

Artificial intelligence is possibly the most important thing that's happening in technology right now. And 80%, 80% of the problem is associated with data. And right now, synthetic data is essentially the only solution that's giving people control over that data set. So like, when we talk about how big could this opportunity space be, it's as big as AI and not just where AI is, but where it can go.

Chad Anderson:

Welcome to the Space Capital Podcast. I'm your host, Chad Anderson, founder and managing partner at Space Capital, a seed stage venture capital firm investing in the space economy. We're actively investing out of our third fund with 100 million under management. You can find us on social media @spacecapital. In this podcast, we explore what's happening at the cutting edge of the entrepreneurial space age, and speak to the founders and innovators at the forefront.

Chad Anderson:

This is the Space Capital Podcast. And today we're speaking with Nathan Kundtz, founder and CEO of Rendered. AI. It's a company developing a common application framework to enable production of physics based synthetic data sets for AI training and validation. We led the company's seed round earlier this year, investing alongside a great group of investors. And we are incredibly excited about what Rendered's building. You may know Nathan from his previous role as CEO of Kymeta, a Bill Gates backed meta materials company that's building innovative antennas for next generation satellite communications. Nathan also has a PhD in physics from Duke and all of this makes him uniquely qualified to help us understand synthetic data, what it is and why its generation is so central to artificial intelligence. Nathan, thanks for joining us. It's great to have you on.

Nathan Kundtz:

Thanks very much. It's a pleasure to be here and thanks for the wonderful introduction.

Chad Anderson:

Yeah. So look, synthetic data is a relatively new technology and I imagine that there's many in our audience that aren't familiar with it. So can you help us start at the basics and understand what it is and why we need it?

Nathan Kundtz:

Yeah, absolutely. So when we think about synthetic data, it's actually useful to just start with artificial intelligence. The community has a broader awareness, but the thing to really understand about artificial intelligence is it's essentially just software, right? It's also software. The difference is it's like software that's written in data instead of written in code. And so all of the behavior of artificial intelligence algorithms is ultimately driven by the data that is used to train it. And because of that, the time spent and the expense of building those algorithms is about 80% dedicated to getting access to data sets and that's 80% today. And so there are still challenges that many of the companies face when building those algorithms and they can't just sort of spend more time on it. It's already the largest expense. And there's several reasons why. One is that the data sets themselves have to be large for effective training.

Nathan Kundtz:

Recently, the government came out saying they need 50 million images for every object they're trying to identify to build accurate, just 60% accurate detection scenarios. And then that becomes very expensive

when you start thinking about the data collection and annotation, having humans come in and actually try to tell a computer what's in there. You still miss a lot of rare events and edge cases, which are incredibly important when it comes to algorithm performance and you're also completely at a loss if you actually wanted to build algorithms to understand how next generation sensors, things that don't exist yet could be built.

Nathan Kundtz:

And so that's where synthetic data comes in. So what we do in order to avoid all of those challenges is we actually focus on essentially using our understanding of simulation, our understanding of physics to build those data sets from scratch. So instead of sort of taking pictures, handing them to a human and saying, what's in this picture and having them draw boxes in order to annotate them, we tell a simulation what's in the environment, what should be in the picture. And then we ask the computer, what should the picture look like? So we ask that inverse problem. And in doing that, we get kind of complete control over those data sets and it opens up completely new workflows and things that you can do with AI.

Chad Anderson:

So look, the artificial intelligence machine learning is sort of permeating its way out into every conceivable business and workflow, right? Countless applications are being developed. It's hey, what are the big use cases? Autonomous vehicles come to mind, but many, many others. And so machine learning algorithms are trained with an incredible amount of data, and sometimes it's difficult to obtain this data for different reasons. So you mentioned a couple of opportunities, rare events in edge cases, cold start problems for new sensors and things, but there's also opportunities in restricted data, right?

Nathan Kundtz:

Yeah, absolutely. So that's another area that we see data restrictions preventing people from getting access to data sets they might otherwise need. And frankly, hey, this is the space podcast. It turns out geospatial data of various types can often be restricted for security reasons. And that can be really limiting in terms of being able to identify some of the content that you'd like to. We just recently did some work with Orbital Insight, where we were demonstrating through leveraging synthetic data, we can improve performance of their algorithms by two to three X over real data alone. And to have tried to do the same thing, simply through a data collection campaign wouldn't have just been incredibly expensive, but it would've led to all sorts of questions around is that data something that is allowed to be on public service.

Chad Anderson:

So there's also data labeling problems. Can you talk us through that?

