1:26

Alexey

**This week, we'll talk about data-centric AI. We have a special guest today, Marysia. Marysia works as a lead data scientist at GoDataDriven. She has a strong interest in education and teaching, both as a part of your current role at GoDataDriven and also as a co-organizer of PyData Amsterdam and PyData Global. Welcome.**

1:46

Marysia

Yeah. Thank you for having me.

# Marysia’s background

1:49

Alexey

**The questions for today's interview were prepared by Johanna Bayer. Thanks, Johanna, for your help. Before we go into our main topic of data-centric AI, let's start with your background. Can you tell us about your career journey so far?**

2:03

Marysia

Yeah, sure. I started by studying artificial intelligence at the University of Amsterdam, I did both a Bachelor’s and a Master’s in artificial intelligence. My early career was focused on specifically applying deep learning on medical imaging, particularly early stage lung cancer detection in 3D CT scans. Within that domain, my focus was on geometric deep learning on the medical domain, which was also the topic of my Master thesis, supervised by Max Welling.

After that, I transitioned into the role of data science educator at GoDataDriven, which meant that I taught and created courses on basically all things data science. Now I work as a lead data scientist at GoDataDriven. In addition to that, I organize meetups and conferences, mostly in and around Amsterdam, with PyData.

2:55

Alexey

**What did you do as a data science educator? You said your responsibilities included creating courses?**

3:02

Marysia

Yes.

3:03

Alexey

**That's cool.**

3:04

Marysia

Yeah. At the GoDataDriven Academy, we teach a lot of courses on everything data science. Some are very generic, like an introduction to Python for data analysts, or an introduction to data science, for instance. But we also create a lot of courses very specifically targeted towards a specific audience. For instance, I created a deep learning NLP course, or an unsupervised learning course. Those are more detailed or more specific topics, and then gave me an opportunity to really dive into that topic and create exercises and assignments and material on that. That was really fun.

3:39

Alexey

**I think I spoke with folks from your company – from GoDataDriven – at the recent PyData/PyCon in Berlin. As far as I remember, you're doing education and consultancy, right?**

3:54

Marysia

Yeah, I was mostly a data science educator for about two years, but I strongly believe that you can't be a good teacher if you don't also have hands-on experience. When I do courses or when I teach courses I like to really tell a lot of anecdotes about my experiences and in my work, because it speaks more to the imagination of why we're doing this than just talking about the concepts.

While I really enjoyed it, I do feel like after two years of mostly focusing on the educational side, I need some hands-on experience. Also I was really missing this decoding bit. So the ideal situation for me is to go do both trainings and education and also work as a data scientist and combine that.

4:39

Alexey

**Do you still teach?**

4:42

Marysia

Occasionally, some courses but I'm more focused on my lead data science role.

4:48

Alexey

**What do you do as a lead data scientist?**

4:51

Marysia

I'm with a company where I'm mostly focused on building a community of practice there. They just went through a transition in the way that they organize their teams. And I want to make sure that all the data scientists still communicate clearly with each other, get to exchange knowledge, but also increase the maturity level of the data science products that we produce. So it’s making sure that everyone is not just doing something on their own time, behind their own laptop, but bringing them together and making sure that we actually get to mature, well-functioning, monitored data science products.

# What data-centric AI is

5:24

Alexey

**Now, this is such an interesting topic. [chuckles] I want to ask more about that – maybe at the end, if we run out of questions, because the main topic for today is actually data-centric AI. Maybe it's related to what you do right now – building a community of practice and improving maturity. So let's go back to data-centric AI. What is data-centric AI? Why do we need to care about this?**

5:54

Marysia

Yeah, that's a good question. I did a whole talk at PyData London about this whole thing – answering this question. But basically, in short, the central idea behind data-centric AI is that the focus has to shift from Big Data to Good Data. So Andrew Ng says that having 50 thoughtfully engineered examples can be sufficient to explain to the neural network what you want it to learn. The reason why we call it data-centric AI is because it's in contrast to the model-centric approach that a lot of us are used to. Because in a model-centric AI, the focus is really on iterating on the model, which means that you create a baseline model, give it the data that you have, you evaluate the baseline, and then you go back to the model, you revisit the modeling pipeline, and you make adjustments there – you change the algorithm, you adjust the architecture in your neural network, tune your hyperparameters, maybe even just some of the data transformations, such as how you compute missing values, or augment your images.

But the dataset is generally considered static. And the idea behind data-centric AI is that we shouldn't just iterate on the model, but we should iterate on the data as well. That means improving the quality of our data by relabeling mislabeled or ambiguous data points. But it can also mean gathering more data or more examples of specific classes or readjusting the train and validation split. So data-centric AI is essentially about focusing on your data rather than just focusing on your model. And the idea of that is, of course, new. I mean, I know that one of the first things that I was taught was “garbage in, garbage out”. That's always been the idea. But the thing is that I strongly believe that data-centric AI is particularly relevant now compared to, let's say for example, 10 years ago.

It's like the hype around deep learning – deep Learning wasn't new in 2010, for instance. The ideas weren’t new – a perceptron, we know about perceptrons in the 50s. The backpropagation has been around since the 80s. Even the first convolutional networks were already around in the 90s. But somehow, deep learning really came to prominence since, let's say, about 2010 and onwards. And that's not because the ideas themselves were new, but because the ground had changed. We have large annotated datasets available, GPUs that became available, which meant that deep learning gained more traction. I believe that something similar is happening for data-centric AI.

