platforms have improved medical research immensely.

However, these platforms are complex to design. Where patient medical data is concerned, privacy and security are of extremely high importance. Compliance with laws that protect anonymity and privacy of the patients is necessary. On top of all this, the platform must be easy to use and reliable.

Castor EDC is a company specializing in EDC for medical research. It's founded in the Netherlands and it's active in many countries around the globe. Today's guest is Derk Arts who is the founder and CEO of Castor EDC. We talk about electronic data capture platforms, how Castor EDC overcame the engineering and design problems of the electronic data capture problem and how they comply with the laws and also how their business model works.

Before we get started, we're hiring a creative operations lead. If you're an excellent communicator, please check out our job posting for creative operations at softwareengineeringdaily.com/jobs. It's a great job for someone who just graduated a bootcamp, or somebody with a background in the arts who's making their way into technology. If you want to be creative, if you want to learn more about engineering, if you want to work remotely, you can check out this job at softwareengineeringdaily.com/jobs.

[SPONSOR MESSAGE]

[0:02:28.9] JM: Today's episode of Software Engineering Daily is sponsored by Datadog. With infrastructure monitoring, distributed tracing and now logging, Datadog provides end-to-end visibility into the health and performance of modern applications. Datadog's distributed tracing an APM generates detailed flame graphs from real requests, enabling you to visualize how requests propagate through your distributed infrastructure.

See which services or calls are generating errors or contributing to overall latency, so you can troubleshoot faster, or identify opportunities for performance optimization. Start monitoring your applications with a free trial and Datadog will send you a free t-shirt. Go to softwareengineeringdaily.com/datadog and learn more, as well as get that free t-shirt. That's softwareengineeringdaily.com/datadog.

[INTERVIEW]

[0:03:29.9] JM: Derek Arts is the founder and CEO of Castor Electronic and Data Capture. Derek, welcome to Software Engineering Daily.

[0:03:36.6] DA: Thank you. Great to be here.

[0:03:38.4] JM: I want to start by talking about medical research. Medical research is conducted by universities, it's connected by drug companies, consultancies. What is the purpose of medical research?

[0:03:50.9] DA: I think the easiest way to explain it is to take an example. Say we have diabetes as a disease, and we want to see if there is a better way to treat patients with that disease. We have, say medication A, which is the current medication that everyone gets, and then we have some new drug that's medication B. Now if you want to prove that medication B is superior to A, you want to run some statistical analyses, right?

To run those, you need data. Basically, in every form of medical research, we need data to be able to run the statistical analyses that's either prove, or disprove our hypothesis. Yeah, that's basically the common denominator you see running through all these branches of medical research, and that data is at the heart of everything to be able to prove or disprove a certain theory, or find evidence for the effectiveness of a new drug.

[0:04:40.8] JM: Medical research has been going on for quite a long time. What are some of the systemic problems in the way that medical research is conducted?

[0:04:50.4] DA: Well the list is quite long, to be honest fortunately. I think currently, there's still a lot of you're reinventing the wheel going on. There's a lot of researchers in the world that focus on their own projects and their own department and are not looking around to collaborate with researchers from other countries, or even other universities. There is actually a degree of competitiveness between universities within countries, which is strange.

From my perspective, if you're an alien looking down on the world, you'd be very surprised to see how we deal with research and medicine, because you'd expect the entirety of humanity to work together and to collaborate to find cures for diseases that plague humanity as a whole, but what's actually happening is this competitiveness of different research groups trying to run their own projects, not collaborating and then ending up with yeah, slightly similar studies that investigate a similar thing, but not exactly in the same way.

Instead of having one large study that studies one thing properly, you have all these independent studies that do things slightly different, which is very inefficient. To make it even worse, after everyone has done their own little study, the data that's been captured is usually not made available for reuse or sharing. About 85% of the resources and the time we spent on medical research don't ultimately contribute to patient health, either because what was investigated wasn't effective, or because data that was captured was never shared or made reusable so no-one could actually benefit from the data that was already captured.

Then finally, a lot of these scientific publications, so if you do medical research, the ultimate goal is to create a scientific publication that gets published by a scientific journal like Lancet, or New

England Journal of Medicine, or some of these other journals. These usually are paywall, so you have to have a subscription to this magazine to read about the evidence that was discovered, which I think is very strange, because it means that two-thirds of the world don't actually have legal access to the scientific evidence that has been discovered recently.

I think yeah, across the board starting at defining the research protocol right up to the point of publication there's big issues going on. Of course, we're focusing all mostly is the data, the data aspect of it. What we're trying to do is make sure that's data that's captured through our platform is actually standardized. I mean, reusable, so it can actually serve its purpose 10 years from now and doesn't just end up in a drawer somewhere, which is what usually happens with data once the original author is done with it.

[0:07:23.4] JM: We'll get to that, but just to connect to some of the points that you brought up, that competitiveness, the competitiveness between different medical research institutions, is that connected to the same culture that makes people compete to – there's that term scooping, like if you're doing research in life sciences, you might be competing with some other university and you don't want to get scooped, you don't want them to figure out what you're doing, or to release a publication in nature, or one of these other paywall journals before you have a chance to release it. Is this the same kind of mentality that you see there?

