A dark, industrial scene featuring a robotic arm in the foreground, positioned over a workbench. The arm is actively welding or grinding metal components, resulting in a bright, intense burst of sparks and light emanating from the point of contact. The background is slightly blurred, showing other industrial structures and equipment, creating a sense of a busy manufacturing environment. The overall lighting is dim, with the primary light source being the sparks from the welding process.

Self-Hosted Open Source LLMs: Best Practices and Strategies

JOSHUA ARVIN LAT & SOPHIE SOLIVEN



Chief Technology Officer of
NuWorks Interactive Labs



AWS Machine Learning Hero



Orange Boomerang: Digital Leader of the Year 2023 Award Winner 🏆



Author of 📖
Machine Learning with Amazon SageMaker Cookbook



Author of 📖
Machine Learning Engineering on AWS

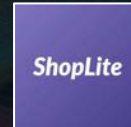
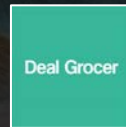


Author of 📖
Building and Automating Penetration Testing Labs in the Cloud



Operations Director
Edamama

PREVIOUSLY



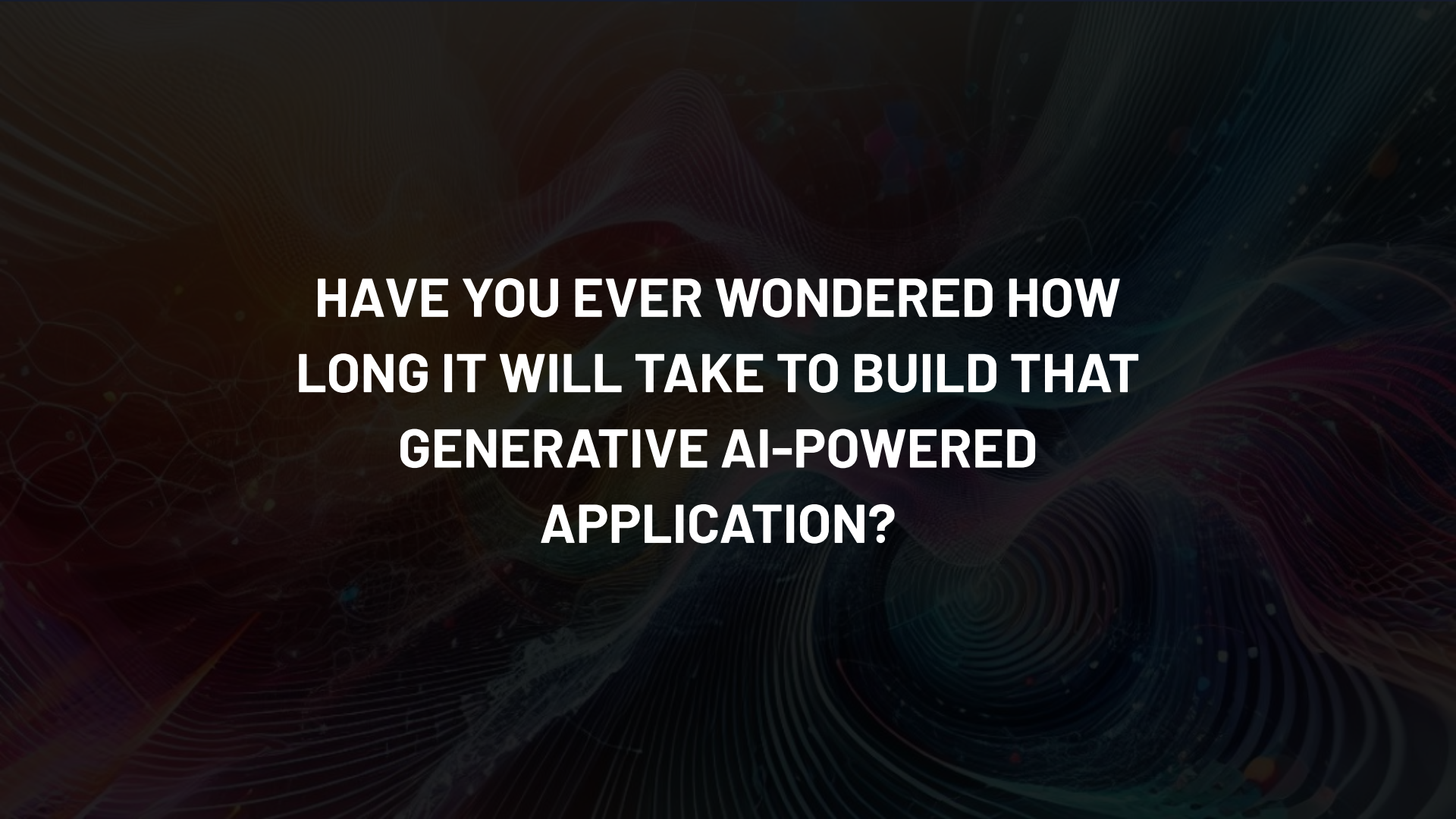
**General Manager of E-Commerce
Services and Dropship**



