1:51

Alexey

**This week, we'll talk about machine learning system design interviews. We have a special guest today, Valerii. Valerii works at Blockchain.com as a head of data science. Before that, he worked in quite a few places. More recently at Facebook – in WhatsApp – as a user data privacy technique lead and before that he worked in Alibaba Russia as VP of machine learning, at X5 Retail Group as a senior director of data science, and quite a few other places. Yandex I think as well. Then also, Valerii is a Kaggle Competition Grandmaster and you are ranked globally in the top 30. That's amazing.**

2:32

Valerii

I was. I don't know. I am trying to take a look now because there is an exponential decay if you don't compete, and what is even more important, if you don't win, your score is decaying. Kaggle is an addiction. So the best way is not to go there because you can suddenly find yourself doing the Kaggle again.

2:57

Alexey

**Yeah. So, I got my Master’s and then for me, it was enough. I thought that it was just too much time.**

3:02

Valerii

I think you made a very, very wise choice.

# Valerii’s background

3:06

Alexey

**[laughs] Okay, so I briefly already told everyone about your background. But before we go into our main topic of machine learning system design, maybe let's talk a bit more about your career journey in detail. Can you tell us a bit about that?**

3:21

Valerii

Well, sure. Let's start from the current time. As you said, I'm head of data science at Blockchain. So a bit about blockchain, first. It's a very old crypto company. When I say very old – it is *very*, very old. It was founded in 2011. Try to come back in your head to 2011 and imagine you're the person creating the company called Blockchain. I mean, come on. It's like they created a company called Amazon in 1997 to sell books online and you're still alive. So the company works with cryptocurrency, but in a very, to some extent, classical way.

Initially, there were two friends that were working on a company named Coinbase. One of the guys was saying that the money has to be in the custody of the company and another guy would say, “No, no, no. The money has to be in the custody of the people.” So it has to be a non-custodial wallet, which means that nobody except you has access to the wallet. After that they parted ways and the other guy founded Blockchain. It started as a wallet and then it turned out to also be an analytical tool for providing on-chain analysis, then an exchanger and trading.

Basically, to some extent, it's a very classic business exchange wallet and analytics, but for non-traditional assets, as we can say since crypto currencies are non-traditional. As head of data science, it's awful – it's a terrible job title, because it's a very broad definition. Head of data science – who is that person? In blockchain, the head of data science is the person who's responsible for data engineering, machine learning operational engineering, machine learning itself (makes sense), data Analytics, BI (business intelligence), product analytics – however, the difference between product analytics and data analytics is so thin that I don't see it. I see no difference, almost. I have spoken about that with a couple of people and they said, “I don't know.” So – business analytics, as well.

5:42

Alexey

**So it’s more like head of data rather than data science.**

5:46

Valerii

To some extent, yes, because it's everything related to data – from infrastructure to applications. From analytics to visualization. Before that, I was working in – well, I joined Facebook and left Meta. I will just rotate my screen a bit – you see those two buildings? This is the new Facebook office on King’s Cross. That's partly the reason why I moved to King’s Cross. However, I had no opportunity to attend this office. But still, I like the area. I started working in WhatsApp to create and found the team called “user data privacy,” which is an important team for Facebook. Because only for user data privacy issues, Facebook has been fined for like $5 billion. You can imagine that Facebook does not want that to happen again.

It was a very interesting change because when I was in Russia, I was working for Alibaba, a retail company, X5 Retail group, a retail company, Yandex.Market, as you can imagine, also a retail company. And then I switched, to some extent, to security or integrity – it was very interesting. So yes, I spent some time at Facebook and then at Meta. After that, I was thinking what to do next and I received this offer from the people at Blockchain. I thought the company was doing great – the mission made sense. We can speak about the mission later, but I don't think it's for this webinar. What do you call it? Is it a webinar?

7:30

Alexey

**Live interview.**

7:31

Valerii

Live interview? Okay. I don't think it's about Blockchain’s mission. That's it. What else? I was leading quite a big team in my time – the biggest team I was leading was almost 150 people: machine learning engineers, data analysts, etc. I was conducting many interviews. I don't know how many. Definitely hundreds, maybe even more. Maybe already in the thousands. I don't know. It depends. Because, for example, right now, I have on average, 30-40 interviews per week.

8:05

Alexey

**It takes an entire week, right?**

8:08

Valerii

Well, it takes a lot of time. Unfortunately, it's not the only thing I'm doing. But having an interview, even if you are the one who asks the questions, is very energy consuming, but they're rewarding and very interesting. So my main area is machine learning. I also know a bit about data analytics, A/B testing and I had to teach myself some data engineering and MLOps, but this is not my strong side. So that's it. I also had the privilege and opportunity to design and implement systems on a large scale. When I say “large scale,” it might be billions of users per day, and hundreds of billions of events per day.

8:57

Alexey

**There are only a few companies that can give you that, right?**

9:00

Valerii

It’s not hard to understand what company that was.

9:06

Alexey

**X5?**

9:07

Valerii

[laughs] Of course, which else?