[0:08:02.8] DA: Yeah, totally. It's exactly did – it's the same thing basically. The whole culture in medical research is completely broken. It should be about having impact, it should be about doing worthwhile research and making it accessible to the world, but the culture that has come into existence is all about impact factor and how often am I cited. Am I the most cited professor in my institute? That's not necessarily doing of the medical research themselves, it's the system that evolved to this state, but now we need to break out of that state.

Then scooping is part of it, so I want to say that some competitiveness is nice, right? That's the reason we also have the economy that we have. We're not fans of communism. It's good that there are some competitiveness in the market, because you want people to try to improve on each other. The current leads, it's not really about that. It's really about getting the biggest implications. I think very often, humanity would actually benefit from more collaboration and working together and sharing in the credit that you get from worthwhile research.

[0:09:06.1] JM: Your company is Castor EDC, that's electronic data capture. As you said, much of what you're trying to do is to collect data, get it in a standardized fashion, will allow researchers to collect data. It's a software platform that allows medical research to be performed in a more standardized fashion.

As an example, so that if you have a drug company that's trying to evaluate the side effects of a pill for example, they need to document everything that happens to the person that's taking the pill. They need to have control groups that are closely monitored and they need to document everything that happens to the control group. How is this typically done? How is this documentation of what happens to two groups in a medical research trial? How is that typically monitored and documented, if we're talking about the world before Castor EDC?

[0:10:02.9] DA: Typically, what you would see is people use paper, which is very cumbersome, and it's still actually very common today for several reasons. What they would do is they create these forms in a word processor and they would just tweak them until they felt it contained everything they wanted to capture in terms of data, and they would just print them out and then send them to these institute's that would participate in the study.

Then research nurses usually, so these are nurses specifically trained to support research activities, they would have to go to the patient, sit with the patient, go through the patient file and extract all this information and write that on paper. Then the ridiculous part is that then they would gather all these papers, send them back to the researcher, the principal investigator and they would have to copy it all onto a computer, because while data was still being captured on paper and it is still happening today, analyses have long since moved to the computer, because everyone wants to use cool things like logistic regression, which is something that obviously requires a machine, because you can't do that on paper, find the coefficients, or a regression formula.

You have first, people writing on paper and then sending it to the researcher, and then them copying it into the computer, which obviously is extremely inefficient and very error-prone. That was the era before Castor. Then there was this intermediate solution, where people would use Excel for instance, so especially in academic research.

In commercial research, big pharma has been better; they just use extremely expensive EDC tools that already exist in the markets that are not accessible for the average researcher. The academic researcher, if they weren't using paper, they would sometimes try to use Excel as a data capture tool, but yeah, that’s almost impossible to use with more than two people and what you will be doing is you'll be typing straight into your sheet, where other patient data is also present so you can understand that that's maybe even more error-prone than the paper-based solution. Paper and Excel are basically the things we see most of our users use before they come to Castor EDC.

[0:12:02.0] JM: Aren't there other platforms for doing this electronic data capture? Hasn't there been software around for a while that is custom-made for this?

[0:12:13.0] DA: Yeah. Like I said, in commercial studies, there is a few big EDC vendors have been around for a while, and even in the academic space there are a few. The most famous one is REDCap that's used a lot in the United States, which is a pretty good platform. When I started Castor as a medical, so I'm a medical doctor myself and now also PhD, but back then just an MD. All the people around me had no access to these tools, so all I knew was all the researchers at least in the Netherlands don't know how to access these tools, already found these tools. That's the reason I started building this platform.

It's still mostly true today that the tools that exist either requires to set up your own server, to configure the tool itself, to help yourself when you're building a study, so there's no real SaaS solution that really helps the researcher get off the ground quickly, without requiring any programming, or server knowledge basically. That's really where we are different from what exists.

[0:13:10.0] JM: Oh, so the old ones were stuff you had to install?

[0:13:13.7] DA: Yeah. For instance, open clinic as an open source solution, but you have to set it up on your own server for instance. For medical research, that's impossible. I mean, you're dependent on your local IT department in your hospital for instance to do that. In the US, we see quite a few hospitals that have REDCap installed and managed locally, so then you don't have to do it yourself. We still feel that REDCap is built by Vanderbilt University and I think it works well, but we really start with the medical researcher in mind and we really take into account the fact that these people have no programming knowledge whatsoever, and really want to get off the ground as quickly as possible.

That's why we make the whole process of creating your forms and setting up your study and setting up randomization extremely easy and extremely user-friendly, because we want the technology to work for our users and not the other way around.

[0:14:00.1] JM: As you said, you're an MD with a PhD yourself. How did you find yourself building a business for electronic data capture?

[0:14:09.4] DA: It was just very organic. I used to do freelance development work during my medical training. I mean, I was always looking for things to build, things to improve, processes, or improve healthcare, where I was the training as a medical doctor. Then I encountered this problem where I saw people trying to capture data in Excel sheets and on paper and stuff, and I thought, “Okay, this is a great opportunity.”

