**CRAIG:** Hi, I'm Craig Smith, and this is Eye on AI.

**CRAIG:** This week. I speak with Bratin Saha, the head of Amazon's machine learning services. We talked about Amazon's growing dominance in model building and deploying AI, about the company's SageMaker platform, and whether anyone can compete with the behemoth.

**CRAIG:** Before we begin, I want to thank our sponsor ClearML, the MLops solution. You can check them out at clear.ml.

**CRAIG:** In the meantime, I hope you find my conversation with Bratin as interesting as I did.

**CRAIG:** Maybe start by describing your background, when you came to Amazon and what you're doing at Amazon now.

**BRATIN:** Okay. At Amazon, I now lead all of our AI and machine learning services. One of which is SageMaker, but we have the broadest and deepest set of capabilities. And then prior to that, I was at Nvidia, VP of software infrastructure there.

**BRATIN:** I did my PhD at Yale university and then also went to Harvard Business School. I've been in research, been on the product side, been on the business side, and now at Amazon am leading all of the AI and machine learning services.

**CRAIG:** And the services are platforms like SageMaker with the various tools.

**BRATIN:** We think of our customer persona in three broad categories. So, there's one set of customers who say give us just the optimized infrastructure and we will build a machine learning infrastructure and we'll build the machine learning models and we'll deploy it ourselves.

**BRATIN:** So, for them, we have what we call ML engines. These are deep learning containers, deep learning frameworks, like the core software components. We optimize them for AWS infrastructure, and then we provide them to customers to use.

**BRATIN:** Then there's the next set of customers, which is where most customers are today, who say building the ML infrastructure is undifferentiated heavy lifting. We would rather have you build infrastructure so we can focus on just doing the machine learning models. These are data science teams. And so that is where most of our customers are. And that is what SageMaker is.

**BRATIN:** It gives you an end-to-end platform, the ML infrastructure, so that you can build train and deploy machine learning models.

**BRATIN:** And then we have customers who we give the pre-train machine learning models as well. So, you can just use APIs. You can infuse intelligence into every app that you're building by just making calls to these functions, doing things like natural language processing, doing things like document processing, doing things like computer vision, image recognition kind of stuff.

**BRATIN:** And we have solutions as well. So, if you look at Monitron, which is one of our solutions for preventive maintenance that uses machine learning to predict when your equipment may be due for maintenance or may actually have downtime. We have things like Panorama, which are solutions using computer vision for things like quality inspection and so on.

**BRATIN:** So, we think of them as three broad tiers. One that people are building their own machine learning infrastructure and their own models. The second is we build infrastructure and they build the models, and the third is we build the models as well. They just invoke it.

**CRAIG:** And they take the models and train it on their dataset.

**BRATIN:** They take the models and train. So, for example, you can have pre-train models, or you can start with some of the models we provide. This is the SageMaker layer, and then they train it using the dataset. So, you either build models from scratch, or you start with some models that we provide, or you get models from the open source and then train it on the dataset.

**CRAIG:** And I've heard from a lot of people that one of the attractions of working with Amazon SageMaker and tools like that is if you're on the Amazon cloud, you're within that ecosystem. Is that important? That kind of interoperability that happens within an ecosystem. Or can you be building on TensorFlow and using TPUs?

**BRATIN:** We have customers today who do all of the machine learning on SageMaker, but we also have customers who do part of it and then do the remaining thing somewhere else. So, for example, you could build your machine learning models on prem and then deploy it on the cloud.

**BRATIN:** You could, on the other hand, train your models in the cloud and then deploy them on-prem. So, we have customers of all kinds and it's important for us to enable that, so that customers have the choice. And our container formats are open source. So, you know exactly how you need to build your stuff. So, customers can do part of the machine learning in the cloud, on AWS, on SageMaker and then do the remaining thing on-prem or somewhere.

**BRATIN:** They have the full freedom to do that.

**CRAIG:** What is the differentiation between the Amazon ecosystem and say the Google ecosystem or the Microsoft ecosystem where you have, in each case you have frameworks.

**CRAIG:** And platforms and cloud integrations and all of that. It's hard from outside to understand whether one is pointed more towards industry. One is pointed more toward research, or whether they're fairly interchangeable at this point.

**BRATIN:** The vast majority of machine learning in the cloud happens on AWS, the vast majority. And so, the next question is why does the vast majority happen on AWS? And that is because of a variety of reasons one is machine learning does not exist in isolation. Machine learning is built on the compute and storage and data services that we have.

**BRATIN:** And when you look at compute storage and database and analytics services, we have by far the most advanced, most broad set of services including security and all of that. So that is one reason. The second is when you again look at machine learning services, the depth and breadth of our capabilities from model building to natural language processing, to computer vision, to industrial monitoring and all of that.