# Who goes through an ML system design interview

9:12

Alexey

**Okay. Let's talk about machine learning system design. This is a part of the interview process and you said you did a lot of interviews as the interviewer. I imagine also, when you were joining Facebook before that, you also had to take this interview. So can you tell us about that? What is machine learning system design, and why is it an important step in the interview process?**

9:36

Valerii

Okay. Before doing that, let's try to review who needs to go through a machine learning interview. First of all, if you're applying to Facebook, Amazon, or Google, I think other big tech companies as well, because these three are the largest ones in terms of number of people working there and market cap. So if you're applying for a data scientist position, what would you do? You'd write SQL code, work with metrics, and dashboards.

If you expect that data scientists have some relations to machine learning in these companies, you are mistaken. People who do machine learning are called machine learning engineers. Right? [laughs] And these people have to pass through the software engineer loop at Facebook, and some additional rounds of interviews. For machine learning, and again, for a software engineer, there are different stages, but there are, I would say, a couple of interviews that are very important in terms of assessing your level.

These interviews are, of course, behavioral, project impact, (that makes sense, right?) and two very important things are the system design interview – which is how to design the system overall – and machine learning system design. These interviews are usually conducted for people starting from level five. Of course, at the very beginning nobody knows what level you are – it might be between four and five, so you might end up being level four, which is still common for this interview.

11:20

Alexey

**Level five is like a Senior, right?**

11:23

Valerii

Yeah, true. Good catch. Yes, level five is a Senior in terms of the level on Facebook, which means that, if you're on this level, it is an honorary thing to be on this level forever. So if you ended on level four, it was probably because of the ML system design interview. This interview tells the interviewer (Facebook or Google, or whatever company) your ability to have an overview of the system. In 45 minutes, you have to be able to tell a story – almost a monolog of yours – about how you will build the system and touch very different points.

I have seen some questions that you prepared, we’ll discuss how deep you should go. But it's tricky thing because you have to do that. You’re solo in front of a person who is silent and you're under pressure. It might be that you've never done that before. Not that many people in the real world have had the privilege and opportunity to build a system from scratch. Even if you've done that, who can promise that the system, which they will ask you to build is the system you really have experience with?

12:50

Alexey

**To summarize – basically, machine learning system design is one of the steps that machine learning engineers have to go through when they interview at Facebook, (probably now I should call it Meta) Google, and similar companies. Machine learning engineers go through this interview and this is a way to assess how well they can design machine learning systems – these are the systems that have to do something with machine learning. Right?**

13:18

Valerii

That's true. But also not just that. The thing is, it's one of the most important interviews. Let's say that you can fail the cold interview – to some extent, since you can fail on different scales – and still, they can push you further. So, it's a critical step.

# System design VS ML System design

13:36

Alexey

**I think this is what happened to me, but this is something that I prepared for later. So, you said that important interviews for detecting, or assessing your level are: behavioral interview, system design interview, and machine learning system design interview. Can you tell us – what is the difference between system design and machine learning system design?**

13:58

Valerii

Okay, let's try to determine the disparity between those two. First of all, when you're asked to do a system design interview, you're usually asked about data structures, about different server-side components, like “What are the databases? What is the amount of data that will be processed? What is the write throughput? What is a read throughput? How would you work with a cache? How would you work with load balancing, sharding, splitting?” etc, etc. So it's basically software engineering.

Meanwhile, on the machine learning design interview, usually, the thing is to understand how you would build it from the machine learning perspective. Let's give an example. Let's say that one of the questions is “How would you build a model that has to catch fraud on the platform?” Let's imagine the best way. If I had a crystal ball that tells me with 100% accuracy if a transaction is fraudulent or not, then the problem is solved, right? I just take the ball, I run the transaction through the ball – the ball tells me one or zero. So that's done. However, we understand that will never happen. There will always be some discrepancy.

Now we can say that we know that we have to put not zero or one, but some score between zero and one, when we have a transaction. When we have a transaction now, that probably means we'd like to have the system in real time. Okay, let's put it in our mind: real time system, a score between zero and one. Okay, it's a fraud. Let's say that we're speaking about money – does it mean that 10 bucks is of the same importance as 100,000 bucks? Probably not. This means that we need to have a probability of this transaction being fraudulent, and not just a score between zero and one.

As soon as we have a probability, we can calculate the expected fraud, which already leads us to the first metric to assess the quality of the model, which is “expected calibration error,” or “weighted expert calibration error.” Okay, we've got that. We also know that the ideal solution would be a binary classification task – one and zero – the crystal ball, right? We know that this will never happen, however, we know that it's a binary expression and that the output has to be between zero and one and it has to be a probability. So that also tells us “What should be our loss function?” The loss function should be from a family of a proper scoring function.

16:43

Valerii

Fortunately, the very basic log loss is good here. So we know that we might start from log loss. We also know that we might start from a very basic linear regression model. Why is that? Because we know that it has to be very fast – in real time, right? We also know that fraud comes from people – people are very creative creatures, very creative, and they are notorious for being very adaptive. Thus, we know that suddenly the pattern might change.

With the linear regression, we can retrain the model in an online fashion and adapt for these changes as well. However, it depends on how fast we will receive our labels. And so you see, we're coming to a completely different question. “How can we gather the labels?” Okay? “What is fraud? And what is not?” Are these labels 100% sure? Or is there some noise there? Because, well – there might be some noise. How would we find it? Let's have the first assumption – there is no noise. We’ll come later to that.

