FEDERAL TECH PODCAST

Ep. 60 Solving a Billion Dollar Paper Problem

00:04
Welcome to the federal tech podcast where industry leaders share insights on innovation for the focus on reducing cost and improving security for federal technology. If you liked the federal tech podcast, please support us by giving us a rating and review on Apple podcast. Welcome to the federal

00:29
tech podcast. My name is John Gilroy and I'll be your moderator. Today we have Chris haar Sales Director, US public sector for a company called hyper science, h y p er science hypersonics.com. I met Chris a couple months back and talked about his technology. And I think it has very specific application for the Federal audience. I thought we'd get him on here and he could explain and at least give maybe our listeners a better idea of what hyper sciences and and maybe bring up an old topic. And the old topic when you bring up is called Oh, see our optical character recognition. Now, an old guy like me, I remember this guy named Ray Kurzweil years and years ago talking about this. Now for a young guy like Chris Haar, he has a whole different idea of OCR. So I thought we need to get the old guy and young guy in the studio and talk this out. And so you can see the Federalist and some money or other Chris K, tell us about your background, maybe a little thumbnail sketch of hyper science, and we'll jump into OCR.

01:22
Thanks for inviting me on the podcast since you launched it. I've been a longtime listener, we've enjoyed a lot of the episodes that you've released. Yeah, so I've been at hyper science for a little over 18 months, but the company has been around since 2014. We're headquartered in New York City. And we're actually a series D backed startup. And we've got just about 300 employees, helping us deliver this technology called intelligent document processing, to the global markets that we serve. I am on the US Federal team serving as a director here, you know, I think we've got a great opportunity in front of us to help the government really eliminate the time tax associated with paper based processes.

02:07
You know, I assume that listeners know certain names, and there may be something listen to this, like, oh, Kurzweil must be second baseman, you know, for the Phillies or something. Now, maybe you could tell us who who the genius of Rei and what he came up with and, and the legacy today what how it impacts the way we do business today.

02:24
Yeah, so Ray Kurzweil kind of considered the godfather of OCR. Well, yeah, optical character recognition, right. And that technology has been around for, you know, three or four decades now at this point. And what that technology did very well was this process of process of digitization, right? Where people would take documents, forms, things that could be evidence for a claim, right, and they would scan that document, and they'd run it through OCR, right. And that optical character recognition, you know, would help extract some key

pieces of metadata out so that you could maybe archive the document or make it more easy to retrieve. But as this process of digitization ballooned, people started to realize, right, there's, there's a little challenge here, this this pesky thing called scale, right? So if you need to process, you know, 1000s of documents a day, right, you need a different tool to solve that problem. We

03:27

know when we're talking about OCR, and I'm taking notes into my little brains kind of going all three brain cells. I'm thinking, Well, wait a minute, here's March, April 2023, what's going on? Well, there's something called taxes coming up. And everyone listening probably has a folder somewhere in their desk with their taxes 2022 on it, and they're trying to scurry around and getting ready. And guess what, it's a big paper base folder. I mean, I just, that's why I bring this up at the beginning of the interview. Because I mean, this whole idea of paper is not some old fashioned topic from 1962. This is something that is important. And I think when you're looking at any federal agency, they have to consider what ways can they be more efficient? What can these reduce costs? I would think that this all this paper fault runs gonna be one way to reduce costs, isn't it?

04:11

Absolutely. You know, going back to the Ray Kurzweil OCR topic, right? When you talk about scale, right? Taxes, tax season, you're not talking about 10s or hundreds of documents a day, right? We're talking about millions, okay, and so scale matters. So when you've got millions of documents flowing into your enterprise every single day, right, the concept of one person opening up every single document, and doing a keyword search over that document becomes incredibly difficult, trying to find out right, what customer this is for, how much does the person owe? How much are they claiming that becomes really difficult. So this concept of intelligent document processing is taking OCR to the next level to solve using natural language processing and machine learning to solve the context contained in all these data documents, especially with tax returns, to help automate and accelerate, you know, whatever business process this journey is, and I gotta tell you, I hope I get my tax return, or the adjudication of my return done more quickly this year than in years past. And from what I've seen out there, it looks like that might happen.

05:18

I have some friends coming in town from Alaska. We're gonna go downtown and maybe see some museums and see what's going on downtown. And I remember once I was downtown, and I had an appointment to talk to the people at the US Chamber of Commerce. It's it's kind of like a museum. The building is like a museum. When you walk by it's like, wow. And the inside the foyer is kind of real, like, like the Smithsonian or something's going on. And I just want to do a contrast here, because they have something new, they come up with a study and they said, let's take a look at this paper stuff. Maybe it's costing you, you taxpayer. Chris, are you taxpayer, John Gilroy, more than you think you want to tell us about this study.

