Interview 7 - Amin

Interviewer1

Okay. Uh, so I'll start the quick overview of the objective of this, uh, study. So what we want to do is to develop a catalog of quality issues in machine learning software systems. Uh, we are interested in any issue you have encountered while building mission learning software system that affected its quality.

Interviewer1

And machine learning software system is any software system with an ML component. So for example, a recommender system, it is a machine learning software system. Um, and yeah, and that's pretty much it. Uh, do you have any question, uh, on the, on this?

Interviewee: No.

Interviewer1

Okay, perfect. Thanks. Um, So, uh, okay. So we can start with some background information from you if, uh, you agree. Uh, so what is your current position? How much experience do you have in machine learning?

Interviewee

Um, I am data scientist at Company X. I started. About exactly X year ago. So DATE. Um, prior to that, well, uh, I've been around for a while. Um, I started to work. Um, well I finished my PhD around DATE, started to work a year before DATE.

Uh, machine learning was not such a big thing in 2006. So I was doing all kind of data, fusion, a algorithm, stuff like that kind. It could fall into the AI. Uh, topic, um, rule based decision based support, uh, support decision aids, uh, system for defense maritime surveillance, and then municipalities then at Company X.

So. In the process I machine learning, uh, came more and more popular. So I don't know, you know, for me, machine learning is one is a tool in my toolbox. So. I have to answer tough questions with data. So sometime it's machine learning sometime it's just base probability. Depends. So, uh, but uh, now all the tooling and stuff

related to machine learning is.

Is quite there. So it's, it's, it would be, it's a good option most of the time, but it's hard to say when I really started to do machine learning somewhere between 2006 and I don't know, 2012 it's whatever. Yeah, yeah. It's way. As I said before that we were calling that data, mining was big thing, data, mining, knowledge, mining information, you know, all these buzzword.

So I I've seen. This transition. Um, yeah, so , that's about it. Yeah. Perfect.

Interviewer1
Thanks a lot. I, I just realized we haven't present ourselves. Okay. And everybody is here now. Uh, so me it's Pierre olivier, I am a master student at Poly under this supervision of Amin in and also a foutse. And, um, prior to that, I was, uh, I was in, um, I did a bachelor in software engineering. So that's it for me. Amin, did you want, continue?

Interviewer2
Thank you. Hello again. Sorry for delay. Um, yes. My name is Amin. I'm working as a research. It's Poly Montreal and in this project, yeah, we are working with PO and Rached and interviews to collect a catalog of quality shoes in machine learning system in practice. Thank you again for your time. Rached. You can

Interviewer 3
Hello. So, so I'm Rached and I'm doing my master, uh, at poly technique and, uh, I'm in SWAT lab and the supervision of Foutse and, and amin, uh, yeah, I think that's, that's it.

Interviewee
Nice meeting you.

Interviewer 1
All right. Uh, so we'll start off with the first question right away. Um, so what are the main quality issues you have encountered with your data model or system? So far.

Interviewee

Oh, that's, uh, that's a big question. Um, I was involved, uh, more than 10 years ago in all kind of project about data quality taxonomy of all data quality aspects.

So it it's a huge topic, especially in defense because I mean, they have to make decision like with large impact on possibly lives. So it was really, um, a thing. So, uh, Personally, I've, I've seen a lot of issue related to data. I consider myself more a data person. So, uh, um, I'd say that's the most part it's, uh, of quality, um, was related to data. And then,

so you, you, the question was quality regarded regarding to the, the models. That impacted the models.

Interviewer 1

Uh, so it can be issues about data model or the me learning software system, like in general, any quality issues. Uh, and since you have expertise in data, it, it can be, uh, more regarding data. So, yeah, but it it's also sometime about, uh, well, okay.

Interviewee

Let's start with data. Um, well, some, some, uh, data, uh, quality issue. Like really obvious missing data, uh, inconsistencies between data sources. This is something that I've dealt, dealt with a lot even built system to be able to pinpoint this inconsistencies between, uh, data sources. Because as I said, I was more into the data fusion, uh, uh, area before.

So it it's all about that. It's making sense of different kind of sources. So, uh, In consistency in terms of, uh, uh, of content, but also in terms of how the data is being served, uh, time in consistency. Uh, one source of data could be, uh, could report data every minute. The other one, every six hours, then you need to.

To make sense of it. And of course that includes the, uh, um, uncertainty, what to do with that kind of uncertainty. So this is one type. Um, also there is, uh, well data quality related to, uh, the cap data capture itself. So data is usually captured with using a, not a system, but a sensor. And you have calibration of this sensor.

This one is easy because usually, you know, it, uh, this, this is one, um, uh, other part of it, uh, a lot I've worked, uh, for the city of X, um, for years and, uh, was with data scientists there. A lot of data was manual manually, uh, entered and in French. So there. All kind of, uh, mistake writing stuff and you have to deal with it.