Now, how do we just gather our labels? How much time will pass until the transaction will be labeled? Is it immediately? Probably not. A day, two days, three days, thirty days? During that, do we need to update our model in real time? So we're coming back, you see? Okay, but let's just say we'll make a very simple design. By the definition of linear regression of a log loss, we know that one of the metrics would be expected calibration error, and would maybe be just weighted expected calibration.

What else? Should we take a look into other metrics? Probably, yes. But we know that the fraud is very class- balance skewed. We know that class imbalance is extremely high there. We also know that it might change. So that means that if we would like to take a look into the metrics, these metrics have to be class-balance insensitive, probably. Because otherwise, yes, class balance changed, metrics change, but the model’s the same. Okay, so what are the most favorite metrics? Is it precision and recall? Recall is class-balance insensitive, while precision is class-balance sensitive. So, forget about precision. Can we replace precision with something? Why not specificity? Also not that. Okay, something else? Maybe. We know that there are some thresholds of expected fraud level, which we can just go with and then we can.

Do we need to introduce some weights? Okay, good. What data will we use? Is it the amount of the transaction? Is it just the history of the user? How fast will we update them? Now let's say we have a model. How can we assume that model is better than the previous one? Of course, we have some offline metrics. We have an expected calibration error, weighted expected calibration error, precision – we don't have precision, forget about that. It's a bad metric because it's class-balance sensitive. We have specificity. We have recall. What now?

We can run an A/B test to see the online performance, right? How will we see that? How long do we need to run A/B tests, etc? So all these things have to be considered. Okay. Now, let's say I told you about the basic features. What about feature engineering? Like I said, linear regression doesn't take nonlinearity into account. Can I do that with basic feature engineering? Probably, if you have enough data, just having a polynomial of the second degree, which helps you find an overlap between features – how they interact with each other – is enough. Because if you have trillions of data points, you can do that, because sparsity is not an issue here. And so on, and so on, and so on and so on.

20:33

Alexey

**That's quite a lot of information. I was trying to process this. That's quite a lot of things. So this was an example of machine learning system design. The interview starts and then the person – the interviewer – asks you, “Let's design a system for detecting fraud.” And then you probably ask this person a few questions and then you do this information dump on that person, right?**

21:10

Valerii

The best way is not even to ask, but to say “My assumption is *that*. Do you agree with that or not?” You see, you asked the question, but actually, you’ve made an assumption. You say “Are you okay with that?” Because you've been given some information. Of course, in the real world, we would gather the context because context can make everything very different. Because imagine, like in the case of fraud – if you receive a label within minutes, it's very different to receiving a label within months. It affects everything. But you could make an assumption, you say, “My assumption is that.” To build, you might be making many assumptions and nobody prevents you from making assumptions, which will make your life easier.

22:05

Alexey

**Yeah, indeed. So, the original question I actually asked you is about the difference between system design and machine learning system design and I think it's very clear what machine learning system design is. It requires some domain knowledge, to some extent, or making some assumptions. Then you need to walk through the process of solving a particular problem.**

**I have an example from my personal experience of being interviewed at one of these companies on system design. I had the question to design a system for finding places of interest. So let's say I go to London – I go to whatever central square you have in London, and the system would need to give me all the points of interest, all the closest interesting places.**

23:01

Valerii

It was system design, right?

23:03

Alexey

**It was a system design, yes.**

23:04

Valerii

I had almost the same question in my interview for Facebook.

23:07

Alexey

**So that was the system design part. There, I needed to think how exactly I would store these things, how I would retrieve them quickly? How I do, sharding, load balancing – all that. And then on machine learning system design, it was a very related question. The question I got there was, “Okay, now we have this system that returns the closest points of interest. Now, let's have a recommender system there. Let the system return 15 of the closest, most interesting places that are interesting to this specific user.”**

**I think this is a nice example to show the difference between the two. In one you need to design a system – you don't think about machine learning at all. Then on the second, you don't need to think about the scalability or load balancing, sharding – all that – you have a specific machine learning problem that you need to solve and then you go through the solution. Right?**

24:04

Valerii

Exactly. Yes, like that. You could also make the same example of the fraud system. In this case, the system design question would be “Can you build a system which will handle 3 billion transactions per day and these transactions are coming from this?” So, you see?

24:21

Alexey

**Yeah, and then on the ML system design, you would talk through the log loss and things like this.**

24:27

Valerii

Right.

24:28

Alexey

**But where does system design actually come into the picture here? Because here, we talked about selecting the right metric, which was the important thing, as you said. You said it was log loss for this specific case. Or even before log loss, I think it was expected calibration error.**

24:52

Valerii

I said that I need a loss, which comes from the family of the proper scoring functions.

24:57

Alexey

**Yeah. So you need to say all these things and then once you say, “Okay, this is the thing we are measuring. This is the baseline model, (like linear regression or logistic regression).” And then you start building on top of that, right?**

25:13

Valerii

Yeah. Then for example – I remember that I was doing that for Facebook, and suddenly the guy asked me, “Okay, you said that a metric would be AUC. What is AUC? Why did you say that it's a ranking metric?” I said, “Well, that's because it does that and that.” And he said, “Okay. You know what you're talking about.”