05:53

So when the chamber released the study, our team, we got a kick out of it, because first off, they conducted a fantastic report. But what the report highlights is exactly what our team has been focused on. And it's at the end of the day, right making work more human, by eliminating the time tax associated with document and data

processing, really through human centered automation, right. So in their report, they talk about how eliminating paper based processes could save taxpayers approximately $40 billion annually. I read that

06:26
I said, What What are you talking about with a B?

06:30
I know it's mesmerizing, right? And they've done a tremendous job in the study, they've got all these artifacts and example use cases. And as we read through them, right, we recognize some of the use cases. And I gotta tell you, it just makes me think about the market for intelligent document processing, right, and machine learning and natural language processing to help tackle this process of eliminating paper based processes. And, in fact, our top four agency clients, they're processing nearly a combined 800 million pages annually today through hyper science.

07:03
So 85 on Jeopardy, and someone asked me that question, no way, I guess that you would you 800 million? Not only that,

07:12
you know? Yeah. You know, when I was interviewing at hyper science, some of the anecdotes that were shared with me were were impressive, right, it caught my attention. And, you know, I think we're at an interesting time, right? Where new technology brought on by natural language processing, and machine learning, thanks, in large part to advancements and other technologies are helping solve these legacy business challenges, right. So I think this staggering number of $40 billion annually is pretty accurate, there might actually be more savings out there for the government and the taxpayer, because I'm telling you from our front row seat, and really our position in the market, right, we've witnessed firsthand how agencies are addressing the topics in that study, as well as John, they're addressing how to improve citizen experience, such as the Biden executive order on citizen experience, right, which calls out, but paper based processes take nearly 9 billion hours of agency time annually, we're seeing technology like ours being applied to solve these two challenges.

08:13
You just use the phrase, natural language processing. And I've been around here there a couple times. And I remember this. This was a phrase that was popular, like 1520 years ago, and it seems to come in and out of style. So So what do you mean by natural language processing, especially with your technology? Well,

08:31
first off, you know, when you and I were talking a couple of months back, I looked into it's kind of the history of natural language processing. Google Trends. Yeah, yes. Well, I did look at Google Trends. And I suggest to your listeners, you know, do this right. Look at Google Trends for natural language processing. And you'll see the topic peak in February of 2004. I remember I was there, yes. All right. And for 12 years, it essentially went

into this dormant status and interests really didn't come back until about 2017. And guess what, that's when hyper science really started taking off. That's when government agencies started contracting with hyper science to help them solve their challenges. And so our team anticipated this interest shipped with natural language processing, right. And so what we've done or what this means to us, is we've delivered custom models to our clients over the past couple of years. And you know, even just recently, right, we've released a new software update, which brings now all of our clients text based document processing in a single unified platform, right. So our clients and partners can now unlock valuable insights that were previously trapped in these unstructured documents, thanks to advanced text classification, on structured extraction, and also data labeling technologies, right. And that's all powered by this thing called natural language processing.

09:57

I'm gonna drop parallel and get real big trouble here. So the early days of databases, they were relational. And what's happened, they've kind of changed over the years. And frequently today's modern databases, and maybe more object oriented, you know, and maybe this is what's going on. And in some ways, though, CR, maybe in the early days, it was just textual. But now, now you have a point, we could actually pull out information that is classified, we'll see go from unstructured to structured and so now you can take unstructured information and scan it in like you couldn't before. Is this a parallel here is I'm I'm off base. So one area

10:30

of opportunity, I think, for this type of technology is around data analytics, right? And even predictive analytics, right? So when you talk about databases, that's where my my mind goes, if you look at my background, I sold data center infrastructure, I sold big data analytics solutions. And here at hyper science, the data that we process sits in that data center. And a lot of times customers that we're talking to, especially in the federal civilian marketplace, they're looking at, okay, how can I take all of this previously digitized data? That's now going through this idea of digitalization, right? How can I take all of the data, all that unstructured data and put it to use? Right, how can I put it into usable schema? Yeah, I think that's one of the that's one of the magic sauces, if you would from hyper science, right? Because the data is trapped in those old records, right? It's only usable if you know exactly what file to look for, if you're using old OCR technology, right. And what we've done is we've now democratize machine learning, using natural language processing, to find that needle in the haystack to put it into a usable schema, right, so that you can extract all of the key critical data out of your records that can be used for some broad agency use, right? Maybe it's mining data to develop additional machine learning models that can maybe predict, you know, an FHA or VA mortgage, or maybe it's, you know, to feed conditions based maintenance, predictive analytic models as well, for major airline carriers or even the DOD.

