**EPISODE 1464**

[INTRODUCTION]

**[00:00:00] JM:** Data labeling allows machine learning algorithms to find patterns among the data. There are a variety of data labeling platforms that enable humans to apply labels to this data, and ready for algorithms. Heartex is a data labeling platform with an open source core. Michael Malyuk joins the show to talk through the platform and modern usage of data labeling systems.

[INTERVIEW]

**[00:00:20] JM:** Michael, welcome to the show.

**[00:00:22] MM:** Thank you. I'm glad to be here.

**[00:00:24] JM:** I want to start with a high-level question. How does data labeling fit into a typical machine learning workflow?

**[00:00:32] MM:** Great question. It really depends on what your workflow and what are your ML pipeline has been so far, how you build that. Because depending on that, it may take one of the central pieces of the pipeline, or that can be something supplementary to the whole workflow of the data label. I think we're moving, as an industry, towards what is called right now the datacentric AI. With datacentric AI, the data labeling solution becomes one of the central most important pieces of the whole workflow.

**[00:01:08] JM:** Can you give more light into how it actually fits into the workflow and what data labeling –what role it plays for making machine learning productive?

**[00:01:19] MM:** Basically, the data labeling in itself is I call it an act of processing the data. This is where people looking at a specific data items, they're given their decisions, potentially, for example, classifying the items that they're looking at, or maybe creating a bounding box on top of the image that they're looking at to doing the object detection task. So, it's a dataset preparation step. Because for machine learning model to be able to become the pars from the dataset, the dataset needs to be what is called labeled. The process of labeling again, isn't the people they're sitting in front of their computers, and they're doing this process manually interacting with their usually with a browser application.

**[00:02:03] JM:** Who are the data labelers?

**[00:02:07] MM:** It depends. If you look at the markets, I would pick three different categories of data labelers. There are data labelers that are professional annotators. That's their day job. They can be outsourced. They can be employed within the company, but their day to day is basically just sit in and label and specific items in the datasets.

The second category would be the data scientists themselves. So, the data scientists preparing the data sets for machine learning, they would seat and label sometimes the whole dataset, sometimes just a fraction of the datasets preparing that for machine learning. The third category that I think we're starting to see more or less, and which I'm personally very excited about is the business users. Because business users, usually, those are non-tech savvy users, but they have a lot of subject matter expertise. And because of that subject matter expertise, that's what can be very, very valuable in terms of the labeling. We're seeing a lot of those business people get more into the data label, because the type of knowledge that they can provide through the data labeling process is really, really, really valuable.

**[00:03:19] JM:** Describe the range of tooling that can be used for data labeling.

**[00:03:26] MM:** Alright, so we have a repository on GitHub, I think, that we created a while ago, that is called Awesome Data Labeling. Inside of it, you have all the types and varieties of the data labeling tools that are available in the open source world. Usually, those tools are different by the data type that they support. Because for tech, you will have one tool. For images, you'll have another tool and basically for every data type, you will have a tool that is specific for the data type.

There are also the tools that are datatype agnostic, and they are universal support in a variety of the data types. There is even the history and some published papers with the research on the data label and tools and how the space develops over time. Because you can imagine with the development of the machine learning models, that data sets, the tooling for the data label, also needed to be developed alongside that. Right now, as a user of the data labeling tools are endless variety of the software that are available to you.

**[00:04:35] JM:** So, you're developing your own data labeling platform, and we've done some shows about other data labeling platforms. Can you just explain what a data labeling platform is, who it serves?

**[00:04:52] MM:** Sure. Any labeling platform, basically, a piece of software that you use to prepare the datasets to train your machine learning models or to improve the accuracy of your existing models, by fixing the predictions of those models. The users of that, basically every company that are built in ML pipelines or ML based products, they all needs a data labeling tool that would support preparation of the dataset, or again, the improvement of the model accuracy. So, what we have seen from our site is the solution that we are working on, it's open source and it's being used by thousands of organizations globally. We are playing hearing the general trend of the ML adoption within the enterprises as well as companies that are basing their core products on their AI/ML models.

