**Transcript**

September 4, 2025, 7:33AM

 **Speaker 2** 0:20  
And a little bit jarring every time that happens.

 **Speaker 1** 0:24  
Yeah.  
Let's see if Speaker 3 is able to get back.

 **Speaker 2** 0:37  
Yeah.

 **Speaker 1** 0:47  
Yeah, perfect. Yeah, you're back.

 **Speaker 3** 0:49  
OK.

 **Speaker 1** 0:52  
So and just as let's say starting question we wanted to ask if OK now for you it it it passes some some months so but if you recall more or less the the the purpose of the study.  
Yeah.

 **Speaker 3** 1:08  
Yeah, I remember the first part mostly. Just one disclosure that most of the answers that I gave on the survey were related to the research that I'm doing in the in the institute where I'm in the GSI in Germany, which is a nuclear physics institute, so.

 **Speaker 1** 1:11  
Yeah.

 **Speaker 3** 1:29  
So maybe it's not the best, let's say use case for for the for your study, but it still has the same problems as everybody else mostly.

 **Speaker 1** 1:37  
Yeah.  
Yeah. And so in the context of this research, you're still dealing with the machine learning systems, right? Yeah. So I think this is is perfect and yeah.

 **Speaker 3** 1:49  
Yes, yes.

 **Speaker 2** 1:53  
So this is in a machine learning in a science research context.

 **Speaker 3** 1:59  
Yes, well, so the basic pipeline is that we're performing experiments on accelerator. We have a detector that detects the data. The problem is that this detection occurs on the level of the picoseconds, sometimes nanoseconds. So we have a huge amount of data that have to have to be.  
Sometimes dealt with in real time, but mostly some of most of the job is done at the execution at the data position time, but we also have some batch processing afterwards that deal with a more say clean data.

 **Speaker 2** 2:22  
2.

 **Speaker 1** 2:37  
OK.  
Yeah. So essentially what we because before the meeting I take, I took a look at your at your answers and specifically to the answers where we asked you to.  
Grade the relevance and the frequency of different machine learning related smells, let's say. So I mean.

 **Speaker 3** 3:03  
Yeah, the data model and pipeline. Yeah, I remember those parts.

 **Speaker 1** 3:07  
Yeah, yeah, it's that. So uh, I can even share my screen probably. I'm not sure if it will help, but let's see because it's pretty small. But yeah, let's see. So anyhow, what I wanted to say is that

 **Speaker 3** 3:10  
OK.  
Yeah.

 **Speaker 1** 3:24  
I saw that basically your answers were generally, let's say positive with grades above as as a average above 3, especially for the relevance. So I I hope that you can read something anyhow. So these are.

 **Speaker 3** 3:39  
Huh.  
Yeah.

 **Speaker 1** 3:43  
Your your your your answer that are let's say isolated. So as you can see above we have the the relevance and you typically give like very high grades above 4 as I said so sometimes.

 **Speaker 3** 3:58  
Mhm.

 **Speaker 1** 4:00  
Yeah, you also go to three, but it's very rare that you go below 3, except for these smells at the end, which are related to the.  
To the organic, yeah, exactly to the organization. So and I'm not sure even if you remember a little bit the yeah, this specific smells. So the prima donna organizational schemes and.

 **Speaker 2** 4:15  
Organizational.

 **Speaker 3** 4:29  
Mhm.

 **Speaker 1** 4:32  
The black cloud what they meant, but like just if you could comment why, why you think that those in general organizational or community related smells are less relevant in your opinion or your in your experience.

 **Speaker 3** 4:43  
It.  
OK, was there a small description of the what what black black cloud smells were?

 **Speaker 1** 5:01  
I don't think that it was, yeah, it was provided in the, yeah, in the in the survey.

 **Speaker 3** 5:02  
Yes.  
And yeah, any answer? So as I said, it's mostly based from my experience and the community in which I work usually is well well organised on the level of when when it arrives at.