I saw so many ways of writing cote des Neiges you have no idea so you, and it seems super easy to deal with, but it's complicated. Um, so this is one type of error I've dealt with a lot. Um, I, I dealt with also, um, I've worked a lot with, uh, geospatial data captured with GPS. Okay. A lot. And, uh, and who this one is difficult because, um, most, most of this data was captured by, uh, phone applic mobile application.

And, uh, there. Error of geolocalization. Let's say you were in downtown. Well, it's, I'm sure you experienced it. You're not able to, to ju localize yourself with Google map, for instance. Well, I was on the other side trying to make algorithm with this kind of data. So that's that's. That was difficult. So again, it falls back on the, on, on the sensor.

Uh, capturing the data. Uh, there was also for, with the iPhone, especially they will, um, introduce in the data on purpose to protect, uh, private life. Uh, and it's more and more. You're gonna see it more and more. So this is one type of quality. I'm sorry. My, my son is very sick today. Just a moment. Okay. Just.

Okay, sorry. Um, I think it, it covers all about data. I think so, as I said, uh, uh, and, and then there's, there's that more difficult data issue is the, it's the meaning of the data? Uh, the definition, uh, what does it mean for you, uh, to say what's a citizen, for instance, Is it the company? Is it someone, a human being, uh, is it some, a human being, having a house in Montreal, human being, renting, or walking on the territory?

It all about the definition. Uh, may your data may be formatted perfectly, but still you may have, uh, difficulty with, uh, and this. And I would say that this kind of, and quality issue is, is, is really, um, , don't know how to say it in English, because it sounds, it looks like it is perfect, but it's not. Um, it's gonna create you a lot of problem.

I've seen this kind of, we discussed this, this kind of issue a lot in the defense field because, um, well, uh, countries exchange data. Have the same definition. And it was, well, there was a, a, a thing with ontologies that were supposed to cover this kind of definition. There was a lot of body of literature related to that.

Um, but I mean, we don't have to go to an ontologies, but this is a different, difficult, uh, difficult one. This is one kind of quality issue. Another quality that would've would impact the models and stuff. It's, it's always like it's not data it's when you try to build something, a system, uh, As a data scientist system, developer, whatever you, you, you listen to the problematic and then you have your own idea of how it should work.

But sometime it's not how the user think it should work. And. That affect a lot, uh, the quality of, of even the model you do, because sometime you, you or the, the user wants something wants what would be translated as a regression model. But I mean, when you look at his, uh, workflow, You understand that just a classification would be enough and that would be great, but you spent a lot of time building a regulation instead of classification.

So to me it affects it. And at the end of the day, also quality of model. I mean, you can have pristine prediction if it's not used. It's, it's poor quality for me. To me, this all, uh, business, um, analysis that comes with the design is really important and can really affect the system because, uh, if it's not there, if the needs are not well understood or not only the needs, because everybody has needs even the user, but really the workflow is not well understood.

Well, you're gonna end up with a model that is not used. For quality to me. So this is one part which I, I think is not really often cover in this quality model. And then after that, uh, also the way results are presented and I mean, it's all part of that. If it's the cost we put into it, the us, if the user, the consult are presented in a way that the.

Do not trust the result, do not want to use it. And it's decision making. Well, to me, it's for quality too. And this trust thing it's, it's not easy. It's well, it's not, and it's not only for AI it's for all kind of system. Uh, Whatever system, even if it's a, a random decision behind or, um, knowledge bases, whatever trust is really impacting for me, the whole quality and to gain trust.

It's all how it's presented. It's you ex I think anyway, from my experience, this is something that really, really. And then you can go into the whole, uh, like modernization aspect of the problem now. Uh,

well, uh, do you have good result the metrics you use? Uh, something that I see very often that will impact the quality is, is when you work on your model in the data send box, or I don't know how we can call it offline, whatever, uh, you have your metrics, which.

Typically, uh, machine learning, recall whatever. And then you say, oh, I have these crazy good numbers and everybody's yeah. Yeah. And then you put your model into production and then, well, bad things happen. You don't have the same. So then you realize people don't like it don't like the results. So for me, uh, the, all the metrics you use to calibrate and to assess performance in the sandbox really important, it has to be the closest as possible to the metric you will use or in real life, or the, you have to create a kind of proxy of how it's going to be perceived by the people that use, you know, so.

This impacts a lot of the qu the quality of the model to me. And then after that, well, it's, it's all in the, like if you have time in your project and if, uh, if, if, then you can go and work like, like more fancy, um, metrics about models, but I won't go there. I think, I think, yeah, it's, it's more these data, uh, human and human and process.