**[00:05:48] JM:** What does the data labeling platform need to include?

**[00:05:53] MM:** It's a good question. I would preface that we’ve seen, that this field even though it has history, it's still very new. So, there are like a variety of the things that some companies take as a table stake. But then, other companies think that these things are not as important. The way we think about it, first of all, the data labeling platform usually needs to be able to work with a variety of the data types, especially if you're a larger organization, you would probably want to process a different types of datasets.

Flexibility should be one of the priorities, because when you're initially starting now, and with the data labeling, you may not know right away how you want to label your data. Sometimes you have, what we call, a policy for the data label and may change over time, because you're realizing that you need to be labeling the datasets slightly different way. Because of that, again, the flexibility is an important feature. And then, I think, for us, the way we see the data labeling platform, it's all really about the people. How do I understand who is doing the label and how well they're doing their labeling work? How long it takes them to do that? How accurate are they when they do the label and how they're consistent between each other, their annotators? For us, it's really all about the people. And basically, all the features and all the functionality that connects to the people, we say that this is something that needs to be included into every data labeling platform.

**[00:07:28] JM:** Can you give me an example of a typical – it doesn't have to be a customer, but maybe a typical user that you've worked with, or that you've talked to, like how their whole day labeling workflow works? Just give me the end to end process.

**[00:07:44] MM:** Sure. Let's say I'll take a hypothetical example here. But it's very close to what we have seen in the production environment. So, for example, when you have a conversation through Zoom, you can have a recording of that conversation. That recording of the conversation can be transcribed, and this is first annotation step that you take. Basically, the inputs for your data label and platform is audio files. The first output would be the transcription of those audio files. That's the first labeling step.

And then the second step might be for the transcription to pick up, let's say things that are more interesting and exciting, and maybe some action points, action items. Basically, whatever you were trying to get out of that conversation, extract out of the conversation. And that will be the second labeling stuff. The output of that is basically just the highlights region within those transcripts that are of a particular interest to you, and that's the output. With this output, you can train the machine learning model, and the input for the machine learning model probably is going to be two different models. The first one would be doing the transcription, the second one would be doing what is called a named entity recognition within the text.

**[00:09:03] JM:** How does that fit into an overall application?

**[00:09:05] MM:** So, we're using Zoom right now to record that. Let's say that I can login into my Zoom accounts, and I can go into every conversation that I have had, and I can see the action items, or maybe the items that are important to me for every conversation. So, for me as a user that brings value, because then I know for sure what are some of the items that I need to work on after every conversation, and this is one of the, I would say, an example of how you can embed that into the consumer-based product.

**[00:09:44] JM:** Got you. So, you're talking about example, where for example, we're going through a Zoom call, let's say the Zoom call is an hour long. And every time I asked you a question, maybe we want that moment labeled as a question. And every time there's an answer, we could have that labeled as an answer. That labeled dataset could be put into some kind of machine learning algorithm that is trying to learn what kind of videos snippet constitutes a question, what kind of video snippet constitutes an answer. My next question is just like, what kind of format does that need to be in? How does the structuring of the data work there? I can understand there's a video file and then you want to have timestamps associated with the video file, do you just have like a tuple of documents, we have the video file, and then you have like a text file with the timestamps?

**[00:10:46] MM:** More or less. You basically, on the data structure level, you usually have something like JSON, we're inside of the JSON, you'll have a Euro pointing to the audio file, and then some object that would encode the transcripts and the, let's say, a start and end positions of the question and answer within the video stream.

**[00:11:07] JM:** There's the regular problem of subjectivity around what constitutes, for example, in this example, a question what constitutes an answer. And that subjectivity is typically dealt with by for example, sending the piece of data that needs to be labeled to multiple reviewers and have that be labeled and then you can use some kind of consensus process. Can you tell me about building consensus around a labelled set of data?