 **Speaker 1** 5:13  
Yeah.

 **Speaker 3** 5:21  
At the level to analyze the data. So I said we do a lot of preparatory work, so mostly of the smells that I disclosed are part of the oral data model or pipeline pipeline smells.

 **Speaker 1** 5:25  
Mhm.  
Yeah.

 **Speaker 3** 5:37  
Because when we arrive at the point, we usually have very well organised workflow that doesn't usually interfere with the with the entire outcome. So that's why I had to reduce it, but that's it's probably related to my to do.

 **Speaker 1** 5:48  
OK, yeah.  
Yeah, to your, uh, yeah, to to the environment. Yeah, it's, uh, I think it makes sense. Uh.

 **Speaker 3** 5:53  
To the environment which I work.  
It is very well established workflows that we don't steer out from.

 **Speaker 2** 6:07  
So, so in general you would say that you did not have.  
You did not note any organizational issues with, for instance, silo. This group doesn't talk well to that group or there's, you know, prima donna, you can imagine as a kind of vain or selfish behavior, these kinds of things.

 **Speaker 3** 6:19  
So.  
Strangely, no, the maybe the biggest problem we had was the cure when we have models that were supposed to be equivalent between themselves. But then it turns out that they're not. Somebody added some additional features that we didn't agree upon.  
And so that are probably the most common cases of some disagreements that might appear. But on the level of of the jar key or way we we communicate with each other, it very, very rarely happens that something some ladies occurs.

 **Speaker 2** 6:49  
Hmm.

 **Speaker 1** 6:51  
Mhm.

 **Speaker 2** 7:00  
OK.

 **Speaker 1** 7:01  
OK, that's good.  
Um.  
Right. So on the relevance, I think that's the most important part that we wanted to ask. And then while answering this, you also touch upon the fact that other categories of smells for you in your professional experience are more important and and and relevant.  
So I don't think we need to dig more on those, but rather we could look at the frequency and we can start from a again a general observation that your answers tend to be lower in general your score for the frequency.  
So it appears as even if you, and correct me if I'm wrong, even if you think that a given smell is relevant, it seems you encounter that smells generally less than not so often, let's say, right?

 **Speaker 3** 8:03  
Yeah, as I said, mostly we deal with the the biggest amount of the problems we have is the reproducibility problem. Generally when when we have to make our data fair, reproducibility becomes one of the biggest problems because of the sheer quantity and the sheer volume of the data.

 **Speaker 1** 8:03  
Yeah.

 **Speaker 3** 8:23  
Because we talk about petabytes of the data and that's why sometimes the hard hard coded parts are the usual problem and we know how to deal with it. So that's why I said it occurs once or twice, but maybe three times in during the project and.

 **Speaker 1** 8:24  
Mm.

 **Speaker 3** 8:39  
Once you solve it, it's usually around everyone behaves based on what we agreed upon. So when it comes, and it usually comes, even if we try to repair, we try to to organise everything. Some for some reason in the consortia where you have at least five or six institutes, it usually happens that.

 **Speaker 1** 8:47  
Yeah, OK.  
Yeah, yeah.

 **Speaker 3** 8:57  
Somebody gets outside the track.

 **Speaker 1** 9:00  
Yeah.

 **Speaker 2** 9:02  
So, so just to give an example, we had a number of smells having to do with data, you know, understandability, consistency, distribution, so forth. Do you have any way of tracking or monitoring?

 **Speaker 1** 9:03  
Uh.

 **Speaker 2** 9:22  
The the occurrence of these kinds of problems? Or is this just your your gut feeling based on your years of experience?

