Andrii Galan

Alison: I'm currently working on the social science trying to understand the differences in job responsibilities of data scientists in startups versus enterprises. Therefore, I am hoping to talk with the experts in the data science field and learn about their experience of working as data scientists in the industry. Thank you very much for agreeing to this interview. I hope this interview will be enjoyable for you as well even though we are not talking the technical side of things. And I just want to let you know that if you want to take a break at any time, you can. And if you don’t feel like answering a question, please feel free to say that you don’t want to talk about it. Um and like I said, I expected to be about half an hour and um I might go off script. And the entire interview will be recorded. Need to write down everything we say so um recording it makes it easier for me to write it down. If there's any like confidential information that we accidentally talk about during the interview and then afterwards we find out oh we probably shouldn’t have talked about that. Just tell me and then I won't write it down. Okay I want to ask that you're still willing to go through with the interview, right? Yeah, I am. Great! Thank you very much. Okay, let's get started.

Q & A:

Alison: First of all some basic information, What is your position at Paypro Global and how long have you been working in that position?

Andrii: I'm working at Paypro Global as a data scientist. Just doing the basics, generating reports, preparing presentations for different departments and stakeholders. And also working with the AI, generating models, validating models, integrating those models into our system. I've been working at Paypro Global for three years already.

Alison: Three years. That’s pretty long.

Andrii: Yeah.

Alison: Like you just said, you go to meetings, you write reports, you build models, and you validate those models. How do you think those translate into business values and in what ways do you think what you are doing is helping the company grow?

Andrii: yeah of course it depends on what type of business we are e-commerce platform so we are working with a huge amount of data and yeah almost all the processes can be controlled by the state is so we have to follow all of the year processes daily and check the main metrics and check the parameters of the processes that we are dealing with so we have to be really careful about everything and take care about the processes that yeah the processes that we are dealing with in our daily job so yeah of course we are generating the reports, the requested ones or even regular reports, some integrated reports with the help of some BI software. It is important to follow, as I said, to follow all these metrices and the processes. If we face some problems, let’s say in ecommerce, the basic ones are fraud detection problems, chargeback rate, authorization rate issues. We can easily follow them using some reports. On the other hand, we can also use modeling to prevent fraud or predict some things like chargebacks, sales for the next few months. Yea so essentially, we can predict things and see what kind of future we can expect.

Alison: Yeah. Actually, since I have been working on some of the reports, I think for our reports, a lot of them are for example for the frontend team and other teams, so you work closely with all of those teams as well right?

Andrii: Yeah. Frontend, backend team and also sales and marketing.

Alison: And the next question is, what do you think in general about your position as the one and only data scientist in the entire company? Are you enjoying the heavy responsibilities? Or can it be stressful sometimes?

Andrii: Yeah there are positive and negative. My job of course. Let’s start with positive things. I cannot say that I am the only one because people from sales and marketing department or we have also the people representing such department as customer success management, so CSM team, actually deal a lot with the data in their daily life so they know something about the basics of statistics so still I have people with whom I can discuss results and so also the backend team. We have also guys that are really educated and they also have some understanding of machine learning. So yes. When I am preparing some models for them, of course, we are going to get through it together, to train and validate it together, especially when we are starting to integrate this model into the system so of course we are discussing these things. So I still feel some support from other people. The minuses. Yeah of course it is always better if you have a better data science team so that you can share your responsibilities and you can focus on more specific things, more specific fields. And of course, you will always have someone who can help you to speed up a little bit with the job. Follow all the deadlines and the requirements that are from other departments. Yeah, I will say like this.

Alison: It’s actually quite interesting for me to know that there are other people workig with data in, I would say separate teams as well. So like you said, they also help you with for example, validate things and discuss for example machine learning models. But you also said that in bigger teams, there are people sharing your responsibilities, so I would, um, I am actually a little bit confused. They are not technically data scientists, right? But they work with data as well? So, how exactly is your job different from theirs?

Andrii: You mean compare me to backend team for instance?

Alison: Sure.

Andrii: Yea. It depends on what kind of specialist you mean. So if you are talking about the backend team, they are of course working more with integration things. Still before they will implement your model, before they will deploy it in production, they of course validate your model. And they can see also all the numbers. Let’s say a specific example. For instance, fraud detection model. So I collected all needed data, then I build the model and I am stating to deploy the model using our backend team. What is happening of course is that when we finalize the deployment of the model. Afterwards we of course are checking what kind of results we are getting. So we are trying to of course, the backend also of course, they are checking and anlyaze the data they can see also to conclude if this model is a nice predictor of fraud. Because we are getting fraud like, not every day, but once a week, for instance, it’s normal. And they can easily see if the model helps somehow to block fraud. So my model is predicting fraud and techniques they are using are using prediction and blocking fraud traffic for instance. So they are more responsible for some technical things like blocking traffics. It was just one example. Of course they are busy with their jobs, but in point where we are validating models, integrating these models, deploying models to the production, so of course they are working together with me and sharing knowledge with each other. I explain something to them if it is not clear and of course they are also explaining me somethings. And in general, when we are deciding to create some model, some algorithm that are going to be used in production in the future, it is not like every time I have an idea in my head and I am going to someone from our manager saying okay this is my idea, I think it will help us. Of course, in most cases, we are discussing what kind of model and what kind of cases could be useful. So, everything is discussed with the management. And actually, at this moment we don't have like a data scientist team. So, I cannot say I am the only data scientist. I would say I am the data science lead. So yeah, at this moment I am working with the back-end team. So, I’m the team lead of the back-end developers. Our communication is quite close. We have weekly meetings. So doing very often, some updates, trying to discuss some new ideas, educating each others.

Alison: Yeah, so like from your answer, I feel like, I would say you have the more interesting part of the job. I mean model building. You build the model and from the outcome, for example, the fraud detection, they can actually start to implement the proceeding steps. So yeah, I totally understand that sounds fun. And having a team sounds excellent.

Andrii: And yeah, we just use examples of, just creation of the model and the integration of the model, but beforehand when you're trying to generate the report and dealing with the SQL trying to extract the data. So, I think that 50% of our back and developers they also are very often working with our database, so they know how to write SQL how to extract data. They know a lot about the structure of the database. So it's not only like me, one person is working with the database, no, we have in this stage some requests. If I do not manage to do all of them till deadline, so it might be that one of the back-end developers or our team members might also generate reports, so he and she can get data from database and extract the needed data. Alison: Yeah, it's quite interesting, because sometimes I think what you guys do overlap a little bit. Also, the skillsets overlap a little bit.

Andrii: Yeah, I think now the full stack specialist are quite needed everywhere. So if you are a backend developer, a Java developer, for sure you are not going to deal only with that in your life. You have to learn new techniques. Some extra jobs, but that's okay.

Alison: Yeah, that makes perfect sense. Okay, so we were talking about your current position and now I want to look back into the history and look back into your education and your previous work experience that made you the current data scientists that you are today. Um, so first let's start with education. I know a little bit about you being a physicist. So, you were studying physics at as an undergraduate student or maybe masters?

Andrii: Yes, I have a master’s in theoretical physics.

Alison: Wow, that's very hard-core science.

Andrii: Yeah, but I enjoy it.

Alison: Nice.

Andrii: We already discussed it.

Alison: Yeah, we can save it for another time. Yeah. So, I'm not sure if there was a point in your education or a point in your life that you decided, oh, maybe doing data science instead of physics is a better idea. So did that ever happen? When did that happen? Or what made you feel like being a data scientist is you know what you want in your career?

Andrii: I don't remember if I had any switch points like you said. That I had to leave physics behind and start something new. Everything was like going smoothly. I was studying theoretical physics, so I had to deal with a lot of mathematics. So it was not like applied physics where you're doing a lot of experiments in the laboratory. And it was not like this. I was working with my supervisor in my last year in my university. So I was working with my supervisor proving some physics processes using a lot of mathematics. And yeah, and then when I finished university, I find out that I can also improve my mathematical skills in Netherlands and I went to a postmaster’s program there. And the name of the program was like mathematics for industry. So again, for me, it was pretty close to what I was doing in university in Ukraine I mean. I would say that in Netherlands, it was not really something new for me because in Netherlands I had the opportunity to study still some things and I have some additional courses like programming, technical writing, improving skills and my skills as a team member and so on. And then I had the opportunity to work on different projects for Dutch companies. It was really nice and some of the projects were really related to mathematics and physics. So those two sciences are actually very often cross and going together to explain some of the processes we will create, let's say some models that you need to understand a lot of the processes are explained by physics. So it was not really something new for me. Yeah. But then at some point in Netherlands, I attend for some specific courses related to data science, and then I started to learn methods more focused on statistics, combinatorics and so on. So, at some point I understand, okay, I'm already a data scientist.

Alison: Yeah, that's cool.

Andrew: Yeah, it was not like making any decisions.

Alison: Yeah, it does sound like a really smooth transition. I mean, I personally would think that physics and math are not very related, but now that you put it, I totally see where all those things overlap and transition from one to another. It is great that Eindhoven had courses that sort of lead you to doing more data science related jobs. I think that's really cool. I'll definitely check out that program. It's also very interesting for me. Um, okay, so it seems that starting from Eindhoven you began to have more internships and work experience in data science, how has your current position been different from your previous jobs?

Andrii: Actually in Netherlands, if we're talking about the university projects, so the short ones I was working more for industrial companies. Yeah, like Phillips, SPG Prints. So these companies are more dealing with some industrial processes. For instance, at that point, Phillips was developing wrist band. A kind of watch. This project was to create you know the traditional wrist band nowadays, they're just measuring your health and yeah, how many kilometers you was walking today, Phillips was interested in creating an APP, but this wrist band was especially for the drivers. It happens very often that people are sleeping behind the wheel. Fall asleep behind the wheel. And because of that, there are a lot of accidents, so they wanted to create a wrist band. I'm not sure they produced it now, but at that point they were creating working on the creation of this wrist band that can follow your sleep cycles. And so if you are wearing this band on your hand for a few days, you can follow all your activities and you're getting sort of in the APP should use the data from the wristband to predict if you are alert enough to drive. Alison: Interesting.

Andrii: So at that point, I was dealing a lot with the data, not so many physics, a little bit, because the way wristband is working is explained by physics, different kinds of processes. The other company, SPG Prints, for instance, that one was working more about the printing technologies, so they had a lot of problems printing on different surfaces. This was also related a little bit to physics because you have the ink that was going through all the systems of pipes. And there are things like pressure, speed up process of that ink and so on. So I think because I had a background in physics, so the director of Eindhoven was searching for a kind of projects that are related to physics. So I was working more in the industrial things, but when I was finishing when I finish all the projects in Eindhoven, when I graduated from there, I went to the company that was dealing with health care things, yeah, and there I had to deal a lot with data and not so much physics. And now I'm working for an e-commerce company, so also not so much physics. So you ask me to compare what I'm doing today and what I was doing there, I would say in university I was still dealing with a lot of industrial things and but now it's like more like far from industry, so like business things, healthcare things, those kind of branches. Alison: It's quite interesting. I have no idea. Healthcare I think it's pretty different from physics. But you're right like dealing with data, the ability to analyze data, is universal across different fields. So I would say that the main difference between your old job and your current position are in the business domain regarding data science, would you agree?

