Interview 36 - Ilan

Interviewee

Okay, this meeting is being recorded.

Interviewer 1

All right. Uh, so to start off, can you give us a bit of information about yourself and how many years of experience you have in mesh learning and also in general?

Interviewee

Sure. Yeah. So my name is Interviewee. Um, I have, um, over four years of experience in data science and machine learning.

Interviewee

Uh, I have a master's degree in computational linguistics. From the University of X. So, um, I primarily did studied and did research on, uh, natural language processing kinds of projects during my master's degree. And since then, um, and before my master's degree, I worked for a year in, uh, just as a consultant at a data science company where I actually did more data engineering kinds of work.

Interviewee

But since my master's degree, uh, I've been in two. Uh, one at a company called Company Y, where I did, where I worked on, um, automated scoring for standardized tests in Country X. Um, so in Country X uh, there's a lot of standardized testing that happens at the school level, um, where, you know, they make kids like write essays and stuff and they need to be scored.

Interviewee

So I developed machine learning systems to. Uh, student responses to tests. Um, and now at Company X where I've been for only about six, seven months, um, I work mainly on HR data. Um, so, um, all kinds of HR data, but mostly human resource, internal human resources survey data. So I work on the pipelines that process, that data.

Interviewee

And I'm currently developing, uh, proofs of concept for, uh, machine learning systems to help analysts to um, kind of search through and, uh, ingest and make sense of, uh, human resources, survey, comment, comments. Um, but I haven't been in this position for very long , so yeah. That's

Interviewer 1
embarrassing. We're happy to have you with us today.

Interviewee
Yeah. .

Interviewer 1
Um, so I'll ask you the first question. Uh, what are the main quality issues you have encountered with your data model or system

Interviewee
so far? Yeah. Um, so I think I'll draw from my experience at my previous company because I have more time. I had more time there. Um, when I was working on, uh, models that did automated scoring, the main issue was.

Interviewee
Um, we would have, so we would develop initial models that would assign scores, say on a range of one to five, for example, uh, to student responses. Um, and every year, or every couple of years or so, we would have evaluators come in and assign scores to, uh, just to make sure. , um, we're keeping up with, uh, human annotations of, of the, of our, just to make sure that our system is matching up with humans.

Interviewee
Um, so we would normally see two issues. Um, one is that, um, over time, because students tend to write about different things and the composition of. The texts that we were scoring tended to change over time. Um, our system degraded in quality, um, because it was built on a corpus of texts from years ago. Um, and both, uh, composition, uh, both the vocabulary that students use on the language that they use has changed and also the content that they use has changed in response to the.

Interviewee

Questions. So, um, our kind of text evaluation models didn't, didn't hold up anymore. And the other angle to that was that the human annotations or the human labeling, uh, that was being done for quality checks, um, they would also, sometimes their patterns would also change over time. , we would have one set of people who did the, provided the original data that we trained our original model on, and then maybe two or three years later, we have a different set of people who might have been trained slightly differently.

Interviewee

Um, and also sometimes educational standards will change, um, the way that they assign scores than the things that they pay attention to sometimes change as well. Kind of from both sides, um, from the like target label side as well as the data side, we would see data drift. Um, so yeah. If that makes sense.

Interviewer 1

Yeah, totally. Thank you. Uh, I'm gonna ask you a couple of questions regarding data collection and basically what I want to know if it, it's if you ever add an issue, uh, with the data from one data collection process, right? Uh, . I, my first question is, did you ever use, uh, data that was manually collected and I guess yes, since you, you, you just mentioned it.

Interviewer 1

Um, yeah. Yeah. What were some of the issues in general?

Interviewee

So, I guess with my, I mean, I guess with my previous job, there was a very structured way to collect data. So maybe, I don't know if maybe that. Not as interesting of an answer to your question because in, in that industry, there's um, a lot of rules around how, um, data is collected because people are trained specifically to score student responses that we will then train on.

Interviewee

Um, and both our company, our company would work with state boards of education to collect that data. Um, so that was, it was a very structured and kind of supervised process. Um, but I, I can say that in my, in my current job, um, you know, I'm working on human resources survey data. So, um, basically we send out surveys to people inside of our company, sometimes randomly, sometimes to everyone.