Nathan Kundtz:

Yeah, sure. So data labeling problems come in two varieties. The first and the ones that we're maybe most familiar with are the ones where a human being sort of doesn't quite get it right. Or maybe just doesn't get it as consistent. And you find this, and you mentioned autonomous vehicles, you find this in autonomous vehicles a lot.

Nathan Kundtz:

So if you ask somebody for instance, to draw a box around a car, that seems like a pretty straightforward thing. Everybody should be able to draw a box around a car and get consistent results. Only it turns out that when you do that, some people include the tires and some people don't, some people include shadows and some people don't. And so you end up with these inefficiencies and then you get like, well, you said car, did you mean truck too? And so you inevitably end up with some inconsistencies just by virtue of the fact that you're using humans and they often will actually just get things wrong.

Nathan Kundtz:

As you start to move into imagery that's a little harder to discern and if we talk again about GIS data or EO imagery, satellite imagery, then it can be tough to tell. And so people are using a judgment or maybe using their knowledge of what's happening around an object in order to try to label it and they get it wrong even more. And so all of that turns into noise in the algorithm training and is one of the leading causes of errors in AI. But that's all assuming that a human could look at the output and tell what's in there from the beginning. And it turns out as you get into other types of sensors, if you think about x-ray and radar sensors, things that are really, really important for industrial applications and increasingly for consumer applications, a human can't look at that data stream and tell what's happening at all. You really do have to know what was there through some other means. That's a incredibly limiting factor when it comes to building AI algorithms with those sort of novel sensor

Nathan Kundtz:

Types as well.

Chad Anderson:

Interesting. And so we need more tools for data generation. We've just walked through four different areas in which synthetic data can be used. What you are doing, you are generating physics based synthetic data sets. Is that right? What does that mean?

Nathan Kundtz:

That's right. So, maybe it's helpful to contextualize that with the alternative. One of the things people do when they lack data is they just try to create more data like the data that they already have, so they say there are applications and techniques principally based on what's called a Generative Adversarial Networks or GANs that can create more of data that essentially you already have. And that can be useful in some circumstances. The most well known is when you have maybe privacy concerns, so you can't actually use the exact data, although there's problems with GANs not actually being effective in blocking privacy issues, but nonetheless, that's an example of a circumstance where recreating a data set and maybe anonymising it can be helpful. But the reality is it doesn't generate any new information, you haven't introduced new knowledge to your data set.

Nathan Kundtz:

So, what we do instead is we really focus on physics based simulation. We're essentially using this knowledge of physics, this knowledge of how sensors actually extracted information to generate data sets and we use all physics simulators. So, computer graphics it turns out is a type of physics simulator, it's a type of light simulation, but our platform is really built with just a variety of different types that can capture different effects, everything from optical to radar to X-ray and others.

Chad Anderson:

All right. So, we've talked a bit about synthetic data, so we need to generate more data, synthetic data is helping out there. We also need more tools to carry out the engineering of that data and so is that where Rendered comes in? What are you building?

Nathan Kundtz:

It turns out it's easy to say the words hey, we're going to simulate what the sensor does. Turns out really easy to say much harder to do. And if you step back and you think okay, well, what do I have to do? What are all the things that I really need to do to be effective at that? So, I need to have a physics engine that captures whatever's happening in the scene, I need a digital twin of the sensor that I care about, I need content, so usually that comes in the form of 3D models, procedurally generated content, I need something for world building. I don't need to build one world that turns out I need to build maybe tens or hundreds of thousands of worlds because I don't want a image, I want many of them. And so I need to have a procedural tool for creating that diversity.

Nathan Kundtz:

All of that takes a massive amount of compute, so you need to have some form of compute orchestration to manage that. Then you need to annotate it in order to fill it into the pipeline. Now, I don't want to just give you the laundry list, the point is to do it all well, the full stack is very challenging for any one company to be able to manage. And so what we've done with our platform is we take as much of that that is horizontal as possible and build it into the platforms that you can plug in. So, you're not having to worry about computer orchestration, you don't have to worry about how you're going to define that diversity of scenes or annotations or AI pipelines and then we start to build content on it and the power comes in both ways.

Nathan Kundtz:

So, you get that existing platform, but then you get standardization around how the sensor can be represented, how the 3D models get dropped into these scenes and by virtue of that standardization, we see sharing and the ability to really pull data, pull content from all of these different sources. And when all of that starts to come together, now you can take what otherwise would be maybe a two year plus investment from a company and turn it into a couple weeks to go from existing content to something that's really solving a problem for them.

Chad Anderson:

How is it being done before Rendered?

