A1

Perfect. My software is running too. We always record with both systems, so we have a backup. Classic IT thinking: no backup, no sympathy. Exactly. And then we can also jump into the interview. Exactly, so you just briefly mentioned your engineering background, but you are also familiar with the technical implementation of AI solutions. What would you say are your everyday work processes in developing and deploying AI products?

I13  
 Yes, okay. So really in the last two years, it has mainly been fundamental analyses in Databricks. So hardly any machine learning models are really used there. That might have been the case in previous years, when I used Scikit-Learn models or did something with TensorFlow. And back then it was more about looking at the toolboxes… so the first problem, which already points toward labeling maybe, is which model should I use? There is such a wide range. Part of it is nicely clustered if you look at the toolboxes like Scikit-Learn and so on. But yes, you still have to read a lot and maybe already have some experience. Because then, yes, you have already heard some things and maybe have a better idea of the direction. I also noticed this with colleagues who were less familiar with signal analysis, that it was even harder for them. They came to me: "Hey XXXXX, which model could we use here or there?" That is difficult. But going back to the question about everyday life, when it comes to ML models, you use standard toolboxes, first Scikit-Learn, get familiar with the whole tree of possibilities, and then just test how well it works. TensorFlow was another story. I did an example there, like with these Variational Autoencoders. There I already knew exactly which model I would use. Then it was about finding a good example and modifying it, adapting it.

A1  
 Yes, exactly. You just touched on labeling a bit. At this point, one quick note I forgot in the introduction. We, of course, also have very concrete research questions in the background that we want to explore with the interviews, but that’s not necessarily what we will ask verbatim in the interviews. It’s really more of an open conversation. Of course, the people in the interview are somewhat briefed, they have at least seen the website, maybe also the graphic. So they roughly know what it is about. But for us it is really helpful. We want to find out where the practical relevance is. Yes, but there are no right or wrong answers in the conversation. Just wanted to say that upfront. Exactly, so one problem is definitely this understanding, the technical understanding: what happens in machine learning? What are models? What do these methods do? What are the advantages and disadvantages? That’s a bit what you just explained in the area of model selection or method selection. In the end, it’s a lot of testing and trying. Exactly, and we talked about this briefly, but do you have any concrete ideas on how we as researchers could maybe make it easier, this technology transfer, and, as you said, also bring along people who have less technical understanding of the subject?

I13  
 That is really a good question. If I think about how I got into it, it was really just attending the lectures. Then a lot was already known, and then it branched into deeper study, for example reading the Scikit-Learn documentation to see how to do it even better. That is always difficult. Because, well, it has to be some kind of clustering. First, which methods belong together? It has to be organized, and in each cluster you need to separate again why method A is better or worse than method B, for which use case. So somehow, you should first cluster by broad use case, regression, clustering for example, then in clustering, unsupervised, supervised, going further down. Then you have the unsupervised methods, and again separate, refine, and show the differences. Maybe, yes, I often see this in my everyday life, you have to think from the use case so it doesn’t get too abstract: what use cases exist, what are the main use cases, maybe take the top five and see how to break them down. This tree I just tried to describe, which branches more finely. And then show the separation. For me, for example, I think everyone has different requirements. If you deploy a large language model that you want to train yourself, I don’t know, XXXXX said that training GPT or ChatGPT costs about 20 million, that’s not my requirement. I’m only doing these small models here and the training takes maybe an hour at most.

A1  
 Yes yes, exactly.

I13  
 That’s why, for example, I don’t need to look at costs. For me, it’s really about which choice of algorithms is the right one. Maybe the choice is clear. You want a large language model, but then the requirements are different again. How long does it train, how politically correct are the answers? So I think you would have to look at what the main use cases are. Which criteria are relevant for each use case? Here, for example, costs for large language models. For me, maybe finding the right model again among this multitude of small models for clustering or regression. Yes, I think you would have to work that out and then present it nicely. One method could be labeling, on the other hand maybe also, I’m thinking again of SciKit-Learn, a tidy website that keeps branching out, would also be helpful.

A1  
 Yes, so a big problem is definitely the versatility that exists both in the learning tasks and in the methodology. So we have, as an example, moving from language models to image models. We have been trying for many years to create good image models, starting with MNIST. Then Random Forests were quite good for a while, then came Convolutional Neural Networks, and four or five years ago everything was completely reorganized again with Vision Transformers, which incorporated attention layers and suddenly outperformed everything else, sometimes delivering better results with fewer parameters. But just understanding the possibilities is extremely non-trivial. The situation is that more and more people want to use this in practice. And of course, training language models involves a completely different effort in data and training compared to, say, having a tabular dataset of some industrial product or some sales data with 10,000 entries, a few features, and just throwing some standard classifiers or regressors on it. That’s a completely different use case than having a million or a billion images and wanting to do something with them, or with language. Would you personally consider the SK Learn website suitable for learning this? I mean, it’s possible, obviously. But do you feel it’s well structured and easy to navigate, or did it have a steep learning curve for you at first?