Interviewee
Um, and so I guess one of the big issues with that is response rate itself is not always very high because it's up to people whether they want to participate in the survey or not. So usually the people who do respond are people who maybe have something to say whether positive or negative. A lot of the time it's negative.

Interviewee
So we have overrepresentation of, um, maybe one kind of opinion. One kind of sentiment, um, that's represented in the data. Um, and also the surveys that we currently conduct. We have one that's on a daily cadence and some that are on monthly, like biannual, some that are monthly, so different cadences. Um, and you know, if something big happens in the company just before, the release of a certain survey, then, then the data that we get from that survey becomes overwhelmingly about that topic.

Interviewee
Um, and so if we have, like I've been developing like, um, a clustering model that, uh, analysts can use to, um, kind of apply to data over a period of time and track how topics are changing over a period of time. Um, and so obviously, If you develop it, say at the beginning of this year, and then you have like every three months over this year, you have a, you have one survey and a few things happen over the course of the year.

Interviewee
Um, people's responses are gonna be affected by those incidents and they will change over a time. And so what you developed at the beginning of the year, um, uh, you know, how you maybe named and grouped the clusters may no longer be valid over time. So, . Yeah, that's one problem. Um, and that has to do with the way that the data is collected, which is people just respond whenever they want.

Interviewee
Yeah. Super far.

Interviewer 1

Thank you. Um, and did you ever use external data? So that includes public dataset, third party APIs, and web scrape data.

Interviewee

No , um, in neither of these jobs do I use that account data? Uh, yeah. We just, perfect. Yeah, we just got Thanks.

Interviewer 1

Uh, and did you ever use data that was generated by another system?

Interviewee

Um, no, I don't think so. I have specifically only worked on data that mostly that was produced by humans. , um, because it's text data that was written by humans. Um, there was one project that I worked on a few months ago, it didn't have to do with text data. Um, it was more about, uh, studying, um, just data about, I don't know if I can talk about this, but data about how people are working in person versus remote.

Interviewee

Um, so that was data provided by another team, but it was still internal to the company. Um, so yeah, I don't know. I guess it's from a different system, uh, like data about how often people are, um, you know, using company resources and office locations and things like that. So yeah, in that context only,

yeah.

Interviewer 1

Yeah. And, and did you have some issues with that data? .

Interviewee

Um,

Interviewee

yeah. Um,

Interviewee

I think one of the issues that came up was, um, it had to do with, because people can themselves specify what location they are. Um, and certain attributes about themselves that was fed into a system. Um, and then the system would process them and allocate them to different geographical locations or regions.

Interviewee

Um, and in our data pipeline, we were trying to use that, um, as a feature or as a, as a data point to compute certain metrics. Um, but because first of all, , uh, this was self-reported data sometime, and it was not controlled, like there were no controlled options. People will just enter whatever they want, and then that, in that system, people came up with a way, I don't know, some kind of rule-based way or something to organize people's inputs into things, categories that made sense, like cities or towns or countries.

Interviewee

So for us to consume. . Um, there were some concerns about whether it was accurate because we wanted accurate information about where people exactly were in the world. But, you know, um, because there's two steps where arbitrary decisions are being made about what people are entering. Um, yeah, that, that was an issue, I guess.

Interviewer 1

Okay. Thank you. Um, so moving on to data preparation. Um, have you ever measured the quality of your data and or tried to improve it?

Interviewee

Um, have I tried to measure the quality of the data? I don't know if that has ever come up for me in that sense, because I work only with, mostly with data. It's produced by people.

Interviewee

So I don't know that there's an issue of quality of that data. Um, maybe like what I was talking about with that particular project where we were working with, um, data about, you know, in-person versus remote working just about making sure that the data is consistent. Um, , you know, um, there's a certain field or certain column that has so many values that are distinct, but many values, making sure that all of them make sense and that, um, yeah, there's no overlap between them, things like that.

Interviewee
Um,

Interviewer 1
yes, yes. Perfect. Thank you. Uh, and is there any other data quality issue we missed that you consider relevant?

Interviewee
Um, I think specifically in the context of text data, um, there's, there's attention we have to pay a lot of attention to, um, how people are, how people are writing. Um, because in my old job it was important to pay attention to.

