Interview 29 - Ilan

Interviewer 1

All right. Uh, so to start off, can you give us a bit of information about yourself again and how much experience you have in machine learning and

Interviewee

in general? Okay. Yes. I'm a, at Company X. Um, I'm, uh, let's say at the current position, uh, for about one, uh, year and, uh, some, let's say three, four months. Um, my previous experience in the same company was as a senior data scientist.

Interviewee

So let's say, uh, after two years I got a promotion. Anyway, so I live now a team, uh, a small team where three people, I'm the lead and I have, let's say two senior data scientists where I mentor them. Um, my exposure is, uh, in my previous, it's not only this, my, the last part, let's say. The last part is not only my experience, I have experience in consulting, in retail, et cetera, but.

Interviewee

My, let's say my total experience as a, as a data scientist is almost, uh, eight years from, uh, June, uh, 2015. I started as a data analyst then, uh, then as a senior data analyst, then, uh, analyst, then as a data scientist, et cetera. Now I have used, uh, let's say machine learning in different, or in various projects, uh, related to different problems.

Interviewee

Now, uh, to, let's say the, our, our last project was related with cloud, um, services or cloud applications. Company X is a selling cloud infrastructure or general speaking, uh, cloud, uh, let's say applications to other, uh, fire firms. Um, is the firm pro is promoting the digital workspace. So, uh, Let's say that every employee of the other, let's say firms or clients to be able to work from Manuel.

Interviewee

So our, our last project was related to cloud out because this is our. This are, we have cloud services, um, so you can understand that there are outages related to these services cause are all in cloud. So we have to have a reliable and, uh, available cloud services. So this, our last project was related to Cloud

Interviewee
And what was the, let's say the, in a summary, the, the description of the project was, Try to identify because when, um, cloud auto is happening, there is, and the management process behind that, which is running, and the engineering and the engineers are trying the product, let's say engineering team is trying to identify and resolve this issue.

Interviewee
So let's say after this process, there is. Um, a summary text, what engineers describing, what happened, what happened then the fixed description were the engineers describing, um, how they fixed this issue. What we were lacking was to, to identify the root code. So let's say to classify this out, this. For instance, um, uh, was a resources issue or was a third party issue related to Microsoft or Google, et cetera.

Interviewee
So we created as a team, role based model. This was, let's say, didn't have a machine learning inside was, uh, based on keywords on, uh, Uh, on, on this for the text. And, and after that, what we thought as a is based on this different classes to try and predict the duration of this, let's say how this, so, uh, that was, let's say, the part where we used the melan.

Interviewee
Uh, we tried to predict the duration, let's say. That, that was our last process, just to give you a summary. And we use the random forest, uh, regression, for instance.

Interviewer 1
Okay. That is interesting. Thank you. Uh, I think we'll have an interesting discussion. Uh, so to start off, what are the main quality issues you have encountered with your data model or system?

Interviewer 1

So,

Interviewee

Quality, I believe. I believe they make quality. I don't know how you mean quality, but uh, yes. Um,

Interviewer 1

yeah, it, it really depends on how you define, define quality. But there is many aspects you can consider. So, for example, maybe scalability is quality for you. Uh, explainability, robustness, um, efficiency.

Interviewee

Yeah, I could be other.

Interviewee

Yeah, yeah, yeah, yeah, yeah. I got it. In our case at Company X, the main problem was, uh, the data collection. And what I mean is, Um, the, there is notary resources, there sources because, eh, we have, let's say restrictions with security. So, ha we had to be secured in order to be able to collect this data. So, for instance, if, if we wanted ASIN to collect, let's say some data local in our machines and run a proof of concept, that was something that was difficult.

Interviewee

Scalability was, I wouldn't say it's an issue in our case because when we have to do with Let's escape, uh, with Cloud , uh, you can imagine that this number is not very large. Cause if it was very large , Company X would have a, a problem and uh, wouldn't have clients. So we can. Cloud, uh, in three years now is let's say a number, like 3000 almost.

Interviewee

And this is declining because we're getting better. I, I say. So it's not scalability. We don't have a million of, we don't have, I mean, for the specific project and in other projects or in engineering projects where they. Um, the service center have log data regarding the availability of the phase and the services, and they usually, um, different, let's say log uh, logging providers.