I13  
 Yes, partially the content is, well, sometimes I would say it’s pretty good school-level, good, yes, good minus between good and satisfactory, something like that. It’s already good that the methods are well clustered, I think, but you are immediately, quite quickly, you select regression and suddenly you have all the methods, they are not further subdivided, I would say. Then suddenly you have 20 methods. Well, yes, I think there are also headings, which helps a bit, but you still have to read quite a lot. And maybe also, I don’t know if it could be presented differently, but what might be a bit confusing is that we have the method description, that is the model description, on one website, and then the mathematical explanation on another, it’s not all in one. I find that sometimes a bit confusing, because you have to jump around a bit, but overall I think it’s quite good. Another thing that comes to mind, besides this whole model selection, what is also important for me as a basic functions user is the operationalization of the topic. We also had that with the Basic Smart Sensor, that would also be important for me as a side topic, how well it can be operationalized, what already exists, so also some entry-level guidance.

A1  
 Yes, definitely, because it is completely different whether you are talking about language or image models or basic algorithms and which library dependencies you have. It’s absolutely not trivial how well it can be integrated into existing systems. How is it when you talk with colleagues about this topic or work with people who know much less about it? Do you feel that there are sometimes communication problems? Is it generally a matter of technology acceptance or more of a rejection? Can you say anything about that area?

I13  
 I think many people definitely see the power behind it, especially now with the revolution of ChatGPT, which is really amazing. I think many were convinced even more by that. But I think it can be applied quite quickly. A few lines and you already have a regression model ready. So I think many development engineers can apply it quickly. But if you go deeper, there is still a lack of understanding of certain things. And I think many might feel very confident. You watch a few YouTube videos and think you already know a lot, but it goes much deeper, the topics. Yes, the understanding of what is behind it, I would say. On the other hand, a lot is really about application in the machine learning world. You might run something through GRIT Search and already get good results. I’m kind of brainstorming with myself. I’m wondering how much one actually needs to know. On the one hand, it’s always good for a developer to understand how the model works. But you would also have to ask to what extent do I need to know it? And to what extent only for application? I don’t need to further develop it. That’s maybe an interesting feature if you want to create a table of contents for ML models. How far do I need to go if the topic is only application? Of course, you need to explain a bit. But I often see that people already get a basic understanding quickly through YouTube. With a few numbers, you can already create first models. I think that’s pretty good. From discussions with colleagues on a project, a colleague also said, the variety of models, I don’t know which one to choose, I’ll just pick one randomly. But that’s what we mentioned at the beginning. Not to lose orientation. But people really see the power. I notice this in the company, that even at higher management levels it has arrived. They explore what else could be possible. I know XXXXX develops less, he wants to apply. He looks at the market, what can be implemented quickly at XXXX? From management levels, but also engineers, look, what potential is there? What else could we develop? So there is definitely an affinity. People see potential. Yes, they want to continue.

A1  
 Yes, okay. Yeah, that’s quite interesting because if you ask the general population on average, not necessarily just people developing something in the company, there is still a very strong skepticism towards the technology. I think, though, that also depends on the daily business that XXXX operates. For those people, it probably doesn’t really matter why the water pump does what it does. Whereas, if it’s about, I don’t know, cancer detection or something, people get much more nervous when they read that AI is being used.

I13  
 Yes, okay.

A1  
 Because there is a different form of criticality there, or also in the area of autonomous driving, there is this moral question of whether a machine should be allowed to make driving decisions. There is a lot of discontent in the population in some areas. But exactly, we take it like that for now. It’s definitely cool to hear. So, I am obviously also somehow proof that XXXX is interested in AI transfer and in using AI in the company, otherwise we wouldn’t have the partnerships.

I13  
 Yes, you are right, you nudged me again. I was thinking very much from my own bubble. You are definitely right. Even friends who are really far from technology, of course, it may also worry them that cyborgs are being created or politically, you always hear about so many fake news. I was just reading a report during lunch that Aussiedlerbote.de, a news magazine, supposedly with 23 editors, is all AI fake. The Tagesschau, or rather a research network, I don’t know, NDR, MDR, and others, confronted the guy with it. Yeah, he then slightly changed his imprint and said, no, we are doing AI research for reports and stuff, but I think only because he got caught. So I think that scares people. So yes, you are right. The influence that can happen because of that, yes.