Andrii: Yeah, I would agree that is the main difference.

Alison: Yeah. And what about in terms of skill sets or in terms of teams or company structure, all of those things? Are there any differences?

Andrii: Yeah, for me, it happens to I switch from huge companies to smaller ones. Because working at Phillips or SPG Prints, I was dealing with more with bigger data scientists teams. So yeah, we had really specific activities. Everyone was busy with specific processes. We had much more communication. Because if you have a team of at least 10 people, so you need to you have daily meetings. You're discussing things, we had a lot of standups and talking about what we're going to work on the next week or months. So everything was like. Let's say I was learning there how to work as a team member sometimes. When I have short-term projects at Eindhoven. I was team leader a few times. So that was nice experience for me to learn how to behave and yeah, when you're working in team, but then slowly will switch to smaller companies, smaller teams. Now I have more responsibilities related to technical things, less meetings and discussions and probably less complicated processes than what you might have if you're working for big corporation so you have to deal with, again, different departments, different people. And sometimes, you have to in the best, at least I used to give more presentations to the whole department because it's not always the best way to share what you're doing at this moment. During the meeting, audience are like the whole department. Then you have to present something monthly, something like this. But at this moment, things are quite “compact”.

Alison: Yeah, I think that's a very accurate description.

Andrii: My life became more compact. We have fewer departments and 10 to 15 people in each of them. That's enough to communicate within departments during meetings, no huge presentations and so on. But when I was working for the healthcare company in The Netherlands. I worked there for two years. I was also participating in the different conferences. So this is what I am not doing at this moment. More focused on current job. I can participate now in the conferences, but just interested in things what people are doing. I'm presenting nothing so.

Alison: That's cool as well. And we briefly talked about how, I would say, you slowly became a team of one. And like you said taking on more responsibility. I have a feeling that SQL and data extractions and report might not be the most exciting things and I would totally agree that model buildings and my work, the work that you help me right now, it's probably more exciting for a data scientist. Do you feel the same way? What in, for example, model building excites you?

Andrii: I can agree with you that probably generating reports is not the best thing. But it is always nice to switch between activities. When you're doing some modeling, building models, let's say. Sometimes it's really nice to refresh your mind and just stop with generating the models and switch to another thing and come back. With a fresh mind you can probably, I don't know, bring some new ideas and change something. You know, models. You already tried to generate some models. You know at this point you feel like you are you don't know what to do the next. So you need to stop a little bit. Stop, think about different things. Think about what kinds of action you can do to make things better, you know, change something about your logic or extract some more data. And this is what I really like in modeling because you are like the painter, decorating the picture and creating your own model. So it's up to you. What kind of techniques to use and help to transform your data, preprocess your data, what kind of techniques to use. Sometimes you are really creating unexpected predictions. So yeah, in the beginning it looks like, okay, the normal things, the data, the input output, your preprocessing, your data, testing the data the first time, checking so okay, it's normal. And then at the end, sometimes you're getting unexpected results. Really something nice. And it is really important in data scientist job, to get some really useful results. And understanding that your model works perfectly and it will be quite useful for the whole company or yeah specific processes in the company. So that's really something that inspires you when you are working more and more with the models. So sometimes you're stuck for months with certain models and then you're thinking, okay, so like I said, what I really like about modeling is that. You're dealing with a lot of things. And you can get unexpected results. And it happens often that you are trying to learn some new techniques. Let's say you're the data scientist. That is using some regression models. And I don't know, random forest. But at some point, you're dealing with the new data. And you're thinking it's not enough. I need to check for some other techniques, some other methods of a model and apply to this data. And then you're again, learning about something. Modeling is pushing you to learn some new things. And that's really a huge plus. And I will say something about the reports. It’s not as annoying as it looks in the beginning. It's not like just extracting data from database and creating reports. Nothing special but in some point, when you have too much request daily from different departments, you're trying to think about some systems of report that will allow you to spend less time in generating daily reports, but also make the report generation more automatic. So then you're using different tools. We've discussed with you power BI, for instance, and different platforms. We have a lot of them and you're creating a system of reports. So at some point you are yeah, people from different departments are coming to you less frequently. And they're already using some automatic reports, some systems of the report. So I mean it's quite interesting when you're creating something systematically. And that's also interesting. You have to deal a lot with the relations between the database, tables. So you get more complicated systems of reports. And then you have to deal somehow with the system and make sure that everything works fine. And all the reports are communicating nicely.

Alison: That's interesting. I feel like in some way the way you describe having a system of report is also like modeling. Because you're automate things, you're making them easier for other people to use. Yeah, I would say I feel exactly the same. I feel like if I can make something that's beneficial to the whole team or the company. That would be a very proud project. So let’s talk about the business side of things a little bit more. You mentioned the other day that you're having GPT-3 meetings every Tuesday and Thursday. Yeah, so you mentioned a little bit about they're probably not talking about the technical things. Sometimes it's a little bit difficult to understand what they're saying. Do you think having what's called business sense or business acumen is very important for data scientists?

Andrii: Yeah, GPT-3 is a nice example. Yeah, probably data science is not really thinking too much about the business side of processes. I think that's your question, right? If I understood you correctly. So we have other departments that are responsible for that. In most of the cases I'm trying to simplify things in my head. So I don't want to really put too much theory is related to business things or something like that to my head. So usually when you have a nice idea or someone from the department is coming to you and say we have a problem here, and probably those problems can be solved using some data science. Then I am trying to understand what is the problem. Of course then you have to dive a little bit into the business things, what the process looks like. But I prefer to do it quite surfacely, not really diving too deep. The first thing a data scientist is should do when he is trying to create something, trying to create the model let’s say. If you simplify the process, it’s always easier to go from simple things to more difficult things. If you would need some more information from business, then you are going to ask more questions on every stage. So I would say the answer to your question is from the beginning till the end of a project, of course you are diving deeper and deeper into the business things, if the project is related to business. But like I said, it is very important to start from simple understanding, and then make things more and more difficult. And hopefully at the end, you don’t have to dive too deep into the business things. Yea just general things.

Alison: Yea that is great. Our last question is how do you think you have grown as a data scientist over the years? It can be about skills, about life or about anything at all.

Andrii: You mean the question is like how I see my further career in data science.

Alison: Ah yea, the question is how do you think you have grown as a data scientist, but it would be cool if you talk about your future career design.

Andrii: Ah you mean, how I came to this point yea?

Alison: Yea sure.

Andrii: Of course. Yea nowadays, you have a lot of possibilities. People sometimes study in Psychology are getting some basic knowledge in mathematics and statistics. It might be that after five, six years of studying Psychology, you decided to become a data scientist. So I believe everyone can build a career as a data scientist. What you need to have is some mathematical background for sure, statistics for sure, some Linear Algebra, mathematical analysis. Everyone should know that, should start from that point slowly, step-by-step, learn more difficult things. I think first people work as data analyst more, when you are just analyzing data. You are trying to visualize data in an understandable way. There you are already learning something about the programming languages, Python, R any other. Just visualizing the data. Then from visualizing the data, you are slowly stepping to the creation of AI. You are dealing more with the AI, machine learning models and deep leaning. This kind of things. And at some point, you will realize that you have enough knowledge to create some models. And then you are starting to build models and what is really important in data scientist life is that you really have to learn things almost continuously. Because everything is changing quickly and there are new methods created and so you need to learn things, as I said, continuously. That’s approximately what happened with me. What I am going to do in the future is that I hope I will have enough time to learn new things also. Apply new models. For instance, I am not really working with the NLP or image processing. I didn’t work too much with even neural networks. Just the basic things, so probably that is something I would like to learn still.

Alison: Yea I find them quite interesting. Very difficult though. But very helpful in application. Thank you for all your answers. That is all my questions I have so far. I am getting a lot of really useful information. Do you have any questions for me? How are you feeling about this interview right now? It’s good?

Andrii: Yea. It’s okay. It was not too technical.

https://www.youtube.com/watch?v=1UjMBw\_7Gi8

Data Science: Startup vs. Large Corporation

00:00 Hello everyone, Ken here. Today I'm

00:02 talking about the differences in the

00:04 data science role at a start-up versus a

00:08 large corporation. I've worked as a data

00:10 scientist and a data scientist manager

00:12 at both startups and fortune 100

00:14 organizations and there's some pretty

00:16 stark differences between the working

00:18 experience in this role. Before I get

00:21 into the meat here, I'm gonna please

00:22 remind you to like this video if you

00:24 enjoy it really helps me spread and grow

00:27 my channel. Also subscribe if you'd like

00:29 to see more content like this. Now the

00:31 first thing I'd like to talk about is

00:33 the actual nature of the data science

00:35 role. In smaller startups, usually data

00:38 scientists are responsible for a lot

00:41 more breadth in terms of activities. So

00:44 you're maybe not going to be doing data

00:45 science all the time. You could be

00:46 helping with even marketing strategy or

00:49 designing the infrastructure of the

00:51 databases there's a lot of things you

00:54 can do a lot of ways that you can pitch

00:55 in. And this is really great experience

00:57 especially if you want to be moving into

01:00 management or managing of other parts of

01:03 the organization aside from data science

01:05 down the road. It's still a great place

01:07 to learn data science because you

01:09 understand the strategy of why you're setting

01:11 up certain analyses and why a certain

01:15 project fits into the bigger picture of

01:17 the company. In large organizations, big

01:19 corporations, the role of the data

01:21 scientist is generally a lot more

01:23 defined you're probably going to work in

01:26 one fairly specific area with fairly

01:29 specific constraints. For example, when I

01:31 was at GE, one of the guys I worked with --

01:33 his whole responsibility was to work on

01:36 and improve the wheel wear

01:40 model for locomotives. So, no I mean that

01:44 was a cool job. He absolutely loved it,

01:46 but his focus was very very specific in

01:49 that area. If a lot of tasks, a lot of

01:53 projects can seem overwhelming to you, this

01:55 is totally a great option. You can

01:57 completely master one subject area and

01:59 that can be the focus of a large part of

02:02 your career. Culturally in a start-up,

02:04 things can happen very fast. You can be

02:07 working on one project. You get it to

02:09 where it's good enough and then you

02:11 start working on another one. You can

02:13 pivot in terms of your strategy and the

02:16 projects that you were working on before

02:18 aren't necessarily relevant anymore. And

02:20 that's something to be prepared for. If

02:22 you like that fast-paced atmosphere this

02:25 is absolutely something for you. On the

02:27 other hand, in large corporations, things can

02:30 generally happen a lot slower because of

02:33 structured corporate culture. There's

02:36 generally a hierarchy and things have to

02:38 get approved in order for them to

02:41 actually get done. So even to get access

02:44 to the data source, get preliminary

02:47 approval to have a dashboard built. This

02:50 has to go through some approval process

02:52 from the higher-ups. And that can be a

02:55 bit frustrating sometimes, but it also

02:57 means that you're gonna be working on

02:59 projects that are very carefully curated.