I basically built the first version of the platform myself and sold it to friends and colleagues, or friends etc., etc. I started growing just by word of mouth and it grew to what it is today. It's pretty impressive – has been a pretty impressive ride so far. It's been six years now since I built the first version myself and it's just because there's a big need, there's a huge need actually for what we offer, so that's why it can grow organically really rapidly without any significant marketing, or sales effort, to be honest.

[0:14:56.6] JM: How heavily regulated is this stuff, as like electronic data capture?

[0:15:01.3] DA: Quite heavily regulated, I would say. It depends on what kind of study you're running. If you're testing a new drug, it's really heavily regulated, but if you're doing a patient questionnaire survey where you say, “I want to measure satisfaction with a new physio-therapeutical procedure for instance,” it's not as heavily regulated. It really depends on what you're investigating, but what we try to do is basically provide all the – provide a fully compliant platform out of the box, even for people who don't necessarily need it, because I think if you can be compliant, it's better to be compliant, especially if it doesn't slow you down.

We try to make it very easy for everyone to use this compliant environment that meets all these requirements that are set forth for medical trials basically, or medicine trials, those drug trials, and those are heavily regulated.

[SPONSOR MESSAGE]

[0:15:56.3] JM: DigitalOcean is a reliable, easy-to-use cloud provider. I've used DigitalOcean for years, whenever I want to get an application off the ground quickly. I've always loved the focus on user experience, the great documentation and the simple user interface. More and more, people are finding out about DigitalOcean and realizing that DigitalOcean is perfect for their application workloads.

This year, DigitalOcean is making that even easier with new node types. A $15 flexible droplet that can mix and match different configurations of CPU and RAM to get the perfect amount of resources for your application. There are also CPU-optimized droplets perfect for highly active frontend servers, or CICD workloads.

Running on the cloud can get expensive, which is why DigitalOcean makes it easy to choose the right size instance. The prices on standard instances have gone down too. You can check out all their new deals by going to do.co/sedaily. 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 a $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.

Thanks to DigitalOcean for being a sponsor. The co-founder of digital ocean Moisey Uretsky was one of the first people I interviewed and his interview was really inspirational for me, so I've always thought of DigitalOcean as a pretty inspirational company. Thank you, DigitalOcean.

[INTERVIEW CONTINUED]

[0:18:03.1] JM: I want to help people understand what the software does. If I'm a researcher, maybe I've got a computer in my medical research room and it's running Castor EDC, and then the patient comes into my room, and then the patient starts telling me stuff about how they're reacting to the drug, and we go through a back-and-forth and I'm just entering this data into Castor EDC. Is that the typical workflow? Maybe you could give me some descriptions for prototypical ways that people are using Castor EDC.

[0:18:36.1] DA: As this is one of the ways. The first step is they have a protocol for the study, so what am I going to research? What data do I need to run the statistical analysis? They're going to define what data points they need. From that, they will be able to build their study in Castor, so they define their own forms. We have about 25 field types, so you can imagine a form builder's type of interface that you also see in things like Google Forms for instance, but then way more elaborate, with all kinds of validation possibilities to make sure that all the data that goes in of the highest possible quality.

First, they define all their forms. Once that's done, they start their study, they invite their colleagues from all around the world to add patients from other hospitals, so that speeds up the recruitment time, or actually lowers the recruitment time, speeds up the recruitment, but also ensures that you have a nice sample, right? If you just take patient from home from a hospital that could be a potential bias. It's always nicer to include a larger population, so then you have multiple people, multiple hospitals using these forms or ECRFs as they're called, Electronic Case Report Forms.

Yeah, sometimes they would see the patient in the outpatient clinic for instance and they would type straight into the ECRFs. Sometimes they would just go through the patient file and copy data. We actually also provide an EHR importer using H0705, which is the standardized way of exchanging medical information. It's quite commonly used now in the US as well. What we do is we take patient directly from the digital patient file and import it into a study, so that saves you a lot of typing obviously.

Then finally, we also allow people to send surveys to their patients, and so then a sick patient actually answering the data on their iPad for instance from home. There's basically three ways to get data to the platform, and then there's the obvious CSV import. If you have some other data source, they can generate a CSV file, you can also import that. Basically, using all these sources of data, you can build your data set.

Oh, yeah. Then I forgot to mention our API. We have a restful API that can be used to connect Castor to say, a mobile app, or some other platform. Castor is really the hub where all the data comes together and helps you define a structured, hopefully standardized data set that you can then easily analyze once you've included, say 500 patients.

[0:20:50.5] JM: Then do people want to do the analysis inside of Castor, or do they want to export it to some other tool for analysis?

[0:20:59.5] DA: Right now, we don't really provide a lot of possibilities for analysis in the platform. We are adding R and Shiny to our platform. People who are comfortable with R will be able to just stay in the platform. Right now, a lot of people would export it, so people want to use SPSS, or SAS, or Stata. We just provide a standardized export that they can import directly. We provide libraries for R to connect straight to the API, so you can actually basically run your analysis in your own R environment and it's pulling the data from our API potentially.