Interviewee
even because students were producing texts that we had to judge for accuracy and language. So, um, you know, are they using punctuation or are they using capitalization or paragraphing? So our data pipeline had to preserve all of this information, um, and not corrupt any of it. Um, so that was important in my current job.

Interviewee
Uh, in current position because we collect text data from people around the. , they write in different languages sometimes and using different, um, like writing or punctuation convention. So all of that we need to make sure that it's getting translated. Um, and we're not leaving out anything because we're not like reading it in English or whatever.

Interviewee

Anything, something like that. Um, so yeah, I guess just challenges specific to text data. , um, I didn't talk about before. Um, so yeah, I guess, I guess that,

Interviewer 1
okay. I'm not sure if I underst send you, uh, what did you meant with, uh, the punctuation and Oh, the difference between, yeah.

Interviewee
Oh, just in my old job, because we were doing automated scoring of student like responses to tests.

Interviewee
Um, if. Because we have to, because some of our models were judging, like grammar and writing skills of students. So, um, like if, if, if people are typing like in weird symbols, for example, um, or like, you know, there are like different versions of quotation marks. Um, That may, and sometimes it's not written in Unicode or you know, it, we have to be able to understand each and every character correctly and not lose any information, um, in that sense, like, um, yeah.

Interviewee
Does that make sense?

Interviewer 1
Uh, yeah. Yeah. B b basically what you're saying is, um, uh, your data ingestion, ingestion, ingestion, pipeline, Must be careful of, um, maybe cons considering the encoding when, uh, yeah, when you're considering

Interviewee
data. Yeah, yeah. Considering encoding, um, spacing sometimes like, uh, the, the application that collects the, that people are typing into, um, it might have.

Interviewee

like when people enter like new lines or spaces, they get represented as like, um, you know, like, like Arrow B Arrow or like, um, slash r slash n and you know, we have to like regularize all of that and make sure we're making sense of it all. So, um, because there are different systems that collect the data, it's still in an automated way.

Interviewee
when people are typing into an application, sometimes they, it comes through in different, the data reaches us with weird symbols or, yeah. So, oh yeah,

Interviewer 1
I understand. Thank you. Um, so moving on to model evaluation. How do you evaluate the quality of models? And as a reminder, quality is not only defined by ML performance or FS accuracy.

Interviewer 1
But there is, there is also other aspects such as explainability, robustness, uh, efficiency, scalability.

Interviewee
Yeah. Yeah. Um, so in my current position, uh, we, we are developing some solutions that, for example, one that, um, helps analysts to track. The topic composition of a corpus of comments is changing over time.

Interviewee
Um, so in a project like that, we uh, first develop a proof of concept, uh, just to see, just to show that using this like whatever clustering method and then applying it to new data, we can develop a tool that analysts can use. Um, and then we develop like a larger version of. . Um, and when we do that, uh, we have to pay attention to, um, you know, how are we gonna deploy it and will it be fast enough, um, in performance.

Interviewee
So when someone enters a query into the tool or makes a selection and the visualization has to change, um, it obviously has to go and retrieve some information. Or it might have to recompute some clusters or something like. How much time that's taking. So we usually will involve some analysts to do a test, um, with a few test queries, um, or a few test scenarios, um, just to see how, if, you know, the speed is satisfactory.

Interviewee

Um, and then, then on the other hand, uh, we also get them to. Develop a set of test scenarios and test queries and, uh, use the system and, and just verify that it's producing the results that they expect. Um, so there is a lot of subjective involvement, um, where people are actually the end users of the system, who are the ones who really matter or will use it.

Interviewee

And then let us know if they see any shortcomings. Um, That, that's, these things are really in the product productionization phase. But I guess before that, the way that we evaluate quality is, um, really by experimenting with different techniques or different algorithms, like different, say if we're using different clustering algorithms or whatever, um, comparing them by some metric.

Interviewee

Um, but I mean, yeah, those are just, you know, it depends on the project, what metric we are using. Um, so yeah. Um, but yeah, at the end, in the productionization phase, that's how we, we just involve a lot of analysts.

Interviewer 1

Super. So from, from what I understand, uh, latency is a concern for you guys. Uh, so if you read an issue with that, what

Interviewee

happens?