We've always known that data is important, but it's more important *now*, I believe, for two main reasons. First of all, because more and more problems are now being solved with deep learning, we are more often dealing with unstructured data, rather than structured, tabular data. So unstructured data like images or audio or texts, whereas for tabular data, it's easier to focus on the quality because you have descriptive statistics and you can easily create visualizations to investigate correlations between features. It's easier to explore your data. And it's also easier to get an idea of the data at hand before you start any modeling.

After some proper exploratory data analysis, focusing on data quality is kind of a natural result of that. But this is remarkably more complicated when you get to unstructured data, because it's more difficult to get good insights of the data that you have at hand with unstructured data. You can sample a few images, but do you know for sure that your data is good and representative? So because we're dealing with unstructured data now more than ever, we are more in need of tooling and techniques to help us out with that.

9:56

Alexey

**It's a bit counterintuitive to me, that we need this data-centric AI now more than 10 years ago, because it seems like 10 years ago, when we didn't have all these GPUs, we needed to be smart about how we approached things. Then it mattered like what kind of data we have. But now it’s just “take a cluster of GPUs, throw in more data, and sit back and wait till magically becomes better.” It doesn't work like that, does it?**

10:28

Marysia

No, unfortunately, not. There's this persistent idea that the quantity of the data will compensate for the quality. So if your data quality is not good, just gather 1000 more examples, and that's fine. But I think something really changed because we're more than ever making use of transfer maps. If you have to train a neural network all the way from scratch, then yes, more data is probably a good way to go. But nowadays, we have large foundation models and we are mostly focused on fine-tuning those. Because, as an individual data scientist, I don't have the resources to compete with something like GPT-3, so I fine-tune it to my own problems.

When you're fine-tuning, that's the situation where the model has already learned a lot about the structure of images, the structure of texts – if you want to fine tune it to your specific problem. If in that situation, you are giving it examples that aren't right, then it's basically fine-tuning on the wrong thing. I think that's one of those cases where the data matters more – the quality of the data matters more – and it's also one of the cases where we, as data scientists, can have a bigger impact. Because I'm not going to adjust the architecture or the parameters of a large model that is provided to me (a foundation model) but I know about my data – I know about my use case – and I can focus on the data more than I can focus on the model itself.

# Data-centric Kaggle competitions

11:58

Alexey

**I'll try to recap everything you said – maybe not everything, but the main idea. So we have two approaches: the data-centric approach and the model-centric approach. In the model-centric approach, the dataset is static and you iterate on the model, maybe tune the model, change parameters, try different architectures, you make some adjustments, and you might also do some feature engineering. But the dataset – the images or the rows of your dataset remain the same. In contrast to that, in the data-centric approach, you change the data instead of changing the model – or in addition to changing the model.**

**You also look at the data, you see the bad examples, you see the good examples, you see where you can improve the data – instead of focusing on the model, you put more effort, more emphasis, on the data part. Right? [Marysia agrees] I think that this model-centric approach is very typical for a Kaggle competition. [Marysia agrees] In a Kaggle competition, you have trained .CSV file, and then you have a test.CSV file, and that's all you’ve got. You, of course, have some room for experiments. You can tune your XGBoost model, you can train as many models as possible, you can put all of them together in an ensemble, but it's a model-centric approach because you're tuning the knobs of the model.**

**The data-centric approach is different. The reason I spoke about Kaggle, because I also heard there are competitions about data-centric AI. To my knowledge, you even took part in one or even in multiple competitions. [Marysia confirms] Can you maybe tell us about these competitions?**

13:45

Marysia

Yeah, sure. Of course, I took part in multiple Kaggle competitions, as many data scientists have. I really recognize what you're saying – that usually, you have your data and you don't go about gathering more data. It's more about the model itself. But as a contrast to that, we have the data-centric AI competition that I participated in with two of my colleagues, Rens Dimmendaal and Roel Bertens. The data-centric AI competition was a competition hosted by DeepLearning.ai that ran for six weeks in September 2021. And the central idea behind this competition was that the model was fixed and the data could be changed. The task was about classifying images of Roman numerals – handwritten Roman numerals.

We were initially provided with a dataset of 3000 images, divided into a certain train and validation split. It was up to us to submit the data and all the compute was handled on the chance side. The model was a fixed ResNet 50. We could submit up to 10,000 images, so there was a cap there. We couldn't just say “Let's just get a huge quantity of data.”

14:57

Alexey

**That's smart to do this limit, right? [chuckles] Because you can just generate so many images and overload the system with this.**

15:05

Marysia

Exactly. There was a cap on that. One of the things that I really liked about that, besides introducing me to this idea of data-centric AI, was that it made participating very accessible. It was a deep learning challenge, because it was an image challenge, but you didn't need any beefy GPUs because all the computers handled it for you.

That means that anyone with about 10 MB of storage space and an internet connection could participate. Like I said, I participated with my colleagues Rens Dimmendaal and Roel Bertens and we ended up winning in the Most Innovative category and presenting a solution at the data-centric AI workshop at NeurIPS alongside the other winners.

15:47

Alexey

**That's pretty great. Was it your first exposure to this idea of data-centric AI?**

15:52

Marysia

Yeah, that's an interesting question. It was my first exposure to the concept. I've not heard of data-centric AI before. But my one major competition that I participated in before was also a Kaggle competition. That was the Data Science Bowl, where we participated in detecting whether someone had lung cancer or not based on CT images. That wasn't a data-centric AI competition (we ended up in third place with the company that I was with at the time) but it turned out that all the top three solutions really didn't use the data that was provided by Kaggle whatsoever.

