FEDERAL TECH PODCAST

Ep. 95 Data Scientist: the Sexiest Job of the 21st Century

00:02
This is John Gilroy from the Phil tech Podcast. Today, we're gonna talk about the sexiest job of the 21st century, hit the music Manny. Welcome to the federal tech podcast. My name is John Gilroy and I will be your moderator in the virtual studio today, we have Aaron Hernandez, he is the artificial intelligence and analytics practice director at a company called Excella. excell a boy tell you something, Aaron, you hit the nail on the head when he started focusing on artificial intelligence. I don't know if you predicted it 20 years ago, but you walk into the seventh lab, and they're all talking about artificial intelligence. I mean, it's everywhere. It's everywhere. It's everywhere. So my question to you is, why did you start there at the beginning? Why did you start in something different? Maybe cybersecurity or something? Why do you have this long history with data?

00:52
So I really got into data because I loved the way that it helps you to answer questions. So I actually started my career as an economist, and got really, really into forecasting. But the quality of the answers that I would get, were just never quite what I like to do. So I wanted to add new skills and new tools and technology to my resources in order to be able to start answering more and more interesting questions.

01:20
Well, I had that silly opener, because there's an article from the Harvard Business Review from 2012. And it says that you can look it up data scientists the sexiest job of the 21st century. And so they they're making some pretty strong, maybe they saw and they forecast the reduction of cost of storage, the proliferation of the internet, high speed internet, I don't know what they predicted. But it's all coming to fruition here, especially in the federal government. And that's what I want to focus on today. So so my listeners are federal, IT professionals, maybe other companies interested in understanding a little bit better. We're just trying to figure out what you can share with them, to help them maybe define some terms like, data analyst versus data scientist, maybe talk about, you know, some of the reports coming in on artificial intelligence, what the state of the art is for artificial intelligence for my listeners, and I've got some fun quotes here maybe to test you got those books behind you here. So yeah, they look like you're smart. So I got some tough quotes here to test you with. So let's come up with something I think that this would be a cop, if I was having coffee with you. I'd sit and go Well, geez, Aaron, data analyst data scientist was same thing, or what's the difference anyway.

02:26
So I would see data analyst is more of kind of your traditional business analyst types, the ones who are able to go in, they know some basic statistics, they're able to run reports, data scientist is where you get more into the machine learning realm. Data scientists are more hypothesis driven testers, and able to think constructively

about the algorithms that they're dealing, if I have a big Excel sheet that has a lot of information in there, and I need a little bit of insight. That's when I'm going to pass it off to a data analyst. If I have a large business challenge, that I'm not quite sure what to solve, or how I'm going to make predictions around it. That's what I'm going to give to a data scientist, a data scientist is going to help to structure the problem turned into a hypothesis driven research question and then test a variety of different approaches in order to be able to glean insightful information from that data. So I guess a good way to think about it, although it's probably not sufficient is that data analysts are much more in the descriptive statistics category, whereas I would put the data scientist in more of the predictive and hopefully even the prescriptive category.

03:45

As a grown adult, I was walking through Georgetown, and I saw these clothing stores and one said bespoke suits. And I went home and asked my wife for that. And what brand is that snack brand? It's custom made. Oh, so that's what a bespoke suit is, hey, I'm learning stuff in a band stage. And so when I look at data analysts and data scientist, I think today the house is more of a generic off the shelf. And then the data scientist is more of a bespoke, because it would seem to me that if you had a data analyst and your LinkedIn profile, be easy to plug in and go to different agencies, or maybe different companies and plug it in, but data science, I think it's at the level where it's a very specialized custom set of knowledge that one has is that is that jumping to conclusions here, that

04:28

that is probably not how I would talk about it. So we like the idea of the bespoke versus the off the rack but that has more to do with tool selection. And honestly, those tools can be used by a data scientist or a data analyst. So if you look at some of the more prevalent offerings out there in the data analytics field, some of them you just kind of plug and play SAS is probably the most well known and oldest example of this MATLAB is another one SPSS they have a lot of these statistical software's. You're just like, give me this and I'm going to do that. Whereas a lot more of the custom coding and you can do the custom coding, whether you're doing it in Python, or whether you're doing it as Excel functions. From the data analyst that's still more of the bespoke it's how are you taking something and applying a little bit of thought to it, as opposed to using just an off the shelf? COTS product?