And, um, this shift. Data sandbox and production this and the, the metrics that you use to, to determine this is, these are the clear for me, the most impacting, uh, For the pulse. Okay.

Interviewer 1

That's quite interesting. I, I think you said a lot of good points, interesting that I would like to, to dig into . Um, so, so, um, the inter so I will go, uh, through each phase of the, uh, Of building a ML system, like data collection, data preparation until the end.

And I, I will have, I will ask you some questions specific to each of these phases. Mm-hmm and I think by asking these question, I will be able to touch upon what you've told me, uh, okay. In the last question, and we will be able to, to go into more detail. Okay. So this way I will be more structuring my questions.

Okay. All right. Uh, so let's start with data collection. Mm-hmm . Did, did you ever use, um, data collector? So people that, that manually Fe create some data for external algorithm?

Interviewee
Yes. Yes. I can think about a, is a survey as a way that manually is, is, is it fall into that people inserting surveys?

Interviewer 1
Yeah. Okay. If it's a trainer model, Yeah.

Interviewee
Yeah, yeah. I, uh, yes I did. And, um, uh, also I guess, yeah, yeah, the city, I would say that is quite common that people, some data yes. Was, uh, altered manually. Yes. I, I can say that. Yeah.

Interviewer1
Okay. And did you have any quality issues, uh, with quality issues with the, the resource.

Interviewee
In general when it's manually, uh, addict and manually, uh, for, for this specific, uh, oh, um, uh, yes, people lie.

I mean, they say wrong stuff. I saw it in the, in surveys. Yeah. People like. Okay. And that's interesting.

Interviewer1

Do you have measure to like detect these kind of issues or prevent them?

Interviewee

No, just by digging, just by. It's by going back, observing weird stuff in the data and said, why, why is that my saying that why doesn't make sense?

And then digging and going back to the initial survey and say, okay, well, this person. Told lie about the, let's say the, the, the place it, uh, the, the place she, he, uh, is living in, um, for instance, is house. Where is this house allocated? What is, uh, place of work, you know, because, uh, I'm thinking spec specifically about a survey.

Uh, it was about, um, we did it every year at the city. Um, it's to record, um, The whereabouts of people and, and they, like you install this app on your phone, you answer a survey, where do you live? What's your typical, uh, commute way? Where is, uh, are you, uh, using your bike bus, stuff like that. And then it records all, uh, your tracks, in fact, and then from that, from this data, we weren't figuring, uh, what was, um, the, uh, Uh, the, the mud, the transport, um, I'm sorry.

It was long and transportation mode. Okay. The transportation mode and were firing another technical part. What was the end in the beginning of, of the track? Because, well, data was fuzzy. So, um, we could see that some, there was mismatched the person twice a day. Go from a point and then go back to that point.

It was obviously his house, but when you look at the data, they put their house, I don't know, in whatever. Uh it's you know, it's lies. So, yeah. Uh, so it was by going back, uh, putting stuff on the map, especially when you have dual reference data. If you put the, the points on the map, well, you see quite fast.

Uh, there are some outlier, so this it's a kind of a reflects, I. I had at, uh, working at the city. Yeah. And even I was, uh, yeah, yeah. At the city. Yeah. So this is one way from, uh, manually entered data. Um, I could see, well, Uh, example at the C of fresher in my mind, because I was there a little bit more than a year ago.

Um, the way people would write, uh, make mistake in writing word, typical mistake, uh, just one L instead of two L uh, one R instead of two French common mistakes. So I knew it was, uh, and, uh, these kind of. I knew about them. So I was able to go and look for them, but, you know, people. Don't stop to amaze me. So the possibility of my heart or infinity.

So yeah. Oh, I Def I, uh, detect them. It's either by experience or either by going back and trying to diagnose. Diagnosis, uh, problem I have.

Interviewer1
Okay, thanks. Thank you. Um, and so we talk about data collector. Um, did you ever use external data, for example, a public data set, third party API or web scrape data?

Interviewee
A lot, a lot.

Uh, I even had a, a project that was to, uh, one security, uh, maritime security was the domain and defense. I was, uh, specialized in. There were a lot of, um, Source private source. I mean, uh, government source about the, the, the, the ships that are going into the ocean, but you have also all kind of open source information.

People. They are called, uh, ship, uh, spotter. They, they take picture about the, the ships and they get the information, uh, in the system. It's so it's also, uh, human, uh, but it's open source and the, the project was to put all these information together and see the consistency of it. So that was exactly that goal of that.

So we had all kind of, uh, open source data inside data and working at the city. Um, a lot of the data, uh, from the city is op is made public. It's it's even a law inside the city that, uh, data should be open unless it's private. So, uh, and as a data scientist, sometimes it. It was more easy for me to get data from the open portal than to call the guys working with this database.