Nathan Kundtz:

So, a lot of... First of all, super new industry. So, synthetic data there's not that much to point to but I'd say there's two paradigms other than Rendereds that we see. The first is let's call it the Tesla paradigm, which is a company goes full stack and builds a solution specific to what they need, their sensors, their environments. And if you are Tesla you can do that, they've applied billions of dollars into those simulations. And so that makes sense. We have not been trying to displace those places where somebody has already put a billion dollars of sum cost into their synthetic data capabilities. The other is and what we see more of in the startup ecosystem are small companies that will essentially build data sets as a professional service. So, they have maybe some 3D models, some procedural content and they'll pipeline that together for you if you ask them for a particular problem, but there's three issues with that.

Nathan Kundtz:

One is it typically is a one off data set. And so there's not if you're trying to detect stop signs and then later you need to detect stop signs with IV on it, it's not like it's in there or extensible there's nothing to really be done with it. The second is for most customers they need to be able to add those features and understand those features as they understand their business problems over time. So, different customers have different needs and ultimately this is a really important innovation tool for companies and so I think it's hard to work with a third party. The third and most important is that you just don't get that standardization across the industry.

Nathan Kundtz:

So, we were just talking about how valuable it is to have digital twins with a lot of different sensors all following the same standard and have 3D content all in the same environment so that people can pull from it, that standardization doesn't happen when it's done in one-off professional services. And so that's where a lot of the power of bring platform together comes from.

Chad Anderson:

Yeah, difficult to scale a model like that it's really a bespoke consulting service, which happens a lot in early sectors like this where the customers need a lot of education and handholding in terms of this is a very new technology and they might not be aware of it or how to plug it in.

Nathan Kundtz:

That's right and how should I say, there's nothing wrong with consulting services, consulting services are important, but they're different businesses and I think if you were looking for a parallel, there's good ones to point to between what we're doing and maybe what those companies are doing and the one that I think is cleanest is Salesforce. There are hundreds maybe thousands of companies that do specific Salesforce implementations or will tailor Salesforce to your particular business needs, but Salesforce provides that platform which allows for that community and that ecosystem to exist and be efficient. And so that's the role that Rendered is really focused on playing.

Chad Anderson:

That's a beautiful thing. So, there's clearly a lot of money to be made here, this is a new technology. We're tracking 18 competitors in synthetic data, 13 of those were founded in the last four years. They've collectively raised half a billion dollars and are collectively valued at five and a half billion dollars. One of those companies is publicly traded. So, definitely a lot of investor interest, a lot of momentum and very quickly over the last few years, I'm curious, what is driving that and how do you think about the market potential for what you do? Is it the entire AI market?

Nathan Kundtz:

Well, Chad I'm going to be a little

Nathan Kundtz:

Bold here, because I think it's important to be clear. Artificial intelligence is possibly the most important thing that's happening in technology right now. And 80% of the problem is associated with data. And right now synthetic data is essentially the only solution that's giving people control over that data set.

So, when we talk about how big could this opportunity space be? It's as big as AI and not just where AI is but where it can go. This is how we get control over our AI algorithms is by getting control over our data.

Chad Anderson:

So, drilling into the market a little bit, who is going to benefit from this then? What industries to help make it real for people who are listening? What industries are early adopters here and are starting to use synthetic data sets?

Nathan Kundtz:

Yeah, absolutely. Well, you mentioned autonomous vehicles and then that was where a lot of synthetic data generation really started. It's not been a focus for us for some of the reasons that I already highlighted. So, the other areas where we see a lot of adoption, I'd say GIS and earth observation broadly. So, whether that's for insure tech or government or crop management, artificial intelligence is really crucial. You simply can't image the earth multiple times a day and pass that in front of humans to get an understanding of what's going on. So, that needs to be done in an automated way.

Chad Anderson:

Just too much data?

Nathan Kundtz:

Way too much data. And then even as I say that I think what we picture is visual information, but when you start to say okay, but we're going to have hyperspectral and we're going to have SAR and it's way too much data to be able to process in a manual way. And so all of those pipelines which are feeding everything from insurance innovation to future crop management, to climate change detection, methane leaks et cetera, all of that needs to be automated in order to be effective. We see the same thing in robotics almost by definition, robotics have some level of autonomy and in those cases, you often have a more controlled environment, but the more that the robotics solutions bleed into uncontrolled environments, the more that synthetic data becomes really crucial for detecting exactly what we talked about edge cases in different scenarios. And again, it's not just RGB cameras on that, so there's a wide variety of sensors that are used in that environment.