05:19

My daughter is an engineer and she's learning Python because she thinks she can help her. I think Python kind of me the maintenance is the commonality between the two I mean, an analyst and a scientist is kind of like the language of the two.

05:30

Yeah, I would say so I would say that Python is definitively almost definitively a requirement for a data scientist, a data analyst is more of a nice to have, they could probably get by with SQL and some basic Excel skills. But it really just depends on the problem space. I think industry wide, there is very little agreement upon the definition between data analysts data science, machine learning engineer.

05:58

I, my job to cook dinner tonight. So I'm going to frigerator and pick out two or three ingredients and make something I don't know what's gonna happen. If I look at this field, I see two, three different ingredients here, I see. You know, someone like a, I have a friend who's a PhD in math, and she works with a software developer, but then they have to interact with a subject matter expert. And so I want to I want to come to, are they data scientists? Or is that host set? I mean, each one of them works together kind of solving a little similar problem. So these three ingredients different three complete and categories? Or how do these categories fit in with a data scientist? Does it just all work together? See is that number four? Number four ingredient getting an answer?

06:39

So I would say the data is data science is a team sport very, very much you need that engineer to deal with the infrastructure, get the data into a place that is actively usable, do things like ML ops, being able to push things to production? The data scientists, like I said, it's mostly the model developer, data engineers tend to hate data scientists, mostly because they pass them these horribly ugly Jupyter notebooks and they're like, What am I supposed to do with this. So they're really the ones that translate a lot of the findings that data scientists have to actual value. I say this as a data scientist who has passed a number of ugly Jupyter Notebooks off of my time. But there's others as well, you also need those business analysts, the subject matter experts, the ones who really, really understand the field, what it is that you're trying to do a data scientist themselves should never be really expected to have the subject matter expertise of somebody who's actively applying things. I think this is particularly relevant in the public sector face, you also need to consider things like data visualization, how are people actually interpreting those results, data scientists can come up with the best model in the world. But if they cannot make it digestible by a human being, then it's really not good for anything. others as well, that we could get into.

07:53

I do another podcast on satellites and space. And I'm telling you, this field is just it's so exciting. So many new things coming up. So many new companies saw many different points, you know, 30,000 satellites in the next five years, I mean, so much so much unstructured data. And I think this is a big challenge and just a starting point. I think it'd be hard to not begin the conversation with talking about noisy unstructured data and how to handle that and and what was his playing the role and and then cleaning a date I mean, is, is most of the data scientist job just cleaning the data instead of thinking about it? What role does this quality of data have quality

08:31

of data is huge. So as we were talking about a little bit before the PROG podcast with the Air Force example, a lot of people really want to rush into AI. As a data scientist, we always hate the term AI, you can pretty much tell the quality of a data scientist based off of how much they hate the term AI. But that could be a whole nother conversation. But really, what it is, is people are rushing, they see these new tools, they're like, oh, let's go get chat GBT or let's go get this thing. It's new, it's exciting, I get to all this value. But what they don't realize with that noisy unstructured data is number one that they don't have really a good handle on the inputs that are going into these models, every model garbage in garbage out, every model is only ever going to be as good as the handle that you have on the model on the other end of that pipeline, so you have to have good data going in. On the other end of that pipeline, you have to have a good idea of what results you're actively driving to. So

a lot of people will see this tool be like, Oh, hey, that'll be cool, but they haven't really done the diagnostics to understand these are the decisions that I'm trying to provide. These are how I'm getting smarter. So one of the often I think lost portions of the data science piece is the science position, that hypothesis driven research. Question of, okay, I have this data, I've gone through some effort. I've made it well structured. I have this vision outcome. How do I utilize the science of data to hypothesize different ways that I can go and make that better? How am I incrementally improving my approach? Right now it's become more of the AI toolkit of okay, we got this thing let's just go throw it into production. Hey, Look, it's working. And then they move on to the next thing, as opposed to truly getting into that refinement, where they can actively start learning from the data, which will drive to better results and outcomes.