 **Speaker 3** 9:30  
So we have some softwares which help us with some. We can raise the issue if something happens like this. For example, when you have data duplication because we have sometimes multiple or inconsistent copies of the data sets across the code base.  
And we can raise the issue and ask, OK, how this happened? How it because one of the examples as we deal on as I said on on the picosecond level of the data input and we usually track values like temperature, voltage.  
The values that don't change so much in the level of the picosecond, we sometimes we have many rows that have basically identical values. So sometimes if they keep the row data as it is, we have to and they usually do because they physicists mostly are afraid of deleting.  
Hitting from the from the from the sets we have to deal with them and we have to get into their mind to get into. OK you know this is is repeated like a few million times you have the same values can we delete them that did nothing on queue this is the probably.

 **Speaker 2** 10:33  
OK.

 **Speaker 3** 10:40  
They're not understanding of the data set by the side that performed experiment and size analyzed the data.

 **Speaker 1** 10:47  
Hmm, yeah.

 **Speaker 3** 10:49  
And as I said, the hard-coded parts, because when we deal with the raw data we do with the petabytes of data. So in order to get the HPC resources to analyse them, sometimes get it can get pretty rough.

 **Speaker 1** 11:03  
So again, if I look at the the distribution of your answers and the relation between the two categories, I, as you can see here, highlighted the some areas.  
But this is I would say is a typical area where basically the the relevance and the frequency is very close to each other. I can zoom in Speaker 2 if you the the score is very is very close and as you mentioned right now I understood that.

 **Speaker 3** 11:20  
Mhm.  
Yeah, it would be.

 **Speaker 2** 11:29  
Yeah.

 **Speaker 1** 11:37  
Essentially this is because, yeah, maybe you encounter very seldomly this, this, this specific smell. Then we have another, I would say not really unusual area related to these smells that you previously said that yeah, in your case are not really relevant.  
So I also assume that as you said, not really frequent because you your organization is is working well, you don't have issues on communication and things like that. So I think that this is not particularly interested, but there is one smell actually that you gave and it's about data, so data.

 **Speaker 3** 12:06  
Yeah.

 **Speaker 1** 12:17  
where you say that is really relevant in your opinion, but at the same time it's like averagely encounter in your specific, I guess in your specific job. So I wanted just to understand if when answering five you thought that in

 **Speaker 3** 12:18  
Mhm.  
Yeah.

 **Speaker 1** 12:36  
In general, data believability is an important smell for this category of systems, or if you even here were pointing out specifically at your experience.

 **Speaker 3** 12:48  
So one of the aspects that we work here is also had on cancer patient therapy and we we deal a lot with private patient data and that is actually the reason why I put here the five because the data that we're gathering are usually data.

 **Speaker 1** 12:58  
Mm-hmm.

 **Speaker 3** 13:07  
From the medical institutions and being protected by the privacy laws, they have to obscure a lot of data that might be very relevant for us and that actually can influence highly on the availability of the data.

 **Speaker 1** 13:09  
Yeah.  
On.  
OK, interesting.

 **Speaker 3** 13:22  
So it might be also one of the aspects, but the the privacy concerning data are not specific just for the medical institution. You have the same for the finance, you have for the defence industry. So the privacy issue is let's say at least 20 to 30% of the data that are analyzed are under the GDPR regulations.

 **Speaker 1** 13:33  
Yeah, of course, yeah.  
They're gone or yeah, cannot be used. Yeah, yeah.

 **Speaker 3** 13:42  
Yeah.

 **Speaker 2** 13:46  
So you you're categorizing that as reducing their believability. Essentially it's saying that you you can do less with them because much of the data.

 **Speaker 3** 13:53  
Or we need more data in order to extract some some frameworks of some.

 **Speaker 2** 13:58  
OK.  
I understand.

 **Speaker 1** 14:02  
Yeah, so in just a follow-up question that is not related actually to the smell, but just out of curiosity, do you happen to work with the synthetic data or maybe you try to augment the data that you that you have?

 **Speaker 3** 14:15  
Yeah, actually we did. We did create some of this diversity data in order actually to overcome the lack of data or yeah.

