**SECON 2023**

**Title**: Introduction to MLOps with AWS SageMaker, running your first LLM

**Abstract**: The increasing demand for large language models (LLMs) has led to a surge in the development of machine learning operations (MLOps) solutions. However, deploying and scaling these models remains a challenge for many organizations. This presentation will demonstrate how to efficiently run a large language model using SageMaker.

SageMaker is a managed service offered by Amazon Web Services (AWS) that enables developers to build, train, and deploy machine learning models at scale. In this presentation, we will discuss how to use SageMaker to deploy a pre-trained large language model, specifically focusing on the challenges associated with deploying LLMs, such as GPU resource allocation and hyperparameter tuning.

Additionally, we will discuss the benefits of using SageMaker, including its seamless integration with AWS services like Lambda and EC2, and how it can reduce the time and cost associated with training and deploying LLMs. We will also discuss best practices for monitoring and debugging the model using SageMaker's built-in tools and how to optimize the model's performance.

Overall, this presentation will provide insights into how MLOps engineers can use Sagemaker to efficiently run large language models, enabling organizations to scale their machine learning capabilities and achieve their business goals.