[0:21:27.7] JM: You alluded to something earlier that was pretty interesting, which is the fact that you want to make the data highly standardized so that people can potentially use other people's data in their studies. You could potentially get economies of scale, or network effects to that research. The main risk I see there is if different researchers have different standards for their data collection practices, isn't there an uneven standard?

Maybe I'm running a clinic and testing for something that is not as rigorously monitored as testing a drug for cancer medication, for example. Then the cancer medication, people might use my data, and if my data was collected in a way that was a little messier than the original, then the cancer people have standards for. Isn't that problematic?

[0:22:28.4] DA: That's a great question and I think a question that all the people who aren't going to be crazy with the big data hype, if that's still going on, but they should ask themselves, because that's a big problem. You can't always solve it, but I think for the most part, these centralized data sets can be used to mostly hypothesis finding. Initially, you would say, okay, we use these combined –

In the ideal world, what we're trying to achieve is all the data is being captured through Castor is fully standardized and machine readable, right? That's what we're trying to do with machine learning, help people add to require metadata to their study fields, to make all the data machine readable. Now imagine, you have a thousand data sets that are all machine readable, you could use something like Sparkle, which I hope the listeners to this podcast know about, especially query language for semantic data.

Ideally you would use these say, 1,000 data sets to discover new potential hypotheses that you would then test in a more structured formal way, right? That's the easy answer to this question. You don't do it to answer your hypotheses, we do it to define new hypotheses that you can then test in a more rigorous fashion.

However, I think well, while we're progressing with helping computers understand the nature of the data and the region of the data, so no just adding metadata to the individual data points, but also metadata on study level, I think more and more we can also start to use these data sets for answering our questions directly. Then we really need to take into account what you just mentioned, so what was the original purpose for this data being captured and to what rigor, what standards were the people held capturing this data? What tools were used? Was there an audit trail? Were their input checks, etc.?

The more metadata we add to our data sets, the easier it will become for an algorithm to determine what data sets can be combined to create a really high-quality new data set and what data set can be combined, but can't really be used for anything but basically at exploring and finding new ideas basically.

I think the answer lies in being way more rigorous in defining what data was captured and how and with what purpose, and what standard was used. I think then, we will also be able to solve this challenge and sometimes we won't, because the data of – some of the data that’s included will be too low to be reliable, but can still contain valuable insights that in a future study might prove to be the cure to a new disease.

[0:24:58.1] JM: When medical research is being conducted in the classic way, or even with Castor, so when people are collecting data sets of some number of people and then they're deciding that, “Okay, based on this medical trial, we have determined that this drug does not cause side effect X, or we've determined that this drug does not cause a side effect X over time horizon A.” Is there well-defined standards for how many people you need, like the sample size you need, how long you're doing these tests for? More importantly, I'm guessing there are some standardized numbers, like the amount of people that you need to test, the time horizon you need to test them. I'm sure there are some standards in place for that, but are they rational?

Do they actually convincingly prove that for example, medication does not have side effects? Because we're having this conversation in the context of, I don't know if you know, but the American medical system, the hospital system is pretty messed up today and there's a lot of issues. I mean, healthcare globally has all kinds of issues. It almost feels sometimes we're still in the bloodletting days, where we're just, we're really guessing, and the drugs that we use are such blunt instruments. We have no idea how nutrition works. All of these things can certainly be alleviated, maybe potentially solved by doing large-scale enough studies. It’s just like, you get to a certain point where if you grow up in the information age, you start to realize how on earth is research being conducted with enough of a sample size and with enough diligence when they're doing it across small populations?

The populations are biased in certain ways, like maybe the entire population that's researched is in the state of Oregon, for example. That's quite a bias. I guess, this is a broad question, but are there some more fundamental issues with the data sets and the way that we're conducting medical research?

[0:27:01.8] DA: Yeah, I think this is a very interesting point. We could probably talk about this for hours. It's one of my favorite topics. I think there's a few things you mentioned; so nutritional research is very problematic and very hard to do properly, and that's why we basically know nothing concrete. Like everything causes cancer and everything cures cancer basically, right? With a few exceptions.

[0:27:21.7] JM: It all works out.

[0:27:23.4] DA: Yeah, it's impossible. That's just because there's so much bias, right? It's so hard to do a proper nutritional trial, because it's really hard to randomize patients into eating chocolate and not eating chocolate for a year, but then I say a year. The problem is and then we get into the effect size discussion, if you're going to investigate if chocolate causes cancer, you need to observe these people for 50 years. Not only should you make sure they eat chocolate every week, but you should also make sure they don't smoke, they don't exercise too much or too little, they don't drink alcohol, etc., etc., because otherwise, you just keep on introducing bias that you can at some point correct for any more.

In nutritional trials, it's just super hard. Having enormous samples will help and then also very powerful computer to try to correct for all the other variables. Basically, you need to measure everything and then have a ___load of data. Data sets of the size that we currently don't know and that again, is I think why – what we were trying to do with Castor is very interesting, because potentially we could have that with the technology that we were trying to develop.