**[00:11:38] MM:** Sure. It's a very interesting topic, really. I have an interest and story there to share. When we were originally developing our software, we already know that all the problems with a consensus with the biases. But what was pretty funny to myself is, I labeled a certain dataset that was based – I was labeling the sentiment, I think it was tweets, or it was textual based. I have done that myself, at the moment when I woke up, and the moment where I just about to go to bed. So, there is basically like at least 14-hour difference. I labeled the same items twice and I was not really consistent with myself even

You can see that like this problem of this subjectivity, it's not that easy to tackle. What do we have seen a good success with, first, yes, you have to send the same items to be labeled by multiple people. So, let's say all three of us, we classify a certain tweet with a sentiment as positive, highly likely, that's really a positive tweet, unless all three of us, we have some weird sense of sentiment analysis, kind of, embedded into us and we just make the same systematic mistake altogether. But you can split that process into multiple steps really.

For example, if three people label the same item, they classify that into three different sentiments. Let's say one of us, classifying that as a positive and other one is neutral, another one is negative, then we can take that sample, and we can send that for review. The assumption would be that a person who is doing review, they have a better understanding and they're more well trained and educated about the sentiment and how to identify the sentiment within the tweets. That might be one of the solutions to that.

Another one might be really dependent on the type of labeling that you do. For example, let's say we're putting the bounding boxes on the image. In that case, somebody may miss a certain bounding box. For example, if you're looking at the image where you have to label, for example, cars, and you have 20 cars in front of you, you may label 18 or 15, and miss the other ones. Somebody else may label the rest but miss the ones that you haven't missed. And then it's really a question how do you merge that? How do you identify the bounding boxes that are more or less attached to the same object? Because you can imagine that it's really hard to pull the bounding boxes into the same coordinates for two different people.

So, for that type of consensus analysis, there are a variety of the different metrics that are available. For example, for the bounding boxes, we're going to use a metric called the IOU metric, which would basically calculate how similar are two different bounding boxes. How well they match. So, for different types of labeling, you'll have different types of metric that you can use to calculate the overall consensus between multiple people, and plus of that, you can have a second or third step where you do the resolution or verification of the label and provided by multiple people.

**[00:14:57] JM: W**hat do you do when you have – like if you're trying to get consensus and you're labeling something, if there are three potential options, and all three people give different answers.

**[00:15:12] MM:** You cry.

**[00:15:14] JM:** It doesn't happen that often?

**[00:15:17] MM:** It depends. It really depends on the type of labeling that you do. For example, let's say that you are doing a hate speech analysis. Hate speech can be very, very work. Because of that, you can either ask more people to label and then you go for the majority vote. And usually you can ask three people, but all three of us agree. Okay, we asked 10 more people, and then we go for the majority vote for that particular item. Or again, you can ask somebody who you think is more of an expert on that problem, to provide their expert opinion and you take that as what we would call a ground truth annotation.

**[00:15:53] JM:** Gotcha. So, you can have somebody – I mean, you can also just like kick it to a fourth person, right? And then just do whatever they say, right?

**[00:16:01] MM:** Yeah. If you know this fourth person, you have, like higher level of trust to their opinion, then you can – well, certainly you can assign that to the fourth and that they're saying like, "Hey, Michael, or Hey, Jeff, I want your opinion of that item, because I trust that you're an expert in the field, and we'll take your label and as the ground truth label."

**[00:16:25] JM:** Now, you have been working on Heartex for a while, and there are other data labeling platforms. How did you want to differentiate Heartex?

**[00:16:39] MM:** Heartex is differentiated on a number of different dimensions. I would say, first of all, we're, I think a single company that is running their product is an open source-based product. So, our new labeling product is called Label Studio. In its nature, it's an open source solution. It's by far the easiest to get up and running with. It's very easy to install. It's very easy to set up. Another key differentiator that we have, and that we realized we want to be that as a kind of part of the product early on was, we want us to support a variety of the data types. I think you can relate to that with software engineering, where you may be working with a variety of different programming languages. You can code in pipes in JavaScript, HTML, Rust and go whatever that is.