**BRATIN:** We have the broadest and deepest set of services by far. That enables customers to get their work done. And then the other thing I would also say is if you look at Amazon as a company at our investments in machine learning, we have been investing in machine learning for more than 20 years.

**BRATIN:** In fact, look at Amazon Alexa and Amazon Go, and Amazon.com with recommendations. We have been deploying machine learning at scale for a long time. And so, what we have done at AWS is taken that deep expertise and made it available to our customers. And then when you look at performance, when you look at the performance of open-source frameworks like TensorFlow and PyTorch, these frameworks run fastest on AWS.

**BRATIN:** So if you look at third party sites and third party analyst reports and all that, they will tell you that, in terms of the breadth of features, in terms of the breadth of capabilities, we have a lot more, and that is what attracts a lot more customers. And our machine learning services, AI services, like SageMaker, they're one of the fastest growing services in AWS history.

**BRATIN:** And so, the key factors are the breadth and depth of our machine learning services. Second, machine learning doesn't exist in isolation. It requires this foundation of compute and storage and database and analytics where we have the most capable services. And then the fact that we are talking to the most customers, because most customers use us, allows us to iterate very quickly and get more capabilities out at the speed of innovation.

**BRATIN:** So, I think these are the most important factors. Today more machine learning happens on AWS than anywhere else.

**CRAIG:** That's fascinating. There is an explosion of MLOps startups right now, who are, they all seem to claim to be end to end. But they typically take one part of the MLOps pipeline and focus on that.

**CRAIG:** Is there room for smaller startups to develop platforms and tools when there's something like SageMaker that already exists. One of the things I hear from the MLOps people is yes, SageMaker is there. people use SageMaker because they're on AWS, but SageMaker is a broad-brush tool.

**CRAIG:** It doesn't have a lot of the granularity for specific use cases that our MLOps platform has. Not to get into the he said-she said, but is there some merit to that? Is there room for these new MLOps companies?

**BRATIN:** This is a very broad space right now. That said, if you look at the breadth of customers that are using machine learning on AWS and on SageMaker, as I said, it's one of the fastest growing services in AWS history. We have more than a hundred thousand customers using these services. And they're being used in every domain, in financial services, in healthcare, media entertainment, and so on. Koch Industries, they use our service Monitron that uses machine learning to detect vibrations and predict equipment failures. And they have been able to save a lot of downtime. GE Pacific uses it to improve paper quality. They use it to predict the pace at which the paper rolls should be operating, and they have been able to reduce paper tears by 40 percent.

**BRATIN:** Then you have healthcare companies like AstraZeneca and Phillips that are using our machine learning services. Then you have Intuit that uses our machine learning services for apps like Mint and Turbotax. And so, given the breadth of customers that are using our machine learning services, they're being used for pretty much every use case you can think about.

**BRATIN:** And that is why having an end-to-end platform is so important because you want to have end to end traceability. You want to automate the entire end to end process.

**BRATIN:** So, sports, media, software, financial industry, healthcare, all of them are now doing machine learning on AWS using Sagemaker or AI services.

**CRAIG:** And when you say machine learning, how much of that is deep learning?

**BRATIN:** Pretty good amount of that is deep learning. And the amount of deep learning keeps on increasing. We are seeing customers moving more and more towards using deep learning and more sophisticated models. It has been increasing over the last several years.

**CRAIG:** It's primarily supervised learning.

**BRATIN:** It's, it's a combination of both. We do see unsupervised learning as well, little bit of reinforcement learning, but I think more people use supervised learning today on our services.

**CRAIG:** SageMaker the infrastructure product, it has a data prep end, and it has deployment and monitoring end. I was talking earlier about these MLOps companies. There's one I'm familiar with, Labelbox, and when I've spoken to Labelbox, SageMaker always comes up and their argument is that they focus entirely on the labeling process, the iterative process, the discovery of bias and mitigation and all of that to improve data quality.

**BRATIN:** Any machine learning process starts with data. You can have structured data, you can have unstructured data. And for structured data, we have data Wrangler. And that is, the tabular data that you have in a databases that you want to be able to analyze for machine learning. Then there's unstructured data and unstructured data could be images, could be text, could be audio and videos and so on. That is where data labeling comes in. And so, for that we have Ground Truth and Ground Truth provides a fully turnkey experience. You as a customer, just come in, hand the data, Ground Truth does all of the data labeling and then hands the labeled data back to you. And we often use machine learning models to automate that data labeling.

**BRATIN:** Now, as part of that, we have quality checks to make sure that the labeling has been done well. And then we have tools like, SageMaker Clarify that lets you do bias detection and lets you do bias detection, not just in your data.

**BRATIN:** It lets you do bias detection even in your models. So, we have a very holistic suite between structured data and unstructured data. When unstructured, that is where you're looking at data labeling. We do the data labeling, we do quality checks, then you can do bias analysis and then once your data is prepared, then you send this on for model building and model deployment and so on.