And it was all messy having data. So I was using a lot of public data, uh, same for when I wanted to have some government data. It was easier to go from the public side than making the call and having a private data. Well, all this fus and it's always complicated. So there's a price to pay some time to have open that up.

But this price is, was less for me than having the, to deal with the people and managing the source. So yeah, a lot, a lot of open data for me, open source.

Interviewer1
Okay. Okay. I see. And, and I understand that the type of ever you had were inconsist.

Interviewee
Yeah, inconsistency. I see the freshness of the data, especially, uh, when you deal with the, uh, open source, because, you know, I, I, if there's, um, one very popular, open data, it's, uh, open source data.

It's open street map. Now, if you heard about it, it's, it's a great map. All the world is map and it's people that really like that and put the map into, so you have the streets and the streets. We compare it with, uh, what we had at the city. And it was a good source, but you have all the shops and, uh, you know, restaurants that people were.

Pulling information there. And that was not up to date because you know, the someday, uh, I don't know, the, the guy loses his job and has nothing to do, want to improve his CV. So, okay. Why not contribute to, uh, open suite maps? Some schools gonna be up my CV. Put a lot of new data in then get a job and never refresh that data source.

So that happens, I would say that's that's number one, uh, consistency and, uh, freshness of the data with open source. And also, I mean also, uh, sometime you have a source and running and then it disappears. And so it's, it's always risky. I, as it's always risky to, uh, To have open source data for that, for these three reasons.

Yeah. So in general, I'm like, uh, I I'd rather use less data sources because of all these yeah. For the game you have, sometimes it's like whatever, better to be. safe. And so, and, uh, yeah.

Interviewer1

Yeah. I see. And, uh, do you have any like, measure to prevent, so you mentioned some issues. Do you have like any way to prevent these from happening or. To fix this problem,

Interviewee

but I, I like to use the metadata around it. Let's say you have open source government source. You see you, you there's always metadata about it. The last time it was refreshed. Uh, people contributing, uh, uh, maybe there's sometimes there's community community around it. Uh, do you have, see a lot?

It's like open source code. I mean, you look at, is it, is it alive? Is it dead? Uh, some sometime private company contribute are contributing. Good science. So I I'd say that this part is a good one. Uh, that's the first thing I look, if it's open source. Okay. So that's that, that's the first thing it's um, and then, uh, we'll open up the data and look, it looks, looks, uh, what's under the hood and see.

Interviewer1

And, uh, finally, um, did, did you ever use data that was generated by another system that can be ML based or not?

Interviewee

Yeah. Yeah. A lot of sensors data a lot, uh, as I said, uh, for, um, related to defense. You have so many sensors, um, capturing information of what's going on on the oceans.

I mean, Canada is the largest, um, uh, water boundaries in the work. So there's a lot of sensors. So even just that I, I did a lot about that. And, uh, also ly, um, at Company X, uh, I have other project using, uh, system generated data. Uh, yeah, that would, yeah, I, yeah, no, let's say I, I think that the, the, the defense, uh, from example from more into the purely, uh, system generated data.

Interviewer1

Yeah. Okay. But, and what are the issues? Oh, sorry. Uh,

Interviewee

Um, the quantity sometimes there's just, just too much of it. There's a lot, uh, reporting every minute. Uh it's yeah, it's a lot to deal with. Um, That's one part of it. Um, there's always this again, the consistency, especially about the time, uh, it is reported. You have data in seconds, other data in hours. So how do you make sense of it?

How do you, uh, I mean, distill one ship, let's say you have one ship going on, but you have 10 source, 10 sensor getting reporting it. Uh, so how do you consolidate everything? So, so consistency again, and, um, calibration sometime. It depends from a, uh, um, I have an example. Yeah, the city, we had cameras everywhere and, uh, sometime it was not well installed.

I mean, the guy had the instruction how to install it, but. Install it like it's was pointing another way. And ah, you look at the data it's not working, but, uh, then with union ska and everything, that was difficult. So it happened the way sensor sensor is installed. That's something I seen and. And calibration.

Yeah, not well installed, um, or install the bad ways with impact. Uh, you cannot reach the spec. Quality that was, uh, the vendor sold you. Uh, this is something I've seen, um, lack of functioning. You have to maintain these, these, these sensors. I was thinking also, uh, at the city, we have a lot of, um, car counting sensors.

Everywhere, pedestrian counting, uh, cycling. You can see it in the, the cycle path, counting people, but a lot for the cars. And, uh, but sometimes they just stop functioning and. We don't have time to fix it, or we see that it's not working, uh, a week after that because you thought you installed something to send you a message when it's off, but it didn't work.

So then you have big holes in your data data. I've seen it a lot. Like, uh, sensors stopped working and people didn't see or wouldn't see it. Couldn't fix it. So you have big holes, so that yeah. Okay.