But you usually use the same code editor. Because you learn the code editor and that's what powers you as a software engineer to do your development work. So, we took that same idea that we wanted to apply that to a detail labeling tool, where you'd have to learn the tool once with all its hotkeys, with all the installation steps, within integrations, with all the flexibility. After that, you can label any type of dataset that you want. This as another key differentiator. And then when you dive into the actual product, there are more differentiators within the product itself on the kind of more granular feature level. But I would say again, building the open source company, building the community around the data label and being truly flexible and multiple data type solution, that what puts us aside from other companies in this space.

**[00:18:26] JM:** Tell me what the hardest engineering problem that you've had to solve while building Label Studio has been?

**[00:18:34] MM:** Well, I think, for me personally, the out route for the hardest one is where it becomes more of an art than a science when you're dealing with a problem. I think, finding that configuration, creating the configuration language that you can use to build the data label and interface, because you got to think about the Label Studio to do more of these data label and tool builder, where you have that configuration language that you can use to build that data phrase that the users are going to interact with.

So, building that configuration language was quite complicated, because you want to make that composable so that you build a different, I would say like, different clauses in the configuration that would be composable between each other, and based on that, our, let's call it a compiler in a way, a compiler would build the data label and interface that users will be able to interact with. You can imagine it's almost like you're inventing or reinventing the HTML language at that point. But you're trying to make that domain specific to the data labeling problems. There was nothing available like that when we were designing that. I haven't seen myself, there is something even now that is comparable to that. I think that took a while to figure are like how we want to map all the variety of the data labeling problems back to that configuration language, and make that extendable enough so that users can do things that we never thought about.

**[00:20:14] JM:** Can you tell me more about that? What do you mean by this configuration language?

**[00:20:21] MM:** Sure. So, you can think about when you're building an HTML page, you're using HTML tags. Then when you open up those HTML tags in the browser, the browser basically compiles them into the user interface that you're interacting like the web page. We have built something very similar, where you use the same – they're not really HTML, they're more like XML formatted tags. You use them and then our application compiles that into the interface that you can interact with. So, you can put – for example, we have a tag that is called an image tag. And we have a tag called choices tag, and when you put those two together, the result that you get in the application in the Label Studio application will be you're looking at the image, and you're classifying that image. Basically, choices translate into the classification. And classification, just to give you a visual is where you would be selecting the checkboxes and classifying that particular image.

So, we have created in with those tags, I think we have about 25, 30 different tags and they can all be composable between each other. For example, you can use the same choices tag and connect that to the tag called text. And then at that point, you're classifying another image, but a text.

**[00:21:46] JM:** Gotcha. So, is this output, the configuration file, is that the output of a user's data labeling practice? If a user goes through and annotates a bunch of images, is there a single file that's output as a result of that?

**[00:22:05] MM:** I would say that this configuration language, it does two things really. First thing, it's based on that configuration language, the interface gets created, that the user is interacting with. That's where they do the actual labeling work. And this is what the configuration, we call the policy for your data labeling. And then when you upload your dataset, and you use that interface to label that datasets, the output would be a JSON. Inside the JSON, you can see the correspondence between the resulting JSON file with the configuration language, because for images, you'll have in the output, if let's say you're putting the bounding boxes on the image, in the output of the labeling work, you will have the coordinates of those bounding boxes, plus the label names that are applied to those bounding boxes.

So, it's kind of like all of these pieces are really connected to each other, and one cannot live without the other one. But again, the input for the Label Studio would be this XML file and the configuration for the labeling, plus the dataset. The output would be the JSON file with the coordinates of the bounding box in case you are doing the object detection on the images.