Then you start to answer these questions. The other part of the question, you can calculate a sample size if you know an effect size, so there's basically some statistical tools and tests you can use to determine how many people should I include in my trial. Usually, that's based on previous evidence for more experimental trials, where you see okay, what we've seen so far in the literature is drug B improves a patient's performance by 10%, so we have a certain effect size. From that effect size, we can have a very nice mathematical calculation. It determines how many people should we include in this trial to prove with or 95% at a significance level, which is a different discussion altogether. I don't want to get into that, but that's our new intervention is better.

I think that's quite solid, but that works really well when you do a control trial and have a very clear idea of effect size. I think in many scenarios, that's not – we don't have previous literature, or we can't create a control setting, like nutritional research for instance. Yeah, like I said, we can talk about this for hours, and it really depends on what you're investigating and what research you're trying to do. Sometimes, it's just not possible to do a larger study, because it gets too expensive.

In epidemiology, you have to be pragmatic as well. Pragmatism is also something you have to take into account. You have a limited budget. There's time constraints. How many people can you recruit into your study in certain period of time? These are all constraints that are put on in real-world research that you have to deal with. Sometimes that determines that study can't run any longer, or that you can’t recruit any more patients. Yeah, the evidence suffers from that.

That's what I was trying to say in the beginning, when you're looking at the human race as an alien, it's just super confusing, because all these problems will be solved by collaborating much, much more and pooling resources to run these larger trials efficiently. I'm not saying large trials are the answer to everything, so I'm not sure how many epidemiologists and scientists listen to this podcast, but there's a few things I'm saying that are a little bit controversial. I think it would solve a lot of problems, and especially we standardize the data, then at least all these people that are reinventing the wheel can contribute to the greater cause and ultimately help answer these questions that we've so far failed to answer.

[0:30:36.0] JM: Let's go a little more controversial. You could imagine a world where I walk into a medical research facility to talk to a researcher, or to do a survey, or even if I'm doing something – some survey at home, you can imagine where you walk into some room, the rooms got cameras, it's got scanners. Instead of asking you what you're feeling, it just derives data from your body, or it takes a blood sample from you, or a stool sample.

To take this even further, you can imagine a device that you just walk around with all the time that surveils you all the time and gathers data, and would be pulled into some repository that people could do mass data analysis over. If you imagine that surveillance for the purpose of research and then you contrast it with how people look at privacy today, and they look at privacy as this all-important human right. Privacy is important. Obviously, we need privacy. I think people underestimate how much value we could get if we were a little less conscious of our privacy, if we were a little more willing to share data and think deeply about maybe how do we –

I know people do think about anonymizing data and I know anonymizing datasets is quite hard, but I almost feel like we're at such the early stages of the conversations that we need to have really around privacy and data, because of – I mean, most people talk about privacy in the sense that oh, this is for protecting me. Fewer people I think, talk about the value and the economies of scale to unleashing those datasets, to perhaps anonymizing the datasets that governments have and opening them up to people, or anonymizing that hospitals have and opening them up to people.

Do you think much about this privacy/surveillance versus the large scale advantages we could have from having these datasets that are currently stigmatized, to be made publicly available, be made publicly available?

[0:32:48.1] DA: Yeah, I think about it a lot. That's actually why we created a tool called My Consent. That's the dynamic consent platform that we're integrating with our UC platform, basically putting the citizen in control of their data. There's a lots of initiatives going around in the world for citizen contributed data, and I think the solution to the challenges that you mentioned is as a hospital, you can't really decide for your patients if you can share that data and if you "anonymize it," because there's always the problem with a one-legged man with diabetes for instance, right? There's not that many patients with that profile. Even if you anonymize the data, it's still sometimes possible to identify who the patient or citizen is.
If you move that question from the entity that holds the data to be, what I would say which should be the owner of the data, usually patients and citizens aren't the owner of the data, but I think it actually makes sense that they are and ask them. Do you want to contribute your data, either anonymously or not? What specifically do you want to contribute to the cause and what other restrictions are there?

If you put them in control, I think that makes that makes a lot of sense, and that's the way I'd want to solve that, because so for me on a personal level, I love to contribute to my data to science, or I wear a Fitbit, or actually a Garmin device every day, an activity tracker I should say, and I love looking at my data and I would love to contribute my data to some project that could help the world. I understand that some people don't want to do that, and I think we should educate them on the value of their data, so I totally agree with you that more effort should be put in not focusing on the negative sides of things, but also focusing on the value and the things that you could bring to the world by contributing the data. Mostly, giving people real control over what goes where.

There's also the spirit of the GDPR. I'm not sure how familiar you are with that, but knowing where your data, or knowing who has your data, know what they're using it for, I think that makes a lot of sense. For me with Castor, of course privacy is enormously important for us, and we do everything we can to secure data and to ensure everyone's privacy. So far, we've been exceptionally good at that and we're trying to do is involve the patient to put them in control of who can reuse your data and who can access it.

I think if you provide the technology, if you provide a platform where people can share data, their own personal privacy-sensitive data with confidence, I think we can start to see a shift in that attitude towards privacy. If you just throw it out in the open or give Google access to your EHR, what they did in the UK, you're just going to have enormous amounts of [inaudible 0:35:29.3] on top of you.