A1  
 Yes, I mean rightly so, it has to be seen critically, any technological development always has to be questioned critically. But I mean, we are doing the survey with users in industry and with people in companies who want to develop and offer AI products. So it’s also correct that you are describing it from that perspective. There is technological affinity and interest and real enthusiasm for it, which is good, also as a researcher. Makes sense. Yes, let’s move on to the main part. Labeling, you roughly know what it’s about, even if we have never really talked about it, but you have seen the website and XXX XXXX XXXX XXXXXX XXXXXXXXXXXXX XXXXXX and so on. So I will share my screen first, that’s probably how I start. Okay, you should, you all should see my screen. Before I explain in detail what it’s about or the purpose of it, can you just give me a spontaneous reaction to what you see? What are your thoughts? Just freely.

I13  
 First, I have to get my bearings. So ABC, that immediately stands out, that’s good. I see an association with things I already know. Labeling of technical products, washing machines, etc. Then at the top, where it says AI label, clear, and below that the items. At first, they seem a bit confusing to me. MobileNetV3, Small. Probably this is already an example for a product. But at first, that would confuse me. And below, I would… yes, I would initially interpret these four points as benchmark features, four of them. Battery, probably power draw per inference. So per model application, how much power is used. Then model accuracy. Corrupted Robustness. Oh okay maybe how robust it is. So for misclassified… depending on the dataset. How reliable is it? Yeah, I find that a bit difficult, especially compared to Accuracy. And Running Time per Inference. Okay, that’s also clear. How fast is it calculated? Okay, that’s clear. Yes. Only the top confuses me. But probably, if I knew which product it is, if it was somehow a XXXXXXXXX, I would have understood it.

A1  
 Yes, exactly. As background for the top part, it is actually about a specific image recognition model, called MobileNet V3 Small. It was trained on ImageNet. ImageNet is one of the most well-known large image recognition datasets. 1000 classes, 1000 images per class, with relatively high resolution, because we have something like 300 by 300 pixels in the dataset. And MobileNet V3 is specifically a model that can somehow run on mobile devices. So it was trained for a certain edge use case. Training is relatively intensive. It actually takes a few weeks on a single GPU to fully train. But it can then run very efficiently with relatively few parameters even on small devices. That’s why it’s called MobileNet. If you scan the QR codes at the top, you can also get to the paper of the model and get more information. This is a bit of an explanation so you can place it at the top. Do you feel that this form of presentation has any kind of connection to your work?

I13  
 I think, for me less so. I could imagine, from XXXXXX’s perspective, who applies the models and doesn’t develop them, yes, then I am farther away.

A1  
 So you would practically create models that could then be labeled?

I13  
 Yes, for me, it’s really more about applying regression, classification, variational autoencoders, I think these are always smaller models, not something big, already finished, that you just buy and apply in a targeted way, right?

A1  
 Yes, that’s completely true, but I actually once wrote a paper where I did this for standard classifiers, because in the end, whether you label a neural network, where do you make the trade-off? A random forest with 1000 trees also has so many parameters that you would reach the same order of magnitude as a neural network. Sure, with a linear regression with 10 features or so, calling that a model or even calling it AI might be a bit stretched, but it is very hard to make a clear cut-off and say, this is a classical statistical method and this is now AI. It’s totally fluid. So, you could provide the measurements below exactly the same way for standard classifiers or regressors. They might have slightly different metrics, but basically you can still do it.

I13  
 Yes, I think these are exactly the things thought from the use case. XXXXX, more of a user of larger models, I think that fits. How long does it run, how robust is it? Many users use it. Robustness, data is different. It runs all the time, how much performance does it use. For me, it’s more what we said before, wow, how do I get around in Scikit Learn? I would need it again as a developer to be simpler, I wouldn’t even call it AI, for me it’s Machine Learning. Simple machine learning models, I would wish for other features. Maybe really classification, then classification of numerical or categorical classes. There are also inputs that are not just numbers, but really, how should I say, categorical variables. Suitable for that, so for me that is more the case. Then maybe also suitable for embedded, not suitable for embedded, maybe how operationalizable it is in Databricks. I think I would need other features. So that is what I meant at the beginning, I think. Depending on what the user does.

A1  
 Yes, other features is a really good point, because that is of course also, we have now just designed these labels, and of course, we are not doing a detailed user study on it, but there are already various studies that say that any form of explainability for AI and so on must always be somewhat dialog-like or at least allow for adaptation, because different target groups need different presentations. That is a bit different from the analogy you drew earlier to the energy labels we have in the EU, because that is only aimed at end users and it only needs one piece of information, and then they make their decisions based on budget, what is the most energy-efficient model and that’s it. There, an attempt is made to reduce it to a single quality criterion, which is of course a bit tricky. It makes a difference whether you run such a refrigerator in an attic apartment or in the basement, but it is this complexity of decision options that is completely abstracted away. As a developer or programmer, you are interested in much more detail. You don’t want an abstract representation, grade A, B, C, D, E, but you want to know, as you just said, where it will be embedded, how operationalizable it is, what is the, I don’t know, when was the last commit? So if it’s a library or something, then it is much more interesting whether it is still actively developed or already deprecated or will be deprecated in a year. But as a developer, you order or present your own solutions that you have developed. So clearly, in the process of developing, you need deeper information, but maybe that is something that would be interesting in reporting, for example, if you want to explain it to people who are not interested in that depth or cannot participate at all?