25:30

Alexey

**Yeah. But where do we actually design systems? Or this is what you mean by that? Do we need to say “This system is doing this and then there is another system?” Or it’s about designing the…**

25:42

Valerii

I don't get the question. But by itself, it's a system. Every machine learning model, it's not like a model – it's a whole system, because you have features coming to the model, the model outputs something, these outputs also have to be taken into account. There might be A/B testing here, and I did feature preparation here. So it's a *whole* system.

There are companies that create just parts – the components – for the systems. Like, take the feature store Feast, – it's closer to the system design. So it might be that you can call that software engineering system design and machine learning system design. Because with both, you have to design a system, it’s just that you’re designing systems with different goals.

Alexey’s interview case study

26:25

Alexey

**Okay, yeah. I was already talking about my experience with interviews. There, I was interviewed for a tech lead position and this question was about designing a recommender system for points of interest. The way I approached it – first, I proposed a metric. I don't remember what the metric was. I think, let's say you have a recommender system – looking at what the user clicks and actually goes there. That could be a nice metric to measure.**

**Then I suggested some heuristics. I don't remember, maybe suggesting clustering people by interests and then selecting the most popular points of interest for each cluster, specifically, and then recommending this to the user. Then I suggested some other heuristics on top of that. At the end, I had a bit of time to talk about actual machine learning. At the time, I thought I really nailed it.**

**I thought I did really well in this interview – the interviewer was nodding all the time, like, “Okay, yeah. Keep going.” So I really didn't think that something could be wrong there. I was really afraid of the coding parts. I was also not super sure about the system design part. Then a few weeks after that, I got feedback where the recruiter told me that I did well in the coding parts. I also did well in system design, but I completely failed the machine learning system design part.**

28:03

Valerii

Completely failed?

28:04

Alexey

**Well, not completely. But they didn't like me, I guess, for a tech lead position.**

28:10

Valerii

British HR would never write you that. They would say something like “Alex, it was wonderful. It was brilliant. There was just that slight miscommunication.” Something like that. They'll never tell you that you completely failed. Never.

28:28

Alexey

**[laughs] I might be wrong with using these words. I think the recruiter probably used different words. But the reason for me failing the process – the whole interview – was machine learning system design. Not the others. I was afraid about the others. But in the others, I did well, but I failed that one. And the reason there was because the interviewer expected me to talk about actual machine learning. Instead, we talked about metrics, heuristics, and then I didn't have enough time to actually cover machine learning. Yeah, so what do you think about this? Is this typical for the process? Is it expected?**

29:09

Valerii

Let's be honest, the interviewer was a human, and humans are subjective. Maybe they had a bad day. However, to some extent, I'm surprised because it's hard to say the interview was nodding. Maybe, again, the way you remember it and the way it was – it's a natural thing for human beings to remember some things. There is even a saying “Lies like a witness.” So that's hard to say. However, usually, you could tell – you could try to secure yourself in an interview by asking “Do you want me to focus on that? Alright, let’s go.”

Also, another good way would be just to sketch like what we've done right now. In five minutes, we almost finished a very, very, very basic design of a fraud detection system. Because we already spoke about loss function, the model, the feature interaction, the metrics. We even mentioned A/B tests. So now we could go, “Okay, we outlined it. Do you want me to focus on something else? I'll go step by step, diving deeper and deeper.” I would make a second iteration, a third iteration. Because usually, how I do that, I tell the interviewer, “I will build a baseline, and then once we have a baseline [we’ll move on]” because usually, what you do in real machine learning, is you either take a heuristic as a baseline, or you take a very simple model.

You're not trying to build a spaceship from the very beginning. But again, it's hard to say. Maybe there were some signals – very, very gentle signals that you were unable to read. Maybe it was just a bad day for the interviewer. You see, it's hard. To some extent, an interview has at least a part of luck in it. So to fight that, you can try to secure yourself.

31:09

Alexey

**Yeah. My question was more about what you think, not about this particular interviewer, but about the way I approached it. So I approached it by coming up with a metric, then a heuristic. I think what I probably should have done instead is, perhaps I spent too much time on that. Right? Of course, the interviewer could have stopped me by saying, “Okay, let's actually talk about the machine learning part.” But at the time, he didn't do that. But maybe this was my fault because I should have asked, as you said. But I’m wondering, how much time exactly should I have spent on talking about heuristics? And how soon should I jump into machine learning and then maybe deep learning, and talking about more advanced things?**

31:58

Valerii

Well, it's an interesting question for which there is no single answer. It depends. My opinion is that the interview has to be as close to the real job – the real work – as it can be. So, to be honest, in applied machine learning, you don't usually dive very deep. You need to understand why and what. If you're applying for a machine learning research position, that's a different topic. But whatever. Usually, you set up monitoring, you pick the loss, the model, the metrics, and then you dive deeper.

You have to be able to just, let's say, provide some arguments. “Why did you pick *this* model? Why did you pick *this* loss function? Why do you pick *these* metrics?” However, I don't think that it makes sense to go into something deeper. What does it mean? Just talking about how gradients flow through their convolutional layer in the neural networks? What for? But that you see, it's my attitude.