[0:35:30.4] JM: Wait. What happened there? I don't know about this Google and –

[0:35:33.7] DA: I'm not sure how this will set up, but Google DeepMind was given access to NHS hospital data in the UK, and it was anonymized apparently. Yeah, of course, the anonymous data discussion is always raging and that happened here too. I don't know all the details, but there was a lot of public outrage over the fact that Google got access to the data and ultimately, I think the consensus was that it shouldn't have gotten access from NHS, and NHS couldn't have decided it on their own. You'd have to look up the details, but it was a still an ongoing scandal I think, so I still see news articles pulling up.

[0:36:07.2] JM: Did it get de-anonymized?

[0:36:08.7] DA: Yeah, it did. It's well de-anonymized, sorry, I don't think so. I misunderstood your question, but –

[0:36:14.4] JM: Did somebody mine the dataset and derive who, because anonymization is like, you basically hide the relevant fields that you need to identify the user. You'll hide the name, you'll hide some personally identifiable characteristic, like this person has three freckles above their nose, things like that, things that are highly identifiable, so if you remove those, then you can anonymize the data set. Did it get de-anonymized, or where people just concerned that, "Hey, I don't want this data out there, even if it is supposedly anonymized."

[0:36:42.6] DA: I have to look up to particulars, to be honest. I just know that a lot of people agree that they didn't follow proper procedure, and I don't know to what extent it was anonymized, or de-anonymize, but I think, a lot of people were very worried that it wasn't anonymized enough and it could potentially be de-anonymized by something as powerful as Google DeepMind.

Just the fact that even an anonymous data, I mean, an entity as powerful as Google was the technology that they have, you could understand that patients would feel uncomfortable that their data is going into this enormous machine, even if it's anonymized, because yeah, what else Google has on you, right? We do the profiling side of things, like you take face with daily Google data and then you throw an NHS data, how hard can it be to start creating profiles and then starting to put one on one together and discovering that this guy with this Facebook profile actually also has diabetes.

I think that's the fear and I think, so just throwing it out there and then giving an entity access to it, I don't think that's the right way to go, but I agree with you that there's a lot of value in it, but I think we should just enable people to give access and make it very clear what the value is and focus on that. Clear explanations and easy to use tool that allows them to easily give access to this valuable data.

[SPONSOR MESSAGE]

[0:38:05.4] JM: Azure Container Service simplifies the deployment, management and operations of Kubernetes. Eliminate the complicated planning and deployment of fully orchestrated containerized applications with Kubernetes.

You can quickly provision clusters to be up and running in no time, while simplifying your monitoring and cluster management through auto upgrades and a built-in operations console.

[INTERVIEW CONTINUED]

[0:39:41.1] JM: It's actually a great anecdote, because if Google was able to de-anonymize that data, or Facebook was able to de-anonymize that data, then they could use it to say, "Oh, this person has diabetes too. Let's sell them diabetes drugs, or let's change their insurance rates," and it certainly could be problematic if you're giving that information to the same company that's serving you ads.

[0:40:06.0] DA: Exactly.

[0:40:07.2] JM: Tell me a little bit about the company. You went from this place where you just hack together this MVP yourself, you sold it to some people you knew and then you just – when did you decide to scale? When did you realized, "Oh, this is actually something that could scale to being a company, as opposed to just being this lifestyle business that I just sell to my friends?"

[0:40:30.8] DA: Yes. I think from the beginning, I already plan to scale the business, so I was never planning to create a business. We're evolving our consultancy, or bespoke custom work, because the whole idea was to create a platform that was self-service, because I saw these researchers struggle with using the tools that already existed, because they were too complex, or required server and programming knowledge.

The whole point was to create something that they could use on the round without external help. Quite quickly, I started to see that it sold really easily, so there was a great market fit from the beginning. While I was doing my PhD, so I was doing my PhD four days a week, and then one day a week I was doing Castor. Quite quickly I saw, “Okay, this can actually work.” We were, I think five people in 2015 and we were already close to closing our biggest contract with an academic hospital in the Netherlands that wanted to provide our platform to all their researchers.

At that moment in time, I felt like, “Wow, this can this can really take off. This can be big company.” I really always invested all the money back into the business to make sure we can grow as much as we can, and actually never pay myself any salary, because I was doing my PhD as well. Use that to grow the company until about last year, when we got this European grant, so there's no investment, but a grant, 1.1 million that helped us grow even faster. Then this year, we got our series A of 6.25 million dollars. That's basically because I saw that the company is doing great, we're a market leader in the Netherlands, doing really well in Europe and in the UK for instance.

We are trying to make a difference in medical research. I didn't quit my career as a medical doctor to just become rich. The point is I want to make a difference in the world. To really have an impact on medical research, you need to also be in the US, because US is the largest producer of scientific publication, scientific output. Quite quickly it became clear, okay we are not here to just build an enormous company, or to make an enormous profit. We really want to have impact. To have impact on medical research, we need to be in the US.