03:02 So from a data perspective, I think any

03:06 organization big or small can have data

03:08 quality issues. Usually larger

03:10 organizations struggle with data because

03:12 of legacy systems legacy things that

03:15 they've been collecting that don't

03:17 cooperate or integrate well with the new

03:19 data that they have that's very relevant

03:21 and useful. So if you are going to have a

03:25 wealth of data, but managing it can be a

03:27 bit difficult from a data science

03:29 perspective. On the other hand, at a

03:31 start-up usually there is not enough

03:35 data. You know there's some data

03:37 collected, but there's not enough time or the

03:40 function that you want to understand

03:42 better hasn't been around for long

03:43 enough for you to be able to use this

03:46 data in analysis. So there's two

03:48 different data challenges at these two

03:50 different levels and you kind of have to

03:52 figure out: one, this is very specific to

03:54 the organization so what your

03:56 organization and what the organization

03:57 that you're looking at is struggling

03:59 with and two, what type of challenge that

04:02 you like do you like working with sparse

04:03 data? Or do you like cleaning, cutting and

04:06 manipulating data so that it can be a

04:08 partial match and can be workable within

04:10 your algorithms. Overall, I personally

04:13 prefer working in the startup

04:15 environment better. I really love data

04:17 science, but I love other elements of

04:19 business as well.

04:20 I like management. I like communicating

04:22 with people. And I found that in a

04:24 start-up I had ownership of a lot more

04:27 things.

04:27 And I was able to really make an impact.

04:29 For me, seeing an impact in an

04:32 organization is extremely extremely

04:34 meaningful. Thank you so much for

04:37 watching and please enjoy your data

04:38 science journey.

https://www.youtube.com/watch?v=0qaS5JyU77E&t=393s

Real Talk with Startup Chief Data Scientist (10 years of experience)

00:00 So when I think of like good data scientist versus great data scientist. Like a good data scientist indexes to value, understands what the business is trying to do and can deliver outcomes. Like a great Day to scientist…

00:20 Hey I’m Molly from springboard and I’m here with Ian Blumenfeld, a chief data scientist at Clover Health. Hi Ian thank you so much for joining us today.

00:27 Hi. Thank you for having me. I am really excited to be here.

00:29 Let’s start out with your LinkedIn profile. You said you do a data science with the product bent. What does that mean for you?

00:34 That means I tend to focus more on the use cases and applications versus the methods and the methodologies. It basically means if I am going to do something, I will start with why might this be valuable, why might this be useful versus is this possible and how will I do this.

00:57 You’ve been doing data science for 10 years now. How has the role changed over the years?

01:01 The biggest difference has really been that you need to be able to, if not write production code, at least interact with production software systems now. The years ago, if you could write to SQL reports and send it off to someone to make a decision. It was good enough. Now it’s kind of hard to make a high-level impact if you can write code that you know 10,000 or more people are going to interact with through some kind of software product.

01:37 Let’s talk a little bit about your own start up. Do you want to give us a little bit of background knowledge?

01:41 Coming out of my first job, I just sort of had an itch of like I want to do something for myself. It was a healthcare data machine learning based start-up. It was a complete and total failure.

01:56 What were some of your biggest learning?

01:58 I mean the big meta learning is essentially like the market has to want something for you to do it. Just because you are a smart person, does not mean you are going to build something that anybody wants. Having that experience of launching a failed product helped me get a lot more critical about my own work and also the decisions that I was making going forward in terms of like what I should be working on and who I should be working with. And so I ended up developing through that owner/investor mindset, so when I went to look for a job after that, part of what I was accessing was essentially like alright if I go work for this company, I’m basically, I am making a large investment in them right? Even though other people are putting up the capital, I am putting up my time. And so is this a thing that I have a reason to believe could be successful? Is it well-setup? Is the environment ready for it? I look at my current job and say like alright, what should I be spending my time on. It’s through that same lens.

03:14 So now present day, you are a chief data scientist at Clover. What are you doing there and what problems are you solving?

03:21 Clover is still a startup. We are on the larger side of the startup at this point. I do a lot of work on our machine learning platform. Kind of looking at what kind of models we can produce to do a good job of really identifying health risk so that our clinical team know where they should be focusing their resources. And I also spend a fair amount of my time, kind of thinking, thinking through how we can address the problem of like what next. Anyone who works for commercial operations basically knows that this point machine learning and the predictive modeling part is a relatively small slice of the problem.

04:05 And once you get the models that starts to perform, the next question you are going to get is okay now what do I do. Um so I am essentially going to spend more and more of my time like working through that part of the problem, which is half understanding what are these models actually do and then half kind of like understanding from our operation perspectives. What are their capabilities and how do we figure out how to match those things together.

04:36 How has your role progressed through your data science career?

04:31 When I first started, I was writing a lot more code than I do now. Part of that is tools like the tools that have gotten way better over the last ten years which means I have to write less codes to solve the same kind of problems, but part of it is there are more people that sit around me that can help. So that means that like I have to spend some amount of time figuring out like who should be doing what. And like how do I pop off parts you know parts of these problems that are really appropriate for people different stages in their career development I do a lot more mentoring than I did and I spend much more of my time thinking through actually solve them versus like actually how to solve that. This is when the breath becomes useful because then you can say something like okay we need to solve this problem like here’s kind of like classes of types of methods that might be appropriate. Here is a few papers or blog posts that go into them so I can send someone on their way. I used to do all the stuff myself. I don’t do that all of it anymore.

06:02 A lot more delegation now.

06:04 I think of it like a last delegation and more like giving other people the opportunity to step up because you know if I keep all the interesting part of the problem to myself like hey great for me but someone else is losing the chance to like try to do something interesting and valuable and useful.

06:29 You spoke about how your current role is more about mentoring. What does mentorship mean to you?

06:34 It is really about helping folks build a framework within their own careers for how they should be making decisions so he can own what their career path and what their career progression is. I think there is a tendency, especially in large companies, for people to come in and say okay the rubrics look like this. I am a level four and I would like to be a level five but you as an individual also needs to have a view of where you’d like to go. And there are places that are going to be aligned with those rubrics and places where it’s going to diverge and where it diverges is kind of like your opportunities to take risk and try to do something but if you don’t have a good framework for sort of understanding how do I make those kinds of decisions. It can become very hard and you just end up playing the video game. Like oh, now I am a level five, now I am a level six and then you look back after 10 years and you say is this the career I actually wanted? You are lucky if the answer is yes.

07:43 Outside of foresight and a bit of risk-taking, what do you think takes to be a good mentee?

07:48 Curiosity. Humility. There’s just a lot of stuff that you don’t know and a willingness to try things. The people that I’ve mentored that I think have gotten the most value out of it are the ones who said basically OK I want this mentorship because I think I should be doing something different then I am doing. And are they actually willing to go out and do that different thing? If you are not willing to do something different than you are currently doing, it is very hard to get something out of mentorship from anyone, no matter how good that mentor actually is.

08:29 What are the two things that sets apart a great data scientist from just a good data scientist?

08:35 So when I think of like good data scientist versus great data scientist. Like a good data scientist indexes to value, understand what the business is trying to do and can deliver outcomes.

08:48 Like a great data scientist, like has creativity, and judgement like wrapped up in that package so that means that number one not only do they understand what the business is trying to do, they can also can map that on to the relative priority of like what should they be working on. So they are able to identify like alright, out of these ten possibilities, these two problems are by far the most important ones and that’s where they are going to spend their time and you know the other eight don’t even think about it again. And they can also like look at a set of methods or a set of proscribed solutions to those kind of problems and basically know OK like this is solved and I’m going to copy here and listen to the expert here. Like in this area, this doesn’t look right to me. And they can take down the reasons from the first principle and figure out like where do I need to be contrarian versus like do all the things that the papers say that I should do. A great data scientist exits the rubrics.

10:02 Alright. Now is the time we’ve all been waiting for. We are going to have some lightening talk questions. So first, where is a good place to go to get a clean dataset online?

10:12 For me, I don’t believe in clean data. If you are going to do a project, go find the dirties data you can find and go like go use that.

10:22 Perfect. Dirty data. Okay. SQL, Python or R, if you can only have one?

10:27 Python.

10:28 If you were not a data scientist, what would you be Ian?

10:31 I would be a policy economist.

10:34 So if you can go back in time and tell your young data scientist self something. What would you go back and say?

10:40 Don’t worry about your job title. Worry about the company that you work for.

10:45 Yeah. I’d love to open it up if you have anything to add?

10:48 for the past six months or so I’ve been doing we call it a podcast data science adjacent podcast that kind of goes through the softer side of data science so like beyond the technical skills like how you reason through problems Al those sorts of things. It’s called small differences. You can find us on Spotify, on iTunes and on Twitter at ofdifferences.

11:16 Well Ian, thank you so much for joining us today. We really appreciate it.

11:19 Thank you for having me.

https://www.youtube.com/watch?v=kZRX4wldSZM

Real Talk with Facebook Data Scientist

00:00 I think the whole league is starting to

00:02 understand that data science is super

00:03 crucial to performing well in basketball

00:05 and as a result, I think almost every NBA

00:09 team has their own data scientist trying

00:11 to help them out.

00:15 Hi, I'm Henry from Springboard and I'm here

00:18 today with Jeevan, a data scientist

00:20 at Facebook. Thanks so much for joining.

00:22 Thanks for having me.

00:24 So first question is how did you become

00:26 a data scientist? Can you take us through

00:28 your journey? I never actually planned on

00:31 becoming a data scientist I went into

00:34 college wanting to do Pre-med and I

00:37 actually ended up changing my major

00:38 three or four times. And as I was just

00:41 exploring classes and taking classes I

00:43 thought were interesting, I ended up

00:45 taking an intro data science class. And I

00:48 just found that that class was so much

00:50 more engaging than the other things I

00:52 was taking. And so from then on, I was

00:54 just trying to take more and more data

00:56 science classes and as I got more

00:58 interested in there, I started looking

00:59 for jobs and I ended up getting an

01:01 internship. And then after having like a

01:04 summer of doing that work was when I

01:05 really realized this was something I

01:06 want to pursue. Nice. And what

01:09 specifically did you like about data

01:10 science? I really liked that in data

01:13 science were like we make a hypothesis

01:15 with some prior knowledge. We try to

01:18 assess how can we test this hypothesis,

01:22 what data can be used, what methods can we

01:24 used. And then even after making all

01:26 those plans, once you start things change

01:29 and just very reactive and it's just an

01:31 exciting process start to end. And what

01:34 was some of the interesting problems

01:35 that you were able to solve? Whether through

01:38 class or at your internship? One of the

01:40 products I did in a summer internship,

01:43 that's Sonos, which is a smart speaker

01:45 company, was trying to predict if

01:47 customers would buy a second pair of

01:49 speakers after purchasing their first. So

01:51 we were able to look at things like does

01:54 the customer have like a premium music

01:57 streaming service like Spotify or or

01:59 SoundCloud. And if they did, they were

02:01 just much more likely to buy it. And we

02:03 use that to help better market the

02:05 speakers. And does Facebook have a

02:07 program for undergrads doing data

02:10 science? Yeah,Facebook actually hires a

02:12 lot of undergrad data

02:13 scientists. Super interesting because

02:15 most of the data scientists that we've

02:17 spoken with have graduate school

02:20 degrees. How does that work for someone

02:22 with a bachelor's degree? I think the

02:25 trend that data scientists need a

02:27 graduate degree is starting to slowly

02:28 die down and the reason is that

02:30 undergrad programs are starting to offer

02:32 data science classes. In the past, they

02:36 weren't as prevalent and so people who

02:37 needed to go to grad school to learn

02:39 these skills, but now you can learn

02:40 everything you need in undergrad and

02:42 with like so many free online tools as

02:45 well. So I feel like data science is like

02:47 not a career just for graduate students

02:49 anymore. What kind of classes do you take

02:52 as a data science major? As a data

02:54 science major, most of the classes you'll

02:56 take are in computer science and

02:57 statistics. But we also have this thing

03:00 called the domain emphasis, where you

03:02 just pick some other field that you're

03:04 really interested in, say biology or

03:06 psychology or economics, and you take

03:08 classes and those too with the hope that

03:10 you take your data science knowledge and

03:12 apply them in these fields. After you did

03:14 your internship at Facebook, you returned

03:17 for senior year and then went back to

03:19 Facebook what was that process like? Was

03:21 it pretty smooth? Or did you have to

03:23 interview again? No the process was

03:25 smooth. We got our return offers at the

03:28 very end of our internship so over my

03:31 senior year I didn't really have to

03:32 worry about recruiting or anything. I was

03:34 just like able to take classes I liked.