33:05

Alexey

**[laughs] Yeah. Or maybe how to do back propagation for batch norm, right? Let’s derive that. [laughs]**

33:11

Valerii

Yeah. I mean, I've been asked that. By the way, I had this question in an interview once.

33:17

Alexey

**So did you remember how to do this?**

33:19

Valerii

Well, I was able, to some extent. I managed this. Because look, I mean, come on. Batch norms – there is some normalization. So? Okay.

# Preparing for ML system design interviews

33:31

Alexey

**Okay. [laughs] So, how do I actually prepare for machine learning system design interviews? It feels as though just being a practitioner is not enough. Because, first, you never know what exactly is expected. I guess you need to ask that. Also, you might get a question that is outside of your domain expertise. Let's say I work in ecommerce and I get a question about recommender systems. Maybe I'm not working with recommender systems right now. So how can we prepare for such interviews?**

34:05

Valerii

There are many ways you can prepare. There are many services on the web, in which people from Facebook who actually conduct these kinds of interviews, can do that for you for a small fee of 200 bucks. And then they will give you a review. However, I haven't seen any credible courses on machine learning design. Well, you could also try to ask for feedback. That's difficult. Actually, I have an idea to make a course on machine learning design. But we decided to start from just system design, because system design covers more people. Obviously, it's easier to sell because the audience is bigger.

34:53

Alexey

**Because it’s also not just machine learning engineers, but software engineers.**

34:56

Valerii

Yeah, everybody – from a software engineer to a machine learning engineer. All these people go through system design. So that's why the audience, by definition, is larger.

35:09

Alexey

**So one way, of course, you do this at work. Another way is to find people who can help you with that. Is there anything else you can do? I don't know, maybe watching some conference talk maybe?**

35:21

Valerii

Well, maybe. On the web, there are some analysis and design overviews on YouTube. I've done my fair share. However, they’re in Russian, so only people who speak Russian or understand Russian can do that. But there is information. Look, the process to get hired at Facebook is standardized. Also, you can have extensive experience.

To be honest, I have made no preparation for ML system design. I was showing that part, because that's the only thing I can do, probably [laughs] design a system on paper. So, extensive experience. There are talks about that. I don't know to be honest. It's hard for me to answer because I made no preparation myself for that.

36:21

Alexey

**Okay. If we take an ecommerce company – a small one – then we can think about what kind of questions they may ask candidates. It could be about designing a search system, designing a recommender system –the typical things that they do. However, when it comes to Facebook, Facebook does so many different things, so you can never know exactly what kind of domain you might get. They might ask you to design a newsfeed, for example. Or they might ask you to design a point of interest recommender system, or a fraud detection system for WhatsApp, right? It could be anything.**

37:02

Valerii

They will. Actually, that's my favorite part. You've seen the ML design interview I conducted, right? So you notice that my favorite thing is – the person comes, I know this person’s background, and I ask that person a question that is *completely* outside of the area of this person. And that's fun. That's hilarious. [laughs]

37:26

Alexey

**[laughs] That's what you did with me, right?**

37:28

Valerii

Of course. I mean, I've been preparing – I fine-tune it for everybody. But that makes sense. However, there are still some patterns. There are still some stages, which are common for everything. You still need to gather data, you still need to understand what should be the metric, the loss function, what's the model? Why this model? What is online versus offline? Should it be adjusted on the fly? Et cetera. To be honest, there are not that many steps. Then come back, come back, come back.

# Machine learning project checklist

37:59

Alexey

**Speaking of this mock interview – a while ago, I had a mock interview with Valerii, where Valerii interviewed me. The question was about designing a fraud detection system.**

38:13

Valerii

Who could’ve imagined that. [laughs]

38:14

Alexey

**[laughs] Yeah. At this interview, you showed a machine learning project checklist. Can you talk a bit about that document? What’s in there and why is it helpful for designing ML systems?**

38:29

Valerii

Back in the days of Facebook, a number of practitioners decided that there were many, many machine learning services. “Probably, we need to write some comprehensive list of checks that we need to pass the service through.” It's actually a very good preparation guide for system design, because it covers exactly these points. Well, it's very comprehensive, like a 16-page document. However, you could also go and find the book from O'Reilly, written by people from Google, about a nail design practice, or something like that.

39:09

Alexey

**Using machine learning to design patterns?**

39:13

Valerii

Yeah, something like that. So you see, to some extent, you might have these checklists – you might just extend it to the whole book – but it means the same. Again, model coupling/decoupling, A/B tests, features, losses, model times, online/offline, batch processing, whatever. If you know the basic points, then you go from A to B, from B to C, from C to D. It’s the same for system design.

To some extent, it’s like cases for a consulting company. They train you to solve any case, even if you've never been working in their aircraft manufacturing company. But somehow, now you're an expert and you can suggest to the CEO of this company how to run his or her business.