I think quite quickly, I saw that's bootstrapping all the way until dominating the US market is going to take a long, long time, because the US is an expensive place to grow your business rapidly. We decided to raise additional funding to be able to do that faster. That's basically where we are today. It's been a very organic process that's mostly driven by the desire to have impact and basically estimating how long it would take a suit to bootstrap, versus raising money. In the end, that that got us here, and so that's the focus now basically bringing the technology also to the US and making sure people start capturing their data in a proper way and standardizing and sharing it as much as possible.

[0:43:17.4] JM: The go-to market for this company is interesting, because it's a fairly niche product. It's something that people definitely want, so I'm really curious about how you expand. If you're saying, we're going to go into the United States because that's the fertile ground, how do you start to get it to people? Do you start with universities? Do you start just hitting people with Google ads? What do you do?

[0:43:45.3] DA: What always used to work really well is just start with the individual researcher. In the Netherlands, we depended on word-of-mouth. The strategy does work really well for us so far, is to make sure a few people in a university start using the platform, and then they basically show their neighbors and colleagues. [Inaudible 0:44:02.6] sharing a room and they're like, “What's that? What’s that you're using, I need that too?”

That's a really powerful way to grow. The strategy is to get the initial people on the platform and that can be through colleagues, obviously, through referrals, but also through directly engaging with them at events. For us, events are a good way to make some first contacts. Then when you have your first contacts and the platform works so well, and the customer support that we provide is valued so much, that most of our users turn into evangelists. We have a very high Net Promoter Score, for instance. When our users turn into evangelists, we can also ask them to introduce it to their colleagues, or to set up a demo for instance.

That's so far been our best go-to market strategy; get that first user in and obviously, they can also be through marketing. Obviously, we're ramping on marketing now with real content that helps them. I wrote for instance, a blog post with 50 tips for people's PhD. You're running a PhD and there's all kinds of things you can do to make that whole process more efficient, so I basically created a blog post that's lists per category, all kinds of tips to make your life easier, or to –

[0:45:10.8] JM: Content marketing.

[0:45:12.4] DA: Yeah, exactly. Content marketing. That content works really well to get the PhD students to the platform. Then of course, we show them the platform and then get them to use it. That's been really successful. Start with the individual researcher, meet them at events, content marketing, maybe interact with them directly or through their colleagues, and then go from there into a department license, and from there to an institute license. That's a strategy that works really well and is very organic also, because it means you have a lots of happy users initiate already, versus top-down approach where you try to implement your solution from through said is the chief information officer while there's no buy-in from the researchers.

It can work, especially with a product that does its job and it's loved by users, but I prefer the bottom-up approach where you already have buy-in from all the people that actually have to use the product.

[0:46:02.0] JM: Tell me more about scaling. When you went to – sorry, I guess you raised 6.25 million, that was how many months ago?

[0:46:11.2] DA: Like two months ago.

[0:46:11.7] JM: Two months ago, okay. Then before that, the largest amount of funding you've gotten was 1.1 million, which was when?

[0:46:19.7] DA: A grant.

[0:46:20.1] JM: It was a grant, right.

[0:46:21.3] DA: It was May 2017.

[0:46:22.6] JM: May 2017. When you get an injection of capital, what do you do? Do you start to map out where the company is going, and then start to think about who you're going to hire, and then just add up how much you're going to have to pay them, and do a timeline over 18 months? I mean, what do you do when you raise – when you suddenly get a chunk of change?

[0:46:45.4] DA: I think you'll never get that amount of money without providing some form of plan. For both the grant and the investment, we really had to come up with a plan, like what's the timeline, what are we going to do to the product, what are we going to do sales and marketing? We actually have a very detailed plan, where they also includes capacity that defines – so for the grant, it was more product-focused. It was, okay we're going to build this module, this module, this module, this is the expected outcome, this is what I will work, this is how we're going to get it to our users, and that's this many person months and that many person months. Basically have a whole plan of capacity for where you're trying to build. This is the grant setting.

For the investment, you really have – basically have built your investor model where I say, "Okay, we're going to make these key hires. These are essential, because of our strategy." Especially in the sales or marketing department, you also try to define what the return will be basically. How is your own revenue also going to grow by adding these people? Because ultimately, we need to start generating money, instead of burning it.

It's a lot of Excel, to be honest. We've gone through and I think, 54 iterations of the model before we basically agree to what we wanted to do exactly. It's very specific, but it's really helpful also, because in the beginning I was like a cowboy just hiring who I felt we needed at the time and not really thinking I had too much. The bigger you grow, the more structure you need. I think it's really helpful to run the business now, because it's very clear what the budget is for each role and do we need first and we need next.

Of course, if we find great talents we hire them, but we at least – we have some prioritization on who we need first. It does require a lot of planning. Then it's execution. You come up with this plan, you define who we're going to hire for what reasons, what kind of revenue they're going to generate, you define when you want to hire them, and that basically defines your priorities, and then you start recruiting.

The first person you bring in is a recruiter for instance, to help you find people, want to source them. For us, scaling is mostly recruiting. It's also spending more marketing, it's also who setting up our servers in the US and then opening your US office. We just incorporate in the US for instance, but mostly it's recruiting. Finding the best possible talent to help us achieve this goal.