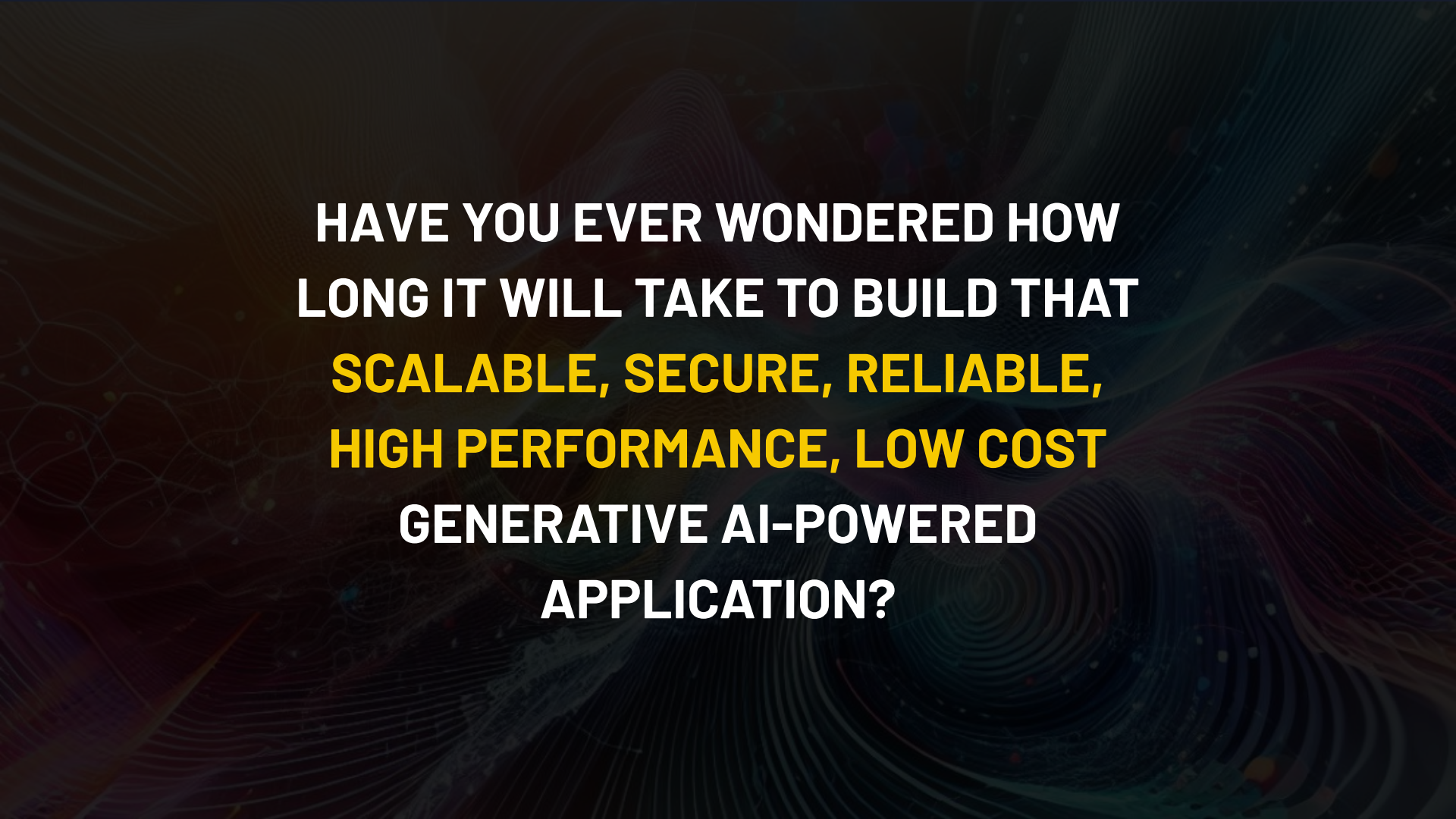
Certifications in:
Cloud computing and data analytics