I13  
 Yes, so up to now I always maybe took some images from Scikit Learn or drew them myself, in variational autoencoders also a bit pre-arranged, self-drawn images of how clustering happens, yes exactly there, or more specifically, I once used a Verge algorithm, that is also for classification, and the good thing was it was also suitable for embedded, because we first thought to run such an algorithm on an embedded system, and there I also simply showed how clustering works, plotted data points in a two-dimensional space, some red, some green, then there is a cluster and then a new point comes next to it, the cluster learns, expands, I explained it to you on that level, took images from Scikit Learn, I don’t know if I could explain that with labeling

A1  
 Yes, it’s about trying to explain how the model works internally. That goes in the direction of, not reporting results, but trying to make the function of the method or model understandable, maybe using abstract examples, clustering is always the go-to thing, you can visualize it well, two point clouds and then, like with an SVM, the line moves so that it separates both linearly, and then you say, well, what if we don’t have a linearly separable problem, then you need higher complexity classes. That is of course not what you would want to do with such a label. A label does not try to explain how a refrigerator works. In this analogy, if you stay there, an energy label or Nutri-Score does not explain why it received this rating or how the model works. The question is whether there is maybe a communication need to stay at this property level.

I13  
 Yes, but actually you are right. Alright, exactly, you are right. Those are two different things. One is exactly, how does the model work? The other is maybe really, yes, it is actually good to have such labeling, to show additional information from a user perspective. I am just thinking, for example, about my boss, who is further away from machine learning, he is an intelligent man, a physicist, but has been active as a manager for a long time. So, he is not deeply involved in the details and is always skeptical. He also knows, XXXXX, you are actually an engineer, how good is your knowledge? That is why back then there was also XXXXXX XXXXXXXXX XX XXX XXX, to evaluate my work, so he could always check how good it really is, because he can no longer assess it himself. And labeling would help there, because then he would know, this is approved by the experts and the features listed are validated, there is something behind them. Only, it would have to be the right ones that interest him. And I am just thinking about what interests him. Now, for example, with these Basic Smart Sensor analyses, he says, all analyses should run in two minutes for 50, 100, 200 Basic Smart Sensors. Therefore, I would say speed interests him. CPU power, I think, maybe also, because then it is again money. Maybe I would need a stronger machine in Databricks if I argue like that. Yes, and also robustness. So, yes, that is maybe additional information that reinforces what I have selected. That would help me. Because it is approved by the experts and builds trust.

A1  
 Yes, or also just comparing the pros and cons a bit. So exactly, instead of hard-to-understand papers or abstract implementations, you could maybe represent the practical properties in a more understandable way. That also allows comparing different solutions. And then you could tell the management department and say, these are the things we tried. Now the decision is yours, what is more important. You can also comment on how well you can personally make this comparison.

I13  
 So here. Yes, what do I see? Exactly, now I even see that below, I thought there were just colors, colors of the features, but now I see they also have a meaning, exactly, how good it is. And now I see on the right, it was even green twice, yellow once, red once, which I think led to a good Score A. On the left, I see, the bottom green is already a bit worse, exactly, the time is also longer, inference takes roughly twice as long, the accuracy is higher, but the power consumption is also higher, robustness as well. But there, if I sum it all up, probably as a regression, the values lead to a worse score, which then is C, because the features deliver worse values.

A1  
 In the end, it is a weighted sum, but the interesting thing is the weighting here is chosen so that the quality goodness should be distributed 50-50 with the resource-related aspects, like energy consumption and runtime. It is about how many resources a model uses. Robustness and accuracy are metrics used to estimate qualitative or predictive accuracy. Of course, you could also say at this point, everything else is irrelevant, as long as runtime is below a certain threshold, or energy consumption does not matter, we just need the highest possible accuracy. And then you can set this weighting. Here, the upper scale of the overall ratings could then allow management to set their own priorities or constraints and rank models based on the metrics below. Actually, more metrics are in the background for the overall ranking, which we examined, for example, instead of just top-1 accuracy, also top-5 accuracy, whether the correct class is among the top 5 classes. In ImageNet with 1000 classes, this is very relevant, whereas for a binary problem it is irrelevant. Here below, we only show four metrics to keep it manageable, but of course, everything can be adjusted again. Exactly, and then the absolute measurements, shown below, are mapped to relative comparison values, from which both the color rating and the relative comparison values are processed together for the overall rating.

I13  
 Relative comparison values, I do not fully understand that. What I understood and found interesting is that if management sets the weighting themselves, with Management XXXXXXXXXXXXX it results in A and D, and with XXXX it results in C and A, because everyone sets it differently. I actually find that quite cool. Tailored to the application. I understand that. And what is it that you mean by local… Absolute values I understand. These are the absolute features.

