**SECON 2023 / KUBECON 2023**

**Title:** Introduction to MLOps: Running Your First LLM

**Abstract:** As the field of machine learning continues to evolve, organizations are increasingly recognizing the need to develop robust practices for managing their specialized and complex machine learning workflows.

In this talk, the speakers will demonstrate how MLOps practices can be leveraged to deploy large language models in AWS SageMaker. They will showcase the use of popular open-source tools such as the Hugging Face Transformers library and the Deeplake vector database, which can be integrated to build an end-to-end inference pipeline that can be used for various machine learning use cases.

By the end of this demo, attendees will have a better understanding of the key components of an MLOps workflow. This talk promises to be informative for both seasoned machine learning practitioners and beginners, providing valuable insights and cutting-edge techniques. Attendees will discover the power of MLOps in developing and deploying LLMs.

**Benefits of the ecosystem**: The rapid evolution of LLM and OpenAI's swift ascent to the top of the industry may create the impression that they are the only viable option. However, this couldn't be further from the truth. We believe that now is the ideal time to democratize this knowledge and make it as accessible as any other cloud resource. This talk aims to simplify what OpenAI has achieved by utilizing a combination of open-source technologies and cloud resources, preventing lock-in at an early stage.

**Biography**: Mariano is a computer geek based in Chicago, originally from Mexico. Over the past few years, he has focused on deploying machine learning models at scale, utilizing tools such as SageMaker, ArgoCD, and PyTorch/Tensorflow, to name a few. With over 15 years of programming experience, he has implemented a variety of applications across various fields, including insurance, IoT, and most recently FinTech. He finds joy in sharing his programming language knowledge with the Chicago tech community.