03:36 Now that you're at Facebook, what are you

03:38 working on now? My team at Facebook is

03:40 called Auction & Delivery. We're very a

03:42 cross-functional team. We work with a lot

03:44 of people in sales and business on ads.

03:47 We work with teams all across Facebook

03:49 and marketing, business, sales to maximize

03:54 the value that users and advertisers

03:55 both get on the platform. In terms of the

03:58 day to day work, what do you enjoy the

04:00 most about data science? Day-to-day

04:03 something I really like is that data

04:05 science isn't something you do by

04:07 yourself. I'm constantly working with

04:09 other data scientists on the team to get

04:10 their perspectives or people in other

04:12 departments that aren't data scientists

04:14 at all because they influence how we

04:16 think about problems and how we might go

04:18 about solving them a lot. So it's very

04:20 collaborative and I think that makes it

04:22 really fun.

04:23 On the flip side, what would you say is

04:25 the most challenging? A lot of projects

04:27 you take on in data science aren't very

04:29 well scoped. So you might take on a

04:32 project that has like a 20 percent

04:33 chance of success and you don't know is

04:36 this even possible to solve until you've

04:38 already spent so much time on it. But on

04:41 the flip side, that's also like it's it's

04:43 pretty exciting to just like be on your

04:45 toes and try to adapt as the problem

04:47 evolves. When you're working on these

04:49 problems, how long are these projects

04:51 usually before you realize maybe it's

04:54 not worth continuing? I think that can

04:56 vary a lot depending on how you want to

04:59 like plan your projects out. Some stuff I

05:02 worked on lasted weeks, months and I know

05:05 of projects that have lasted years. What

05:07 advice or tips would you have for

05:10 someone who is currently in college that

05:12 who wants to become a data scientist but

05:15 don't have the data science major to

05:18 pursue? If you're new to data science and

05:21 looking to add some stuff to your resume,

05:24 I think Kaggle is a great tool you can

05:26 go on and make your own little notebook.

05:29 You don't have to do a machine learning

05:31 problem although those are there.

05:32 You can just do exploratory analysis

05:34 on their datasets and they have so many

05:37 datasets that you can find something

05:39 that like you have some domain knowledge

05:40 and you're interested in. And then from

05:43 there, you know, make a couple notebooks,

05:44 put it on your resume, try to

05:47 showcase different techniques in them.

05:48 One great thing you can do is look at,

05:50 for a particular job you're applying for,

05:52 if they want say they really are

05:56 interested in exploratory analysis then

05:59 you can go use that skill in a Kaggle

06:01 dataset and then put that on your

06:03 resume and then just like match the

06:05 skills that you can do with what they're

06:07 looking for. In terms of your personal

06:09 growth, what do you see yourself running

06:12 next? I think a lot of new data

06:16 scientists are sort of given a project

06:19 and expect you to execute and the next

06:22 big step to take is figuring out how to

06:24 scope your own projects, figuring out

06:26 what areas, what problems, can we solve

06:29 with data science. I think a great way to

06:31 improve on that is

06:34 you look at the projects that are going

06:36 on and try to talk to the people who

06:38 came up with it and get through their

06:41 thought process. According to your

06:42 LinkedIn profile, it seems like you

06:44 play basketball right? Yeah, I do.

06:47 Can you hypothetically use data science

06:49 to be a better individual basketball

06:52 player or help a team won more games?

06:55 Yeah absolutely. I think the whole league

06:58 is starting to understand that data

06:59 science is super crucial to performing

07:01 well in basketball and as a result, I

07:04 think almost every NBA team has their

07:06 own data scientists trying to help now I

07:09 actually have friends that were data

07:11 science students at Cal and they went on

07:13 to become data scientist for NBA teams.

07:15 And they helped them out with you know

07:17 picking the most efficient plays, trying

07:20 to value players properly. Who was like

07:23 the pioneer in basketball analytics?

07:26 I think the pioneer basketball analytics

07:29 is Daryl Morey. So there's actually a

07:31 term coined after the pioneer of

07:33 basketball analytics called

07:35 Morey ball, which is a Daryl Morey's

07:37 method of playing basketball and he had

07:40 the Houston Rockets shooting

07:41 three-pointers just way before everyone

07:44 else valued them and soon after they

07:46 noticed how effective it was, just a shot

07:49 with the better expected value each play,

07:51 all the other teams started joining

07:53 along. And now you see that data science

07:55 has changed the entire landscape of the

07:57 league. Now his teams are moving so much

07:59 towards the three-pointer and the open

08:01 layup and you don't see teams taking

08:03 long tubes anymore. There have been

08:04 stories in practice where a long tube is

08:06 actually like negative points because

08:08 it's just such a poor efficiency play.

08:10 That's the way data science has

08:11 influenced the game so much. So I had a

08:13 friend who was a data scientist on the

08:15 Brooklyn Nets and part of his work was

08:18 trying to figure out a way to value

08:21 their players and compare that to how

08:23 the players are evaluated in the free

08:25 agency market. You know if there's some

08:27 discrepancy there. If you value a player

08:30 less than the market values them, then

08:32 you can get positive assets from that. In

08:34 today's NBA, if the player wants to

08:37 maximize that value, what kind of skills

08:39 should they be working on?

08:42 that's actually an interesting question

08:43 because the best player right now

08:47 doesn't necessarily mean they perform

08:50 the best on how we evaluate players

08:53 right now. Because one thing is it's very

08:55 hard to measure how well a player is

08:57 defensively. One really important metric

08:59 to evaluate your best player is their

09:01 true shooting percentage. It's an

09:03 advance that became a lot more popular

09:05 recently as players started shooting

09:06 more free throws and three-pointers

09:08 because those are so much more valuable

09:10 than a regular two-point shot. What's

09:13 your favorite team right now? My favorite

09:15 team is the Houston Rockets. Why do you

09:17 think Carmelo Anthony is out of the league?

09:20 I think Melo is out because he doesn't

09:24 have his hoodie on. And if he puts that

09:26 on, he'll be back in MVP status. I think

09:28 Carmelo Anthony is sort of a great

09:32 player in the old league where we really

09:35 did value those two-point shots a lot

09:36 but now as we found out three points and

09:39 free throws are so much more valuable he

09:41 just doesn't have the efficiency that a

09:43 lot of other players bring. And last

09:45 question is what is your life motto? I

09:48 think a motto that's been really

09:50 important in my life so far is just to

09:53 not fear failure. You need to be able to

09:56 like take on some bigger project that

09:59 like there might be a high chance of you

10:01 not not accomplishing it, but that you

10:05 can't let that stop you and you just

10:06 have to go for it anyway. Awesome that's

10:08 great advice. Cool thank you so much for

10:11 sharing your experience about data

10:14 science. Thanks

10:15 Had a great time. Hey I'm Eliza from

10:17 Springboard. We're an online school that

10:19 gets you hired. All of our courses come

10:21 with a job guarantee one-on-one

10:23 mentorship and real-world projects. We

10:26 teach UX design, data science and machine

10:28 learning engineering. To learn more, check

10:30 out the links below. Happy learning!

https://www.youtube.com/watch?v=KOU2Zu3h7zM

Real Talk with Uber Data Scientist

00:00 Bikes and scooters are really fun. First

00:02 of all, it's a brand new product to just

00:05 not us, but also to the world. Hey there,

00:12 I'm Molly from Springboard and I'm here

00:14 with Shae a data scientist at Uber. Shae

00:16 thanks so much for being here today.

00:17 Thanks for having me. So let's start from

00:20 the beginning. How did you become a data

00:21 scientist to Uber? In college I studied

00:24 statistics and then computer science and

00:27 after I graduated, I knew I wanted to be

00:29 a data scientist somewhere.

00:30 Uber was my top choice because I used

00:33 the product a lot and I felt like I had

00:35 a really good understanding of what it's

00:37 like to be a customer. And number two I

00:40 just was always intrigued by the product

00:42 itself.

00:42 I loved sometimes seeing the cars little

00:45 cars move around and when you booked a

00:47 ride they're always so so on point with

00:49 you know the ETA and with the amount

00:52 charged. And everything was just working

00:54 so smoothly and I wanted to know what

00:56 was the mechanics behind it and I felt

00:58 like joining the company would help me

00:59 get there. So within Uber what teams do

01:02 you work on? Right now I'm working on the

01:04 bikes and scooters team and previously I

01:06 was on driver.

01:06 Oh cool! How do they differ? The biggest

01:09 difference between the two teams is that

01:11 on bikes and scooters, we not only have

01:13 mobile data, but bikes of scooters all

01:15 have hardware on there that capture a

01:19 lot of IOT device like data a lot of

01:22 firmware data. All of these different

01:24 data could add another dimension to how

01:27 you can understand the users' journey.

01:28 When they're actually on the scooter or

01:30 bike, what exactly is their experience.

01:32 Because the entire scooters and bikes

01:34 experience extends from booking it and

01:38 the entire trip and ending it and then

01:40 reusing it and being able to

01:42 consistently find it on the street.