The second solution, by Julian DeWitt was all focused on creating a system so he could easily evaluate whether the scans or the data that he had was any good, and making selections based on that, and creating a tool that allowed him to easily annotate additional information, which made it easier for him to get to that second place solution. So while that wasn't named as a data-centric AI competition, it was very interesting to me that all the top solutions ignored the data that we were provided with and decided to focus on gathering additional data, creating a model in a completely different way, based on the data that we did think was more useful. So it was kind of data-centric in that sense, as well.

17:20

Alexey

**You just didn't know it was called “data-centric”.**

17:22

Marysia

Yeah, exactly.

17:24

Alexey

**It's pretty typical for Kaggle competitions, I think, to use external data. You have to disclose that you use this particular dataset. But I guess it gives you an edge in image competitions when there is an external dataset that you can use to make your model better.**

17:44

Marysia

Yeah, I agree with that. But I also think that there's a difference between simply gathering more data, and for instance, what some of the winning approaches were based on – focusing on finding out what the most useful data to add was. Like, “What things were we missing in our data? What things weren't we missing? Which examples had to have a higher weight because they were more important than other data samples?” Those kinds of decisions about your data, which I think goes a bit further than just getting some extra images.

18:17

Alexey

**This is a nice answer to the question, “Why should they take part in data science competitions?” You might stumble into an idea. For you, you said it was September 2021, so not so long ago. But now, you're given talks about data-centric AI, you're talking on a podcast about that and you also won this Most Innovative Approach award. I guess your life has changed a little bit after taking part in that competition.**

# The mindset shift to data-centric AI

18:46

Marysia

Yes. I think data-centric AI for me is – when we talk about something like this, it's very often focused on the tools and the methods like, “How do we do something like this? What kind of packages do you use? What kind of tools?” But for me, the most important thing was a mindset shift. Focusing on the data and not seeing the dataset’s aesthetic has really helped me throughout my career since that moment, because I think that's a very important insight – that, yes, this may be the data that I have, but I can also make decisions about that and change it throughout my modeling process.

19:24

Alexey

**Yeah, that's interesting. The point about changing – earlier today, you mentioned that your train/validation split also doesn't have to be static. Right? [Marysia agrees] Immediately I thought, “But wait, if our validation set is not static, how do we compare two approaches and say that one approach is better than the other if our validation set isn't consistent – if we change it between two runs?” You see what I mean?**

**Let's say you have two models, usually the easiest way to compare these two models is to evaluate these models on the same validation dataset and whatever model gets the higher score is better. [Marysia agrees] But the moment we change the validation dataset, we cannot compare these models because they are evaluated on different datasets. [Marysia agrees] For me, it got me thinking, “Okay, if we start changing the validation dataset, how can we be sure that it's actually an improvement?”**

20:26

Marysia

I think that's a very valid question. First of all, when we were participating in the data-centric AI competition, I think that our insight that the train and validation split that we were provided with wasn't a good split for us – our validation set turned out to be not very representative of all of the data. There was a huge part of the data that was not represented in the validation set that *was* represented in the train set, so we decided to rebalance that. In our case, in the data-centric AI competition, we didn't have access to an actual test set, because that was handled on the competition side. So there was another holdout set that we were eventually evaluated on. The validation set in this case was used in order to determine when we were done training – for early stopping was when it was used. In that sense, it still matters.

I have a very technical background with artificial intelligence – I like to focus on the numbers. But I do also believe that we shouldn't only focus on the number of the metric. Sometimes you can make a change to your test set to your validation set, which makes the numbers go down, but gives you more confidence that that number is correct. If you notice that, for instance, in your test set or your validation set, the thing that you eventually validate on is missing a part of the data that you do expect to encounter in practice – do you not add that data to your test set, because you're no longer able to compare the bundles? Do you not add it because your metric will go down? I think it's better to change it but be confident about the change that you have made. That makes it more trustworthy regarding what your eventual results will be.

22:13

Alexey

**Then at the end, you can just re-evaluate all this. I know we're talking about the data-centric approach, not model-centric. But then at the end, if you change your validation dataset you can just reevaluate your approaches on the new split – on the new validation.**

# Data-centric does not mean you should not iterate on models

22:25

Marysia

Yeah, exactly. Yeah. I also want to emphasize, of course, that data-centric AI doesn't mean that we shouldn't change the model. [chuckles] I think it's often named in contrast to model-centric AI. But for me, data-centric AI means that we iterate on both the model and the data. So yes, it doesn't mean that I just take a baseline and never change my model anymore, because in practice, that won't get me the best results. But I do think that very often, there's more to get from improvements on the data than, for instance, changing the entropy in your decision tree to a tuning.

# How to implement the data-centric approach

23:02

Alexey

**Right. You said that for you, the most important realization was the mindset shift from not just how we do this, but also that the dataset is not a static thing and you can change it. But I'm still wondering, how do we actually do this? What are the tools? What are the approaches for that? How do we implement this?**

23:24

Marysia

Yeah, that's a very good question. So there's not one toolbox that I can recommend that has everything. I think it's a very broad subject. It's also what you focus on. There are a lot of tools out there that can help you with labeling, or finding the data points in your data that need labeling. But of course, there are different tools for text-based or image-based or audio-based. But there's way more to data-centric AI than just relabeling. There are also tools for, for instance, generating synthetic data. How do you create good synthetic data to augment your current dataset?