10:12

I just want to jump in and say if you would like to work with people like Aaron, you can go to Excel e x, C, E, Ll a.com. And all kinds of information away plugged into the dev SEC ops community way plugged into software development and, and can answer all kinds of complex questions. I'm always amazed by the talent. That accelerant I don't know how they they bring smart person, smart person, smartphone, I don't know how Jeff does it maybe I don't know what dangles money in front of you. Or Lamborghini, maybe it's a Lamborghini. I don't know what it is. But really good people excel haven't haven't really been impressed with the people I've met there. Speaking of talent, what about our Duffys? What about the federal government? How can they attract the specialized talent that someone like you have him?

10:56

So it's often hard for the federal government, especially you brought up salary to start competing on levels like that. I think what the public sector has that private industry does not and honestly, what keeps me working as a public sector contractor, is a nuance of problems. So if you look at the data science or analytics problems that could exist within the private sector, they're generally a little bit more routine, hey, we're optimizing this business promotion, hey, we're trying to get as many eyeballs on this particular ad as possible, hey, we want to be able to segment those customer base. There are a lot of kinds of the same thing, hey, we need to optimize the search algorithm. If you look at the public sector, there's a lot more nuance on the way that they're approaching things simply because their mandates are not quite so metrics driven. One of the projects that I scoped out, we didn't end up delivering on it. But we did end up scoping out. That was just fun from like a mental exercise thing was utilizing data analytics and data science to be able to reduce the incidences of rape in prisons, we put together this proposal for the Bureau of Prisons, not a fun subject matter, obviously, as I see that, as I see the cringe. But the idea of how to actually leverage data to make that type of real world impacts, for me is a lot more fascinating than I think you get in the private sector, in terms of how the federal government can better be able to situate itself. Again, I think this goes back to providing that exploratory and curiosity mindset in the way that they are conducting their analytical initiatives. So if they're able to give people an outlet, in order to be able to start asking the questions that they care about in order to be able to start exploring those, I think that's a benefit. Number two is you don't always need a data scientist in order to make better improvements, you can utilize a government contractor, like it's Allah, to kind of get you off the ground, bring in some of that specialized talent that you don't need for the long term when you're developing out a solution. And then when you actually get the solution developed and are doing that incremental improvement, that makes it significantly easier and significantly lower skill level to try to recruit for.

13:02
I, I listen to podcasts, believe it or not do a lot of talk and a lot of listening. And I listen to a podcast with a guy from the US Customs and Border Patrol. And his name is Sonny and Baca Walia. And, and I knew they handle out of date. Okay. And so he started and I wrote this down because I can't believe this, they handle 10 billion transactions every day, 50 billion data exchanges every day, 175 petabytes on its network. I mean, when it comes to volume, it almost seems like someone like you can jump in there to go, okay, okay. Now, we're not going to handle all of this. I mean, how do you handle the volume challenges with data science.

13:42
So by not thinking about the data at all to begin with, honestly, you start again, data science is meant to build up decisional support tools. I am a person, I am in the customer, Customs and Border Patrol. And I need to make this decision, I need to decide whether or not this person is a threat, I need to decide how I'm allocating resources, etc, etc. Get narrowly focused in on the decisions that humans are making, or that are going into any type of process, and then figure out what data you actually have to support that. So as a Customs and Border Patrol, I'm just kind of going to make up a mission statement. I need to know whether or not this person coming has a fake ID or not. There's decisions that I can actually put in there, there is data I can start looking for there, I can start looking at passport control records, I can start looking at other records and start building up from that decision, a chain of information that can help support that and start putting modeling underneath that. So yes, the the size and complexity of the data is is intense, but that's if you look at it hugely. And as one big thing, if your slice off on those individual decision context, it becomes a much more manageable issue.