12:07

Yeah, I think at your website, I read this phrase this morning, unstructured extraction. That's what hit me with unstructured data, unstructured access. That's, that's the tough part. I mean, you know, if you have a name, address, phone number, street address, that's one thing. But some things are much more subtle than that.

12:23

Yeah, exactly. And, you know, our team sees this as a major area of opportunity. And I'll tell you what, there are some federal agencies that see this as an opportunity as well, you've got the agencies that obviously focus on protecting our borders and helping with immigration, for example, right. And some of those agencies, they've been in front of the chamber study, they've been in front of the Biden executive order on citizen experience, where they've actually outlined in their strategic planning, and even with recent RFIs, how they're going to harness the power of machine learning and natural language processing, specifically through intelligent document processing. And so agencies that are looking to not only reduce their business backlogs, right, whether it be FOIA or even, you know, citizenship, or even tax processing, right, they should look at intelligent document processing. And a lot of them are, right, because they recognize their unstructured data can be accurately extracted, and in most cases, right, automated very quickly,

13:25
what's uh, you kind of stumbled on an executive order and talked about mentioned this US cybersecurity strategy policies coming up. And so because it's in the news, and if you're at an event and someone walk up to you and say, Well, what does your scanning have to do with cyberspace strategy? Maybe you can bring up two or three points here and how maybe something was supply chain or something? Or how does it How does it tie in?

13:46
So I appreciate that you call it out supply chain. Right. So I think it's the fifth pillar of this new cybersecurity policy actually discusses secure supply chains and John, secure supply chains have been under the microscope for years. I think the first time I remember it really taking center center stage was probably about about exactly 10 years ago, you saw organizations like Parasoft, DLT, even image group now arrow, right, looking at secure supply chains and agencies, they've been focused on how to ensure the goods that they're purchasing do not have malicious devices, or code embedded in those systems. And by the way, part of that process comes with significant documentation. But now I see something else happening and some additional legislation coming down the road, John, and that is this idea of secure supply chain from a cybersecurity perspective, but also from a from a human rights perspective. And I'll tell you, you know, hypersonics just announced our brand new AI ethics steering committee, of which I'm the Public Sector representative on and so you know, we're taking this topic seriously right, how can our machine learning and AI technology helps support Are these types of topics. And I think the US government is doing a really good job at putting forth this legislation to give some direction to agencies so that they can protect our critical infrastructure and, you know, secure our databases. But there's, there's a really important piece here. And that is, again, they're largely document driven. And again, natural language processing can help effectively do that supply tracing.

15:25
And I think what people are more and more discussions I have with people about artificial intelligence and machine learning, it usually goes back to well, where are you getting your information from? And, and really, is this as their bias their, or, or their conclusion is going to be fair, and I think this is a critique that many people have this Chechi Beatty is like, well, this is some ethical problems here. And I never thought that say truth, I never connected the dots between, you know, that massive amount of information you have to pour into the machine to get machine learning out of it. And it's not all O's, and ones, maybe it starts off as paper

documents, you know, maybe there's some sensor information that was driving a laboratory on paper and has to be scanned in. And that can bring a bias to artificial intelligence as well. So maybe that's where the ethical part fits in. Is that

16:15

right? Yeah, you know, I love this conversation. Right? So you're talking about chat? GPT, right, you go to the website, they're taking AI ethics Seriously, just like we are. And when you think about, Okay, how are we going to use tools to solve our business processes? You know, one thing I think your listeners can consider about hyper science is that accuracy matters for the business processes that you're undertaking. And I think having the human and the machine teamed together to ensure accuracy of the data that's extracted that ultimately feeds your downstream business processes, matters. And so the example that you use about some document in the lab going into a process that could affect downstream deliverables, right, the one thing I'm always going back to here is with human and machine teaming, you need to have a traceable, auditable record of who touched the data. And we built that into our platform, right, you can see if the machine extracted the data, or if a human altered the data in any way. And so traceable data is really key as we think about AI ethics in practice, in the field with our clients and partners.

17:30

There's a concept that I'm not real clear on data labeling, what is that? Was that for the discussion here?