**[00:23:18] JM:** Gotcha. So, let's again, go through the overall workflow. If I'm using your tool, Label Studio, and let's say I've got like a bunch of images of radiology scans, and I want a set of domain experts to label those scans as being problematic tumors or benign tumors, and then I want to send that labeled data through whatever is the typical means of getting that label data to the end engineer who's going to use that labeled data to train a model. What's the full workflow for Label Studio?

**[00:24:08] MM:** The full workflow would be for the data science engineer to create or set up a project. So again, they will use a configuration language. Actually, for the example that you were given, we already have the templates that are pre-configured. They will just select a pre-configured template, they will create the label names, the annotators are going to be using, and they're going to upload the datasets. Inside the Label Studio, we also have connectors to the Cloud Storage. So, in case those scans, they're in the S3 bucket or Azure Blob Storage, we can seamlessly connect to that. At that point, you have the fully configured data labeling project.

So now, you just need to send in the link to that project to your annotators, they create accounts, and when they create accounts, it's a very simple setup for them. They just have a single button called Start Labeling, they click Start Labeling button and they start using the interface to put the bounding boxes over the regions that they think might be considered as problematic and be a signal to a tumor.

After they're done with that process, we the labeling process, potentially, you want to have a verification step. So, you want somebody else and other professional to get into Legal Studio and go through all of their annotations or part of their annotations, and do the verification step. If you're happy with the verification results, now we can get the output, and we can send that outputs to the training phase. So, the training phase actually happens outside of the Label Studio. But Label Studio can trigger that training phase based on the different parameters, for example. Okay, we have annotated and verified 50 images. Now, it's a good time to retrain our model. So, the time when you retrain is actually specific to the business logic of your application that you're running, and so it can be different depending on the type of product that you're building.

**[00:26:13] JM:** Do you have a lot of tools for managing the annotators? Finding ways to figure out if some of your annotators are not doing their job well? Or perhaps kicking them out if they're doing such a bad job that they need to be kicked out?

**[00:26:34] MM:** Great question. Unfortunately, no. And I think our solution is one of the first one to give that type of reporting and analytics to the end user. The reason for that is really historical, because for quite a long time, all the data annotation work has been performed outside of the like offshore, where you would take your dataset and send it somewhere to the regions in the world with the the cheapest labor possible, and get back the results. Because of those companies, they would not want you to have too much understanding of the quality or who is doing what and how well they're doing. And you as a company, as a customer of such company, you will not sometimes even carry yourself, because you just need labelled data. And you may not be paying at the first steps, not paying too much attention into the quality of that.

But then what turns out that, in a lot of the cases, you can get the results and the quality are not going to be really high. So, what are we going to do now? I think now, what we are seeing is really the transition for the companies to be paying more and more attention, and have a requirement for this type of functionality within the software that they're using, because they really want to understand if somebody's not doing really good work. Okay, we don't want to be paying that person. Why would you pay them? And we let transition, as I mentioned, we are building our solution around people. So, we want to give as much on these type of reports to the end user as possible, so they can make a more informed decision around their data labeling process.

**[00:28:26] JM:** Taking a step back, we've now had probably, call it five years of development of really high-quality machine learning tooling. How have machine learning workflows changed with the avalanche of new tools that are available?

**[00:28:44] MM:** So, I think really, if you zoom out and you look at the machine learning pipeline, I would say that it consists of four major parts. The infrastructure and hardware, the algorithms, the datasets, and the people. I think over the years, what we have seen is companies and the industry as a whole were invested into the infrastructure and algorithms, and we had – first, we have some advancements into the hardware, then algorithms, then hardware, then algorithms. And where we stand right now it seems that the infrastructure and the algorithms are becoming more of a commodity. And the reason for that is really, because if you look at the research community that is developing the algorithms, its nature is to publish that into the open. And what that means basically, most of the companies they are using the same algorithms that are available for free out there, we all run on the same frameworks more or less. We all use TensorFlow, we all use Python. We all host on AWS. We all host on Azure or GCP or whatever your cloud render might be.