01:43 What's it like working with bikes and

01:46 scooters? Bikes and scooters are really

01:48 fun. First of all, it's a brand new

01:51 product to just not us, but also to the

01:54 world. And I would say that you have to

01:58 deal with data from a lot of different

02:00 sources. For example, right now it's

02:02 you're not just dealing with customer

02:04 data anymore. You're looking at an IOT

02:06 device and you're looking at mobile. And

02:08 you're looking at the you have to use

02:10 all this data to tie together

02:11 users journey from where they're trying

02:14 to go somewhere and they're trying to

02:16 find a way to get there. And they're

02:18 choosing to use this product and then

02:20 the entire on trip experience till

02:22 afterwards and how do you repeatedly

02:24 engage them. You have to synthesize an

02:27 understanding of what is actually

02:28 happening to a user and what do they

02:30 really want out of your product. So

02:33 that's a really challenging part is

02:35 combining these different types of data

02:37 together. What are some exciting data

02:39 principles that you're able to apply

02:40 every day at work? I think the most

02:42 exciting part is when you run

02:44 experiments sometimes you run into

02:47 network effects. Sometimes you run into

02:49 cannibalization and you have to think of

02:51 creative ways to address these problems

02:52 while keeping your experiment results

02:54 very robust. One example is say you want

02:57 to run promotions on a set of bikes and

03:00 scooters for a certain area and you want

03:04 to see if that increases how many trips

03:06 people are actually taking. Now by

03:08 applying the promotion, you take away

03:11 more supply for from the control group

03:15 if the treatment group actually do end

03:17 up taking more bikes and scooters which

03:19 might lower your conversion rate for

03:21 your control group. Basically how do you

03:23 make sure that your environmental

03:24 controls are actually controlled. What

03:27 excites you about working with data

03:28 science everyday? I think the most

03:30 exciting part about being a data

03:31 scientist is being able to derive

03:34 insights and make decisions from

03:37 incomplete information. Every day you are

03:39 making decisions and you never know if

03:42 they're right or wrong. And you never

03:44 have complete information. But data

03:46 science actually quantifies that sort of

03:48 uncertainty and I think it's a very

03:50 empowering experience to be able to help

03:52 businesses make decisions and move

03:54 forward, even when they have incomplete

03:56 information. And then in contrast but

03:58 something that you might not like as

03:59 much about being a data scientist? I

04:01 think one thing that all data scientist

04:03 hate the most is messy data that they

04:06 have to clean up and somehow debug. I

04:09 encounter this in my day to day work and

04:11 you have to write a lot of sequels and a

04:13 lot of that piece of the job is

04:16 unavoidable but I think it's also nice

04:18 to always see that as an opportunity to

04:21 improve the platform and infra. It's a

04:25 good opportunity for people to say hey

04:26 maybe I want to contribute a piece of

04:28 open-source code to write something that

04:30 could help make other people's lives

04:32 easier or suggest something to your

04:34 company to build internal tools or even

04:36 build it yourself. So would you mind

04:38 speaking to your experience first being

04:41 in a formal educational setting and then

04:43 immediately going into the workforce?

04:44 What was that like and do you have any

04:46 advice for people who want to have a

04:47 similar career path as you? I think it

04:49 was very exciting and overwhelming at

04:51 the same time. The things you learn in

04:53 school are all printed on a textbook and

04:56 a lot of times they're outdated sort of

04:58 old material. And in the workforce you

05:01 realize that people are using new

05:02 methods. They're using experimental

05:04 methods and people are constantly on top

05:07 of new papers that are published.

05:08 Especially in the world of data science

05:10 where new algorithms are developed every

05:12 year at conferences, I think one

05:15 piece of advice I would have is to stay

05:17 on top of news and stay informed, go to

05:20 conferences, meet people because data

05:23 science is not a super mature field

05:25 and there's always opportunities for

05:27 improvement. You should never be afraid

05:28 to try new algorithms yourself or even

05:31 try to develop a new method. Cool!

05:33 What's your favorite podcast? My favorite

05:36 podcast is Seneca it's actually a

05:40 podcast that describes present-day

05:42 Chinese politics and economics and

05:46 everything about China that you need to

05:48 know. I like to stay informed since I

05:51 left China when I was 14 so it's a good

05:54 way to learn about how the macro is

05:56 going in my home country. If you could

05:58 travel back in time, what decade or

06:00 period of time would you go back to? If I

06:02 could also choose where I can be, then I

06:06 would say that I want to travel back in

06:08 time to work with Alan Turing. Love the

06:12 movie imitation game. I think that was a

06:14 main game changer for me in college when

06:17 I was trying to decide what I wanted to

06:18 do. Well I think that's all we have for

06:20 today. Thanks so much for joining us

06:21 Thanks for having me.

https://www.youtube.com/watch?v=stz12TB72XI

Real Talk with Reddit Data Scientist

00:00 And then I got to apply it to some

00:01 really fun stuff when I was like testing

00:03 the metric out like I got to answer

00:04 definitively do people care more about

00:06 dogs or cats.

00:11 Hey I'm Eliza from Springboard. I'm here

00:13 with Katie a senior data scientist at

00:15 reddit. Thank you so much for being here.

00:17 Thanks for having me. How did you become

00:19 a founding data scientist at Reddit?

00:21 Take us through your journey. The story

00:23 is probably not super exciting. I

00:26 mean it could be. But I had been working

00:29 as a data scientist in EdTech for a

00:30 while and a recruiter reached out to me.

00:32 I interviewed I actually interviewed for

00:35 a different position and I ended up

00:36 getting I was originally gonna be

00:38 working on like trust and safety type

00:39 stuff but then ended up moving on into a

00:42 more consumer product role and been

00:45 there ever since.

00:46 When I started we had all of our data in

00:49 hive, which if you've ever worked with

00:51 the hive database, it's it feels old. You

00:55 run a query and it takes three hours. And

00:58 after three hours, you might have like

00:59 spelled something wrong and you get

01:00 nothing back. So like that that was

01:03 something that when I first started felt

01:05 like a major challenge, but we pretty

01:06 quickly started migrating all of our

01:08 data into the Google cloud ecosystem. And

01:11 like that was a big challenge in and of

01:12 itself, but like once we got there we had

01:14 to start building our actual analytics

01:16 infrastructure. So there was a lot of

01:18 like putting stuff into ETL, figuring out

01:20 what are the business questions people

01:21 actually care about. The process of doing

01:23 that revealed like data quality problems

01:25 that we had that then resolved or figure

01:27 out how to like smooth over that and

01:28 make data easy to use. And then there's

01:31 also just a lot of like coaching people

01:33 and being able to think well with data.

01:35 It's it's definitely not intuitive. If

01:38 you've ever studied statistics, there's

01:41 so many like mental optical illusions I

01:45 guess that doesn't really you know like

01:46 the right way to describe a purely

01:49 mental thing, but like it's really easy

01:50 to have your intuition hijacked and to

01:55 think you know something and then find

01:57 out in the data that you don't. So there

01:59 was just a lot of work around getting

02:01 people to think data well and it's

02:04 something that we still do today and

02:05 probably will always have to do cuz

02:07 that's just what the life of a data

02:08 scientist is. Yeah it's a lot of

02:10 troubleshooting I imagine.

02:11 There's always something. What's been one

02:15 of the coolest projects that you've

02:16 worked on? I've gotten to do a lot of

02:18 really interesting stuff. I've touched

02:21 most parts of the product at this point,

02:22 but probably one of the most notable

02:24 ones was something that I did right when

02:25 I first got to Reddit, which was build

02:27 our time on site metric. That's a pretty

02:29 foundational metric in web analytics, so

02:32 it's a very high impact high profile

02:33 thing to get to work on and it was very

02:36 challenging the actual methodology

02:38 we used was pretty simple, but there was

02:41 a lot of edge cases, a lot of things that

02:43 as we dug deeper into the problem, we

02:45 found out like it's actually really hard

02:47 to figure out. When something happened in

02:49 a distributed system, which timestamps

02:52 do you trust? What do you do

02:54 when something an event like gets

02:57 through your pipeline like four days

02:59 after it's supposed to and it looks like

03:02 your session lasted four days. Like

03:04 that's probably not a human thing but

03:07 you have to figure it out somehow.

03:09 So it was a really interesting challenge

03:11 just in terms of like figuring out what

03:14 were all the boundaries and edge cases.

03:15 And then I got to apply it to some

03:17 really fun stuff when I was like testing

03:19 the metric out like I got to answer it

03:20 definitively do people care more about

03:22 dogs or cats. What was the

03:25 answer? Cats. Hurts me a little bit as a

03:27 dog lover myself, but I also I can I

03:30 controlled four different types of

03:31 animals subreddits and what's wrong with

03:34 your dog was far higher in time on sight

03:36 than what's wrong with your cat. So maybe

03:38 people are just worried about their dogs

03:39 malfunctioning. I don't know. Cats are

03:41 all individualistic. They don't need us

03:43 anyway so that's true.

03:44 Dogs have friends. Cats have staff. So what

03:51 are some data scientists specializations

03:53 and how do they differ? That's a very

03:56 good question to ask. Data science is

03:58 it's a newer profession. It's definitely

04:01 come a long way in terms of like people

04:03 have a pretty good idea of like what the

04:05 skill set of data scientist is and like

04:06 data science job description they're way

04:08 less insane than they used to be even a

04:10 couple years ago. But again like it means

04:13 different things in different places.

04:14 Sometimes the data scientist is more of

04:16 like a turbocharged analyst like they're

04:19 doing a lot of like very just ad hoc

04:21 analysis but like with sophisticated

04:23 methodology. So they might be like

04:25 modeling the like data generating

04:27 process of some like retention metric or

04:30 something. Or they might be like doing

04:33 like really gnarly A/B testing analysis.

04:36 I would consider that kind of like a

04:37 decision science I guess

04:39 specialization for data science. That's

04:42 that's one of the major ones I would say.

04:44 And then there's also people who are

04:46 really what I think is starting to get

04:49 called machine learning engineers. And

04:51 that's a lot of like you're putting a

04:53 system into productions like that's very

04:55 engineering heavy data science work and

04:57 it might involve actual modeling. It

05:01 might not. It might be like you're more

05:03 of like someone who's building in for a

05:05 bit like you need to know how the models

05:06 work to make sure you're implementing it

05:08 correctly. And yeah those are those are

05:12 probably the two main ones that I see,

05:14 but then there's also data scientists

05:17 who are kind of generalists. There are

05:19 data scientists who really are sort of

05:21 like lightweight data engineers like

05:23 mostly doing like ETL. It really varies a

05:25 lot. Would you consider yourself to be

05:28 more a generalist or a specialist? Am I

05:30 hurt I want to say a generalist. They're

05:32 definitely like some areas where I've

05:33 gone pretty deep I know a lot about

05:35 natural language processing. For example,

05:37 I wouldn't necessarily say I'm like a

05:39 natural language processing specialist,

05:40 though I definitely see myself as more

05:42 of a generalist. You currently manage the

05:44 data science internship at reddit.