There's a very broad spectrum of things that are all data related. And there's also a lot of development at the moment being done on these tools. Because I think that at the moment, we all experience, especially working with financial models, working with unstructured data – we experience a need for tools that help us out with these kinds of things. A lot is currently being developed – a lot of high tech tools, a lot of low tech tools as well. I'm personally a big fan of not using one thing and having everything work out of the box. I wouldn't like a magic “fix all your bad labels” tool if there was one, because I think that my value as a data scientist is in understanding the data and talking with subject matter experts. I like to have a lot of control over that process.

I think the most important lesson that we learned is that, at the end, it's not about what tools you use, it's about how easy you make it for yourself to iterate on the data and how you keep track of that. That could mean that you use DVC to version your data and you have a good overview of all the datasets that you've used. But in our case, for the data-centric AI competition, we were still naming things with “\_v3” which maybe wasn't the nicest versioning approach, but it was easy for us to re-label our data and that was a very important thing. That wasn't a bottleneck.

25:31

Alexey

**Funny that you mentioned this approach for data versioning that you had. In DataTalks.Club, we recently launched a competition. This competition is about classifying images of different kinds of kitchen stuff – be it a plate, cup, glass, fork, spoon, and so on.**

25:52

Marysia

I think I saw that, yes.

25:55

Alexey

**For me, I was preparing the data for this competition and the folder’s name that I ended up with was like “new2\_Kaggle\_final” or something like that. [laughs]**

26:08

Marysia

I think we've all been there.

26:09

Alexey

**Yeah. It was terrible. [laughs] Maybe for the next one, I'll use something like one of the tools that you mentioned, like DVC, or something else. Because at the end it was really hard to keep track of what was changed between like these 1000s of folders.**

26:26

Marysia

Yeah. It's the same thing – when I started out as a data scientist, when I was trying out different hyperparameters, we'd be writing things on a post-it note next to my laptop. [chuckles] There's a better way to track your experiments than just writing everything on a post it note. I think the same goes for your data, probably. So that's the first thing, but also just changing your data. I think that's an important thing. For a lot of people, it's difficult to do because there's a whole mindset – it's considered static.

One of the things that made it really easy for us during the data-centric AI competition, was that when we basically started labeling in Google Docs, we found out that when you have the URL of an image, you see the image itself, so we could very easily change the label that was associated with it, and then turn it back into a pandas data frame. And then we could do some magic scripts so that every file will be in the right folder. It took a bit of time to create those scripts, but it made all of the other work a lot easier for us in the long run – to adjust things. I think that's an important one. Make it easy to make adjustments.

27:28

Alexey

**So the process you had – you had a Google spreadsheet with one column with the URL and other column was class, right? [Marysia agrees] Then you could just go there and change a label, or multiple labels, and then you had a script that would pull data from this Google spreadsheet and train a model. Then it would say, “Okay, for this version of the spreadsheet, this is the score you have.”**

27:55

Marysia

Yeah, exactly. We also had some little tricks to make it easier for us to work with the spreadsheet, because again, there were 3000 images. So that means that you have 3000 rows. We did things like, which was relatively easy to do, put the data through the model, already get some predictions, and then order the data points on confidence of the model, for instance. Or we made extensive use of the embeddings, visualizing the embeddings, and seeing that some data points were very far away from the distribution of that class. Then that's one of those data points that you pay attention to. We made use of little tricks like that to filter on what to focus our attention on.

28:37

Alexey

**I recently had a chat with one of your friends – PyData colleagues – Vincent Warmerdam. I think he's really into these tools that help you with finding bad data, right?**

28:51

Marysia

Yes, that's true. He actually did a talk in PyData Eindhoven last week on labeling – a lot of tricks – which was a really good talk that I recommend.

# Focusing on the data vs focusing on the model

29:02

Alexey

**You said the most important thing is focusing on the approach, how you iterate over this, and how you make it easier for you to iterate, instead of focusing on high tech and low tech tools. I guess the low tech tool that you used in your competition was Google Spreadsheets. But I'm really curious about the approach that you took.**

**How would you actually implement this in practice? Let's say you join an organization as a consultant, or maybe as an in-house data scientist and you want to follow this data-centric AI approach. How would you structure your project? What kind of tools would you use to make it easier to implement all the things we discussed?**

29:55

Marysia

That's a good question. I don't think I have a specific answer on what tools I would use. But I think that’s mostly because I don't feel strongly about certain tools over others. I'm more interested in the process. I think one of the most important changes that I would make and have made in the organization that I'm with, is that one of the most important lessons that I learned from the data center AI competition is that data is easier to talk about than the model is. It's very hard to go to subject matter experts and talk about your results and say, “Well, I used weight normalization instead of batch normalization.” That's not a very viable conversation. But you can show examples of the data. You can talk about the data.

You can talk about the odd examples that you find and go to someone who knows more about the source of the data that can explain things to you. I think this was particularly relevant when I worked in a medical imaging company where we actually had a doctor employed, who, whenever I found odd things in the data, I could go and talk to him, show him, and he would explain things to me, which would really adjust the way that I approached the modeling as well.

I wouldn't have any specific tools to recommend. But I would recommend having a very close connection to the person who knows more about the data and realize that if you get the same performance, but by changing the data, or by changing the model – by changing the data, it's much easier to collaborate. It's also much easier for the person who you're presenting the model or the end result to, to have a bit of faith that it's working correctly, rather than just an abstract metric.