Interviewee

Um, yeah, if there's an issue with that, then we typically, I guess we, um, try to see if. There's a way that we can make the underlying model if there's an issue with the speed, maybe we try to find, see if there's a way to make the underlying model lighter. Um, maybe it, it has been, uh, like if it's, um,

Interviewee

uh, you know, if it's using a lot of, um, Parameters, maybe. Then we see if we can make it lighter by, by training like a, like a more condensed version of the model. Um, or we try to, uh, if we have to, like, if it's about, okay, user has to enter a query and receive and receive some results, if you know that's being too slow, then uh, we try to see if we.

Interviewee
um, uh, reduce the corpus that is being searched through something like that. Um, uh, just ways to, uh, minimize the, the, the time, minimize the load on the system, uh, by, um, trying to select, uh, what, what the search is being performed against or the number. systems that have to be query to, to, to return an inference to then use.

Interviewee
Um, but if it's a quality issue, I mean, if it's an issue with, um, the, I guess, you know, if like, um, a certain set of, uh, all right, data points are being assigned to the wrong class or something like that. . Um, we typically take note of that and try to collect data and then just improve on it at the next iteration.

Interviewee
Um, usually in that kind of scenario, one of the problems is that maybe for that particular, uh, class or that, um, in that, that, um, particular kind of data, there's just not enough data points. Uh, adequately train a good, like, adequately get like good recall or good performance in that class. So, uh, we just take note of it and we, I guess, um, try to collect some more data while the system is being used, um, so that the next time we can just train and try to improve the performance of that, that particular class or the categor.

Interviewer 1
Perfect. Thank you. Um, so moving on to mo, uh, to deployment. Sorry. Yeah. Uh, so how and where are your models deployed?

Interviewee
Yeah, so, um, here at Company X we use Azure for everything. So, um, our. , our models basically live in, uh, Azure Blob storage or data lake, and uh, we usually we're using the Azure machine learning works.

Interviewee

There's a service called Azure Machine Learning, um, where you can create pipelines, load models from where they are stored, and then, um, perform, inference and return. . Um, and you can create endpoints that are then called from, uh, other workflows. So we construct like data processing workflows in a service called Azure Data Factory or Azure Synapse.

Interviewee

And then these components will make calls to Azure Machine Learning that then retrieve the model from the model storage, perform inferences, return it to. Data Factory pipeline and then return that to the ui, something like that. Um, okay.

Interviewer 1

Perfect. Thanks. Yeah. Um, what are the challenges you have encountered during the deployment of a machine, machine learning software

Interviewee

system?

Interviewee

Um, yeah, challenges with that, I guess. Um,

Interviewee

Speed and time, uh, of, well, actually most of the systems that we use are not really, that, there's not really a very strict requirement on, okay. Um, if a user makes a query, they have to get an answer in like a few seconds, or it's not, we don't have systems really like that. But I guess one challenge is, um, Sometimes, uh, inference takes a long time on, especially when we have a lot of data and some systems need to be refreshed on a daily basis.

Interviewee

So, um, we end up having to run the pipeline to do data processing and inference, and that takes a long time. Um, and sometimes the way to improve that is, Optimizing the code that does the data processing. A lot of that code is in, uh, using PI Spark or Scala or something like that. Um, so we can either optimize that, um, or uh,

Interviewee
yeah, it's usually that only that, that we can change or we can try to use more expensive compute, I guess, to do the, in. But yeah, we're limited in that. We, we can't really, we try to avoid using expensive compute. So yeah, I guess that's one challenge. I don't have that much experience really in ML Ops, so maybe I, uh, and I also haven't been in this position for that, that much time, so maybe I don't, I don't know enough about the issues that could come.

Interviewee
Yeah, yeah, no

Interviewer 1
worries. So, so just to summarize, so when issue, let me repeat it. It's Friday and my English, it's getting worse. So one issue, , one issue you have is, uh, the latency and one way to solve it is, uh, to, to refactor the pipeline basically and

Interviewee
make it more efficient. Yeah. Um, Yeah, make it more efficient.

Interviewee
Make the data processing site more efficient. Um, and the other way is, yeah, just to use more expensive compute engines for performing inference. Mm-hmm. . Yeah. Yeah. Or make the models lighter. But I mean, when you're productionizing, you can't really do that. .

Interviewer 1
I see. And and just to be clear, when you, you, you say make, make the model lighter, do you mean, um, reducing the number of layers?