But what is not that easy to commoditize is the actual data sets that are specific to the company and the people that are employed by this company. And because of that, I think the change that we're seeing is really around the software that's been built, that utilizes the datasets and the people, and helps companies that are building their ML pipelines to make them as competitive as possible with these two pieces. That’s, again, the whole idea, I think, or at least the way I see that, behind the datacentric AI, where solutions like RS or numerous others, helping companies to leverage their things that are unique to their companies, to be in the datasets and the people to build the most competitive machine learning models.

**[00:30:50] JM:** Has dataset gathering gotten more sophisticated? Are there better ways of gathering datasets?

**[00:30:57] MM:** Oh, for sure. I think, again, with the frameworks that we have right now, the idea previously, at least what I have seen was that let's capture as much data as possible and figure out later, what are we going to do with that? Right now, it's way more sophisticated with the whole pipeline. You have all sorts of verifications, checks, clue and procedures along the way, before you actually want to store the data that you're getting. So, I think if you look at all the steps over the pipeline or the typical machine learning pipeline, on every step, there have been some advancements in terms of what type of software and what type of frameworks we use to go through that step.

**[00:31:44] JM:** To me, it seems like machine learning is still fairly unapproachable, unless you have some domain expertise. If we wanted to make some kind of tool to take in podcast interviews, and be able to take a podcast interview and automatically label the questions and answers based off of a thousand labeled interviews, that's still a non-trivial exercise. Have you seen any work around people building tools to make machine learning more approachable? Have you seen this actually be feasible? Or is it just going to be something for the realm of people who can actually engineer a classifier?

**[00:32:35] MM:** Yeah, I think we're – again, it's the industry itself, it's still pretty new. And for these types of applications that you're thinking about to be built, you really have to first build what I would call like more of a foundational layers, meaning that you need to have a good infrastructure in place. It's usually, there is a change in the direction between the infrastructure of applications and going back to the infrastructure. To answer your question, I have not really seen much that I think can work really well. Going to your example, with a podcast, I think it can work even we may be able to find some solutions that can process majority of the podcasts that you have. But my assumption would be that as soon as you get somebody with a little bit stronger accent, then you'll have a trouble processing that. I think we'll – over the next year or two years, we will see more solutions that are focused on what I would call a low code or no code in the MLAI domain.

**[00:33:51] JM:** Let's come back to engineering Heartex. So, as far as I can tell, this is an electron app.

**[00:33:58] MM:** No. It’s React. So, it runs in the browser. It's like a typical – we have a Django backend and then it sources the React app, and then you open it up in the browser.

**[00:34:08] JM:** Gotcha. And then do you just give it a URL of an S3 bucket, and then it grabs the data from the S3 bucket and people label it from there?

**[00:34:19] MM:** Yeah, more or less. You give the location of it wherever these stores will be support multiple integrations for that. You tell the app like, or you label names, and from there, yeah, you can start the labeling work.

**[00:34:33] JM:** Gotcha. And so, the after labeling, the labels are output and what the label file gets put into some output source or output destination?

**[00:34:45] MM:** Yeah, you can either download that to your hard drive or you can put that back to S3 buckets. From there, you can send it to the model retraining phase or model training phase.

**[00:34:58] JM:** Gotcha. What are you working on right now on the platform?

**[00:35:01] MM:** We are working on a lot of the features connected to the automation, because you can think about the data label, and especially when you have large datasets, it can be very complicated to get enough resources to process a large data set. So, any sort of automation that helps the end user to process faster or to athletically process the items is of great value. Majority of our work is focused on making data labeling as automated as possible.

**[00:35:31] JM:** Is that mostly around building tools for the annotators?

**[00:35:36] MM:** Actually, both. So, there are different categories for that. You can either try to save time on the annotator label, and at that point, you would, what we call a pre-label the dataset, pre-label the item. So, when they're looking at the item, let's say you're putting the bone again, going into the example you're putting the bounding box on the cars, maybe you can use the model that is already out there to do that. And then the job of the annotator is to only confirm what the model is showing to you, or just that prediction a little bit and then that can provide you with some savings.