# The importance of defining a goal and ways of measuring it

40:11

Alexey

**So about this checklist – let's say we need to design a system, not necessarily for an interview, but just design a system. What is the first thing we need to do? Do you remember what is in this checklist?**

40:21

Valerii

I don't remember what the first thing that’s there, but I think that the first thing is “What would you really like to do? What is your goal?” And “Is it really achievable? Why are you doing that?” Because “What is your end goal in this fraud system? What is your end goal in recommending some interesting places to people? Is the goal that they will find it as quickly as possible? Is the goal that they will rummage through your app? Is the goal that they will have to spend more time on the platform? Which, mind you, is the goal for many companies – their main metric is how many minutes, or how much time, does the person spend on the platform? Now understanding the goal, you have to think,” Okay, can I directly run for this goal? Or I can't, for many reasons, and I have to approximate it? So I have to use a proxy goal to do that.”

41:20

Alexey

**Like for measuring if you're moving towards this goal or not, right?**

41:25

Valerii

Yeah. For example, let's say you need to create a system, like ads on Facebook. Why do you need to do that? You would probably like to increase your total income – your revenue – right? Okay. However, what can you do? Can you train your system on the clicks? Is that good enough? Well, probably not, because the person who just bought an ad expects that the person who clicked will buy. The click by itself leads to clickbait. So, now “Okay, can I train the system on buys?” Well, to some extent, that’s more difficult because clicks are rare events. However, to purchase something is even less frequent.

So then you try, “Okay, maybe I can try to take a combined loss. However, I will never be able to really assess it in offline. The only thing I can do is to just assess it in real time, like in an A/B test. But if I do an A/B test, now I have an old system 95% And a new system of 5%. Is it still that they're not affected?” Or “If I will run this on the whole traffic, the money will somehow just move from one pocket to another. Are they really independent?” Sometimes it happens, like market budget allocation problems. So there are many things which might shoot you in the leg.

43:02

Alexey

**Okay. So we need to define the goal. It could be people spending more time on the platform, or earning more money. Then, we need to find a way to measure if we're moving towards achieving the goal – define a metric.**

43:17

Valerii

Yes. To approximate, “Can you move directly to your goal? Or can you approximate moving to your goal?” Also, the thing is that – if a metric becomes your goal, with some time, it usually ceases to be a good metric.

43:36

Alexey

**I imagine in the case of “more money,” you can just fill your entire feed with ads, right?**

43:41

Valerii

Yeah. For some time, it will work. But again, you need to see the long run.

43:46

Alexey

**So you need to also have some other metrics, right? Not just the main one, but also something like “Are people still spending time on our feed or not?”**

43:53

Valerii

Right, right. Like spending time, the attrition rate, the churn rate, retention, what else? There are many, many things.

# What to do after you set a goal

44:01

Alexey

**Okay. So we do this, and then you also mentioned A/B tests. We define a metric, and then we say how exactly we are going to measure this metric. What do we do next?**

44:11

Valerii

Let's say we know what we would like to do. We know how we can try to optimize it in this way. What does that mean? That means that if my model improves, there is a high chance that my metric of interest will be better. Now, I need to think about the labels, but that's obvious, right? There’s a proxy metric, you can say it's a label. I will construct my labels. We know that you can say that labels are y’s, now we need to think about access.

What are the features? Okay, what features do we have? We have this, this, and that feature? They might make sense, right? We have x and y, now we need a model. What kind of model? We have a target, we have labels. What about the loss function? Can we just put in the loss function directly or not? Now let’s come back to the features – we have basic features – do we think they interact with each other? Do we need to do some pre-processing? Okay, think about that. Now let's say we can put the model, we have x, we have y, we can train it, right? So what happens here? Let's do that.

Now, we've done that and we receive some output. Okay, how do we know if this output is good? Let's think about validation. Because we didn't speak about that on the fraud system. But actually, we spoke about offline metrics. For offline metrics, you probably need to have a data set which you have evaluated. Then A/B test. How would you run the A/B test? How long? How many samples do you need? What metrics of interest? Etc, etc, etc.

46:02

Alexey

**Perhaps if you cover all these parts during your system design interview, you're already in quite a good position. Right?**

46:09

Valerii

Yeah. But to be honest, if we speak about the real system, there are more. Because let's say you have an A/B test, output, tra-la-la – everything is good. But in a real system, many things might appear. Distribution shift for features might appear, and we need to be able to detect that. Target or class imbalance might appear. The model might become broken. Do we have a fallback?

We need to monitor the model performance. What will we do if the performance is much lower? Do we have a fallback? You see, like a system – because there are many more checks for the real system. Let's say a perfect model for our ads ranking. And the model is somehow broken or turns to be crazy.

46:57

Alexey

**By “crazy” do you mean it outputs random stuff?**

47:01

Valerii

Yeah, yeah. Or because there is feature shift distribution. So we need to detect this – the feature shift distribution, target distribution, model performance. And we need to have a plan B to switch that. But I need to take a look into these documents before I can tell you. [laughs] Smart people were doing that for quite some time. It's not like I can pull it from my head immediately. But there are many things which might shoot you in the leg.

47:27

Alexey

**Yeah, maybe before you do this, I realize we don't have a lot of time left, and there are quite a few questions. But before we go to these questions – we talked about this distribution shift, class imbalance, the model breaks, fallbacks. We should also mention that during the interview, right? It also shows our experience and exposure to these things breaking in production, etc.**

47:48

Valerii

Of course. You see if you'll do that, you'll be ahead of 95-99% of the other candidates.