Interviewer 1

Yeah.

Interviewee

Okay. Yeah, I just mean like retraining with a lighter model. Like, um, yeah, not, yeah. Reducing the number of layers or using a simpler model rather than a more complicated one that runs for a longer time. . Yeah. But that's more with, in the experimentation phase, like sometimes you do the experimentation and then you decide that one, one method is really good, but then when you try to scale it, it, it, um, is a little bit slow to refresh or use in production.

Interviewee

So yeah, I guess in those situations, either you try to optimize somewhere else, or at least in my limited experience in this position, that's what I've. Try to optimize in the data processing or, yeah, usually we still stick with expensive, with the bigger solution because like I said, in this work, I'm not, we're not really restricted so much by latency issues.

Interviewee

It's, it's just like something to keep in mind, but it's not like a very strict requirement or anything, so.

Interviewer 1

Okay. Perfect. Yeah. Thank you. Um.

Interviewer 1

sorry. Uh, yeah. Um, how do you ensure that the quality of a machine learning software system does not decrease over time?

Interviewee

Yeah. Um, we try to incorporate, um, um, monitoring for data drift. . So, um, specifically when it comes to text data or well, not just text data, any kind of data, even if it's like numerical data, um, every time that you get like a new batch of data, you can compare it with your training corpus and see how different it is from the mean or some measures of medianess or, or average.

Interviewee

Um, and if you observe a trend over time that there's a lot of divergence, then, um, look into that and I guess, uh, pay more attention. See if, you know, take some data and then take some inferences that the model is pro producing and, um, kind of do an evaluation and see if they're still good. Um, Yeah, I think by monitoring data drift, that's the most important thing because usually like half of the issues with quality and machine learning are because the composition of the data has changed.

Interviewee

Um, and so it produces results that are different from what is expected.

Interviewer 1

Yeah, sure. Thank you. Um, have you encountered issue with data source during the maintenance? .

Interviewee

Um,

Interviewee

yeah, but I mean, our data source, as far as I know that I've worked with, it's um, usually already like, vetted by a different team or, um, like. It comes from somewhere else. So I don't know if I have directly experienced issues with the actual data source. It's more about like, oh, if the, somehow the connection isn't being made anymore to the, where the data we achieve the data or something like that, some kind of technical issue.

Interviewee

That's the only kind of issue that I've really experienced. . Otherwise, otherwise, no. Um, the flow of data has always been there. I, I haven't experienced like, oh, suddenly we're not getting the data anymore. That doesn't really happen. Oh, okay. I see.

Interviewer 1

Thank you. Um, is there any other issue with maintenance or deployment that we did we, that you'd like to mention?

Interviewee

Um, no, I don't, I don't think. Perfect.

Interviewer 1

All right. Uh, I'm gonna list a few quality aspects and, uh, ba basically, if you had an issue with one of them, uh, feel free to to to to mention it. And, uh, some of them we already covered, but if you want to add something else, uh, feel, feel free to to, to mention it. Yeah.

Interviewer 1

Uh, so fairness, robustness, explainability, explainability, scalability, data privacy, and model security.

Interviewee

Um,

Interviewee

I think, I guess we talked about some of them, but, um, fairness and data privacy are important. Um, fairness wise, in my previous position, um, it was something. I mean, it was not just a concern of the data science team. It was a concern of the organization as a whole and of our customers that we be, that we should be fair to, um, different groups of students whose data we are scoring.

Interviewee

Um, so part of the process to get a model approved was, um, sometimes having blind tests, um, to. Basically they would give us a set of data to score and return, and that data would be composed of like different demographic groups, um, or people of different students of different backgrounds. Um, and then they'll take that data and they study, they study it, and basically make sure that there are no big differences between the way that one group was scored against another.

Interviewee

we were blind to that. Um, that people doing the inference or training the models don't know the composition of the data. So, um, yeah, that was one thing that we did. If, if it so happened that, um, a model failed that test, then we would have to retrain, uh, possibly take more data or do a different kind of sampling.

Interviewee

Um, . Yeah. So that, that used to happen back then. Um, with regard to data privacy, uh, yeah, that's important. Uh, especially in my current position. Um, we, I mean, the way that we enforce that is just by making sure that, because all of our things are hosted on Azure, on the cloud, so, um, just restricting access to our.