05:46 Can you tell me a little bit more about

05:48 that? So the first time we did it was

05:49 last year. It was in partnership with

05:51 the University of San Francisco master's

05:53 of data science program and it was

05:56 really cool. We got to interview the

05:59 candidates. We got to spend several

06:03 months with them. It was really more of a

06:04 fellowship, so we had them from I think

06:07 October to May of this year so it was a

06:11 very very long project to work with them

06:14 for all that time. And it was an

06:18 interesting experience in just like

06:19 making a company a hospitable place for

06:21 a junior employee. A lot of companies

06:25 they they hire senior people not because

06:27 they have senior problems but because

06:29 they can't or maybe don't have time to

06:32 create structure to allow a junior

06:35 person who maybe doesn't really end

06:37 and how to negotiate like requirements

06:39 or deadlines. They can't

06:41 make that easy so they hire senior

06:44 people who know how to do even if the

06:46 work is not for senior people. So having

06:48 an internship program is

06:50 very much oriented around cultivating

06:53 that type of environment so when you

06:55 actually have interns with you. You spend

06:59 time like clearing obstacles for them

07:01 like making sure like they can just do

07:02 their work. Communicating for them like

07:05 telling people what they're doing so

07:07 like they might be able to help them or

07:08 pick up their work and use it. And when

07:10 they're not there and sort of the

07:12 offseason it's about like finding

07:13 projects that you can curate into intern

07:16 acceptable projects, so finding things

07:18 that are high impact or at least like

07:21 medium impact but also like not so hard

07:24 that like someone who is not super

07:25 experienced can't do or also like if it

07:29 doesn't get done like it won't be the

07:32 end of the world. It's like a very hard

07:34 balance to strike but making sure to

07:35 keep your eyes peeled for projects like

07:37 that makes a company a great place for

07:39 interns and junior employees more

07:41 generally. I think a lot of data

07:43 scientists end up with grad degrees and

07:46 I was wondering what's been your

07:48 experience with a bachelor's degree? how

07:50 did you break into the industry? I've

07:51 often joked that the most direct career

07:54 path into data science is being a jaded

07:56 physics PhD. I don't think that'll be the

07:59 case forever especially because a lot of

08:00 colleges are starting to have undergrad

08:02 degrees for a data science so this might

08:04 not even be true for people in the

08:06 future.

08:07 But for me, breaking in I think it took

08:10 a little bit of convincing. I had been an

08:13 analyst previously so like I certainly

08:15 had like relevant skills and like I

08:17 studied a quantitative social science in

08:21 school, so like I know the stats stuff.

08:23 But there's just something about

08:24 credentials that I think a lot of people

08:26 they trust. It's just like a filter that

08:30 they can use, so I had to be very

08:32 aggressive in my applications. I would

08:35 apply for things that it's like I don't

08:36 even know if I'm qualified for this. I

08:38 don't even know if I want to do this. But

08:40 I will see what happens by applying

08:43 and like I would have recruiters reach

08:45 out with like other analyst positions

08:47 and I'd be like I'm overqualified for

08:48 that I want to go to a data science position

08:50 and eventually it worked. It took

08:53 a lot of persistence and I was

08:55 frustrated a lot, but eventually I got

08:58 someone to give me a shot and now that

09:00 I've been a data scientist like it

09:02 doesn't matter. The work experience

09:04 matters a lot more than a background in

09:08 any particular field. I think that's true

09:10 really for any job in tech. People want

09:12 to know if you can do the job. What is

09:15 some advice that you would give to

09:16 someone aspiring to be a data scientist?

09:19 I would say that it is a job that will

09:25 always make you learn and that can be

09:28 both good and bad.

09:29 Sometimes it's exhausting where it's

09:30 like there's a new technology I have to

09:32 learn like man I have to like learn the

09:34 architecture of BERT now. I just learned

09:36 what an LSTM was. There's so many new

09:38 things constantly changing that might

09:40 slow down eventually, but right now the

09:42 field is so new and it's growing so fast

09:43 that it's like it feels like you can

09:45 never learn enough. And like as a part of

09:49 like be prepared to always learn new

09:51 things like first of all like learn

09:53 things you're excited about because

09:55 you'll want to do it. And that'll make

09:57 work pleasant if you're digging in deep

10:00 to something that you already care about.

10:01 But also like don't stress about knowing

10:05 everything because it's impossible and

10:08 technologies and techniques and

10:10 frameworks they all pop in and out of

10:13 existence all the time. Like you could go

10:14 super deep on something and then it

10:16 would just get deprecated so like focus

10:18 on what you're interested in what you

10:20 need to know for your job. What do you do

10:21 for fun in your free time? I really like

10:24 baking. I make a mean loaf of sourdough

10:27 bread and I'm currently doing a great

10:30 British Bake Off bake along challenge

10:31 with some different people. And I usually

10:34 do not win on aesthetics but my flavors

10:37 are pretty good so I'll take it.

10:38 Have you had like a what is it's like a

10:40 sourdough starter right that you have

10:42 have you had it going for a while? My

10:45 sourdough starter is four years old. Oh

10:47 it's like going off to kindergarten. Yeah maybe I

10:52 don't know. Maybe it's precocious. It's

10:55 made some good loaves of bread. What do

11:00 you want to learn next either in your

11:02 career or

11:03 just in life? Something that I really

11:05 want to do is strategy for for data

11:09 science and machine learning. How people

11:12 figure out like what's the business

11:13 model what makes sense for what we're

11:14 doing here.

11:15 And how can we align it with like the

11:18 people who are actually working on it so

11:20 that like they feel enthusiastic to do

11:22 it. So you just said you were on vacation.

11:24 Where do you want to travel to next? I

11:27 would love to do a road trip across the

11:29 US. I did it when I was a super little

11:32 kid so I don't really remember it and I

11:34 feel like I would feel really

11:35 differently about it now so I'm just

11:37 drive across the US. Like dude the

11:39 national parks are just kind of all over

11:41 the place I mean it would definitely be

11:43 better like if I could step off

11:44 everything. Just like I need to go across

11:46 the US in seven days. No yeah no I

11:49 would I would love to visit national

11:51 parks. I love being outdoors and hiking

11:53 and just seeing all the different things

11:55 that are out in the world so. Thanks yeah.

11:58 The last question what is your favorite

12:00 motto? Never attribute to malice what can

12:03 be attributed to incompetence, which

12:07 maybe seems a little dark but no I

12:10 usually I feel like people are not

12:11 actually trying to screw you over they

12:13 just like kind of forgot what they were

12:15 doing or weren't paying attention or

12:17 something if you just talk through

12:19 misunderstandings, usually you can

12:20 resolve them. So I think that's all the

12:22 time that we have for today. Thank you so

12:24 much for coming in and telling us all

12:25 about your data science experience. Yeah

12:27 thank you. It's been super fun.

12:29 Hey I'm Eliza from Springboard. We're an

12:31 online school that gets you hired. All of

12:33 our courses come with a job guarantee

12:34 one-on-one mentorship and real-world

12:37 projects we teach UX design data science

12:40 and machine learning engineering. To

12:42 learn more, check out the links below.

12:43 Happy learning!

Randy Au

Succeeding as a data scientist in small companies/startups

It’s nothing like at a big mature company.

This’ll probably be an unbounded series of posts that spawned from this question that came across the awesome community that is the data-nerd twitter cluster:

Some Background

I’ve spent almost 12 years now at companies sized between 15–150 wearing various hats of data analyst, engineer, and occasionally, scientist. Wandering into mega-corp Google Cloud as a UX researcher is a bit out of character, but with new products constantly being churned out, it feels like a startup in terms of questions and chaos, despite the billions of gross revenue involved.

I’ve worked in a mix of businesses from interior design (office design), ad-tech, social networks, link shortening, e-commerce, and now enterprise cloud. I come from a social science background, with a dash of NLP, applied math, and business administration. I mostly live in back end systems, logs, and SQL.

In a word, I’m a generalist, the sort that people seem to recommend for startups, and that’s where I’ve thrived my whole career.

Those are my biases, and the experiences I draw from. If you’re in a startup where AI/ML is literally the core foundation of the business, I’m likely irrelevant to your needs. YMMV.

Being the 1st “Data Person”

I’ll go ahead and say it, small companies don’t need a data scientist, but they need a “data person”. They might call the job “data scientist/engineer/analyst/ninja”, whatever.

My experience is that somewhere between 20 and 60 employees, there’s enough customers, accumulated data and role specialization that the need to bring in someone who can use data to give useful business insight starts to justify the cost of hiring someone. Until then, they make do with the skills available.

The job title can be almost anything, but the job descriptions tend to be various mixes of:

Make sense of the data we have

Help build out our data systems

Help us be data driven/ run experiments

Grow the business

Education/Certifications that may or may not be relevant to anything

It’s usually quite likely that they don’t really have a full understanding of what they need. There’s just a generalized sense of “we have data, it seems useful, but we don’t have anyone who has the skills to make it useful.”

In practical terms, there are 2 big things someone taking this position needs to do in parallel:

Help the company succeed TODAY

Set up the company to be data driven TOMORROW

Helping the Company Succeed TODAY

Startups are surrounded by uncertainty. They’re not sure who their customers are, production systems can be dodgy, they don’t know what their customers are doing with the product, they don’t know how to make decisions using data they have, they don’t know if the data they have is useful.

Smart answers to the questions lead to smarter decisions and hopefully that mythical hockey-stick growth everyone dreams of. The problem is that most of those issues don’t lend themselves to fancy methods. The useful ones are often a century old and/or based on qualitative methods instead of quant.

Most DS methods are most powerful when optimizing an existing process, they’ll get you, 5%, 10%, even 25% growth on things like customer acquisition, conversion, retention and spend. A/B testing, recommendation systems, ML classifiers, all of them help to optimize. The gains are real, quantifiable, and can be significant, but early on there’s likely bigger fish to fry.

The biggest impacts early on often involve insights. Insight changes what the company fundamentally does. They come from very pedestrian things like research into user preferences/behavior that uncovers a new marketing concept for sales folk, or helping product teams realize that the most vocally hated feature on Twitter is actually used by 90% of paying customers and they shouldn’t drop it for no reason.

My view of this “help the company now” role is that the data person is a force multiplier. People within the business have problems, the job is to help them solve them.

Being the 1st “Data Person” = Being a “Scientist with data”

Being a scientist to me means that you have a problem, a research question, and you use whatever methods you can to come to a solid answer to that question.

As data scientists, we tend answer questions using quantitative methods and data collected from systems, but that’s not the only path to insight. Sometimes you flat out observe or ask users (qualitative methods), or you go out and collect data (experiments and surveys), or you watch others (competitive analysis). A good scientist doesn’t define themselves by their methods, and neither should the 1st data person (or any subsequent data person).

The goal is to answer the pressing business needs: “Why is no one using our product?” “How come our returns are so high?” Should we be running this expensive sale or not?” “What drives customer churn?” “What’s a customer’s lifetime value, what drives that?”

Becoming Data Driven TOMORROW

A common trap I see are people who come out of Data Science programs joining these positions expecting to be using sexy things like Spark and applying RNNs to their work. But sadly, they want to live on top of a mountain of foundation work that needs to be done first, both from an engineering standpoint and from cultural standpoint. The mismatch is brutal.

Fancy “Data Science” methods rely on a ton of things, DO NOT expect each layer to be “done” before moving to the next. Think of the colors as “time spent”.

Being the first person specifically hired to handle data, it’s very unlikely any pieces of the pyramids are sturdy. It’s a multi-year, cross-functional, full company effort to get all the pieces in place. Nurturing those all those pieces in parallel is a big part of the job.

Note, in a typical business, you’ll be attempting to do things up and down the pyramid at the same time, regardless of the stability of underlying layers. I’ve built plenty of dashboards and classifiers against fragile new systems, and you will too.

There’s room for tons of posts about aspects of all this stuff, but here’s a rough overview of my thoughts.

Solid Production Systems and Engineering Practice

This is a thing that’s firmly in engineering’s wheelhouse, but you can’t really measure the behavior of a system if the system is buggy and broken, so expect to help out here as needed.