# Typical components of an ML system

47:52

Alexey

**Okay. So let's go to the questions. We have quite a few of them. The first question we have is, “What are the typical components of a machine learning system? And what percentage of it are machine learning algorithms?”**

48:07

Valerii

I think algorithms are just one of the smallest parts, 1-5%. Well, I was speaking with a candidate recently and I told him “Look, imagine that you're a machine learning engineer in the company for two years,” He said, “Okay, okay. I can imagine that.” “Imagine that you spend an immense amount of time creating an algorithm – finding the best algorithm, setting up the loss function, the metrics, all the rest. It took you a humongous amount of time – two weeks. And you’re in the company for two years. What do you do?” Right?

You probably say as an answer “So if you have the right output and right input, then the model is not that important, if the model can handle that.” Of course, you probably wouldn't use linear regression for images. Look, you might argue “Should it be resonant? Should it be a visual transformer? Should it be…” Whatever, I don't care. But if your features are very good and your labels make sense, then it's a second order of improvement. But if you have a best model, and your features are mediocre, or bad, and your labels are wrong – you're screwed.

49:35

Alexey

**The typical components of a machine learning system – this is the first part of the question – are things like data pipelines, data preparation, things to calculate features?**

49:45

Valerii

Features and labels, of course. That's the most important. Features. So I think features are very important.

49:54

Alexey

**And then the things that monitor these…**

49:57

Valerii

Let's do a mental exercise. Let's imagine that you have a computer vision, deep learning model. Very sophisticated – 175 layers. And then there is a classification model. And on top of this model, you have what? You have a linear classificator. What does it mean? It means that, actually, this model classifies with their linear model. And all that is done before is just representational learning, transforming the original features to the features, which might be fed to the linear model very successfully. See – features. Just with this mental exercise, you can see that. So that's why you can take embeddings, put them in whatever model you would like to, and you have a proper output.

# Applying ML systems to real-world problems

50:57

Alexey

**Thank you. Let's go to the next one, “How to make machine learning algorithms work with other parts of systems to solve real world problems?” I guess the question is more about, “Okay, we have this model that we just discussed. This model for classifying images – how do we integrate it with the rest of the system and what do we need to do for that?”**

51:20

Valerii

The model is nothing by itself. That's why you have a machine learning engineer. That's why. I don't like the job title “data scientist”. Because what is a data scientist? The person who does something in Jupyter Notebook? Who needs that? Yeah, people need a model integrated in the system. That's why they need machine learning engineers. That’s why in Facebook, there are machine learning engineers – you engineer.

You're accounting for the software engineer plus machine learning. So yes, the company needs machine learning engineers. Then again, what was the first task for us? “Understand what we want to achieve.” As soon as you understand what you would like to achieve, it's much easier to achieve that. Without understanding, of course, randomly, you might achieve a desired goal, but the chances are not high.

52:14

Alexey

**So the most important thing, when we start with building a machine learning system, is to think about the goal. This is something that was first in your checklist. Then the rest will come.**

52:25

Valerii

“Do we really need machine learning here exactly?” Maybe we can be lucky and we can just avoid it.

52:32

Alexey

**I think there is an article, or more like a mini-book, from Google, which is called The Rules of Machine Learning and I think there the first rule is, “You don't need machine learning.” Or something like that.**

52:45

Valerii

I don’t know. I have not read this book. You see, I passed the ML design interview, so that's why I can just now lay on my back and do nothing. [laughs]

52:56

Alexey

**[laughs] That’s cool. The question is about the book you mentioned – the book was Machine Learning Design Patterns, right?**

53:02

Valerii

Something like that from Google. Yeah. I mean, it's a good book. Unfortunately, it didn't reveal anything to me. But it's still okay. It's a good book. It makes sense.

53:15

Alexey

**I guess for practitioners who work with machine learning, they will think, “Okay, I knew all that.” But what the authors did was categorize all this.**

53:24

Valerii

Yeah, it's a good taxonomy. It's a good taxonomy. It's a good book. If it didn't reveal anything new to me, it doesn't mean it's a bad book. It just means that it's my problem.

53:37

Alexey

**[laughs] But I think for many people, it will be useful because for each pattern there, they talk about when exactly you need to apply this and how to apply this. They also talk about what kind of tools there are. And since this is a book from Google, there is a lot of focus on Google Cloud, but they also talk about open source solutions like Kubeflow, for example.**

53:59

Valerii

Sure. Well, of course. Google Cloud is not the worst cloud, definitely. We use Google Cloud at Blockchain, for example.

# System design and coding in interviews for new graduates

54:07

Alexey

**Yeah, so another question from Alvaro. Alvaro is graduating soon and he is a machine learning intern at a startup. He's starting a job hunt, hopefully [inaudible]. So how much system design should he expect as a new grad?**

54:23

Valerii

I think no system design at all, probably. I mean, look, who would expect from a fresh grad to design a highly complicated, distributed system of high/low with sailed out machine learning? I mean, it's ridiculous. As far as I know. But again, I didn't apply as a fresh grad for Facebook, but as far as I understand, there would be no system design at all.