Nathan Kundtz:

We see strong demand from things like security imaging. There's a huge close circuit, television market out there, same problem, vast volumes of data being generated, but you don't actually want to be observing everything, you just want to be alerted when there's actually a problem. And so we've seen that pop up over and over again. The area that I think is up and coming for synthetic data and will become really exciting is medical, because if you think about an industry where the diversity of imagery is so broad and the impact of ineffective or incomplete data when it comes to things like bias or poor decisions is going to be so painfully felt, I think the healthcare industry is right at the top of the list. It's a little slower moving than some of the earlier adopting industries we just mentioned but I think there's big opportunity there long term.

Chad Anderson:

Fascinating. And so some of these are near medium term and longer term things. We talked a little bit about the need for customers or potential customers are starting to understand how they can benefit from synthetic data sets. But the geospatial intelligence market, the satellite market is already eager for

the capabilities that Rendered offers, this seems like a really interesting beachhead market for you and one that you know really well?

Nathan Kundtz:

Absolutely. So, two things, one the market is needing it desperately for all the reasons that we just highlighted. That was a big reason why it's been a focus for our company and why we were so excited to have you guys lead our seed round. Like a lot of other technologies, does the Space is leading the way. We're really learning a lot by using synthetic data in a space environment that we can then bring back to earth. My own experience with Space and how it led to Rendered is that I was working with Kymeta. I left there in 2018 but kept seeing companies that had really clever sensor design plans and 201 they would say now that we've got that sensor designed, we're going to put it on a bunch of satellites, we're going to collect a bunch of imagery and then we're going to turn that into automated insight generation, we're going to sell the insights. Chad, I'm sure you've seen at least 1000 pitches like that.

Chad Anderson:

Absolutely.

Nathan Kundtz:

And the question would always be okay, well, how do you... And every entrepreneur needs to know how to do this? How do you sell that before you've actually spent whatever it is, 1,500 million on satellite constellation. And what you have to be able to do is demonstrate that you can generate those insight, show exactly how that can be done, use what is presumably simulated imagery to get your customer base going with a complete solution before you actually start getting that raw data back to earth and there was no solution to help them do that. So, the actual genesis of the company was me writing a white paper on how we needed a platform to help people solve that problem, really the systems engineering problem at the end of the day. And that led to a broader understanding of what was going on from an educate standpoint and to really seeing that this is a broader issue than satellite imagery. It was something that was affecting many different fields.

Chad Anderson:

Who are your customers? Who exactly are you selling it to? Is it data scientists? Is the engineers at a company? When you're approaching a customer, it's one thing to get the CEO of the company to say or the product person to say, I can see the benefit of synthetic data but who's actually buying this?

Nathan Kundtz:

Absolutely. So, for the most part, it's computer vision teams that are the buyers at the end of the day, it's the technical buyers, typically the head of a computer vision team. The important thing to reflect on though, the important thing to understand in making that sale is that you do need to think about not just the computer vision user, there's another important user here which is the synthetic data engineer, which is the person who's going to take that existing content, but then extend it and continue to push it forward either within a customer organization or on behalf of a customer organization. And so we've built out a tool set that addresses both sides of that equation.

Chad Anderson:

I'm interested to get your take on something that was reported on probably a few months back and training data for video surveillance and deep fakes. So, deep fake satellite imagery poses a really not so

distant threat. When we think of deep fakes we think of people and Tom Cruise videos and things, but we don't often imagine deep fake geography, but people have been lying with maps for a long time.

Nathan Kundtz:

Yep.

Chad Anderson:

Fake satellite imagery could be used to misdirect. I'm just curious your take on this and how synthetic data can help?

Nathan Kundtz:

So, I think you want to think about our role here as a little bit like white hackers, because at some level the synthetic data is its own kind of deep fake. We're creating fake images of the earth and principle that tech could be used for ill, but what actually happens when you start to get your arms around it and as we provide the tools to the good guys is they can then use those

Nathan Kundtz:

... those tools to really better differentiate between the two types of imagery. And so we integrate actually a variety of different post-processing steps that are common to deepfakes into our platform in order to reach hyperrealism with the output. But what that also does is it empowers our customers to sort of be able to do deeper levels of comparison. Because now they have an AB, right? They can look at maybe original synthetic imagery, and deepfake imagery, and start to find the connections between them. We think we have a role to play there, and can help people better identify where those gaps are.

Chad Anderson:

Great. So you are ... Like I've been mentioned at the beginning of this thing when I was introducing you, very unique background. Seemingly purpose-built to lead a company like this. But you've also built an impressive team around you of some really fantastic people. You've recently added a COO and a head of product. I'm wondering if you'd like to talk a little bit about them and how they're helping.