 **Speaker 1** 14:23  
Yeah, yeah, that was, uh, that was why I was asking, actually, yeah.

 **Speaker 3** 14:28  
Also because even when we're getting the patients data, some of the patients actually agree, but it's like maybe 7 or 8% of the patients agree to share the data and based on those we try to to expand and to try to synthetically fill the other data. So it's.

 **Speaker 2** 14:32  
That's it.

 **Speaker 1** 14:43  
To expire, yeah.  
Yeah.

 **Speaker 3** 14:46  
Introduces the availability as well, because we are still guessing, for example, the age, the sex of the patient, the day of the first occurrence of the of the of the incidence of the, yeah of the incidence of the of the of the.

 **Speaker 2** 14:54  
Yeah.

 **Speaker 1** 14:54  
Yeah.  
This is or whatever. Oh.

 **Speaker 3** 15:03  
Illness. So yeah, some of this data can be very relevant and we are obscured by the privacy rules, but we are trying to overcome them as well with the federated learning. Federated learning can overcome this part of the of the privacy issue, but still introduces a noise in the system that can still.

 **Speaker 1** 15:07  
Hmm.  
Sh.  
Worst.

 **Speaker 3** 15:22  
I'm seeing the model model. I can't do the back tracking. I can't go really back and analyze the data from some other institution. So it's still kind of influences heavily on the availability of the data. Plus we have a huge problem with the data poisoning. If something happens in one hospital, it can expand all over the place like.

 **Speaker 1** 15:32  
Yeah.  
OK.

 **Speaker 3** 15:39  
In a matter of seconds.

 **Speaker 2** 15:42  
What? What would cause the poisoning?

 **Speaker 1** 15:42  
OK.

 **Speaker 3** 15:45  
For example, still the we are dealing with the cancer patient data and we have some areas that are heavily influenced by the let's say either local doctors that are more prone to diagnose or more prone to give certain treatment to the.  
To the patient than the others or something like this. Or for example, we have some areas in Ukraine that are after the Chernobyl disaster that we're having influenced by the variation. So we need to reduce the importance of these data. So it's a bit bit of a.

 **Speaker 2** 16:06  
Yeah, yeah.

 **Speaker 3** 16:18  
Weighting the scales and how how to arrange the data and how to give importance to each data.

 **Speaker 1** 16:19  
Hmm.

 **Speaker 2** 16:23  
OK.

 **Speaker 1** 16:25  
Yeah, yeah.  
Interesting.

 **Speaker 2** 16:33  
Yeah.

 **Speaker 1** 16:33  
All right. So I think that from my side, the analysis of your answers, I am done. I don't have anything else to to ask. I don't know, Speaker 2, if you have any thought.

 **Speaker 2** 16:48  
Do do you do you run statistical tests to find data outliers, for example? So you mentioned Chernobyl. Is that something that everybody knows? Well, everybody knows, but you know some other maybe.  
Some other less well publicized disaster everyone doesn't know and and you need some statistical tools because you have such enormous volumes of data.

 **Speaker 3** 17:09  
Yeah.  
Well, we actually had to do it because in in ex Yugoslavia, because of the NATO bombing, we had a heavy impoverished uranium poisoning in some locations and we had also to deal with those because we had like 30% increase in osteosarcoma or in blood related.

 **Speaker 1** 17:25  
Hmm.

 **Speaker 2** 17:33  
Mhm.

 **Speaker 3** 17:34  
Blood related cancers in the years following up like in 2000 from 2001 to 2012 in particularly. So we had this in this we had somehow to.