A1  
 Exactly, these are the absolute measurements below. These relative values result from the fact that, practically speaking, this is how it’s done in the background with energy labels as well, which is where the inspiration comes from. In MobileNet V3, for example, compared to many more models, we evaluated a total of 15 different models, MobileNet V3 had the highest speed and the lowest, no, highest speed, lowest runtime, and lowest power consumption. That’s why it now gets an A rating and is assigned a relative value of 1. And all the models that are slower, for example on the left, as you just correctly said, were about half as fast, roughly, would then get the corresponding relative index value of 0.5.

I13  
 Ah, yes.

A1  
 Of course, you can become arbitrarily worse, you could be, say, 100 times slower, then you would be around 0.01 instead of 0.5. And that’s how it can be done, and then these bins are determined for the color coding on all relative comparison values.

I13  
 Yes, I understand. So always relative to the best model or whatever, to existing results.

A1  
 But of course, in depth, it’s not shown or explained because, again, what is that, these are detail information that end users ultimately don’t care about. But this comparability is one of the main points why I feel such labels could have practical relevance. Do you feel that, regarding the metrics below, from your work in this field, are there any measurements missing for you? Are there things where you would say, wow, it would really be important to see this at a glance? We’ve already talked a bit about, like, this characterization, classification, what kind of model it is, what it does, what technology it uses in the background. That would be something you could do. Aside from that, any other things?

I13  
 Yes, exactly. First of all, when we talk about it, and I’m not very familiar with it, I think they’re actually very obvious. So computation time is important, how fast it is. Then also how much power it needs, because power is money if I operationalize it. Do I need a bigger machine on Databricks, do I have to pay more? Robustness on different datasets is also good, and in general, how well does it perform? That’s already good. I think again from my… The problems I have… Where I have pain? That’s always with XXXX. Thinking from the pain: which model should I apply? How do I navigate? That would be interesting to represent somehow. Then also, thinking from the pain: how do I operationalize it? I’m always on Databricks, I don’t know the wider world, Azure Cloud and so on, Amazon Cloud and AWS and all these other things, I’m not familiar with whether it’s different there. So you could say, is it well operationalizable in Databricks and maybe bad in AWS or vice versa, so whether to show something like general operationalizability. Is there something I can embed it in easily? Is there some kind of wrapper where I put the model in and with little effort it runs? That is a pain. Wait, pain for me is still runtime, always trying to reduce it. Pain is cost, also in terms of computation power. What else is pain for me? At the moment, I can’t think of anything else.

A1  
 Yes, so it’s good if we’ve covered everything for now. As I said, what you always have to keep in mind is that when you look at different learning tasks, they all come with different metrics, and of course, you could also look at precision-recall instead of top-1 accuracy and so on. So in terms of quality, it always depends on what learning tasks you have. Are you doing regression, classification, clustering? What I also find interesting, I just noticed it again, although I already knew it, it strikes me here with the labels. Many, so it seems natural to assume, if the model is faster, it also consumes less energy, and that’s usually the case. But we can see very clearly on these two labels that runtime almost doubles, but energy consumption almost triples. Yes, that’s true. In the end, that’s because the model on the left is also significantly larger and uses the GPU more. That means the GPU on the left will simply draw more watts because more processors are being used when the model is deployed. We haven’t talked about this in detail in other interviews, it just struck me that this is a good example of runtime not being equal to energy. Yes, correlated but not equal.

I13  
 Yes, so the only thing is, you could also think about money. How expensive is it to operate, but I think that is equivalent to performance.

A1  
 Yes, you are also in this operationalizability problem. In the end, it’s just Python code, and how operationalizable it is depends entirely on what infrastructure the company uses. I mean, you could also somehow put it on a Raspberry Pi, then it would get data from a sensor locally, process it locally, and then send the rest somewhere via some POST API. That would also be operationalized. So the question is always, I find it very hard to imagine, how operationalizable is a TensorFlow model like in this case or an SK-Learn classifier. At first, it’s not good or bad in terms of operationalizability. The issue of embedding it and connecting it to data and so on lies more with the company. I don’t think that’s something you can say about the method.

I13  
 Maybe it’s like that, XXXXXXX, as you said earlier. I actually liked the idea if you have input parameters. One chooses Raspberry Pi, another chooses Databricks, another AWS Cloud somehow, and then suddenly a different labeling comes out, right?

A1  
 Yes, definitely.