Nathan Kundtz:

It's so important to have a phenomenal team. And we've been really, really lucky at Rendered. Yeah. Chris Andrews just joined us. [Esri 00:25:09], he was leading a lot of their ... well, essentially all of their 3D content, and 3D team, and heading up some of their most important, most innovative products there at Esri. And so he's been a tremendous, tremendous add. We actually continued to pull a little bit on that vein. We've recently added one of the engineers who did a lot of the original AI work at Esri to our team, so we can help our customers learn about how to integrate AI pipelines into their GIS use. Those have both been phenomenal adds. We've got some additional ones that are coming on pretty soon, but it's been great to be able to pull together a team of real world leaders to get this job done.

Chad Anderson:

Great to hear it. I mean, the team is so important for any organization. But particularly at the early stage, like Rendered is. But in addition to the team, you've also assembled a world class customer advisory board. Can you tell us about them? You know, how it came about and how they support the company?

Nathan Kundtz:

Yeah, absolutely. Both Jimmy Crawford, who we talked about briefly earlier, and John Wolfe, the founder of Twilio. And the guy who really headed their product through their growth have joined us in the advisory board. It's been really helpful to hear from them. Really, Jimmy reflecting on the customer set that we have been focused on in this call and much of what we're doing right now, which is in this geospatial analysis market.

Nathan Kundtz:

But then Twilio, I think, is such an interesting story for one of these really powerful developer tools. For those in the audience maybe who aren't familiar, Twilio essentially powers all of the automated sort of text messages that you get. If you're doing things for maybe the two-factor authentication, that's powered by Twilio. And they built integrations with a lot of the cell phone industry, and made those tools available to developers. And so we've been learning a lot from John about kind of what it looks like to build those tools and make them effective in the hands of developers.

Chad Anderson:

Great folks to be involved. What is next for Rendered? You're rolling out a new version of the platform?

Nathan Kundtz:

Yeah. To date, we've sort of been keeping the platform to ourselves. We've been building applications on it. We've been supporting our customers with the platform. As we end the year here, we're actually going to start opening it up to more people to build their own applications on top of it. We're excited about this transition. We've made a huge number of infrastructure advancements to take what was otherwise really just in an in-house workflow to manage those applications and make it something that can be public.

Nathan Kundtz:

And so this coming year, and especially in the first six months of next year, we're not just going to be opening that up. But then we're going to be open-sourcing actually a number of different applications to teach people how to use synthetic data. My impression is there are a lot more people out there that need synthetic data than necessarily have the programming skills to generate it. We can provide the platform, but we also want to provide them with a lot of content to get started. And so those things are going to start making their way out into the ecosystem as we get into the new year.

Chad Anderson:

Awesome. Is synthetic data going to replace real data, then? How much real data do we need to train our algorithms?

Nathan Kundtz:

Yeah. So, there's no physical law that says you have to have a certain amount of real data. I would say right now, there's certainly a lot of benefit to using some real data. And different people come to different conclusions about the percentage. I mean, we've found that sort of even just 10% real data can be really helpful in some of the work that we've done. But I think there's a lot of economic and technical pressure to get to more and more synthetic data. For obvious reasons, right? Much lower cost.

Nathan Kundtz:

And then you just have control over it. You know what's in it, which you don't actually do when it comes to real data. I think we're going to see those natural market forces and technical forces push us to increasing levels of synthetic data use. Gartner came out with a report recently saying that in about three years, 60% of the data we use for AI would be synthetic. And that by the end of the decade ... So, by the time we hit 2030, essentially 100% would be synthetic. So, we'll see. We'll see. There's nothing preventing that from becoming true. And I think Rendered is providing some tools which will help accelerate that adoption.

Chad Anderson:

How can listeners learn more about Rendered?

Nathan Kundtz:

Sure. You can go to our website, of course. Rendered.ai. We're becoming increasingly active on social media. We've been kind of quiet for a while, and so we're going to be sharing quite a bit more as we go forward. And of course, you're welcome to email us if you want to get in touch. So, info at rendered.ai, or you can go to our website and just fill out a simple form there, and we'd be happy to engage and let you know more.

Chad Anderson:

Yeah. It's been great to have you on. Thanks for coming on the show.

Nathan Kundtz:

My pleasure.

Chad Anderson:

Thanks for tuning into the Space Capital Podcast. If you enjoyed this episode, please leave us a review and subscribe to make sure you never miss an episode. And if you're interested in learning more about investing in the space economy, I invite you to visit our website, spacecapital.com, where you can get access to more industry-leading insights and learn how you can join the entrepreneurial space age.