Interviewee

um, for the data, for the models, uh, even the, the, the user interface where the, any model, anything is consumed, any results of the machine learning system are consumed, restricting access to that. Um, yeah, that's pretty much it. That's all we do.

Interviewer 1

So, so for data privacy, you're, you, you just restrict the, the. The attribute you can access to. Am I correct? Yeah.

Interviewee

Yes. Um, yeah, pretty much. Oh, well, yeah, we also do this thing. Yeah. I mean, yeah. We also do this thing where, um, we, right, because of data privacy, we need to make sure that the identities of people are protected.

Interviewee

So, um, there's a system of hashing. , uh, none of the data that even that we train on or that analysts consume, they cannot actually see like the IDs of the actual IDs of employees, for example. Um, they see like a hash id. Um, and that applies even for like team IDs or like manager information. Um, or Yeah, any personal information.

Interviewee

Uh, there are hashes to hide people's actual IDs. Um, and then sensitive information like, uh, demographic attributes or gender or race, things like that. Um, those are com usually a completely hidden, so we cannot even use those. , um, use those attributes for any models and, uh, we are not allowed to show them on the

user interface either, so.

Interviewee

Mm-hmm. , yeah, we do do that.

Interviewer 1

Okay. Makes sense. Perfect. Thanks. Uh, I got two more questions for you and, and then we're finished. Uh, so in your opinion, what is the most pressing quality issue researchers should try to solve?

Interviewee

Um,

Interviewee

The most pressing quality issue that researchers should try to solve? I think, I mean, I don't know that I have that much experience over different areas of machine learning research, but from what I have seen a lot, most of the time, the thing that makes the most difference is. Data quality or how well data is representative of, um, the classes or categories that you are trying to maybe classify or that you're trying to get more information about.

Interviewee

So I think just focusing on getting a representative set of data usually helps with almost everything. Um, . So, but I don't know if that, maybe that's not really a quality issue. It's more about just having good representative data, but No, no, it's fine.

Interviewer 1

I, I mean, uh, whatever comes to your mind. Yeah.

Interviewee

Yeah. I guess that's what comes to mind as being most important, uh, more important than like, using state of the art methods or whatever.

Interviewee

Yeah. Yeah. That,

Interviewer 1

that's a good point. And finally, do you have any other comment about the quality of machine learning software system?

Interviewee

Um,

Interviewee

I don't know. Um, comments about the quality of machine learning software

Interviewer 1

systems, or actually just something you forgot to mention that that's more Oh, what I'm

Interviewee

asking. Yeah. Uh, . Yeah, I don't know. I, I don't know if there's anything I forgot to mention. I feel like you covered a lot of topics, . Um,

Interviewee

yeah, I, I don't know. We talked about, um, collecting data and, uh, I guess maybe one thing we didn't talk about is the experimentation. As much where you are trying to use different, trying to see which approach will work the best. Um, so the way that I go about that is I usually have a very minimal benchmark.

Interviewee

Um, and I like the benchmark to be even like a rule-based system or something very, very basic, like, um, a very basic. , even a logistic regression model, or even, not machine learning, but like, um, just like, like if you're developing a search engine, for example, um, just trying to use like keyword matching, use that as the baseline and then.

Interviewee

yeah. Explore other methods on top of that. Um, I think that's usually helped in deciding whether something's really good or not. Um, sometimes I've seen people will start with, start somewhere very high or start to compare some different methods that are already kind of complicated. Um, and then, yeah, you can compare among them and pick one.

Interviewee
it seems to give you the best metric, highest metric in performance over the others, but maybe it's not that much better than a much simpler system or a much simpler approach. So yeah, that is something I always keep in mind. Yeah, that's a good point.

Interviewer 1
Yeah. All right, so we're finished, uh, with a question today.

Interviewer 1
So I'd like to thank you for, for coming for this interview. I think what you mentioned will be, For our study, and I really appreciate the, your time

Interviewee
here. Okay. Yeah. Yeah. Thank you. Thank you. Um, , I really, yeah. Thank you for, um, this interview and I, yeah, I wish you all the best with your research and hope it was useful,

Interviewee
Yeah,

Interviewer 1
it, it is. Thank you.

Interviewee
Thank you. All right. Have a good weekend. Bye bye. To she. Bye bye.