The more of a “data engineer” hat you wind up wearing, the bigger a role you can play in helping build solid systems. People will naturally ask your input for questions like PostgresSQL vs MySQL, AWS vs GCP, Spark vs Redshift, etc. Helping with those decisions adds lasting value. Expect to have to set up systems and run them yourself if there’s not enough Eng resources.

I also find that pushing engineering to instrument systems for business purposes has a nice side effect of finding interesting bugs. Once, I found duplicated IDs in a critical table that shouldn’t have dupes, that set of a bug hunt that fixed some things in a revenue-generating system. Poking at production tables often can find really weird bugs.

Reliable Instrumentation

Garbage in, garbage out.

Getting reliable instrumentation is THE most critical thing a 1st data person does. It is an endless journey, from picking frameworks (multiple) to collect system and user data, making sure engineers learn how to implement things without counting errors (which are so easy to make), making sure the databases and logs are doing the Right Thing(tm), and making sure you’re counting what you think you’re counting.

At the same time, systems will be added specifically to collect and report on data. Those will need to be put together and managed by someone, maybe you.

On the culture front you’ll constantly have everyone, from the new guy to the CEO, asking how reliable the data is and what to do with the information. They’ll ask for clarification, explanation, deep dives as they themselves learn about the system they’ve built.

This cultural training alone is a long journey in itself as you make reports and dashboards, and people find inconsistencies vs other systems and get results that don’t jive with their view of reality. Sometimes their reality is wrong, oftentimes they’re correct. Just having these conversations makes everyone involved smarter.

I honestly don’t think this phase is ever “complete”, it just gets to a point where you only have to worry about it when a new feature or big change goes out.

Reporting and Acting on Data

Dashboarding and reporting is not a sexy job, but unfortunately, it’s often the first/only way for people throughout the company to observe the health of the company on a daily basis so the investment is necessary. The goal is to put (potentially) actionable information in the hands of people who can take initiative to act on it.

In the beginning, most dashboards and reports will be manual. It takes a lot of iteration to hit upon metrics that people care enough to see multiple times. Automation is a nice-to-have, but getting insights out takes priority.

The technical aspect isn’t super complex. There are many services and platforms for generating reports and dashboards, you can even just do it with custom code. The trick is to get all the data systems to play together (Ha), have good performance, and minimize the (significant!) overhead in maintaining dashboards and reports as the business grows.

The cultural side is where things are interesting. You’re training people to become more data driven here. This also can take years of work and practice.

There is a ton of education involved here. You’ll be teaching people how to read the results of an A/B test, what significant difference means, explaining what a confidence/prediction interval is, explaining why that chart “is only an estimate because we rely on a 3rd party report and they’re flaky”. You’ll field questions about sample size (never enough), and constantly need to teach people good methodologies.

And what should people do if they’re concerned about a number on a dashboard? Personally, I tell them to come talk to me. Their concern is potentially a research question (or a bug), and that’s literally research GOLD. These people are domain experts in that part of the business, while I’m just a nerd slinging SQL.

Automation and Experimentation

Over time, you’ll set up metrics and dashboards for when new features go out. Inevitably people will be disappointed in how a feature is doing and want to know why. If you haven’t picked up pre/post analysis methods and memorized all the national holidays of your primary market, now’s a good time to do so.

Once people get used to having information and maybe run a few A/B tests to disappointing results, expect to spend a ton of time verifying that the numbers are correct when test data comes back contrary to people’s assumptions.

Despite those painful moments, we’re now finally doing Real Science(tm). Having a hypothesis and collecting data to test things out.

By now, useful dashboards should be fairly automated and people are used to using data to make decisions. You know you’ve arrived when a new feature is proposed and “how will we judge the success of this?” comes up without you pushing for it.

One other thing interesting I’ve noticed is that around this phase, a company can get TOO comfortable running experiments. They learn to design tests that they know they’ll succeed at (low risk taking), or they’ll “test” a thing that they 100% know they will launch regardless of the results (“we’ll optimize it later”).

It’s now your job as the scientist to call people out on this behavior. They can totally launch regardless of test results, radical changes often test poorly, but that should be explicitly state that intention.

Finally, Data Science

After that the long, long, long journey above and with many detours, the company itself has transitioned to being data driven. They have hypotheses, can reliably collect data, and make deliberate decisions based on the outcomes. They’re also more self sufficient, can read (and maybe create) dashboards with some guidance, and have learned when to worry and how to raise issues.

There’ll always be things to do further down the pyramid, but at least now things aren’t on fire all the time. The business has hopefully become better at understanding its place in the market and the world, and interest shifts to optimizing exiting processes for incremental (but significant) gains.

Now you can think about breaking out the fancy algorithms…

Or maybe there’s a data warehouse that needs building because you’ve got too many systems now and analytic queries can’t be sped up any further, oops.

Comments/Feedback Welcome

I wrote this whole piece in one extended sitting because the content just needed an outlet. I’ve glossed over a ton of things for length reasons. So any feedback and topic requests for future posts would be most welcome. Hit me up on Twitter.

Eryk Lewinson

My experience as a data scientist in a startup

What are the main things to expect when working as a data scientist in a “smaller” company?

We live in times of rapid data expansion, as basically any activity we perform using technology (and not only) generates some kind of data, which can be then further analyzed and used to draw insights. No wonder that Harvard Business Review named “data scientist” the sexiest job of the 21st century!

Given this brief introduction, it only makes sense for companies to use the available data to make better, more informed business decisions, while at the same time improving the customers’ experience (at least in principle). That is where the role of data analysts/scientists comes into play.

However, the definition of the data scientist — or rather the tasks a data scientist actually handles — differs a lot depending on the company, industry, etc. In this article, I wanted to provide my perspective by describing what it looked like to work as a data scientist for a FinTech startup in the Netherlands.

A product data scientist

My position could be more accurately described as a product data scientist, both because of how the teams were structured within the company and what tasks I used to work on. My role was deeply embedded within the product team, which consisted of a product owner, designers, front- and back-end developers, and more. You could say that each data scientist worked in two teams: the general Business Intelligence Team and then in the respective Product Team. In practice, that kind of a hybrid-approach to team structure worked really well and enabled efficient collaboration, while keeping everyone in the loop about what is currently happening in the teams.

After this short intro, I wanted to touch on some of the points that stood out during my time there and which I believe can be helpful for other people searching for a job in startups.

Being close to the decisions

Definitely one of the highlights of working as a data scientist for a smaller company is that you are closer to the decisions and stakeholders. This means that your projects— be it a single Jupyter Notebook, a multi-sheet dashboard, or a deployed model — can be directly used for decision-making or shaping the product (in my case, it was a mobile app). That can be especially rewarding for data scientists, as they see that their work is actually used (as opposed to some report buried in an avalanche of emails) and provides clear value-added to the business.

Domain knowledge and stakeholder management

Being that close to decisions enables the data scientists to understand the business logic and gain valuable domain expertise. This way, they can better understand the potential requirements and expectations of the stakeholders, which makes the cooperation easier and more fruitful.

As an example, it happens that stakeholders do have some goal in mind, but instead of expressing it, they already describe the means to achieve that goal, which they came up with. While such an initiative is always welcome and can be very helpful, it also happens that another approach should be followed, either because the suggested one has some disadvantages or maybe there is no such data available, etc. Because in the end, this is exactly the data scientist’s job to point the best way to achieve the goals of the stakeholders.

Data scientist = jack of all trades

When you hear the term data scientist, you might associate it with a person who is mostly training machine learning models, constantly tuning the hyperparameters to improve the accuracy/precision/recall/another score in order to impact the business. While this may be the case for bigger companies or companies with dedicated machine learning products, in my company (and similarly for other startups I heard about) the role was more of a jack of all trades.

In practice, this meant that we were working on a variety of different projects. Naturally, most of them were related to typical data scientists’ work, but some of them could easily fall under the data or software engineering label.

Regarding the diversity of the projects, while some of them did indeed involve building predictive models (for example, a conversion prediction model), others could be substantially different. Sometimes, we had to build and deploy an importer that downloaded marketing spend from social media platforms and stored it in the internal database for further reporting,

Another time, we built a detailed onboarding funnel, which could be used for optimizing the user journey, localizing the bottlenecks, and reducing the drop-off. We also planned and designed experiments and analyzed the results of multiple A/B tests. And those projects were just the tip of the iceberg.

I would say that working for a startup enables a significantly higher project diversity than in a dedicated team, for example, focusing on customer segmentation or time series clustering. Whether this is something that you would like to experience — the potential width vs. depth trade-off — is entirely up to you. For me, it was a lot of fun and I definitely learned a lot from different areas of data science. And then I documented some of those learnings in the form of Medium articles :)

Pragmatic approach

I believe that startups are one of the best places to learn the pragmatic approach to data science. That is because as young companies, they need to quickly grow, develop, and gain traction. To achieve these goals, it is often best to focus on the low-hanging fruits. Let’s say you have a conversion prediction model indicating whether a given customer is likely to convert this week or not. You can probably spend months trying to improve the metric of interest, either by creating new features, trying the latest state of the art classifier, or tuning the hyperparameters using any of the ten available libraries. Yes, that number is made-up and probably seriously underestimated.

But the real question is, will this extra percentage point or two above a perfectly good Random Forest classifier bring lots of value to the company? Assuming the company is not a behemoth like Google or Facebook, probably not. That is why achieving satisfactory results in a reasonable amount of time is so crucial for smaller companies and — contrary to what you might think right now — it is a thing that data scientists (myself included) often struggle with. When to call it quits, settle for the current solution, and proceed to another project. That is exactly when advice from an experienced manager with a bird’s-eye view can really help.

Not just the shiny things

Connected to the above, but in a bit broader context, for many people, training and tuning models is the place where all the fun is. Also, to take it one step further, everyone heard that a data scientist spends 80% of their time cleaning and wrangling the data and 20% on actual modeling.

But in practice, every day’s work might not even involve training models. Sometimes there is already a lot of value to be extracted from doing some smart aggregations in pandas or prepping a quick and efficient dashboard in Tableau to tell your story. Personally, for some ad-hoc requests, I often found that actually just working with SQL was faster and just as good as loading the data to Python for further analysis.

Is it the most glorious thing to do as a data scientist? For many, probably not. But I would argue that a lot depends on the actual definition of a data scientist. Most likely, in just a tiny % of data scientist positions out there the person only works on training and tuning the models. And for many companies, especially smaller ones, a data scientist is a generic umbrella term for a person who takes data, does some magic, and generates crucial insights for the business. For me, it was rewarding to be presented with a problem, or a task, and then having the possibility to decide which approach (simple aggregate statistics, a fancy model or somewhere in-between) would be best suited to solve it.

The data

It might seem obvious, but there are no data scientists without the data. Before starting actual work, many students and aspiring data scientists mostly play with clean datasets available online (be it the UCI Machine Learning Repository or Kaggle) and unfortunately, this is quite far away from the truth. In my company, we already had a great data engineering pipeline built by my manager, which removed many potential pain-points.