14:54
It does seem to me a guy like Jeff kalamar would say well, what's the business case? For this specific instance? It's not a matter of data. And I think that you that's what you just articulated. Well, what, what what is the business case for this? Let's go switch. Let's go across over to the Air Force. Okay. Well, Air Force has been testing out artificial intelligence and what are they just found? Well, if you're Jared server just came up with this study, just yesterday, the National Academy of Sciences did a study and they said, they are not getting the expected outcomes, they have the kind of weary of applying artificial intelligence to serious. I don't want to get leaf to lethal situations. And so they're very wary of it. And so I think, in the federal government, the approach to artificial intelligence in days got to be one for Border Patrol, but another for the military kind of different applications for you can't continuously change and alter a lot of things the military can, you

15:47
know, and I would say the so there's a couple of points that would come from what you just said, number one is particularly in these life and death situations, and I would say, particularly in most public sector situations, where even if they're not life and death, there is an individual on the other end, that is going to receive some outcome that may be negative. I think the human in the loop is key ethical considerations that are gone into model development are also things that you need to start thinking about. But I think one of the reasons that the federal government struggles with getting value from data is number one, they really just kind of they they hear about these new tools, and they're like, let me go do that. Let me go do that. Let me go do that. They haven't put enough time and effort into actually setting up that foundation. What does the data need to look like? How

is the infrastructure going to support this? Have I actually gone and done enough research on these specific decisions and the inputs to them that I need to be able to derive value from these things? I remember one client, which I will, I will, I will keep nameless. But when I got there, they had this very particular business issue, they had actually done a lot of the infrastructure work, they had done a lot of the other things, but they just really wanted the big tool. So I got there, and they were like, we have this problems. And we really, really need deep learning in order to be able to solve it. And I was like, Well, I'm not sure about that. And they're like, no, no, we need deep learning. About a month later, we realized we can solve that problem by joining two tables and SQL and running a SUM function. So they didn't need deep learning. They just needed to think about the problem. And a lot of times, again, people will be like, Oh, check GPT it's the newest thing. Let's go get it. Yes, it's useful. Yes. It's transitory Yes, there are probably some issues with hallucinations that need to be taken into account as you're doing it. But you need to start being thoughtful about the application and where you're applying it. And doing so with a lens of that decisional human in the loop support tool.

17:38
Now, what the humans who are listening to this, many of them work for the federal government, and they have data handcuffs on, you know, they have to be very careful about data access and data security. And and then we'll maybe old Aaron can't see this particular data file. Yeah, oh, rathole can't? And, and so I mean, if he went into Progressive Insurance in Ohio, you'd have some restrictions there, but not like you'd find in some agencies here. So. So how does this contribute to the federal government understanding how to use data, given all, you know, security and access restrictions,

18:14
I mean, I happen to be a fan of most of the security and access restrictions to a lot of the data, not all of the data, but a lot of the data that exists within the federal government simply because a good member of it is sensitive, it's PII driven, it has a lot of details that really should not even be shared among individuals. I'll give you an example. When I did my first public sector trust, back in 2013, with the Federal Energy Regulatory Commission, I had to disclose some financial details that I probably didn't want the rest of my peers knowing. Simply because I had just gotten out of school, it was a little bit of debt, there's just information there that well, not necessarily harmful, should desperately definitely be protected. How you can do that is a number of different ways. First off, only use the data you need, how much of that personal information is going to actually be mission driven. If it is mission driven, how are you able to respond to it in the appropriate level of abstraction? So there's a number of different masking methods, there's a number of different ways to systemically be able to allow somebody to get the value from insights from those data's without actually having access to the underlying data itself.

19:24
Yeah, and that's a whole separate topic that a lot of companies forever about that. But But that's what's important is that it can be done and it shouldn't be a handcuff on a person at that interior trying to figure out something or whatever. It's just, it seems like there's, there's a lot of these theories and a lot of people talk about this and that they don't realize that there's limitations to this. We've talked about large volumes and security. And I think we had talked more about this. You mentioned earlier about the ability to communicate to non technical stakeholders and taking this information and And going from your lofty data scientist position

down to the, you know, mere mortals to understand things. And so So what is the approach that you use? I mean, how can you take some of the sophisticated stuff, and I hired a PhD in statistics to do a project for me a year ago. He gave me some stuff. And whoa, I wrote the check and moved on. But it's, that's, that's, that's the last hurdle, I would think.