17:34

Yeah, so there's been a lot of talk about data labeling these days. And that's because to feed a algorithm for machine learning purposes, you know, you need good data in. And so data labeling is the practice of telling the machine, what piece of data exists and where it is, right, classifying it. Simple examples. It's the practice of teaching the machine that this image is a picture of a ball or a dog with hyper science is the practice of saying this piece of data is the invoice number, this piece of data is the date the client was seen, right? This piece of data is the claimants details on why they're submitting a claim, right. And so data labeling is a really serious practice if you want to have accurate, robust algorithms to help predict future processes. And so within hyper science, we've now released this idea of guided data labeling, to accelerate the humans teaching and machines learning process within hyper science.

18:40

Hey, the truth when he began discussion, we talked about Ray Kurzweil. I don't know if you know how big those machines were. But back in those been the size of a Chevy, these huge machines, and it's gotten smaller and smaller. And so so the product that you have, is it a service is part of a data center. I'm just trying to put my eyes on Oh, that's a that's a hyper science over there that I keep thinking of big OCR scanner, you know,

19:03

yeah. So you know, this, we talked a little bit about magic sauce earlier, I'll tell you. Another thing that I think is magical about our software is that we're containerized. And all of the machine learning and natural language processing or all the magic that happens in our within our software happens within our clients enclave. So

there's no third party data call happened to our full capabilities, right? There are other capabilities out there that might require a third party data call, but we do not right because of that container. We operate on Kubernetes and Docker. And so our deployments are on prem. We've even got some detail, tack deployments, right and no cloud or internet connectivity at all. And then make no mistake, we have a SaaS offering. We're partnered heavily with AWS. We've got some deployments to Gov cloud today. So John at the end of the day for people But listening that or thinking, Hey, how can I use natural language processing or machine learning, you know, reducing these paper based processes, we've got a deployment model that works for them. Because every model, right has security at the front and center.

20:12

I'm just thinking of several agencies that may be interested in the on premises solution. Not I couldn't name any three letter agencies, but it could work. Crystal Ball tie, we're not going to talk no more three letter agency talk talk about crystal ball. So billions and billions of your job is to reduce the cost for taxpayers. I mean, five years from now, I think we're going to catch up on this, or is there gonna have to be an instant and what's gonna, what's gonna be motivates people take this, this technology seriously?

20:36

Oh, man, this, you know, we just had our company kicked off. And we had a breakout where everybody was asked to predict the future, right? No, yeah, five years. Right? What's gonna happen? There were some fun ones about tax returns, right? Because everybody's looking at their tax documents right now. But what I'll tell you, right, in five years, I think you're gonna see the federal government actually taking the lead getting out in front of the commercial sector in terms of adopting machine learning, and intelligent document processing to digitalize their processes. What does that mean, in terms of future reports coming out of organizations like the US Chamber of Commerce, I think it'd be pretty neat for them to do kind of a retroactive look, right? How much how much money has been saved. I mean, just from one of our deployments that we've seen so far, once they go enterprise wide, we're gonna save them over 50,000 hours of time. And that's not even one of the larger use cases where, you know, we're processing nearly 800 million pages annually. So I do think you're gonna see the paper based processes being eliminated, right, we're gonna see people's tax returns being, you know, processed more quickly, right? We're gonna see backlogs for whatever the use case coming almost to a complete zero

21:52

here. So I've tried to take this interview and kind of wrap it up with a bow. And I talked about doing my taxes earlier. And then you mentioned the phrase kind of casually time tax. So wait a minute, this is all about tax. It's not about nothing to do with the IRS, a time tax involved in handling paper documents. And so so maybe that's the lesson of this is that it's all about taxes, but nothing to do with those taxes that you send off every year. It's got to do with with the amount of time it takes to process and many different federal agencies, it could be FEMA, it could be my goodness, and ih anyone out there. So they all have a certain amount of extra burden in handling these documents manually. It's gotta be, maybe it is the billion numbers seems, seems hard to believe, but let's find out. We'll come back in five years and we get a big cake. We'll have a big board like mana money saved in the last five years. No number hours say that's it? Wouldn't it be?

22:42
The headline? Yeah.

22:43
timetec Saving last five years look great. You've been listening to the federal tech podcast with John Gilroy. Like think my guest Chris, our sales director, US public sector high for science.

22:52
Thank you, John.

22:55
Thanks for listening to the federal tech podcast. If you liked the federal tech podcast, please support us by giving us a rating and review on Apple podcasts.