# Resources to help implement the data-centric approach

31:34

Alexey

**I guess what I wanted to hear from you was more tactical. For me, what you’re saying sounds like a strategy, “You need to be close with subject matter experts,” which is super valid, but I'm still wondering, how do I actually make it happen? I have a project, I have a bunch of subject matter experts, and I have a dataset. I want to make sure that I don't go crazy, that I don't have like 1000 folders with names like “new\_v2\_Kaggle\_final” and so on. How do I make that happen? How do I organize this? Do you have any tips and tricks or best practices or talks that I should check out or anything like that?**

32:22

Marysia

Yeah. The reason why I find it very difficult to give an answer to this is because I think there's a lot of great tools out there. But there's two resources that I find very useful. One is by Haiti Research, and one is by WhyData. They have a really good overview of awesome data-centric AI tools that are structured in a way, “Is it about profiling? Is it about synthetic data? Are you working with images? Are you working with text?” Because the tools are very specific to that.

Those are two resources that I would recommend to look for the right tools for the use case that you're working with. I would, as a data scientist, still start with a model-centric approach. I would still create a baseline as a model, but then use those model results to not only go back to the model and how to adjust those hyperparameters, but also use the results of that to see if there's any gaps in my data that I'm seeing.

33:16

Alexey

**It's basically doing error analysis and understanding where the model was wrong. [Marysia confirms] And then trying to understand why the model was wrong and talk to people who know data well to figure this out. Because maybe for you alone, it could be difficult to understand why, for this particular dataset, this was the final label or these were the predictions. It helps to talk to subject matter experts to figure this out and conclude that maybe the label on this example is actually not right and it should be a different one. [Marysia confirms]**

**Okay. This is what the process looks like, right? You train a model, you analyze the errors, you analyze the mistakes of the model, you talk to subject matter experts, and you iterate, iterate, iterate until the model is good enough.**

34:11

Marysia

Yeah, basically.

34:13

Alexey

**Sounds simpler than I thought. [chuckles]**

34:16

Marysia

I think it's always important in these kinds of things to keep humans in the loop. And that could be subject matter experts but it could also be just the data scientists who've learned, obviously, a lot about data as well and knows what they're doing. I'm not a big fan of the type of tools that automate everything away. I know there's a lot of optimization going on that can help us, but always keep a human in the loop to be sure that you can truly trust your results.

# Data-centric AI vs standard data cleaning

34:42

Alexey

**I must admit it sounds terribly similar to “standard” data cleaning. You have errors, then you go to the dataset and you see, “Okay, this row doesn't make sense. This is an outlier, so I’ll just throw it away.” And then maybe you even have a rule, “If this value in this feature in this column is two sigmas away from the mean, then you just throw it away, or you add it (or whatever),” which is a pretty standard data cleaning step, probably. What's the difference between these two approaches, or is data cleaning the same as a data-centric approach?**

35:24

Marysia

I think data cleaning is a part of data-centric AI. But then data-centric AI itself is more broad. It's easiest, I guess, to talk about data-centric AI in terms of “What is a good label and what is a bad label?” Or “How do you choose to deal with your missing values?” But it's a lot broader than that. It's also, for instance, about (what I think is a very important thing) “What is my dataset representative and is there any bias in that? Is my dataset complete?” Those are not things that are typically part of data cleaning. There are tools being developed, for instance, for images to see if you have a representative dataset – if you have a complete dataset – and those tools can really help there as well.

# Making sure your data is representative

36:04

Alexey

**How do we actually check that? I'm really curious. I have a dataset now with spoons, forks, cups, glasses, etc. How do I know if it's complete?**

36:14

Marysia

Yes, that's a good question. [chuckles] I think it's very hard to know without any domain knowledge. For example, a toy project that I did once was classifying penguins – classifying penguin species based on images – and I sourced the dataset, basically, just through downloading all the Google image results. Of course, lots of the data is right, there's a few mistakes in there and it was easy to focus on getting those bad labels out of there and re-labeling those. But I think that's one of those cases where it's very easy. I mean, if there's one thing that’s classified wrong, maybe we can gather a bit more data, and that kind of overcompensates for that.

To make sure that it was representative in that particular case, I thought about “What situations are there where I can encounter penguins? They can be on land, they can be in the snow, but they can also be on water, for instance. And I need to have groups of penguins, and I need to have individual penguins, and I need to have penguins from the side and maybe not necessarily from the top.” So all the things that I can encounter. I noticed just by going through my data that I was missing a lot. I didn't have a lot of examples of baby penguins for different species, for instance. So that was one of those examples where my dataset was not complete. Of course, this is something that I could have figured out by just scrolling through all the images and noticing this. But in this particular case, I came up with the different things that I thought I should have in my images and I decided to visualize the embeddings of my images.

I basically put my data through a neural network that was already trained, that didn't take the head of it. But I took the embeddings and I reduced the dimensionality to UMap. With UMap I can visualize it and I use an interactive tool to be able to view my images. And I saw different clusters of types of images because, of course, all the penguins in the water were kind of together in terms of embeddings and all the on land were kind of together. I used that interactive tool to get a bigger insight into my data. Then I noticed “Yes, I did see a few baby penguins,” but I didn't see a lot of data points around those. There were only three in my datasets.