Interviewer1

You said a lot of interesting thing.

Interviewee: I'm not, I talked a lot, sorry.

Interviewer1
yeah, don't worry. Uh, so the first thing, uh, you mentioned, uh, sometime data quantity can be a problem because you have too much data because of the sensor.

Yeah. Um, it, uh, is it a problem for you or for the model? Like. Oh, yeah. Is it? Cause it seems less from the model.

Interviewee
Yeah. Oh yeah. Yeah. I see. Because, uh, I was more thinking about like, you have to manage this data, so we have database, you have to put data into database. So it's, it's less about model, but more about, uh, the, the data, uh, management around it.

Interviewer1
Yeah. Okay. I see. Thanks.

Interviewee
But because model, yeah, you, you will have to sample it. Uh, still it's, it's an art to simple data, the right way. You can induce bias by doing it. Yeah, it could be another, it could be, um, I'm thinking right now about, um, too much data. At tech problematic, I was I'm working on and we thought about sampling it, but we said, yes, now we have two problems.

yeah. It's, it's, it's, there's always a risk when you sample something you of course bring a bias into your model and yeah. Yeah. True fix. So, so there's this, it could be for, yeah. Yeah, I see. Thanks.

Interviewer1
Um, you also said that sometimes sensor are not well installed. Yeah. Um, so, um, how do you, how do you fix this kind of issue in your data set? How can you prevent, um, well,

Interviewee

most of the time you have to discard data, uh, you have to, uh, So you, you, you have to fudge this well, well, of course you call the people say, try to fix this, fix this and everything, but it, you have to forget the, the, the, that time, the Easter heat that was not, you know, taking and you cannot, I've asked, um, Personally, um, we thought about ways to like, uh, regenerate that data because it's a, okay.

The sensor is at the corner. Let's say the example of the streets, counting street, the, the counting car, uh, sensors it's it's on the corner, the streets that really look like another corner we could, blah, blah, blah. And then we said, okay, no, let's not do this because it's, it's another problem itself, you know, to.

To mimic that. So we remove this data and try to make decision using other stuff, other kind of data. Um, I've seen also, uh, yeah, so, so that's, that's, that's one thing. Yeah. There's no, no. Going back. Okay.

Interviewer1

I see. Thank you. Um, okay, so, so we're done with, uh, data collection part. We are moving on to data preparation. Uh, have you ever measured the quality of your data and, or tried to improve.

Interviewee

Yeah. Uh, I mean I've built system trying to just put data together and present it without any, uh, well, it's a model to me, but it's not a AI model, uh, to, um, to, but you, I think it's the data sciences job to. Make sure to try to find ways to improve quality of the data because there's no such thing as perfect data.

I mean, I've worked for army and they didn't have perfect data. So, well, I wonder where they have perfect data. I think it's part of the job to, you know, do something about it. Uh, data fusion, as I mentioned is. To improve the data quality because you take data from one source, which ISN imperfect another data from another source, which ISN imperfect.

But together, I mean, there are more than the sum of these two data sources. So there's algorithms to try to, uh, to add fastness in there and you know, this is one part, so yes, I did work to do that. Um, Yeah, that's part of that. Okay. Interesting.

Interviewer1
Uh, do you mind digging a bit in your data fusion, um, system that you have treated?

Interviewee
Yeah. Um, it's um, okay. That's I would have a very, uh, concrete case, so that's, that's, that's more clear. Um, Go back again to these, this problematic of having one boat, going to ocean and not full source of information kept that. So, uh, there are, um, operators in, within the army that look set clean every day, trying to make sense of what's going on on this sea, but pharma.

So if we present to that, All the source, all the data point that are being collected. I mean, this brain's gonna expose, so we need to pre-process it and try to make a picture that makes sense so that it can decide, okay. Is this ship a. It's looking like having a weird behavior. Uh, okay. This ship has a, a flag, but it's, it's like, like it's fishing, but it's a falling ship, so let's do something about it.

But before that, and before the human makes that kind of decision, we had to, to improve the quality of the data. So, uh, we had to do that kind of fusion. So we basically had points that we thought belonging to. Like we receive information from one source, uh, declaring that, okay, we have a shift there. This is the point another, uh, source from reporting the exact, the information from that, that boat.

And we had to put it to put it together. So it was, uh, capabilities. That every point belong to the same entity. So it it's probabilities be behind that. There's a bay framework, but there's also, I go ahead. That is called ster. She, which is another kind of, of, of things. So basically it says, okay. Am I sure not that I'm seeing the same thing, if yes, we were fusing that information for, uh, the operator telling him, okay.