Technical Reviewer of the book:
Machine Learning Engineering on AWS



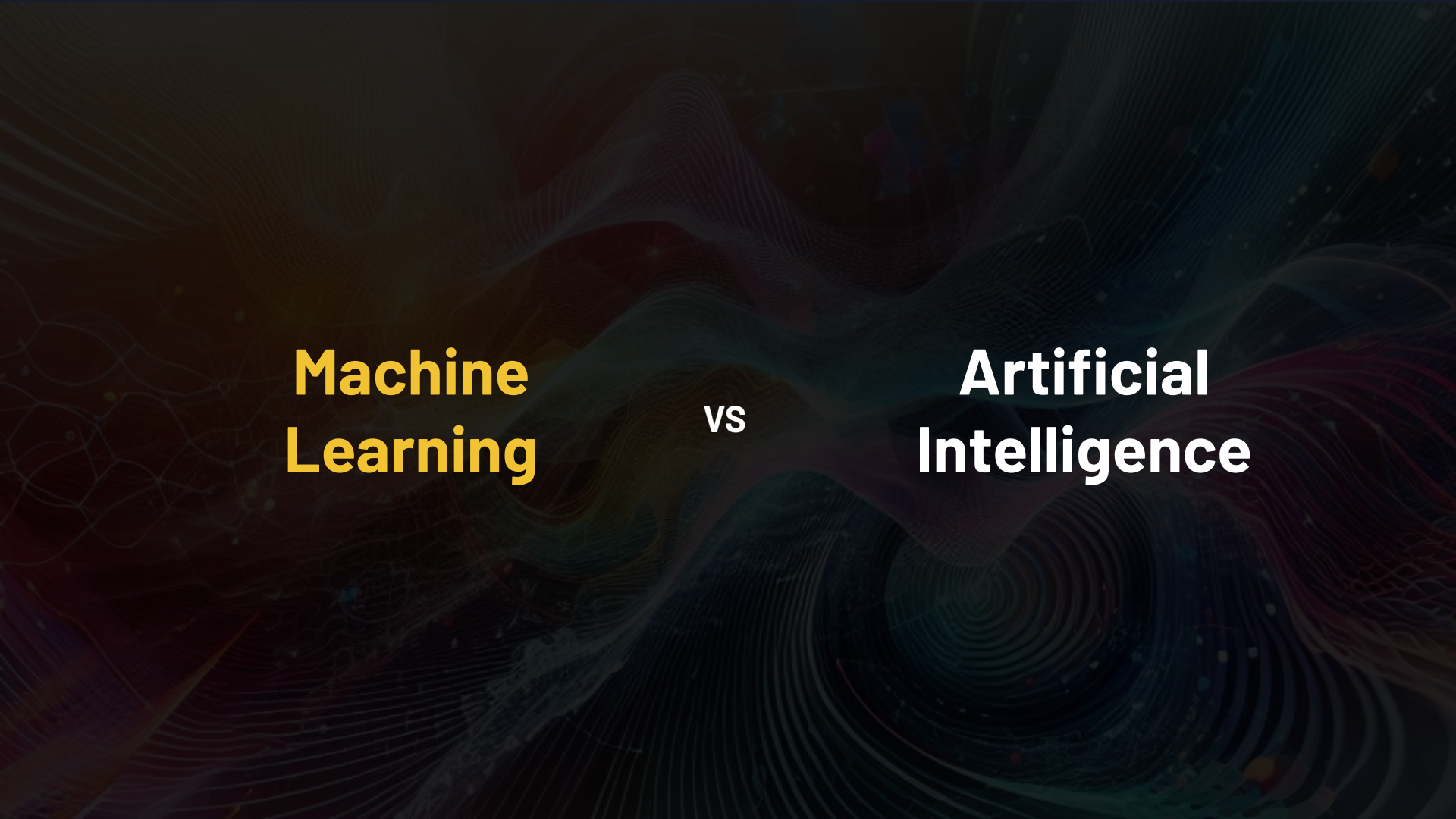
**HAVE YOU EVER WONDERED HOW
LONG IT WILL TAKE TO BUILD THAT
GENERATIVE AI-POWERED
APPLICATION?**