20:24
Yeah, so we use a lot of human centered design techniques, when we're actually building up those decisional support tools. Sometimes it's in the form of a dashboard, I think dashboards get a little bit too much of the spotlight, it can also be in terms of the under the hood, background engineering as well. But again, by really getting a little bit focused in on what people are looking at how they're making the specific decisions that they're driving for, you can have a lot of fun, I'll give you an example of something that I've done in the past, which I've called dashboard for your dreams, which is really we get, we get a bunch of government contract government workers in there. And we literally just start driving around, if you could have any dashboard that would help you do your job, what would it look like? What would be the types of information you would have there? How would it be displayed? When you click through it? What would you be able to get more access to, and then being able to take that and actually turn that into analytical solutions. It's always a give and take, because when you start that way, and then only build that, you're kind of sometimes missing the value of data. So it has to be a continually evolving conversation, get a first pass what they think would be useful, but say, Hey, wait, we could also give you this, we can do this, you're not necessarily aware of this, because you're not a data science expert. But we could actually abstract it like that, and continually work with the actual users of the data that people are going to be making use of that information to drive the decision. Without that there's really no value in the analytics at all.

21:52
You know, what you just said, I think that as a tool set, incorporated by a data analyst, you know, this whole idea of using Tableau or something to come up with some kind of a dashboard that works. But you said earlier, it takes a team. So you probably need a data analyst in there, you probably have to have someone who has deep and thorough understanding of the science itself. And then what about you know, the coders and developers they have to, they have to make it work in the system? And so it's a, it's much more complicated, I think, I mean, we get dazzled by so much stuff out there. And and I've talked to people who adamantly don't want to use the word AI for anything yet. They'll say LM, but they won't say the AI words. And so you got to you got to juggle all these things. And I'm guessing from, from someone who is maybe in college now, I mean, if they desire to have a career similar to yours, I guess I should focus in on quoting software tools, maybe math and statistics. And in general, I imagine there's got to be able to logic and don't forget ethics in there, too. So is that the mix you'd recommend for young people?

22:53
Yeah, absolutely. And it's really just about that curiosity. So like, as a data scientist, you need the coding skills, definitively you probably at this point, when I started, I was able to be a little bit more generalized. Even though I got kind of specific in NLP, you need to be able to figure out what type of data science you want to do. Since the field is getting vastly too complex for people to know everything. And the statistics and math is huge, but not as much as you would need for a regular statistician. I've always kind of it those two statisticians I see is

like, you know, your, your, your really big statistical root cause analysis, go in there and figure things out. data, scientists need enough statistics so that they can understand what kind of the output of the model is looking at. So I would focus more on the computer science and just given where the industry is at today, because that has a lot more to do with the optimization. How is this actually going to turn into a production ready? system, statisticians don't have to worry about taking something and turning it into something that's going to go through petabytes of data, and do so in a fraction of a second data scientist.

24:03
Yeah, makes sense. We got to end this interview with a prediction. If you looked at the transcript of this interview, you'd see the word tool pop up several times tool tool tool. So what kind of tool would you or what kinds of tabs is magically appeared of what tool would you like to have? I'll put you in the situation of I'm at the whiteboard, and you're in the room and I'm saying, Okay, what tool would you like Gilroy software to come up with out help your world and the next five years? So what would you like to see

24:30
that help would help my world in the next five years? That is a tough question. I don't know if I have a great answer to it. But I think it would be something there's a lot of divergence in how you actually develop out AI software as of right now. So you can look at things like JIRA support tickets in order to actually develop out the models. You can look at Model Management regimes. There's nothing really that ties them together that does a good job of of being able to provide that decisional support kind of one view thing you have Tableau, where you can go in and see a dashboard, but there's really no way to see the overall health of the data science industry or the data science practices within your specific company.

25:16
Yeah, standards is something that kind of even just emerging right now. Maybe in five or 10 years, we'll have better ideas of where they're all ends up. That's real good. Well, thank you very much for your time this morning. You have been listening to the fiddle tech podcast and John Gilroy would like to thank my guest, Aaron, who Hollandaise AI and analytics practice director at Accela.