Interviewee

And they have some issues because depending on the queries they do, in order to identify some, let's say, issues or anything. Uh, we pay different money to these providers. So there, there is a scalability, but from engineering perspective, perspective, from data science perspective, I would say the accessibility is the main, at least at, in our case, um, related to quality.

Interviewee

I don't find, I don't think, uh, anything else from what you mentioned.

Interviewer 1

Okay, perfect. Thank you. Um, so we'll move on. You, you mentioned data collection problems and I would like to to go deeper on that and, um, yeah. So, so uh, basically what I'm gonna ask you is if you ever use one of the, one of the data collection technique I will mention, and if you ever use it, I am looking for quality issues with the data that was collected.

Interviewer 1

Mm-hmm. with a data collection process.

Interviewer 1

Uh, so do you, did you ever use data that was manually collected and if yes, what were the issues with the data

Interviewee

manually? Uh, yes. For instance, uh, you are referring, let's say, to someone, let's say files were, have been created manually, something like that. Yeah. Uh, yeah. The main problem is that there's no consistence in this data.

Interviewee

I mean, that. Let's say specific file might start from an next person and, uh, um, let's say might, uh, um, he might do some changes. And then, uh, you don't have, let's say the, the, the, the controller about the versioning, about which are the differences between, uh, this, let's say, uh, how can I say? The different

versions of this file.

Interviewee

So what is difficult when you have this, let's say, type of files, is to have consistence of who or what, or which is the change or what change, et cetera. I, from my, from my experience, I think this is the most difficult part.

Interviewer 1

Okay. Uh, just, just to make sure I, I understood everything. Mm-hmm. . So you said that it's not consistent.

Interviewer 1

Am am I. Okay. And you also mentioned that there is not enough, um, metadata, like, uh, diversion. Versioning,

Interviewee

yeah. Versioning, yes. I think versioning or, uh, controlling the, let's say the part of, uh, what changed in the file, because if there are different users and do different, let's say changes in a specific guide file, if you don't have this version controlling, you cannot understand what change, who changed it, et cetera.

Interviewee

Okay. You cannot, yeah. Yeah. You cannot track. Okay. I see.

Interviewer 1

Yeah, I see. That's, that's, that's interesting. Um, and so if I understand correctly, you're meaning, uh, what data said that has been modified by different person through time? Like, uh, yes. Okay.

Interviewee

And how did, yeah. Sorry to interrupt you, but we're talking now about the local.

Interviewee

Not, uh, let's say data that, uh, are available in a database. This is something different. Um, I mean, if you have, let's say an Excel file where, um, a colleague or a different user has shared with you and this file, let's say lives through the course of the year, you have to do changes or anyway, work with this.

Interviewer 1

Okay, understand. And how do you address this problem usually? Uh, do, do you have anything to keep the story of data sets or not really?

Interviewee

Uh, I think that we are changing, uh, the system , because what, what the ministry depends on what we are you are using in, for instance, in order to avoid, let's say these, the things we have, uh, we are, we are using, uh, database.

Interviewee

So we don't have to keep track of anything. The data sync are updated. I, uh, I mean automatically. So everything is, uh, automated. Um, or, uh, for instance, I'm changing a little bit, uh, sub subject, but in order to have a, a version controlling in, in your code or, or the teams code, you can use GitHub, what you can track these changes and what changed in the code.

Interviewee

So I mean that all these. Is mature enough in order to avoid these things, you can use different systems, uh, more automated, uh, better, let's say, than the previous practices. Perfect.

Interviewer 1

Makes sense. Sense. Thank you. Uh, so did you ever use ex external data such as public data sets, third par, third party APIs, or web script data?

Interviewee

Eh, Not at Company X, because as I said, we have very. Let's say secure environment. So we avoid to use, uh, data from other, uh, from uh, uh, let's say other sources or external sources. Um, so we're using only our data internally. Data from my experience, uh, I mean generally speaking, or let's say myself, I, yes, I have used or, uh, for instance, Or at Company X for, for the needs of a, of a project, uh, I had to, to connect to, to elastic cells.