I13  
 [Unintelligible] that everyone gets something different displayed in these four fields. That there is a website

A1  
 Yes, so the environment here is now also somehow a single cluster node with 8 GPUs. Otherwise, there is no abstraction in between. There is a Python instance on it, the model is loaded into it, and that’s it. That is, of course, yes, definitely, this, I always call it the environment when I write about it, you have to keep that in mind. And of course, the measurement values below will also fundamentally change if you change the environment. Of course, it somehow needs, like response time of one to two seconds for a complete batch of image data. That might be realistic on such a heavy GPU machine, but certainly not on a notebook or something. Yes, exactly. So of course, you have this dependency on the environment. And then the question arises again, it would of course be cool to have a website where this was measured on all kinds of things and you could query the results. The alternative would be that people conduct their experiments on their infrastructure, then get the labels spit out afterwards for the things they tested. And that’s a bit the way I am doing it at the moment. So the library to create these labels is open source. You can find it on GitHub and you can just load a Pandas DataFrame with this model, has these measurement values on this dataset, and then you can dynamically generate such labels from it.

I13  
 Yes, that would actually be cool too. There is this website, the CPU benchmark or whatever it’s called, where every user does it and it is then collected centrally in this database and then you have a ranking from it. You don’t need you or anyone to do all the tests, but the world works for you, but somehow there has to be a benefit.

A1  
 A big problem is, for example, I looked at Papers with Code and OpenML. These are online repositories to just share machine learning experiments practically open source, so people can see how good it was. But resource consumption is absolutely underreported. About 80 to 90 percent of reported results are quality-related, and runtime and energy consumption are hardly reported. But yes, in itself that would be cool. That’s also a bit the vision I have. No bachelor student needs to run their classifiers locally on their standard datasets anymore because everyone has that. I don’t want to know how often MNIST has been mined. Or ADDLT/Adult as a tabular dataset or something. It’s actually nonsense that people still do that because it’s already been done. But yes, still, there is no good repository to collect something like this online and somehow learn from it. Exactly. We have already talked a lot about SKLearn and software documentation and so on. Do you know any other systems to get information about the behavior of machine learning models or methods, with the words used interchangeably in that context?

I13  
 Yes, so I also learned a bit about what I like, that the community is really open worldwide. All papers, many papers on X… Arxiv or whatever it’s called, are available there. I mean, I haven’t read much, but I know they are even available with source code. But what is also great is that there are open lectures, complete lecture series. And there are really, like, some of great quality from Leadmind, they had them long on their website, now it’s all migrated to YouTube and is now all on Google, but there are great lecture series.

A1  
 So right now we are more in the mode of learning about Machine Learning in general, but not about specific models, or well, maybe a little about specific methods. There is an Introduction to Unsupervised Learning, and then in 45 minutes you get a bit of an overview. But if you are looking for concrete stuff, more than just getting a general overview, can you think of anything else?

I13  
 Wow, then actually I usually go directly to the websites. So to Scikit-Learn, TensorFlow, I look at the examples there. Yeah, but I also don’t really know, for example, Variational Autoencoder. I need to first understand the theory, what does it do? There are different levels of abstraction, up to mathematical descriptions with all the details, just as an example. And then I know how it works, then I want to build it. And then I miss how to actually build it, how to implement it. And then an example in TensorFlow helps me, I build on that and then I read more and more about the individual layers, or how to connect them, or which function I need to put in the class. So that’s how I proceed. So actually, I am always directly at that point, meaning on one hand there is the theoretical description in lectures, and then the practical application using the toolbox with its documentation. That’s how I have gone about it.

A1  
 Okay, that’s cool. The background of the question is that we, of course, also compare ourselves a bit with other systems that are already on the market. The things we came up with so far actually miss lectures, but that obviously goes in the direction of scientific publications, because most videos or stuff that present these things are somehow again, a bit of science communication, making scientific results more usable. Then Library Documentations, point five on the list, of course already much discussed. What is a bit closer to what we have done are the Model Cards, originally from Google, but now also available, for example, on Hugging Face for pre-trained language models or image-generative models. Similarly, the Fact Sheets from IBM, which are only published for IBM products and made available on demand, or even on financial payment. And Papers with Code, if you don’t know it, maybe also a helpful resource for you. Now for MobileNet V3 here on the left, the actual paper, about 40 pages, you can read it or not if you don’t have the time. On the right the Model Card as it is on Hugging Face, where you can see, for example, benchmark results for devices, model size, and a bit of the interface, what kind of images the model processes. Then here on the left on Papers with Code, which is really cool, all kinds of papers and the corresponding codebases where MobileNet was used in some form or maybe re-implemented. There you also have a nice timeline for datasets, showing how good the models are. So for overview stuff this is also very helpful. On the right a blog post, here the Keras documentation, and a generic IBM Fact Sheet, because MobileNet was not released by IBM. Looking at all these things in comparison, can you somehow name advantages or disadvantages? What do you notice, what are your thoughts?