I think that's one of those cases where, yeah, it's part of my data gathering process, but I did think about this upfront, I didn't think about “What kinds of things can I encounter and do I see these in my datasets?” I didn't want to manually go through the images because that would take a lot of time, so I used a neat little trick of visualizing the embeddings to make that process easier for myself. Then I gathered more data by Googling the type of penguin and then the word “baby” after it. That's how I gathered more data and the right category.

39:18

Alexey

**I think one of the tools that we can potentially use is a tool from Vincent, which is called Embetter, right?**

39:36

Marysia

That’s really cool. I actually did this project before I knew about Embetter, so I would really like to try it out again, whether that makes my life a bit easier.

39:46

Alexey

**For anyone who is watching this right now or listening to this, there is a video from Vincent on our channel. It was published this week, I think. It's called “Open Source Spotlight” and “Better in Bulk”. So Vincent built two tools. I'm really amazed by how he turns his ideas into these small little projects and then just publishes them in open source. So the approach for you is – you need to think about all the situations where it's possible to encounter a penguin or, if we generalize, all the contexts, all the situations where we can see the objects were detecting (understanding) and then see if we miss anything.**

**For example, if I go back to the dataset with cups and glasses, perhaps I need to think about the conditions, like the light conditions, because I need to have images that are well-lit, when it's dark, when it's bright. Then there are different angles – sometimes maybe in some situations, I see the handle of a cup in some situations head on. So it's just sitting and thinking – maybe taking notes.**

41:10

Marysia

Right. And I think it's important that you do this at the start of your process, when you first gather the data. But this can also be part of your error analysis. When you see that your model is specifically making mistakes on cups where you can't see the handle, that might be a reason for you to think, “Okay, maybe I need to gather a bit more of that data and verify that hypothesis.” So that's a hypothesis that you can make, “Why is my model having trouble with this particular image?” Verify that with your initial data source, see if you can gather more data, and then have a new version of your dataset and try, let's say, exactly the same model again and see if it works better now.

# Knowing when your data is good enough

41:47

Alexey

**And when do we stop? How do we know if it's good enough?**

41:50

Marysia

Ah. I think that's always a very difficult question in data science.

41:53

Alexey

**It depends, right? [chuckles]**

41:54

Marysia

It depends. [chuckles] When your results are good enough for your usage. How much time you have also matters. For instance, it did for me when I was doing the penguin project, because I was just sourcing images through Google, it was very easy to just Google, “baby penguins”, “addley penguins”. It was very easy to gather more data. But if you have to actually go out there to your kitchen and photograph a bunch of additional pictures of all your cups, that does make it a little bit more complicated. So when is it good enough? Given the time that you have and given the project that you have, if your results are satisfactory. And that can be because of model tuning, but that could also be because of data tuning.

42:35

Alexey

**Then I guess, talk to subject matter experts, stakeholders, and ask them what they think. Is this 80% accuracy satisfactory or do you need more?”**

42:47

Marysia

Yeah. Though, in my experience, that's also always a very difficult conversation to have because when you just talk about metrics, to lots of people, it's just a random number. “80% sounds nice. Highest number is best.” So that's always a very difficult conversation, in my experience, to have. But you can give examples of your data as well, like “These are the data points that it classifies correctly. And these are the ones that it still has some troubles with. Is that okay for you?”

43:15

Alexey

**Okay. I think you've been advocating for this throughout the entire interview, “Don't focus on metrics.” Right? [chuckles] “Focus on data.”**

43:24

Marysia

Yeah, I guess so. Yes.

43:25

Alexey

**So it means “Okay, if I know that my model is making mistakes when it's dark (there is not enough light) I can just talk to my stakeholders and show them. ‘Okay, there's a picture of a fork, but the lights are turned off. That's why the model thinks it's a glass. Are you okay with this? Or do we need to collect enough pictures of forks in complete darkness, and then the model will be better. ’” This is the approach you will take? [Marysia agrees] Okay, cool. Interesting.**

43:57

Marysia

I think that's also because of my background in the medical field, where we were doing deep learning, but explainability was very important because trust in the system was a very important thing. You don't gain trust by just showing a graph or just showing a metric.

# The importance of user feedback

44:13

Alexey

**Another thing that occurred to me while we were talking is that we can take this a simple model, and if our conditions allow – maybe for the medical field, it's not good enough, but if it's a simple thing classification – we can just deploy our model and see how users play with this and collect feedback from the users.**

**Let's say if we want to deploy… There was a project I did a couple of years ago, which was about classifying garbage types. In many European countries garbage of a certain type needs to go to a bin of a certain type. In Germany, you put plastic in the yellow bin, and you put paper in the blue bin. I don't know how it is in the Netherlands – probably something similar.**

45:02

Marysia

Yeah. I heard the municipality of Amsterdam is doing very similar projects where they send around cars to notice the garbage on the street and notify the right people to pick it up.

45:12

Alexey

**And then you have this model. You can create an app and then see what kind of things users send and see if there are any mistakes there. For example, if the model says that paper should go to the black bin, which is for everything else that doesn't fit the other bins, maybe you can try to understand why the model is making these mistakes. “What kind of things are we missing? And is it because our dataset is wrong? Or is it because our model is not so good? How can we fix this problem?” Then probably the reason for that is data.**

45:52

Marysia

Very often it is, yeah. I think an important part of data-centric AI is focusing on the data. It’s also about how you actually label your data. How do you know that it's good? Exactly what you just said – collecting user feedback can be a very, very good way to get more knowledge about the data that you have.