Interviewee

So what I did is that I used Python in order to be able to connect to this, uh, api and then let's say created some code in order to be able, uh, to, to collect, uh, the specific, uh, OCS that I wanted to search, et cetera. So this was a little bit automated because, um, I used also.

Interviewee
S I mean, mark, uh, mark Bus and Unix, et cetera, in order to be able to, every day to this, to automate this and, um, be able to collect, um, uh, on a daily basis. The specific data, uh, this, this is something different. Um, I have used, uh, web scraping, but for personal, let's say reasons for personal interests. Um, What other, uh, what else you mentioned?

Interviewee
Uh,

Interviewer 1
Public data sets and

Interviewee
third party APIs? Yes. Public dataset? Yes. Uh, again, for, not, for, not at Company X, I mean, uh, per, uh, I'm referring to myself, uh, personally. Uh, for personal, let's say prep projects or for instance, as you do in your case where you are doing a master. For instance, when I had to, to write my thesis, I had to, to do, let's say, some, uh, master learning with.

Interviewee
A part of my thesis. So I, I trusted the public data, uh, from my machine learning repository. I think that you might already know it. And I trusted this data in order to, let's say, to create a machine learning model, uh, and third party APIs.

Interviewee
Uh, specifically or generally speak. Whatever, whatever, whatever API you could collect data. Your question.

Interviewer 1

Yeah. Any API that is not maintained by your team.

Interviewee
By my team. No, we don't, we don't, uh, have any api. But, uh, what is happening, uh, is that we have engineering engineers who, let's say are responsible for maintaining API for, let's say, creating the APIs, et cetera. And let's say to.

Interviewee
Database and have this data available for us. But we don't do this engineering stuff. We, we are responsible for the collection part and how we'll be able, or what system we will use to be able to collect this data from there. Yes, of course we have used, uh, have used, uh, third party ebs, but uh, from the data science perspective,

Interviewer 1
Okay.

Interviewer 1
I see. And in any of your, of your experience, experiences, is there any, uh, like general pattern of quality issues with the data from external sources or not really?

Interviewee
The main, the main issue is that there is a lack of, uh, let's say communication between engineers and, uh, data scientists. So what I mean is that engineers think with, uh, different way than data scientist.

Interviewee
So they, what, uh, might happen is that, uh, they will provide an API with some data, but they cannot. Think, let's say more than this, in order to be able to, let's say, to think more and, uh, make a sense about how this data will be used in order to be consumable or to be presented to a management level, et cetera.

Interviewee

So I believe this is the, the main problem, the communication between what has to be done and what, uh, is being. . Okay. And

Interviewer 1
can you give me an example? Uh, you mentioned that data is not in optimal, uh, uh, format or,

Interviewee
uh, yes. Uh, for instance, when we designed, we, we had to collect some data, uh, from different cloud services and, uh, lose something with this data.

Interviewee
Okay, so, The, the format, what they designed is they created DB attached to an api and to we, let's say we will be able to connect via power and collect this data in order to visualize this data and present this to the management or to the executive, uh, let's say level. Um, but. They could not understand that this, let's say format has to be standardized.

Interviewee
Okay? And B, the same for X, let's say. Um, not months for period. Okay. But this has to be standardized also. We have to have these columns, these names, in order to help the other system understand that this data are the same and didn't change and have, let's say, mal malfunction with our, uh, reports. Okay? Um, if we do, if we, we wouldn't, sorry, if we didn't want to standardize this, this data or the format of this data, we, we could have something more flexible and be able to do different things.

Interviewee
But what happened is that we didn't have, uh, nothing of this not standardization or the flexibility. We had something in the middle, so, That was, let's say, a main issue. Okay.

Interviewer 1
So if I understand correctly, uh, what you're saying is the issue was that the, the format of data changed? Like, uh, like the, the features of the data set?

Interviewer 1

Yes,

Interviewee

yes, the feature. Okay. Then, In standard. So what I mean is that we require from the beginning to follow a strategy and standardize the format or the features that you said of the data. Okay? In order to be able to automate this. So every time I don't have to. See if there is, um, the same column in the data, et cetera.