I'm sure. 30% or 100% is the same thing. Sometime we would ask him to just, uh, say, okay, yes, it's the same. No, it's not, uh, when it was fuzzy. So that's an example of data, uh, case, uh, there's a lot of also that. About identity now, I was using more, uh, for this decision, uh, geospatial information. So basically the contact where it, it went into a show, but you can do that also for, uh, identity of the ship.

You have all kind of, uh, visual aspect and also, uh, information like, uh, the country, um, all kind of identification that may not write may not be. Uh, may miss maybe missing and stuff. So same as, as I said, the, the, the, the work we we're doing with the dual location, we could do that kind of, of fusion based on nominat of information.

That's, that's quite common also. Yeah. Okay. Interesting. Um, yeah, I don't know why it's not more, uh, it's it's only used using the more used to the defense area. I think it's super interesting. It could be used everywhere anyway. That's

Interviewer1
all right. Uh, so, um, the next question is, um, is there any other data quality issue we miss you consider. No, I don't. Well, it's more like a, but, um,

Interviewee
not like that, not on the top of my head.

Interviewer1
Okay, perfect. Thank you. Uh, so I will quickly go through the last, uh, section of the interview. I will ask one or two questions. There's like three, three, so maybe four question left. Um, So model model evaluation. How do you evaluate the quality of your model? And as a reminder, uh, by quality, we don't only mean ML performance, like accuracy, blah.

We also mean we also consider other aspects such as expandability, uh, robustness, scalability. So really anything.

Interviewee

Yeah. Yeah. Um, As I mentioned earlier, I'm not really fan of ML metrics, as I said, accuracy, because when you launch your model in real life, people don't care about confusion, metrics and stuff like that.

It, the, it will always depend on how you use it and in the, the context you use it. So I'm a big fan of, uh, proxies that you build and that will mimic your real business metrics. So there's two things there's. The metrics that you use to, to see the performance in the sandbox data sandbox, and then you have your business metrics.

It is launch. We have our dashboard and okay. So, um, I'm I like trying to find PS that will mimic mix these business metrics and stick to it. Uh, news only that, so that's one point, I think it's really important and this is one way of, uh, working, uh, One way to, to, to look at the quality of the model. Um, another way while it's always has been used or not.

I mean, you sit with the user and you see if it's make, if it makes sense for her, him, her. And this is not really easy. And most of the time, it's not part of a project. We ship something like that. But this all UX experience it's it's and I, I want to and turn toward a user centric development that way even further models.

You can't work like three months full time to improve your quality, your accuracy, whatever 2%. Okay. It's beautiful paper. But at the end of the day, it's not, if it's not used for quality for me, so, and. I've seen it so many times. So many models ended up on the shelf. I mean, majority of it, so it's always wipe my heart, but I know it's part of the game.

Yeah. So for me, it's, I would like to tend forward and also my work at Company X it's to center, um, um, the, um, the development around experiment with humans. And trying to find metrics to make it, uh, to capture it and on not only, I mean, you talk to John and say, I like it and Paul it, or hate it. So something like more, uh, you know, objective than that.

And it's, it's not easy, but, uh, I think it it's. Yeah, this is where this is where I see that going. So this is where, what I want. This is not always what I did. You know what I mean? But I, I think that's the, that's the right way to go. But other than that, um, Well, uh, uh, there's always this, the I've I've put system and production, uh, models, but I didn't, you know, work, uh, like with, um, you know, data gift and stuff like that.

In the long term. I never monitored models for a very long time. I never done this. I know this Kiwi's there. That's difficult. And, but I don't have any. Experience to, to report that. Okay. I see.

Interviewer1

Thank you. And you talk about the proxy metrics. Mm-hmm so this could be, um, KPIs if I underst some correctly. Yeah, absolutely.

Interviewee

Yeah, yeah, absolutely KPIs. Yeah. Okay. KPIs that could reuse to, uh, so when we train our model, we look at, at these KPIs. Instead of, okay. I look on accuracy and then cross my finger and then put the then production. Oops. Uh, KPIs are not working, trying to figure out to make that bridge between KPIs and, uh, your training and stuff.

So that's, I think that's a good, and it was gonna be easier to also communicate the results of your models. You know, if you find a common language,

Interviewer1

Okay. I see. And how, how do you, can you improve a KPI based on a Mo how can you change the behavior of a model to improve a KPI by, well,

Interviewee

you're gonna work, let's say you, you do, you do hyper, uh, tuning, uh, Parameter while you can input any kind of, of metrics in there and you, you implement your KPI, you select one or two, uh, or just one, ideally.

Uh, and then you calibrate on that. That's okay. I see. I see.

Interviewer1

Thank you. Um, let me just see, for which question I should ask.

And you answer most of the questions. That's great. Okay. Okay. I will move on to the model deployment section. Okay. How and where are your models employed?