46:15

Alexey

**Is this a typical approach of how we put this in production? Maybe there are other approaches? We just roll it out and see how users react? I guess if we talk about the medical field, we cannot just deploy this model to lung cancer things and then let people just use it and correct data later. It's simply not applicable. There, we need to use a different approach. In the case of garbage classification, it's okay if one piece of paper will end up in the wrong bin.**

# “Shadow Mode” deployment

46:52

Marysia

Yeah. That's actually interesting, because we did do this with the medical imaging software. We actually did deploy it in hospitals but, of course, not very broadly. It was first with a few people who were interested, and they volunteered – these radiologists volunteered to have our software run next to their day-to-day job. So they were still responsible for judging scans, the software was not making any decisions. I think that's a very important thing when you're still developing. But they did see the model results of the software and we talked to them extensively about the feedback. We got a lot of feedback, for instance, that certain mistakes were being made, or certain types of mistakes were being made. That point already led us to gather more data of those kinds of examples.

47:44

Alexey

**I think that's called “Shadow Mode” deployment or something like that. You deploy this thing in addition to whatever process is there and then you just use this to collect data. Then you compare whatever decision the model is making with the decision of the subject matter expert, in this case a doctor. Then you see where the model is wrong and where it's right.**

48:05

Marysia

Yeah, exactly.

48:06

Alexey

**I think we used something like that for moderation. Where I work, at OLX, which is an online classifieds site – a place for selling second hand stuff. In the Netherlands I think you have Marktplaats. [Marysia agrees] So it’s similar to that. In one of the projects, we just let it run in parallel to moderators and then we compared the output of moderators with the output of the model and we concluded that the model is good enough. There were, of course, some issues, but after one or two iterations, we concluded that it's good enough. Interesting. I was asking you about that. [chuckles] How do you know if it's good enough? For us, it was talking to these moderators and thinking, “Do you think it will help you or not?”**

48:52

Marysia

Yeah. I think that's the question that you need to answer eventually. “Do you think this is a help?”

48:58

Alexey

**And after that, we just rolled it out. I guess it summarizes what we've been discussing so far pretty well. Right?**

49:06

Marysia

Yeah, I think so too.

# What to do if you have a lot of bad data or incomplete data

49:09

Alexey

**What if we have a lot of bad data? What do we do? Maybe it’s not so easy to collect new data.**

49:19

Marysia

That's just a very difficult situation. If you have a lot of bad data and you can't collect new good data, and you can't re-label the data, I guess you could do a lot of manual work to make your bad data a little bit less bad? But if you don't have enough data, then some problems just aren't solvable.

At some point, I was talking to someone who offered me a project and said, “Yeah, also, I want you to classify 13 different classes, but I have 54 examples. Also, it's 3D images.” That’s just not going to work. [chuckles] Unfortunately, there's not a clear-cut answer there. If you have a lot of bad data, it might require a lot of manual work to make it good data, but if you don't have enough still, then it's unfortunately not a feasible project.

50:07

Alexey

**Okay, so you might even need to open your favorite image editor and edit some of the data, right?**

50:16

Marysia

Maybe, but I wouldn't be very enthusiastic about the project, if that's what I ended up doing. [Alexey laughs] I studied artificial intelligence because I find this whole topic, this whole field, very interesting. I do try to automate those things away, because if I ended up just doing data cleaning by opening image editors and removing stuff there, I don't think that's the reason I got into this job in the first place.

50:41

Alexey

**So maybe it's better to have a model that is doing the editing, right?**

50:45

Marysia

That would be really nice, yes. Or tools that can help you out to automate a lot of this stuff away.

50:50

Alexey

**But then if it's just 50 images, then maybe you cannot really do this.**

50:54

Marysia

Maybe not.

# Marysia’s role at PyData

50:56

Alexey

**Another topic that I wanted to talk to you about was your role with PyData. I know that outside of your work, you're quite involved in the PyData community. You are the co-chair of PyData Amsterdam, and in general, you're quite active with this, as I said at the beginning. I think I came across your talk in PyData Berlin this year and this is how we decided to reach out to you. Was it this year? I think I found you there.**

51:28

Marysia

I spoke at PyData Berlin twice. I think I did a tutorial this year.

51:33

Alexey

**Yeah. What was the tutorial about?**

51:38

Marysia

It was called Serious Time for Time Series. Time to take time series seriously.

51:43

Alexey

**Okay. I do remember seeing something like that.**

51:46

Marysia

Quite a mouthful. [chuckles]

51:47

Alexey

**Yeah. [chuckles] So what's your role there? What do you do in the PyData community, apart from giving tutorials and talks?**

51:54

Marysia

Yes. So I joined PyData Amsterdam in 2019, I believe. I basically joined it because I enjoyed going to meetups. I studied artificial intelligence, I was focused on this very specific topic – deep learning for medical imaging – and I was just interested in everything about the field. I really enjoyed going to meetups and learning more about the experiences of others in the field as well. By joining the committee, I was able to organize this as well.

We organized the monthly meetup and we organized a yearly conference. So I've organized a conference for PyData Amsterdam, I've organized an online PyData festival for PyData Amsterdam, and I've organized PyData Global last year. This year, we’re full-on back on organizing PyData Amsterdam and making sure that we can create a really cool conference that brings together users and developers of basically open source packages in the data science ecosystem.