Interviewee

Okay. So we have to standardize the format or the features, and then R BI will be able to understand that this data have been updated without human intervention and the report will be available. But they did agencies in this data without. Informing us, let's say. So this was, uh, the result was to have, let's say, a report that could not actually be automated.

Interviewer 1

I see, I see. Perfect. Thank you. That's interesting. Uh, and was there any issue with the, the, the values themselves? Like the, the, the, you know, the feature change, but Yeah,

Interviewee

go for it. And the values? Yes. Uh, Because, for instance, we had an issue because there was a different level of aggregation. So when this changed, value changed because if you aggregate the level the data is different levels, then you don't have raw data.

Interviewee

This, eh, might, uh, result to, to wrong results.

Interviewer 1

Okay. You at different level of

Interviewee

aggregation? Yes. What I mean? Yes. When you have a, um, let's say the, the data you need the, the level. Ok. So you need the minimum level of information because if you have product, okay. And the cloud service and the value, if. Agree is if this, let's say, is aggregated to a higher level, like the product, not the, the product and the, the, the service you'll have larger, normally you have a different number than if you do this ion in lower levels.

Interviewee
So they did some aggregations before fund. So this data, when then. Let's say we're aggregated in a different system like Power bi. The results obviously were wrong because we we're already, had already been aggregated from, uh, as said before, beforehand, from the, from, uh, the third api Yeah. Or Mongo. Okay. I see.

Interviewee
Thank you.

Interviewer 1
Um, I'll move on to data preparation. How, uh, sorry. Have you ever measured the quality of your data and or

Interviewee
tried to improve it?

Interviewee
Uh, yes. That, uh, was happened in my previous, let's say, uh, not experience, but anyway, in the first part of, uh, uh, my employment with, uh, Company X when I was senior, um, what we had was a nap times report. This, as I said before, we have cloud out. For our cloud services. So, uh, you can create a metric, which is called uptime in order to measure, to measure how much available, uh, the service was during a month or whatever time interval.

Interviewee

Uh, um, you decide. So, uh, we were reporting to the management some numbers. Well, from my, uh, experience and what was happening, uh, um, in, uh, during, let's say the cloud diagnosis or the incident, sorry, pause. Um, from . Yes, because yes, it's not, uh, does, doesn't make sense. Anyway, um, so what was happening is that we, we presented better numbers than the actual, uh, reality.

Interviewee
So we were saying that we, we are available for, for instance, to, for a cloud service, 99.9%. I'm giving an example, but also this, let's say cloud service had 10 out that this during the month, so something was wrong. Um, And, uh, when I started to digging into this and, uh, try to understand why we have all this, let's say why we present that, uh, everything is fine, but we have a large number of incidents.

Interviewee
These, these are controversially between them. Okay. Uh, I found that, uh, they had, uh, what they had created something like a weighted at time. Okay. They did, they created a weighted measure, um, depending on, uh, specific features. Um, so this weighted feature, eh, let's say, gave a. Picture one, uh, comparing to what was happening in reality.

Interviewee
So yes, uh, I had to dig as I before, so this data were, were presenting something more, um, pretty, let's say, than the actual picture. Okay, I see. Thank you. So, so I ha so I have to change this and this all, uh, actually,

Interviewee
Uh, sorry. What? No, no, no. So what I said is that I had to, to communicate this and, uh, change this, uh, way of calculation because it was wrong, it was wrong because, uh, it, uh, showed the better, uh, picture that mm-hmm. , uh, comparing to what was happening in reality from cloud outta this perspective. So, We communicated this to the teams and we changed the way of, uh, how we calculated this, uh, uptime or this metric.

Interviewer 1
Perfect. That's clear. Thank you. Um, is there any other data quality should we missed that you consider relevant?

Interviewee

No, I think these are the huge, let's say, issue. Perfect.

Interviewer 1

Thank you. Um, how do you evaluate the quality of your models? And as a reminder, quality is not only defined by the performance, ML performance or accuracy one score, but there is also other aspects such as robustness, like I mentioned earlier. So

Interviewee

how did I manage this or how, if there is a name, uh, the question is how I, I typed these issue. Or if I had experienced something, uh, let's say like a short a model that I, as a team, we created a model that then, uh, didn't, uh, perform well.