And then on another level, you can figure out how to preprocess the dataset, so that the annotators, they only need to label the items that we would say that they're most beneficial to the model. So, just to give you a visual again, let's say that are your labeling images, or you're classifying images of cars, and you have thousands of images of Mercedes, you have thousands of images of BMW, et cetera, et cetera, maybe you don't want to label each thousand, maybe you want to label 3, 5, 10 images of each individual car model, and then propagate those labels to all the other images that are very similar to the one that you have labeled.

So, that's another one where you can basically figure out how to do what is called an active learning, where you're always focusing on the items that would provide the most information to the model to be able to label the items for you in a way. And then there is even, I would say, the third a year is when you create some rules for labeling, a simple mental model that you can build is, think about regular expressions and product margin. For example, if we are labeling tweets for the sentiments, we can say that, in case you have a word happy inside the tweet, then we want to apply the sentiment positive. So, you can create a library of those rules and you can apply those rules to every item inside your data set. It becomes more complicated, of course, when you start thinking about domains other than the text, like images. But there is also some research on how that can be applied to the different data types. It's all around the automation in the end.

**[00:38:07] JM:** What are the outstanding bottlenecks in the automation of machine learning processes?

**[00:38:12] MM:** I think you have to split it up into different – at least how we split it up, we split it up into two different categories. The one that we call common knowledge problems, and another one that we call subject matter knowledge. With the common knowledge, you can think about the type of knowledge that you need to have as an annotator to be able to label certain items. And you want those to be, this type of knowledge to be available anywhere in the world more or less. So, you can think about labeling cars, right? Everybody, despite their geography can be labeling cars. We know what a car is, we know what a street sign is, you know what a tree, what a bird is.

So, I think for these types of problems are the common knowledge problems, we'll see a greater, greater level of automation available. They're always going to be the edge cases. And you'll always need to have people in the loop to be resolving the edge cases. But for the most part, it’s going to be automated. But then for the subject matter knowledge, the one that you were talking about, for example, the doctors and scans, or even your podcast can be the type of the main knowledge, if you're labeling that yourself because nobody other than you probably – and your friends can label your podcasts really well. For these type of knowledge problems, I think it's going to take a while for us to build the automation tools. I think there is a lot of relaying into this approach with building the library of the functions or operations that you can use to label certain items. But I think isn't going to be an open problem for quite a while.

**[00:39:50] JM:** Just to close off, do you have any predictions about how machine learning workflows will be different in 5 years, 10 years?

**[00:39:57] MM:** Yeah, I think I might be biased here but I think it's all going to be around the datasets and the people who process the datasets. I think a lot of that – because the workflows, the pipelines, they're going to get advertised, they're going to get commoditized, and then it's all going to boil down to the people sitting in process and the datasets and pitching, really pitching the model on what type of prediction it should be doing. So, from that perspective, I really see that as just another way of doing the computer programming, where you teach the model by showing them in a way an example of how you as a person would process certain item, and machine learning the pipeline, the infrastructure basically helps scale that way of processing the data.

**[00:40:42] JM:** Actually, lastly, any newer tools that you've seen recently that have impressed you in the machine learning space?

**[00:40:555] MM:** I think from the tooling, I would not be – it's all relevant on the timescale. I think, nothing that I have seen for over the last quarter that I was super impressed with. I may not be also following some of the new releases as closely as they used to. So, I'm more impressed with some of the worries that are some of the existing tools that we already know about are doing. I can mention one thing well for the deployment of ML models, ways and biases for the experiment track and **[inaudible 00:41:31]** phase for their model registry right now. This are the ones that continue to impress me just how much they're creating and most of them also open source.

**[00:41:42] JM:** Well, Michael, thank you so much for coming on the show, and anybody who is looking for a data annotation platform should check out Label Studio.

**[00:41:51] MM:** Thank you.

[END]