I13  
 Yeah, you really think about how quickly you can see something at a glance. Your label, I think, is the most abstracted, most condensed, least text and images, and you see it pretty quickly. Yeah, paper, typical paper. You think, you have to read a lot, invest a lot of time to understand it. Then the one above, I think that was the Factsheet in the middle top.

A1  
 Yes, I think that was actually the Model Card.

I13  
 The Model Card. Yes, I think it already goes in the direction of what you have, so also less text, more numbers, more overview, a table, already more abstracted. And bottom left it is also similar, table, code. So in general, I would say, you see, all are descriptions of models, and now there are more or less strong levels of abstraction, condensing everything as an overview, yes, or that is actually what you did, yes, now I’m repeating myself. It actually shows the most condensed and maybe is the friendliest for the user.

A1  
 That’s the question. From your perspective as a user, a technical, somewhat deeper documentation with direct interface, code, API, how to use it, might be more helpful. But that is something you cannot really show to the management level. They want a more abstract presentation of the results. And it also comes in that we do not see this label as a replacement for other systems, but rather as a complement. And what we feel is missing at the moment is this more abstract representation, so that the technology can be made more understandable, transparent, maybe also more usable, because through the QR code you can also access more detailed information. For people who do not understand the subject very well yet.

I13  
 Yes, exactly. Actually, that’s what I said at the beginning as well, right. Definitely a good addition. It always depends on the application. Exactly. My boss, of course, he would maybe actually set the weighting now. Or if I knew what my boss wants, I would go to the website, set the weighting, press Enter, and then I would probably pull your label, if I were to go that way, if it existed. And then maybe I would also like to go to that website via the QR code, where the code and the paper description are, to understand a bit more about what is happening. And that’s how I might proceed. It really always depends on the user, what they need, how abstract or detailed it has to be.

A1  
 Okay, great, thank you. Let’s move a bit to the final part of the interview, where we can talk a little about the keyword trustworthiness. Do you feel, as you just described, a website where you go, upload your results, and then get an interactive display where you can adjust how you want to prioritize things. Do you think that would be accepted, or are there trustworthiness issues? Can you put yourself in that situation or say something about it?

I13  
 Okay. So, yes, I think that really varies from person to person. I’m actually always the kind of person who doesn’t really distrust. I’m usually open and so on, but actually, in light of Russia, hacking, and all that, the idea that something distorted might come out or that you might download malware or something, that could be critical, or that you might reveal company secrets. So if I put on a critical lens, then that could be a problem.

A1  
 Would that be a problem? I mean, if you really had a data frame, maybe in Pandas or something, just a CSV, with only measurement values, basically a table without specifying the purpose, which machine, or anything. Do you think that would be okay for XXXX, in the context of your work at XXXX, if you just uploaded it and got labels back, which you could then discuss with your bosses, or would they immediately ask where it came from? Did you give away our data? What happened?

I13  
 Well, I would say it wouldn’t be good. Sometimes at XXXX, we have our chatbot, and we’re supposed to always use it. Sometimes the answers are not very good, so I use another one and abstract the problem. But it wouldn’t be great. It would be tolerated if you could say, okay, I needed the labeling, it also moves us forward.

A1  
 Well, theoretically, you could run the software locally because what I do is open source, so you could just run it locally on your computer and wouldn’t have to give out the CSV, that could also help.

I13  
 Yes, then that’s unproblematic.

A1  
 Yes, well, relatively speaking, you don’t know if my software in the background might send it back to me somehow. That’s a bit of the problem. You don’t have the time to check every line of code in every open source tool. Sure, you could run it in a secured VM without internet, but in practice, people don’t do that.

I13  
 Yes, I mean, you always have to trust to some extent. We also trust Microsoft as XXXX, we are a Microsoft company, the others are a Google company. It might depend on that. If it really comes from XX, then you know it’s not from Russia or something. That might be okay. I would say at XXXX, they are not that strict. Now, IT is stricter, but in development it wasn’t that strict. Before strong IT restrictions, I could download any Python toolbox, there were no restrictions. But in the last few years, it has all been restricted. It’s not so simple anymore.

A1  
 That’s understandable, right? Microsoft, we had just a few days ago that absolute global failure with a library, which Microsoft rated as semi-safe, and Microsoft themselves are not exactly experts on IT security. Sure, everyone uses it, and our university is moving more to Microsoft, but there are extreme cases.

I13  
 Here with me, I was hacked once, in my cloud, in November or December. It was my Microsoft cloud, Azure, on my phone I always use Azure Storage. Luckily, I had no data there, but suddenly I saw a new folder called Documents in Cyrillic. A month later, there was news that Microsoft was hacked. So, yes, you can never be completely sure. It also depends on the data. If you suddenly want to do business models that include business data, actual company results, it could be more critical. At our company, it’s not really welcome, but in development, you can try it.