**BRATIN:** And then in model deployment, we have model monitoring. So once your models have been deployed, you do that.

**BRATIN:** Now the benefit of having this end-to-end platform, which is very comprehensive, data labeling platform included as part of this whole thing, is that once you have deployed the model into production, you want to monitor it.

**BRATIN:** And then if there are some mis-predictions, you want to be able to come back and retrain your model.

**BRATIN:** Now, because it's all in a single platform you can look at, where did the model go wrong, pick up those kinds of data, relabel it, and then retrain it. And that is why customers say, want to be in a platform where I can get the data labeling through the model deployment. And in data labeling, we actually have lots of modalities.

**BRATIN:** So, we have audio, we have video, we have text, we have LIDAR. Customers can build their own data labeling workflows. They can choose from three different kinds of workforces. So, we have our own public crowd workforce, then they can bring their own workforce and then we have a number of vendors, so they can contract out to those vendors.

**BRATIN:** So, it's a very comprehensive data labeling solution that includes not just labeling, but also quality checks so people get good, labeled data. And then, of course doing bias analysis and so on with SageMaker Clarify on your data.

**BRATIN:** The other thing I also say is we are seeing customers do a lot more what we would call multi-modal data analysis.

**BRATIN:** For example, in the financial industry, they want to be able to combine both tabular data and text data, for example. Financial industry has been traditionally using tabular data, but now they want to analyze things like SEC forms and so on.

**BRATIN:** And so, this ability to be able to process both tabular data and unstructured data, do data labeling and this other one is something that's very valuable to customers.

**CRAIG:** Is there a cottage industry of people that are experts at SageMaker, who are hired out as consultants, or does Amazon have that service?

**CRAIG:** So that they don't have to start from scratch and do tutorials and figure out how SageMaker works. And if you have a team of consultants that would come in and help.

**BRATIN:** So, we have a variety of resources, we have solution architects that help customers.

**BRATIN:** We have ML solutions lab where we have data scientists work together with customers to bring their models to production. We have a lot of documentation available. We have a number of other customer SIs and GSIs, system integrators, global system integrators that help customers get started.

**BRATIN:** And also, we are constantly focusing on making it easier and easier to get started. So, we recently, launched a part of SageMaker that we call SageMaker Canvas. That's a no code solution for doing machine learning. You don't have to write a single line of code. You bring your data and the system automatically does the data preparation for you, automatically creates models for you, and then, it'll deploy the models for you.

**BRATIN:** So, we've been very pleased to see how many people have started using that. So there are a number of things we are doing on the one hand, making it easier to use our services at the product level, through all of these innovations, then having a lot more documentation and collateral, a lot of tutorials, lot of examples, use cases, then our own solution architects, ML solutions labs.

**BRATIN:** We have the machine learning university where all of the machine learning courses that we use for training Amazon engineers, are now available for free to everyone. Then we partner with online education sites, like Coursera where we have the practical applications of data science course with Andrew Ng. We use that to help customers get started on SageMaker. So, there's multiple things that we are doing, from the product education, collateral, ML solutions lab, solution architects to help customers get started.

**CRAIG:** In your job, you described the three layers. Where do you spend most of your time? Where is the most innovation happening? Is it at the infrastructure level?

**BRATIN:** We have a lot of innovation happening at all three levels and a lot of growth happening across all of these services. The AI services are being used by customers in very interesting ways.

**BRATIN:** So there, you don't need a lot of machine learning expertise to get started. So, these are more like solutions. So, we have customers like Anthem, who have been able to automate almost 80% of the claims insurance process. We have customers like Discovery who are using our Amazon Personalize to provide a much more personalized experience. What kind of shows you would like? We have Koch Industries using Monitron for industrial maintenance, then we have Intuit, that's using Transcribe for call center intelligence. We are seeing broad growth across all the three layers. Now SageMaker is one of the fastest growing services in the history.

**BRATIN:** That's what most of customers have used, but we are seeing growth in all three areas and very robust growth in all three layers.

**CRAIG:** Okay, my last question is on the models and the APIs available. You have computer vision and natural language processing and things like that. What's the newest model that you're making available.

**BRATIN:** We are adding a lot of machine learning for DevOps, if you're running cloud services automatically detecting anomalies. So that is new, DevOps Guru. We have services for forecasting, so we have a very broad collection, and we are constantly innovating. We are looking for what customers want looking for what customers need and just constantly iterating on those.

**CRAIG:** That's it for this episode, I want to thank Bratin for his time and invite you to read a transcript of the conversation on our website, eye-on.ai. I want also to thank our sponsor ClearML, the MLOps solution that you can try out for free at clear.ml.

And remember, the singularity may not be near, but AI is about to change your world.

**CRAIG:** So, pay attention.