Of course, scaling is also setting up the processes to make those things run smoothly, so we had no process whatsoever around interviewing or how to do assessments for a developer, right? Basically half our companies, these developers, or product related roles, so what assessments do you do? When you start scaling, you want that to be processed, because you want to be able to compare candidates to each other and you want the candidate to feel that they're working with a professional company, where they get prompt replies and where there's a certain time line that we stick to basically.

That's also part of scaling, defining processes so things run efficiently, and not everything is in one person's head, but you can actually have a team as follows the same process to create a very consistent experience for candidates. I think that that's extremely important when scaling, because otherwise, everything is just going to crash and burn sooner, rather than later. Then it's funny actually to hear me say that, because I'm totally not a process guy. I'm really –

[0:49:50.3] JM: You’ve had to become one.

[0:49:51.2] DA: - that's winging it. Yeah, I've brought in people who are good at that, but I've also had to become one. I do value processes a lot more than it used to when we were just five people, then it was just me coming up with ideas and executing together with the other four.

[0:50:05.1] JM: I believe that.

[0:50:06.1] DA: Now it's a bit more process-driven. Yeah.

[0:50:07.5] JM: Yeah, I believe that. I mean, I think you’ve learned to fall in love with process, because you learn that the lack of process is flying blind and that's how you end up suddenly out of money, or with some outage and no idea how to deal with it, or some huge gap in features or miscommunication.

[0:50:27.1] DA: Then happy people, mostly also. I'm surprised at how quickly people reference and burnout if there's no communication, no processes, nothing. It just becomes super stressful. With five people, communication is easy. You have 15 people – is an exponential function, it suddenly stops working, right? Depends on the group of people, but at some point communication starts failing without some process, or thought behind that.

[0:50:53.2] JM: Yeah, I know we're almost out of time. I want to get your picture on the future a little bit, because you've got a front row seat to a lot of change that's happening in medical research and in technology. From the shows that I have done, there are some changes going on in medicine that seem exciting. Just the basic application of software tools and changes in hardware to the world of medicine, I mean, you've got the drop-in cost of device creation. It's becoming much easier to rapidly prototype hardware. You've got widespread access to supercomputers basically, so that people can do interesting data analysis.

I don't know how far along things like, protein modeling for coming up with drug ideas, I don't know how far along that is, but I'd love to get any thoughts you have on the future of medical technology, whether you're talking about device creation, or the changes in the pharmaceutical industry. What is exciting to you, or is nothing changing? Is it still gated by something like making CRISPR viable?

[0:51:56.9] DA: Let’s see. That was a very broad question. I think there's so many – it's almost unfair if you see the innovation medicine, versus medical research. That's crazy. Like how little innovation there is in the medical research process and how we conduct medical research, as opposed to the field of medicine, where there's so many exciting things going on.

There's nothing specific that I feel is here jumps out at me, to be honest. I'm just happy to see that there's so many people in the world doing amazing things to make sure people live healthily for longer. I don’t want to necessarily say longer, because I think we are quite old in the Western world at least already, but it is very nice to live healthy life. For me, of course what I'm interested in most is the thing that you mentioned before. Capturing data through sensors and capturing data without a specific purpose and standardizing it and giving people the option to contribute to studies.

My dream is to come up with a Reddit April Fool's project, so I'm not sure if you're familiar with those, but Reddit always does something cool on April Fool's, where they involve the entire community. What I would really love to do at some point is come up with this experiment and where we involve all the Reddit users into measuring something and contributing the data into one huge data set that's completely standardized, and suddenly answers one pressing question that we've always tried to answer.

Gather data from five million people in 24 hours. I mean, something like that would be amazing for me. I think slowly, we're getting there with the smart phones, with more sensors, with like you said, more affordable devices that you can for instance connect to your smart phone. For me personally, with my background in data, that's the most exciting for me to see that trend, where people can capture all kinds of advanced data in their homes and then decide to contribute it to some cause to solve one of humanity's biggest problems. That's really where I also want to – I want to be part of that when that starts happening.

[0:53:54.5] JM: Okay. Well Derek Arts, thanks for coming on Software Engineering Daily. It's been great talking to you.

[0:53:58.5] DA: Well, thank you very much. It was a lot of fun.

[END OF INTERVIEW]

[0:54:03.5] JM: GoCD is a continuous delivery tool created by ThoughtWorks. It's open source and free to use and GoCD has all the features you need for continuous delivery. Model your deployment pipelines without installing any plugins. Use the value stream map to visualize your end-to-end workflow. If you use Kubernetes, GoCD is a natural fit to add continuous delivery to your project.

With GoCD running on Kubernetes, you define your build workflow and let GoCD provision and scale your infrastructure on the fly. GoCD agents use Kubernetes to scale as needed. Check out gocd.org/sedaily and learn about how you can get started. GoCD was built with the learnings of the ThoughtWorks engineering team, who have talked about building the product in previous episodes of Software Engineering Daily, and it's great to see the continued progress on GoCD with the new Kubernetes integrations.

You can check it out for yourself at gocd.org/sedaily. Thank you so much to ThoughtWorks for being a long-time sponsor of Software Engineering Daily. We’re proud to have ThoughtWorks and GoCD as sponsors of the show.

[END]