A1  
 If you think about the scenario we were just talking about, there wouldn’t really be any important secrets. These would just be measurements, like how fast the model runs on the data, but you don’t have to release the data, you don’t even have to describe what was done, it’s just some calculation that happened. You’re only reporting, just the results of this calculation, not the real data results, but metrics about the results themselves. So in that sense, it wouldn’t be security-critical, there wouldn’t be anything there that someone could profit from, which could then harm XXXX.

I13  
 But I understand, the CSV includes, those are my input data.

A1  
 No, you don’t have to upload the input data. You basically only upload the CSV where it says model XY on dataset name has these measurements. So you’re not sending data that goes through the model back and forth, just data about the behavior of your model.

I13  
 Yes, I understand. But then I also wonder, if XXXX develops a model for a business case, not always in the technical field, I’m imagining accounting, controlling or something, then they do it for their dataset and I don’t know if they would bother taking a known dataset that is similar, like MNIST or another known one, and run it through to have comparability. So that would...

A1  
 Maybe it wouldn’t even exist. But by the way, we are running a little out of time. For me that’s not a problem. But if you have a follow-up appointment...

I13  
 We can do a little more, until ten past at most.

A1  
 Yes, alright, we can manage. I just wanted to say that once, exactly. So I think the scenario would be more like, you have a dataset that you got somehow from a business use case, then you tested 10 models on it and you want to report which model performed how well and then leave the decision to the management level which model should be deployed in production in the end. You don’t need to get an external dataset or anything, you would develop the code to test models, which is something you’ve already done, do hyperparameter search, different configurations, test everything and then figure out, okay, these are the models we tested, they have these properties, these comparison values, and now you want to communicate that in a more understandable way. I think that would be the use case.

I13  
 How do I do that then? For example, I have software, we download it, it’s then at XXXX and then I send the models there but I also need my input dataset to do it.

A1  
 No. You basically have code that runs all the experiments once and writes the measurements as a DataFrame or CSV, and then you can give that to my software and it just creates the labels and saves them as PDFs. So it’s just a reporting thing, it transforms what your software outputs into something you can visualize. With the software I wrote, you also have an interactive application. Then you can, for example, compare PowerDraw versus RunningTime or PowerDraw versus Accuracy, which is probably more interesting. You get a nice 2D scatter plot and can hover over each point, each model, and see the corresponding label and so on. So it’s not like I’m saying you have to do it, but that’s what I’ve been working on and what could be done.

I13  
 But what I don’t understand is, for example, the runtime. If you say I only need the results...

A1  
 Exactly, you’d have to measure the runtime and somehow profile it in your code.

I13  
 Aha, okay. So there are certain interfaces. I need to make sure, for example, that the output tracks these metrics with the features. And if these columns exist, I can upload them to your software and it evaluates them. Yes, I understand.

A1  
 Exactly, right. So in the end, I’ve done this already. For example, the code to do it for ImageNet, you can take inspiration from that and see how XXXXXXX XXXXXXX did it when he measured everything and then do something similar yourself. And I also have in mind, I plan to have an experiment API where you can say, run this code and report these metrics, so you can integrate it as a service. Then you quickly get into something like RapidMiner, or various companies have already worked on this to provide an AI testbed.

I13  
 I can really see, yeah, the effort, I don’t know if… On one hand, it’s of course nice to show this to the boss, but I think for my boss it would be enough: Ah XXXXX, cool, your model… Mh, if I test 10 models with hyperparameters and then just show him with my metric how good it is, I think that would be enough for him.

A1  
 So if the quality is right, then the rest is fine too.

I13  
 Yeah, I could just imagine, and this is again, I know, I lack the experience here, if I were in a real data science department that also develops quite a bit, whether maybe the boss would want something like that, because we are a bit… My boss doesn’t really put much value on processes, he just wants to see the result, that it works. And that’s it. And how it is presented, well, if I already have my metrics, that’s actually enough for him.

A1  
 Yes, functionality is always the highest priority. As long as it works, the rest is not so important. It’s probably also because you are not working on AI products that you sell in the sense that you provide the models, but when you make AI products, it means that in the end it’s a device that someone can buy or a service, but they don’t directly use the AI themselves, so to speak, it is all embedded in another system. But yes, to tidy this up a bit here, in any case, thank you very much for the insights. Super interesting and definitely very cool answers as well. To conclude, I will send you the slides here in case you want to look at them again in peace, in case you took something out. If now any people come to mind, I mean, of course we have already done an interview with you and with XXXXX, but if you still think of people from your acquaintances or also at other companies or at XXXX who you think, wow, I talked to them about this before and I think they would also have interesting things to say or so, then please don’t hesitate. Always happy to forward, the website is linked here as well and definitely helps us. We have done about ten interviews now, we want to do around 15. So we are on track, but yes, if any people come to mind, always happy to forward. The study will probably last until the end of the year roughly, but as soon as we have results, you will get them by email as well. Exactly. I think we can end the recording now.