**HAVE YOU EVER WONDERED HOW
LONG IT WILL TAKE TO BUILD THAT
SCALABLE, SECURE, RELIABLE,
HIGH PERFORMANCE, LOW COST
GENERATIVE AI-POWERED
APPLICATION?**



CONCEPTS

The background is a dark, abstract composition. It features a network of thin, light-colored lines on the left side, resembling a molecular or neural structure. The rest of the background is filled with fluid, wavy patterns in shades of blue, purple, and brown, creating a sense of depth and movement.

**Machine
Learning**

vs

**Artificial
Intelligence**

**Machine
Learning**



**Artificial
Intelligence**



**Artificial
Intelligence**

A Venn diagram illustrating the relationship between Artificial Intelligence and Machine Learning. It consists of two concentric circles. The outer circle is defined by a white border and contains the text 'Artificial Intelligence'. The inner circle is defined by a yellow border and contains the text 'Machine Learning'. The inner circle is entirely contained within the outer circle, indicating that Machine Learning is a subset of Artificial Intelligence.

**Machine
Learning**

ARTIFICIAL INTELLIGENCE

ANI
NARROW

AGI
GENERAL

ASI
SUPER

MACHINE LEARNING

Supervised Learning

Unsupervised Learning

Reinforcement Learning

DEEP LEARNING

GENERATIVE AI

LLM
TEXT

IMAGES

AUDIO



SELF-HOSTED OPEN SOURCE LLMS

YEARS AGO...

**BUILD EVERYTHING
FROM SCRATCH**

VS



PRESENT

**HAVE A SELF-HOSTED
OPEN-SOURCE LLM
SETUP**

VS

**USE A PAID
GENERATIVE AI API**

The background is a dark, abstract composition. It features a network of thin, light-colored lines on the left side, resembling a molecular or web structure. The rest of the image is filled with large, flowing, wavy shapes in shades of blue, purple, and teal, creating a sense of movement and depth. The overall tone is mysterious and technological.