Interviewer 1

Uh, yeah. So, um, how, how do you evaluate your, your models? So, oh, okay.

Interviewer 1

Try to compare, like models or to see how good they are,

Interviewee

right? Uh, yes. This, there are a lot of things that you have to. Take into consideration, um, the amount of data, I mean the volume of data or because some, some models, uh, do not perform well with, uh, uh, let's say small volume of data, and they need the large volume of data with something which is like crucial.

Interviewee

And also if there is a dependency between the feature, If you have a dependent or independent features, some models might work or not. Uh, with multi, I mean, um, what else? The scalability of the data. You might to, you might have to do scaling for some models because if you have scaled data or large difference between, um, the values of the data, Some models might not work, perform well.

Interviewee

Um, and yes, I think and the dimensionality. So some, if you have, let's say large dam dimensionality, you need to, to focus on specific models because some others do not perform. Uh,

Interviewee
Yeah.

Interviewer 1
Totally makes sense. Thank you. Um, I will move on to deployment and maintenance of machine learning software system. If you, if you think you do not have experience, I'll skip the questions. Um, so what are the challenges you have encountered during the deployment and the maintenance of machine learning software system?

Interviewee
Uh, yes.

Interviewee
As I said before, I don't have much experience with, uh, a Recomme commander system or something similar. But, uh, from, let's say what we, uh, have done as a team in, uh, ML Azure, um, the, the main problem is that, uh, this is a black box. So you have some, uh, automated models that are adding. And, uh, there is an output.

Interviewee
So you, you cannot intervene to this black box. This is the main issue. But this is not related. Related, sorry. With, uh, the, the maintenance of, um, let's say of the modeling create, this is how is functioning, so it's something different. I'm not sure if this is related. From Maintenance's perspective that you also, if for, for instance, if you have the ownership of this model, you have to retrain this model.

Interviewee
When, when there is a black box where is running and you have an out output, you do not know if, or if the model is the same. What changed? Uh, what are the, if the, the same features are important, et cetera. Perfect. Yeah.

Interviewer 1

Thank you. Um, so, um, to finish, I will ask you a couple of questions regarding different quality aspects and basically if you ever had an issue with one of the quality aspects, uh, please tell me.

Interviewer 1

And there are some of them we already covered. Uh, so yeah. Okay. Uh, so did you ever add an issue with fairness, robustness, explainability? Scalability, pr, data privacy and

Interviewee

model security?

Interviewee

Uh, model security? No, because, uh, yes. I cannot, uh, think something that, uh, um, happen with, let's say the security of any model. We, I have more as a thing we created in the past and robustness. Uh, yes. This is, uh, let's say, uh, an aspect that has, uh, significant meaning. But, uh, you, you are trying different models in order to find the more, uh, robot.

Interviewee

So this is. Say a game that you played during the training of this or during this learning path. So you're trying to be robust from the beginning. Uh, otherwise you could use one, two, or three, let's say, models, different models, and you would compare between them and keep the, the, let's say the best. But this is not what is happening.

Interviewee

You're trying to un to understand the data, to see which model. And performs better in this my, in this data. So this is a part of this, let's say, robustness at at least, uh, uh, this is my thought. Uh, what else? Sorry, can you repeat the others?

Interviewer 1

Uh, I mentioned fairness, explainability,

Interviewee

scalability. Ah, and yeah.

Interviewee

Uh, yeah. Scalability.

Interviewee

Uh, from a store's perspective, because when you, let's say when you have to store incoming data, uh, and this is climbing, uh, there is an issue with, uh, the cost, so, I believe that the scalability, what, what I could, uh, think like an issue is that scalability is also related to cost. So this is an issue from company's perspective.

Interviewee

Um, fairness, if you compare, uh, if you create, uh, not create, but if you do not compare or treat the models with, um, With fail, something like similar? I dunno. Uh, no, because as before, you have to know what, uh, type of model you, you, you can use. Or let's say this. Type of models perform better to the specific data you have, eh, explainability, eh?