Interviewee

well, um, on the shelves, I'm kidding. Uh, no, no. Um, because, uh, here, prior to when I was working defense, I was mostly working in R and D uh, so of may, but some of what we did were deployed, but it's not, as I mentioned.

Uh, machine learning models. It was system that we develop where there were algorithm processing data and presenting data processed to a user so that it can make a better decision. So this part of algorithm could have been machine learning, but I don't think like deployed in production in the past.

There was, it was not machine learning, but, uh, at Company X and at the city, yes, we have deployed. We have deployed, uh, at the city, some models that were, uh, depends, um, at I'm thinking about two models at the city. One was for, uh, predicting the count of, uh, cars that were passing at the volume of fact of traffic.

Um, so it was predicting. Hourly, uh, was deployed into our OnPrem, um, environment, uh, production and another one. Was train only once a year, deploy that on plan accessible via app API. So it was very, um, custom because we had this on-prem thing, uh, at Company X it's more cloud. Well, it's all cloud, uh, cloud infrastructure.

So, um, And the I've, I've worked on two project at Company X and they were both on AWS. So using all this AWS suite. Yeah. So is, I mean, when you say, how did you mean. Like technologies or the human people using it? Uh, I, I assume it was technology, but maybe I for while. Okay. Okay. Okay. Uh, and for me,

Interviewer1

from your experience is the deployment of a machine learning software system, a problematic or complicated problem?Uh, challenge.

Interviewee
Well, uh, it, it is it's, it's less and less, but still, I think, you know, a lot of people. Think that it's easy and they look at their neighbor and say, okay, they're doing it. I'm not doing it. What's wrong with me, blah, blah, blah. But I mean, I think, uh, and not a lot of people, uh, like companies doing it right.

And they all. Being very scared about it. So there's, you know, there's this perception like by vendor, that's so easy push button and everything is there, ML ups, blah, blah, blah. But I think it's harder than what it's being sold. Maybe in a couple of years it's gonna be mainstream, but it's not there yet that there's still problematics about that.

Interviewer1
Yeah. Okay. And why is it difficult?

Interviewee
I'd say that one of the most difficulty that I see it's, uh, it's all related to the data source. I mean, these, these are big pipelines and if you have one of the table that is data is missing, uh, it's it all falls and you can do nothing. So you have to prepare all these fall.

Plans, you know, that I think the data, the upstream data like this, all, it, it it's, it's a pipeline. I don't know how to say it differently, but the, the source of the pipeline, the entry point is sometimes. Problematic. And this is where I think most of the problem happens because it's outside of your control.

you know, okay, you're looking, you, you, you're taking data from certain data source sources. And if one of them is missing well, shall buy your, your system. And even if it's code, I mean, the more you can do it's, you know, sending alerts and stuff like that. But at the end of the, the year, if, if, if your system has so many alerts, they may just shut it down and say, okay, whatever, let's forget about ML, you know, that's, that's something.

So really the upstream data. Okay. I see

Interviewer1

. That's a, that's a great, interesting point. Uh, you have, because, uh, I was going to ask a question to you on it. Okay. So have you encountered issues with data during the mainten of MLS system and yeah. You said yes. So all the time could you, yeah. Could you, could you gimme some example?

Interviewee

Um, yeah. Yeah. I'm thinking about, uh, a system right now at Company X. Uh, it's always something failing it's about there there's latency into, uh, all the data latency into filling. Some of the table we use what the data is. Uh, and these late latencies are, can be caused by a lot of things and cannot do well. We, we, we define the limits where, okay, this model can be trained on that amount of data.

You know, ideally it's, I don't know, 24 hours, but in the worst case, we made all kind of tests in the worst case, it can be on 12 hours and it's gonna have okay. Result still. So we did all this kind of definition, but still, you know, there's always latency in the, the system, even though, uh, we heard, okay. Uh, never more than three hours.

Well, it happens. And. You know, it's always exceptions. Oh, this, oh, so that happened. Okay. So yes, it happens.

Interviewer1

Okay. And when you say, I see. Yeah. And when you say, I say latency you mean that the data, uh, you receive it too late to refresh it.

Interviewee

It's not useful anymore. Yeah. I mean, there's one database I'm thinking about this database is one table that we use a lot.

Uh, our system take data in that, uh, arrange it, uh, build a model and stuff and also do the inference also on that same table. But this table. Is filled by other systems that are putting data systems and also all kind of, um, data, uh, information, well, data treatment that is done before that. What happened if one of the system has an issue?

Uh, on that is filling that one. So this is where the Latins see happens. Multiple reason why, uh, we could have Latins in this data. So our pipeline goes and look, oh, there's no data. Come back an hour later. Oh, no data, no data. And sometimes it's more than 12 hours, you know? Okay. Or often it's, it's less than 12 hour, but still at the long, on the long term it can affect.