54:45

Alexey

**Do they ask for coding, like LeetCode-style coding?**

54:50

Valerii

LeetCode-style coding, behavioral. Probably, that's it – like two or three coding and one or two behavioral. That's not much to ask from [an ML Engineer]. Maybe for machine learning, they might ask about algorithms, how do they work inside? It makes sense, right?

55:07

Alexey

**Then at what level would they ask this – I think you were saying level four, which is the Middle Level and level five, which is a Senior.**

55:13

Valerii

Level five. But there is no clear way, nobody will tell you “You’re level five. You'll be trained for level five.” Of course, there’s always some margin. So you might end up being level four, but still go through this interview, because you were on the brink between four and five.

55:30

Alexey

**Basically, when you interview they do this automatically and probably at this round, they use it to assess which level to put you.**

55:39

Valerii

Yes, this is one of the most important stages – to estimate the level. I mean, you can’t estimate exactly – Okay, you solved the LeetCode Medium. It doesn't mean you’re level four or level eight. Come on. LeetCode is just to show that to some extent, you can write code, which to be honest, in my opinion, these LeetCode-style interviews are not very much related with the real ability to write code.

56:08

Alexey

**They show how we can solve puzzles. [laughs]**

56:10

Valerii

Also just how you can train yourself. To my surprise, I've seen people who just told me, “Look, look! I've done these 400 LeetCode exercises, but I failed an interview because they asked me a new task I've never seen before. So now I'm doing 500 more.” And I’m thinking, “Wow, come on. There are just six or seven patterns, even fewer.” What is there? Dynamic programming, backtracking, divide and conquer, and there are a couple others you have to know and data structures. And that's it. But still, that means that you can just train yourself in this LeetCode style, and you can be very weak in writing real code. And vice versa also might happen.

57:01

Alexey

**So if you're a fresh graduate and you're interviewing for a Junior position, you will not have this. But if you apply for a regular, let's say, machine learning engineer role (doesn't even have to be Senior) you will have this and then they will decide what kind of level to put you in.**

57:20

Valerii

I believe so.

# Humans in the validation of model performance

57:23

Alexey

**Okay. I don't think we have a lot of time for more questions. There is an interesting question from Vijay, which is about, “What is the best way to validate the model performance in production? Do we need humans for that or are there other ways?”**

57:37

Valerii

I mean, the best way is having A/B testing. However, if you need humans to have labels, then yes. You then label it. If you don't need a human to label the output, then you don't need a human. So it’s A/B testing that says causal inference, right?

57:57

Alexey

**So let's say in this example that we talked about points of interest. There we can validate, based on the feedback, how exactly people use our system.**

58:07

Valerii

Yeah. We run the A/B test there, and what is the metric of interest? Again – you see, this question pops up every time. “What is the metric of interest? What are we actually trying to achieve?”

58:21

Alexey

**Yeah, and in some cases, I guess like in these fraud systems, it's trickier. Sometimes you need people, like fraud specialists, to look at the transactions and say “That’s fraud.”**

58:31

Valerii

Yeah – how fast you can receive labels?

58:35

Alexey

**Yeah, exactly. Okay. Maybe one last question. It seems like you have a very solid data science profile, from Grandmaster at Kaggle. That's pretty solid.**

58:47

Valerii

Did you use the data scientist profile, because I told you that I don't like “data scientist” in my job title? I find it awful and terrible. So you’re just nudging me in my pain point.

59:01

Alexey

**Yeah, so the question is, “With this profile, you're very good at doing data science stuff. How did you transition from data science to being good at system design?”**

59:11

Valerii

I mean, there was an issue, to be honest. Because I was in the right place and at the right time in order to have the opportunity to do that. But again, it's system design. It's very simple. You have these pieces – not that many pieces, to be honest – and you just [put them together] and that's it. I don't have a good answer.

59:37

Alexey

**Yeah, I guess the answer might be just being a practitioner? Because models don't live in isolation, right?**

59:43

Valerii

Yeah. In fact, if you know how to do that, and you've been hired – you feel very good. I felt very good at Facebook – very easy. Had great results on performance review, me and my team. So, pff – it was easy. I left at the right time, if you take a look at the stock right now.

# Finding Valerii online

60:03

Alexey

**[laughs] Okay, I think that's all we have time for. So maybe last one – How can people find you?**

60:12

Valerii

Well, you can find me on LinkedIn. Just type in my name, you use a y instead of ii. With the new rules, it should be ii at the end.

60:22

Alexey

**I copied it from Slack.**

60:25

Valerii

Well, I think that people can still find me on LinkedIn and find some questions there.

60:31

Alexey

**Yeah, there are so many different ways of spelling “Valerii”**

60:34

Valerii

Oh, yeah. Not that many different ways, but there are some.

60:39

Alexey

**More than one.**

60:40

Valerii

Yeah, true. Some ways.

60:42

Alexey

**You can also use w, right? Maybe for German.**

60:45

Valerii

For Germany, right,

60:47

Alexey

[laughs]

**Okay. Thanks a lot. Thanks for joining us today. Thanks for sharing.**

60:51

Valerii

Thank you very much, Alex. And you have a great evening and great weekend. Take care and see you.