 **Speaker 1** 17:43  
Or worse, yeah.

 **Speaker 3** 17:49  
Value the importance of the data because we were the entire the idea was to get the 10 countries of the Balkans and get the data from them and we noticed the statistical anomalies and very early in the in the analysis. So we had somehow to accompany this and to reduce them.  
Or because very learning allows actually to mostly model the database on local data in the importance of the features to allocating by the local data. So if we are dealing with monte neuro, monte we will give presenters will give the match to the.  
Multinerin data respect to the others in terms of deciding the future importance or deciding the level of treatment because the the entire scope of the project was actually to to pre-select the patients for the therapy because we can select up to 500 per year and we have 60,000.

 **Speaker 1** 18:28  
Hmm.

 **Speaker 3** 18:44  
Incidents in this area. So yeah, so it was a bit of a, let's say, morally very touchy subject and we had to be in constant and most of the statistics we got already from the medical institutions.

 **Speaker 1** 18:46  
Reported, yeah.  
Challenging. Yeah, yeah.

 **Speaker 2** 18:55  
Oh yeah.

 **Speaker 3** 18:59  
So once we compare them, we're already done, but we then repeated them in order just to confirm the the outcomes and to see if there are any other outliers that we might actually encounter. But yeah.

 **Speaker 1** 19:13  
Mm.

 **Speaker 2** 19:14  
So one final question from me. So you indicated there are a few of these smells, primarily the social smells that you did not see, did not experience. Were there, are there any smells that you experienced that you felt that we did not cover?  
And he is, yeah.

 **Speaker 3** 19:32  
I don't know. Well, you kind of packed together some of the smells that maybe I would separate like I know some modular related smells or some pipeline and code smells that maybe should be separated in terms of.

 **Speaker 2** 19:53  
OK.

 **Speaker 1** 19:54  
OK.

 **Speaker 2** 20:05  
Right.

 **Speaker 3** 20:05  
Maybe a broader picture for this?

 **Speaker 2** 20:08  
OK. So the category was there, but you're saying you might, you might have liked some subcategories in in certain areas?

 **Speaker 1** 20:12  
Yeah, more fine grained. Yeah, yeah.

 **Speaker 3** 20:13  
Yeah.

 **Speaker 2** 20:15  
OK, that's fair.

 **Speaker 1** 20:17  
Mm.  
Yeah. And also I think that the the guys, they took the categorization somehow based on the previous published papers. So yeah, that may also have some a little bias towards, yeah, higher level categorization, let's say.

 **Speaker 2** 20:19  
That.

 **Speaker 3** 20:27  
Mhm.  
Yeah.  
Yeah, that probably influenced some of my frequency level answers. For example, I mostly I mostly encounter smells in the pipeline and then cold smells like for example the the glued cold.

 **Speaker 1** 20:37  
Yes.

 **Speaker 3** 20:51  
Smell or sudden when you have some hidden feedbacks from the loops. This is something that occurs mostly in my area, so that's why we cram them together. More related than data smell. Data smells we happen once or twice or three times, that's it.

 **Speaker 1** 20:51  
Yeah.

 **Speaker 3** 21:07  
But let's say pipeline is usually the one that we have to work the most.

 **Speaker 2** 21:11  
Mhm.

 **Speaker 1** 21:12  
No.

 **Speaker 2** 21:14  
I see.

 **Speaker 1** 21:15  
All right. Then I think we we are pretty much happy with this and of course we would like to thank you once again Speaker 3 for the liability for both the survey and the interview.

 **Speaker 3** 21:31  
Of course, no problem.

 **Speaker 1** 21:31  
And hopefully we will back at you with the published papers once, yeah, yeah, once, once it is all done.

 **Speaker 3** 21:36  
Please let me know. I would love to read it.

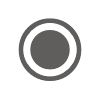
 **Speaker 2** 21:39  
Yeah, we really appreciate.

 **Speaker 3** 21:41  
OK. Thank you very much.

 **Speaker 2** 21:42  
Appreciate your contribution. Thank you so much. Enjoy your day.

 **Speaker 1** 21:44  
2.

 **Speaker 3** 21:45  
OK. Thank you. Bye.

 **Speaker 1** stopped transcription