BARRIER TO ENTRY

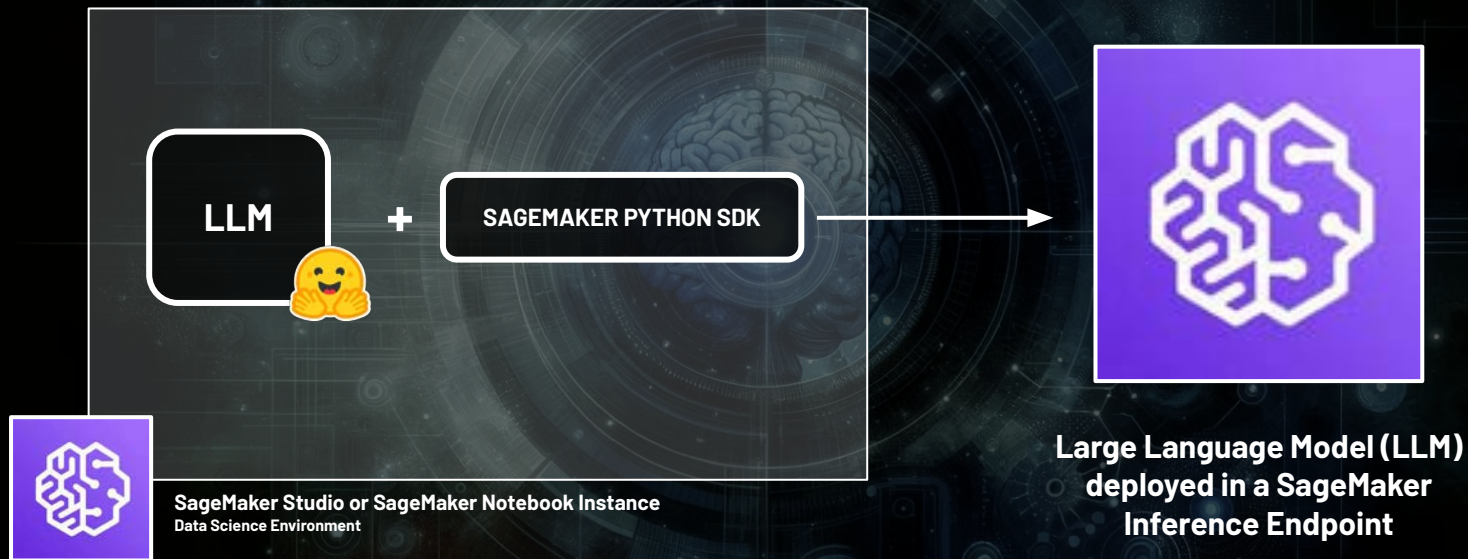
The background is a dark, abstract composition. It features a network of thin, light-colored lines on the left side, resembling a molecular or web structure. The rest of the background is filled with large, flowing, wavy shapes in shades of blue, purple, and teal, creating a sense of movement and depth. The overall tone is futuristic and technological.

INFRASTRUCTURE COST

The background is a dark, abstract composition. It features a complex pattern of thin, wavy lines in shades of blue, purple, and green, creating a sense of depth and movement. In the lower-left corner, there is a faint, light-colored hexagonal grid pattern. The overall aesthetic is modern and technological.

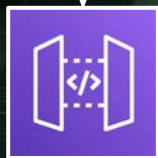
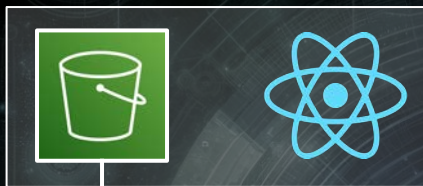
FLEXIBILITY AND LEVEL OF CONTROL

SELF-HOSTED LLM SETUP



SELF-HOSTED LLM SETUP

S3 Static Website Hosting (Frontend)



API Gateway



AWS Lambda



Large Language Model (LLM)
deployed in a SageMaker
Inference Endpoint

DATABASE

MODEL MONITORING

DEBUGGER

SETUP

SECURITY

VERSIONING

**DEPLOYMENT
OPTIONS**

AUTOMATION

A close-up, low-angle shot of a large industrial robotic arm in a factory setting. The arm is dark grey and metallic, with various joints and cables visible. It is positioned over a workbench where it is welding a piece of metal. A bright, intense light emanates from the point of contact, with a large spray of orange and yellow sparks flying outwards. The background is slightly blurred, showing other industrial equipment and a glass-enclosed area. The overall lighting is dim, with the primary light source being the welding process.

THE END