Interviewee

Yes, but this, because the, the engineering team, let's say, eh, haven't explained what type of data we are receiving. So the features didn't have, um, clear names, so we're couldn't understand what. This feature is, uh, um, uh, what's the meaning of this feature? So yes, explainability could be an issue, uh, problem, not an issue.

Interviewee

Uh, yes.

Interviewer 1

Okay. Interesting. Yeah. And if we go deeper on, um, scalability, I think you, you briefly covered just a bit, you mentioned that, uh, something related to cars with your company. Uh, can you give me an example when it was an issue for you?

Interviewee

Uh, not, not for me, uh, because, uh, as a team we don't manage a huge, uh, amount of data.

Interviewee

Uh, but, uh, as, as company, uh, I mean that the firm, uh, had to manage these things because, uh, we're, we're, uh, using, um, uh, third party providers for, uh, logging data. Okay. This is climbing. Uh, so depending on how you are going to store or query or whatever you are going to do with this data, this is related to course because providers, um, uh, the course of this, the provider is related.

Interviewee

These, uh, three things, if you equated the data. And how often can, let's say, how much memory do you use when you query the data? Um, the storage, uh, of this, uh, data and, uh, uh, what else? Yes, this.

Interviewer 1

Perfect. Thank you. Um, and I have one or two last questions for you, uh, in your opinion. In your opinion, what is the most pressing quality issue researchers should, should try to solve

Interviewer 1

at the moment?

Interviewee

Yeah.

Interviewer 1

Whatever pops in, in your mind.

Interviewee

Uh, data integration.

Interviewee

I think this is the most, uh, let's say difficult part of, uh,

Interviewee

Of the recent, uh, uh, years because, uh, we have different systems. We can use different systems to do many things, but. These systems are, uh, um, the own, the ownership of these systems, um, are, is different, okay? Because Amazon has web services, for instance, or Google has big query, et cetera. So the integrion between these three or these, these three, anyway, this end type of system is the difficult part because you have, you might have to use or to be, let's say, no, not to use.

Interviewee
The portfolio, which, uh, is, uh, which is given by a specific cloud provider, for instance, in order to be able to automate things. But you don't have to be, you have my perspective, you have to be more flexible. So be able to combine different tech stacks.

Interviewee
Okay. Uh, let, let there is, sorry. Yes, I might, yes, there is some improvement on this because there are, uh, let's say some, uh, systems that are in the middle and try to join things, but I believe this is the diff the from my per from my thinking. I believe this is the most difficult part to combine different.

Interviewee
So by the other word, there are some compatibility issues. Yeah. Yes. Compatibility. Yep. And when you are trying to integrate something, develop in other systems or with another frameworks into others or some previous developed Yes, yes. I understand. Yeah, exactly. Yeah. Thank

Interviewer 1
you. Okay. And just make sure I understood everything.

Interviewer 1
Uh, so you use system from different providers, Amazon, Google, and Azure, for example. Mm-hmm. . And I think you mentioned cus at some point, right? Yes. So you, you're trying to pick the one that is the most efficient in term of cost, depending of your requirement. Is that correct?

Interviewee
Yeah. Okay. That's, uh, statement.

Interviewee

Yeah. Okay.

Interviewer 1

And this,

Interviewee

yeah, go for it. Yes. What I'm saying is that, uh, the difficult part is how you can, uh, integrate. Uh, for instance, you have to develop, all engineers have to develop applications in, uh, Amazon for instance, but you can, you, you have to do the same. Think for, uh, Google or the something else from, for Azure.

Interviewee

So the integration between different or cloud, not cloud providers anyway, uh, how you develop, uh, things between the different systems and how these systems are, um, integrated is the, okay, I see most. Perfect.

Interviewer 1

Thank you. Um, and do you have any other comment about the quality of machine learning software system?

Interviewee

No, I don't have any other comment.

Interviewer 1

Super. Well, thank you Tans for participating in, in this interview. You're welcome. I think it, yeah, I think it was really interesting and we'll have a good material for the paper.

Interviewee

So, oh yeah. Thank you for having me. Of course. Uh, I'm excited to see the research results

Interviewee

Sure. Appreciate you. Okay. Yeah. Thank you. Thank you again. Thank you. Bye. Thank you. Good weekend. Bye.