So we had to define. These worst case scenarios before that we, so it was filling out the time and then we had to. Test to see how missing data is impacting the quality and stuff. So we had, but, but this is, this is, this happens a lot, like, yeah, I guess okay.

Interviewer1
I see. So basically, okay, so your, your data. You, you have data that is supposed to come, like let's say period periodically.

Yeah. And at some point you don't get it and your model is retrain on that data Uhhuh. But if it, if the data is not there, then the model is, cannot be retrained. Yeah. Yeah. So the model becomes sales. Okay.

Interviewee
Yeah. So yeah. Yeah, that's it. Uh, we, we let's say this model, we train, uh, needs 24 hours of data. Well, the day before to train, to make prediction for the next day.

Um, and it was able to get on the, um, five hours. Well, it's not enough because, uh, it's not enough. So we need to keep the model from the deep before in place to make the friends. Uh, but we made some tests and let's say that 24 hours is the best, but we can train a model in that decent result with, I don't know, 18 hours, 17.

I don't remember. What's the number of hours that, that can be used to train the model and have a valid model under than. Forget about it reuse the old model, but yeah, you need to think about these things within, at the beginning and then, oh, oh, okay. Wow. It happens. And then you need to, to have all these fallback plan because it will happen.

yeah. There's always a problem in the upstream data.

Interviewer1
Okay. See, and your way to avoid this problem is fall back. So a previous model or maybe a U yeah, maybe a six.

Interviewee
Yeah. I like to have realistics but the way the pipeline was set up was not possible. So, uh, we, we used the node model.

Interviewer1
Okay. It's very interesting.Uh, it's fun because, um, we already answer a lot of my questions. Okay. But I have to

Oh, yeah. Uh, maybe, yeah. Okay. I will, I will ask two more question or two or three. Uh, so we finish that on time. Uh, have you ever faced problem related to the scalability of the trade model? For example, scalability regarding the number of machines deploy onto or anything.

Interviewee
Yeah, well that is usually, I don't know maybe, but, uh, there's a point where I, uh, I see it's I cannot be helpful in that topics were, I'm not, uh, I can say humanities at the right time.

But that's about it for scalability, you know, I've deal with big data and stuff, but not related models and deployment and stuff. So I'm not the right person to that.

Interviewer1

Okay. I see. Perfect. Thank you. Uh, yeah, maybe, maybe this one, I will try, have you a very investigated the explainability of your train models?

Interviewee

Uh, some, you know, some have, uh, I don't think, I don't know if it falls into explainability. Uh, but you know, uh, like the, the feature importance, sometimes a model as internally, that kind of metrics. Uh sharpish. I don't never know how to pronounce it. I played lit bit with it. Uh, what I did is.

Yeah. That's and makes sense of it. Like, like digging small into small data, you know, you have, you can explain what big data, like the whole data gives you a picture, but I like to go sometimes picking examples. Why it did that for me, it's important. Helped me, like, I, I take one example. That is weird. So I dig in.

And try to understand why the model predicted that. And so, and also it gives me a new notion when I talk to clients, I say, okay, give that. I don't want it to fall into of course, exploring every year data point. But sometime it helps also for me, for my brain. Now it works. It helps me to, to fall back on concrete cases.

Interviewer1

Yes. Yes. Okay. I see. Thank you. And, uh, so last question, um, in your opinion, what is the most pressing quality issue researchers, uh, should try to solve?

Interviewee

Well, um, I think it's that gap, the famous gap and famous gap between, um, What you do when you know, we read paper and what you do when you, you train your model and send books and when you launch it and an example that is, um, I've seen some research about that.

Uh it's um, recommendation systems, um, Test is always you train your model and then, okay. You have, uh, on your website, some kind of recommendation. Okay. User, like you likes to buy this stuff, uh, based, but you train it and then you launch it and it never, and the mid, and they don't have even metrics to, I was talking about plexi.

It's super hard to. Cocky about, um, the crate and stuff like that, that you can mimic make in your data sandbox. And it's, it puts a lot of pressure when you, you, you try to deploy model and stuff. So I think that gap would be really interesting. I don't know if and how research can do something, but it's definitively somewhere I would look at and.

I would love to have some researcher input yeah. To reduce that risk offline versus online. That's that's something that is, so yeah, I think this one is a good one, Yeah. thanks a lot. Uh, so yeah, that, that's, that's it for, for.

Interviewer1
Okay, thank you. Thank you a lot for your time. Really great input. And thanks. I hope it, yeah.

As I said, it's, it's less on machine learning, but anyway, still. Oh, thank you. Very, very interesting. Okay. Thank you. Thanks. Yeah, it was super relevant. Thanks a lot. Yeah. Yeah. True. bye. Bye. All right.

Created with the Delve Qualitative Analysis Tool (http://www.delvetool.com)