The confusion that a lot of people have about PyData – I know that the name is maybe a bit confusing – it's not just about Python, it's also for Julia and R users.

53:00

Alexey

**It is confusing, I must admit. [chuckles] Maybe it started as a Python conference, right?**

53:06

Marysia

Yeah, I think so. [chuckles]

# How Marysia joined PyData

53:08

Alexey

**You said that in 2019, you joined the committee. But I don't think it happened that one day you woke up and walked in the committee and said, “Okay, do you mind if I join you?” It was something else, right? How did it happen that you joined them?**

53:28

Marysia

I think it was actually 2018, come to think of it. But basically, I was attending a couple of meetups. At some point, the committee is on stage (it was actually Vincent who said on stage) and they said “Hey, is there anyone who would like to join the committee? In the break, come talk to me.” And that's how. I like organizing these kinds of things. When you organize it, you have the luxury of also determining what you organize.

So I get to organize the meetups that I like to attend and the conferences that I'd like to attend. I decided to volunteer and that's also how we got our entire new committee. We have about 16 people in our committee at the moment, to organize the conference and all of them joined, basically, because we just did a shout out at a meet up, like “Do you enjoy this kind of thing? Do you want to help shape this? Come and join us.”

54:14

Alexey

**I guess it's pretty useful to have a community and then you just say “Okay, does anyone want to help us?” [Marysia agrees] So this whole process doesn’t sound like it took a lot of time for you, actually – from the moment you started attending meetups to the moment you joined the committee.**

54:33

Marysia

No. No, that wasn't a long time, I think. I don’t remember when I first started going to meetups, though.

54:41

Alexey

**Well, I guess when you live in Amsterdam, there are so many meetups and communities that it’s just easy to be a part of one.**

54:49

Marysia

I think that's one of the reasons why I really enjoyed helping out with PyData Global as well, because I realized I'm very privileged. I live in Amsterdam and one of the reasons why I attended a lot of meetups was because they were simply very close to my house. So if I didn't enjoy it, I could just go home. That's really nice about the city that I'm in. But, of course, a lot of people live in places where it's not as easy to attend meetups, and therefore share that knowledge and gather that knowledge.

That's what I really like about PyData Global. I personally do enjoy in-person meetups and conferences more, but I do think it's very important to make all of this information as accessible as possible. That's the idea behind Global – it's for everyone, all over the world. Anyone can join and that's why I like helping out there.

55:31

Alexey

**I guess it started in response to the pandemic, right? People couldn't just go to in-person meetups. But then, in addition to being able to connect during the pandemic, it also allowed people from any part of the world to join and also take part. If somebody does not live in Berlin, or Amsterdam, or New York, or any other big tech hub, then they can just connect to PyData Global from their village and take part in this, right?**

56:00

Marysia

Exactly.

# The difference between PyData and PyCon

56:01

Alexey

**Cool. And what's the difference between PyData and PyCon?**

56:06

Marysia

I think the major difference is that PyCon is for everything Python, and PyData also is Julian and R. I'm not sure how PyCon feels about that. [chuckles] But we are focusing more on the data side of it. But it is actually the educational flag of NumFOCUS and all the proceeds of the conference that we organize go to support the open source ecosystem, specifically the packages that I, as a data scientist, use a lot – like NumPy, Matplotlib, Pandas, Scikit Learn – those kinds of packages. That's also what you'll see that most of the talks are about. Whereas PyCon, I would say, is generally a little bit broader than just data science and data analytics.

56:50

Alexey

**I still think there is some bias towards Python tools in PyData. A little. For historical reasons, right? [chuckles]**

56:57

Marysia

Yes. We have to make a conscious effort to make sure that the R and Julia folks feel included as well. I am mostly a Python user. I started with the Java Matplotlib, or Java and Matlab and then switched to Python. I’ve only played around with Julian and R. Most of these talks aren't really about the tools or about the language or about the code. It's more about the concept. So I think that translates well into other languages as well.

57:28

Alexey

**I remember one of the talks in PyData Berlin this year, was a general approach from a company who is selling cars – similar to what we do at OLX, which is why it was very interesting for me to check what the competitors are doing. And they use Java, for example. Which is not any of these three languages, yet the talk was quite nice. Maybe you should rename it to something like… you know how iPython notebooks got renamed to Jupyter?**

58:02

Marysia

Julia, R and Python. That's what Jupyter is for, yeah.

58:05

Alexey

**So maybe it should be Jupyter data, instead of PyData?**

58:08

Marysia

Yeah, maybe. But I don't… I'm very involved with the Amsterdam chapter, but I think there's like 100 chapters. I don't think I have the authority to change all of those names. [chuckles]

58:18

Alexey

**Yeah, probably not going to happen. Right? [chuckles]**

58:21

Marysia

Maybe not, no.

# Finding Marysia online

58:24

Alexey

**One last thing I wanted to ask you is –– how can people find you if they have any questions?**

58:31

Marysia

Yeah, that's a good question. I always really like it when people reach out. I'm reachable through LinkedIn. I have my website, which is Marysia.nl. Just my first name and “.nl” for the Netherlands. My email address is on there as well, so feel absolutely free to reach out. Also, if you want to get involved with PyData, or maybe speak, or maybe get some tips there, or talk about data-centric AI. I’m really happy to talk about this topic.

59:00

Alexey

**Okay, thanks a lot for joining us today. Thanks, everyone, for attending too. It's Friday today, so have a great rest of the week and have a nice weekend!**