However, we still encountered quite a few issues when we wanted to add some more data (like user session info or data from external parties, like Firebase events for A/B testing). An example could be a weird sequence of events happening for some users (which in theory should not be possible), that only came out while trying to create a linear funnel or applying process mining techniques.

But I would not say that this was a bad thing in the end. Of course, it added some extra work in order to locate those weird patterns. But first, it was satisfying to actually spot the issue (sometimes it required cooperation with developers to narrow down the issue). Secondly, each case left us smarter and we knew what to look out for in the future and how to prevent similar issues from happening again. So all in all, definitely a very educational experience.

Photo by Headway on Unsplash

Fostering creativity

While I already mentioned the diversity of the projects in startups, my company was great when it came to supporting your own initiatives. When there was no pressing sprint work to be done, every now and then we could spend a day playing around with a new library or researching some concepts that might be applicable in the company. In the end, it’s a win-win situation, in which the data scientists grow and develop new skills, while the company can benefit from using a novel approach. Oftentimes, such small exploratory “things” turned into bigger and more significant projects.

In my opinion, it is great to find a company with a high focus on personal development. So you not only complete your tasks and learn on the way, but you can also try to find something worth pursuing on your own. And then, the company supports you by providing you with time or resources to follow it through.

Responsibility and ownership

Connected to the points above, in my previous company the data scientists have quite a lot of freedom in how they can approach the tasks, what tools they use, or even coming up with new ideas and exploring them to provide an MVP or proof of concept.

A person working directly on the project might have some insights into why a certain approach might be better or what kinds of analyses might be beneficial for the company as a whole. Ultimately, the company trusts that the data scientist will always do their best and provide deliverables of the highest quality. This atmosphere of trust further fostered creativity and made you want to put your best effort into each project. Because it was not only one of the twenty projects in the pipeline assigned to you this half a year but something that you could actively shape, build, and deploy.

Tech Stack

What is definitely more common for startups than big companies (especially those under lots of regulations such as banks) is using the latest and greatest tech available. Because startups often start from scratch, without years of legacy, they can quickly adopt the new standards and iterate on them.

That was also the case for my data science team. We had quite a lot of freedom in choosing the libraries which we wanted to use, even if they were just released. But something you should definitely take into account though is the library’s/framework’s maturity, as it matters a lot, especially if you want to put something into production or make a crucial decision based on it. Also in terms of software (Python vs. R, etc.), we followed the pragmatic approach and choose the best tool for the job, even if it sometimes lied outside of our comfort zone.

Conclusions

In this article, I described my experience of working as a data scientist in a startup. Of course, those opinions are subjective and are based on a sample size of 1, so naturally, there will be a lot of variability from startup to startup. But even so, probably some/most of the patterns will be applicable to many companies at a relatively early stage of growth.

For me personally, I learned a lot over the 3 years I worked in a startup, not only the skills strictly connected to data science but also other ones such as pragmatism, stakeholder management, responsibility for the deliverables, time management, and probably some more. And I am sure they will be important in my future career.

If you are interested in finding out more similarities/differences between working as a data scientist for different kinds of companies or you are just starting out your adventure, I can definitely recommend reading Build a Career in Data Science.

Did you have a vastly different experience? I would be happy to read about that in the comments, or you can reach out to me on Twitter.

In early 2020 I published a book on using Python for solving practical tasks in the financial domain. If you are interested, I posted an article introducing the contents of the book. You can get the book on Amazon or Packt’s website.

Soner Yıldırım

The Dark Side of the Sexiest Job of the 21st Century

What is it like to be a data scientist?

In October 2012, a Harvard Business Review article described data scientist as the sexiest job of the 21st century. This article is not the reason why data science is so popular now but I’m pretty sure it motivated some people to become a data scientist.

Before I start on the downbeat, let me state that I’m glad that I made a career change to work in the field of data science. I love learning, practicing, and implementing in this field. It might be the only job I enjoy in my professional career.

However, there is a dark side which is hard to see before you get in the field. The box is all shiny and beautiful from the outside. Once you open it, you see some things that might lower your motivation a little.

When I first created a jupyter notebook that contains a machine learning model, I was super excited. The model achieved a pretty high accuracy. I felt like I’m already tackling down some problems.

The problem was that it was a simple and ready-to-use dataset. I could achieve high accuracy by using any off-the-shelf machine learning algorithm without understanding what is going on under the hood. I did not have to do any feature engineering or extraction.

To be a little more pessimistic, machine learning is only a small part of the data science pie. You can aim to be a machine learning engineer but not every business can afford to have a separate machine learning engineer.

Big tech companies usually have data science teams and separate positions that focus on different part of the data science pipeline. However, those positions are limited.

Medium or low level companies that want to adapt data science in their business tend to hire one data scientist or two and expect them to handle the entire workflow. Thus, it will dramatically increase your chance to learn about each step in the workflow.

This is what I mean by the dark side of being a data scientist. You have to learn a lot more than you could anticipate.

If you follow a self-taught process, the learning process is more dynamic. The more you learn, the less you feel like you know.

Data science is an interdisciplinary field that combines statistics, math, and programming. On top of those, you need to have domain knowledge in some cases.

There is a wide variety of topics and tools you need to learn to become a data scientist. I will try to briefly explain what they are and why they are important.

Data scientist roles tend to be full stack.

Data is the fuel of any data science related product. Collecting and maintaining the data is fundamental. You are likely to engage with a SQL and NoSQL database a lot. You will probably not be in a position where you can just tell “let me see the data”. It is for your best to be able to get your own data from a database.

The next step might be the most important of all. You need to explore the data. I’m not talking about calculating the mean or creating simple distribution plots. In order to discover the structure or relate variables in a real-life dataset, you need to have to comprehend the statistical concepts thoroughly.

Having a decent knowledge of statistics will make it easier for you understand the machine learning algorithms. Without statistical concepts, you wouldn’t be able to explain why linear regression is or is not appropriate for a given task.

You also need to cover some topics in linear algebra and math to a certain extent. The computations done by a machine learning or deep learning models involve matrix multiplications. In order to understand how the optimization algorithms used in the models, some fundamental math knowledge is necessary.

It is not enough just to know these topics. You need to be able to implement them. Thus, it is inevitable to learn programming skills. You don’t have to be a software developer but all these algorithms and data analysis tools are used via a programming language.

There are many alternatives but the most commonly used programming languages in data science are Python and R. There exist many packages that expedite the data analysis and machine learning process but a basic level of programming skill is needed to use them.

Let’s say you identify a problem and design a solution to the problem that involves data. You collect, clean, and maintain the data. A useful and accurate model is created.

The next step is to deploy your model. If your work stays in the a jupyter notebook, it is useless. It cannot create any value. MLOps is a whole different world. There are many alternatives. It is hard to even decide which one to use.

If you work on a medium or big scale project, you are likely to use a version control system such as Git. You should not be unaware of such tools. Moreover, it will make your life a lot easier to be comfortable with working on a Linux environment.

Last but not least, you may also need to have hands-on cloud computing experience. More and more companies start to maintain their data on the cloud. They do not possess physical servers anymore.

I have tried to touch on almost anything that I think you need to learn. There is, of course, no limit on what you can or should learn. The more skills you have, the more appealing you become to the companies.

The dark side becomes more clear after you complete a few data science certifications. You feel like you are ready to tackle down a business problem. However, when you encounter a real-life problem, you face the dark side.

The certifications are helpful but will definitely not make you a data scientist in a few months. Please keep that in mind when you set your goals. Improving your skills in all these areas will take a long time. It is a challenging yet promising task.

Conclusion

You may argue that it is not necessary to obtain all these skills. You are right in some cases. However, considering the popularity and potential of data science, having all these skills will increase your chance to get in the field.

If you browse through job postings on LinkedIn or any other portal, you will see what most companies expect from a data scientist position.

There are specific positions such as machine learning researcher but they are limited and require a high level of experience.

I don’t want to sound pessimistic. My goal was to shed light on the journey of a data scientist. You should set realistic goals and be ready to sacrifice a quite amount of time and effort.

Thank you for reading. Please let me know if you have any feedback.

https://www.quora.com/Whats-the-difference-between-being-a-data-scientist-at-a-large-company-vs-a-startup

Brian D'Alessandro, Director of Data Science at Zocdoc (2016-present)

I’ve been a data scientist in both. For bigger companies it really depends on the company and industry (think Facebook vs Pepsi). Consider this assessment trends vs. absolute judgments too.

Larger Company - expect your role to be much more specialized. I.e., you may just be focused on a particular business vertical (product feature vs marketing), and the set of technologies you use day-to-day may be smaller. For instance, you may only need to work on an aspect of Search, a Recommender System or marketing attribution. You should also expect (or hope at least) to have more robust data engineering support and thus spend more time on analysis than structuring data. You’ll also likely deal with more process and bureaucracy, and your direct marginal impact to the org will be much smaller (certainly as a %, not always as an absolute).

Startup - you might have to do a little of everything, including things you don’t think are ‘data science.’ You may be called into a sales meeting to help explain the system to potential clients. If the service goes down, you may be joining other engineers to debug. Expect to do your own data engineering, even to the point of spending months developing data pipelines and not analysis. The company may change strategy over night (often after a board meeting), and all of your prior work gets dropped to go after the new pursuit. The plus side of the chaos is you’ll likely develop a larger range of skills, get better business acumen and have the potential for a much larger impact. Another aspect is growth. Becoming a Director or VP at an established large company is very difficult. But if you’re an early employee at a startup that eventually grows, you’re going to experience that growth in your career progression.

My advice to newer data scientists is to start at larger companies where you’ll get the mentorship and structure to really hone your craft. Once you have more maturity and confidence, join a startup to accelerate your career. After you have a family and kids, go back to a large company so you can avoid the added stress :-)

Vyasraj Vaidya, lives in Hyderabad, Telangana, India (2017-present)

Working as a Data Scientist:

In large companies: Infrastructure is ready but for smaller ones - you may need to build one.

In large companies: Data storage and extraction is easier. Assuming that you will be working with large datasets, the data management is easier. Application of ML algorithms on such large datasets will become extremely complicated. You cannot do this on local machines.

Deploying a data science experiments as a web based services is complicated - In large companies, you will have ready made services or tools for deployment like for example Azure ML studio and AWS (which is why they are world leaders!). There may be some companies that use these tools as 3rd party services.

Deploying these services may have to done by building their own service for smaller companies unless you want to use services like AWS or Azure ML studio. In startups, you have a choice to use SAS and SPSS. In case you do not want to use services mentioned above, data training and real time data ingestion would be the greatest problem you will face.

To be concise and precise, a one line answer for the question would be that - you will need to think of end to end solutions for everything when working with smaller companies which may not be required in larger ones.

Jason T Widjaja, Analytics and data science team lead

In a startup, the title says ‘data scientist’, but it is really:

data infrastructure management AND

BI & reporting analyst AND

database and data warehousing analyst AND

high performance data engineer AND

data scientist

In a large company, the title says ‘data scientist’, but it is really:

Enterprise Hadoop platform - Mahout team (Eastern Europe) OR

Machine learning model management, consumer marketing team OR

Data mining project manager, finance time series forecasting team OR

Data scientist (manufacturing operations), stochastic optimization team