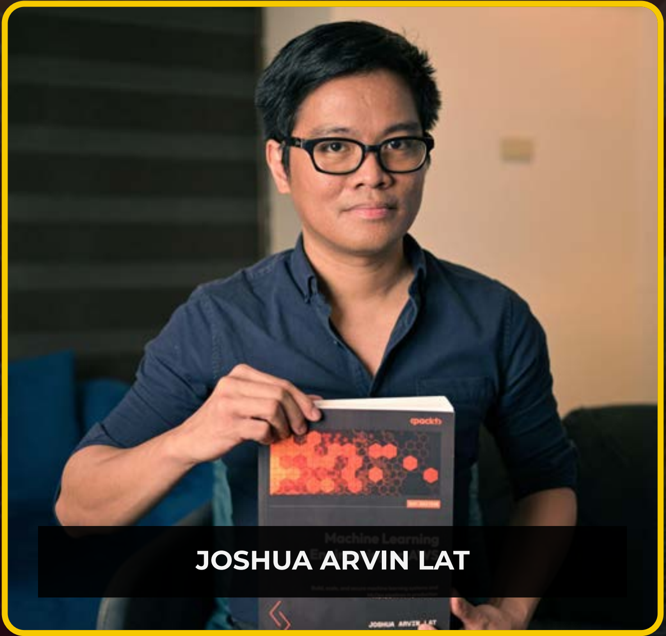
Self-Hosted Open Source LLMs: Best Practices and Strategies

JOSHUA ARVIN LAT & SOPHIE SOLIVEN



JOSHUA ARVIN LAT

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

SOPHIE SOLIVEN

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

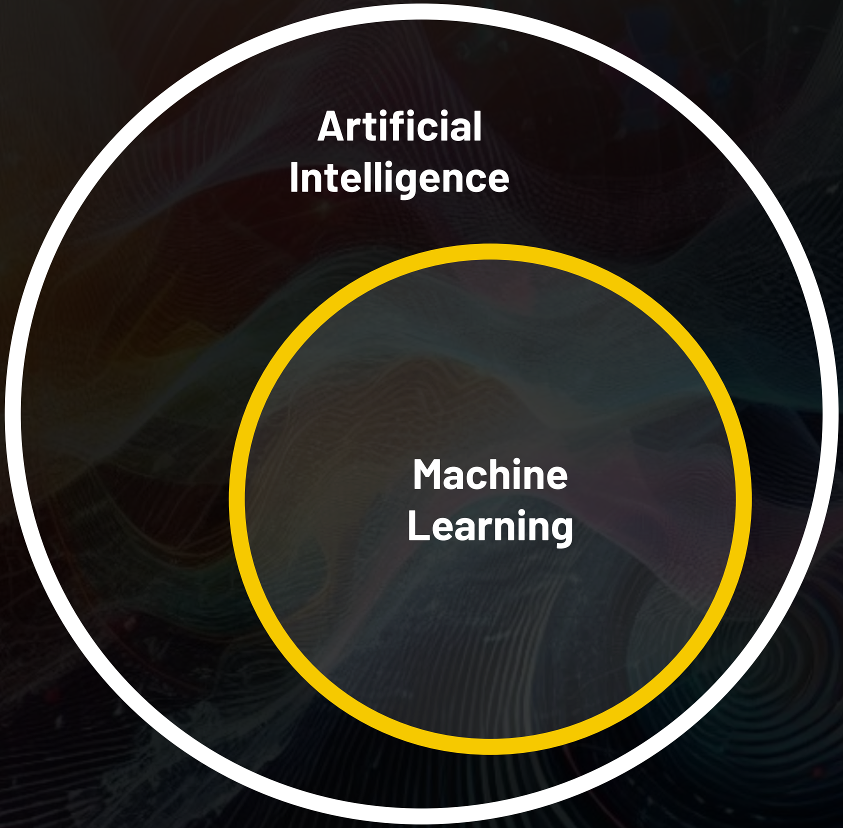
HAVE YOU EVER WONDERED HOW LONG IT WILL TAKE TO BUILD THAT GENERATIVE AI-POWERED APPLICATION?

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

CONCEPTS

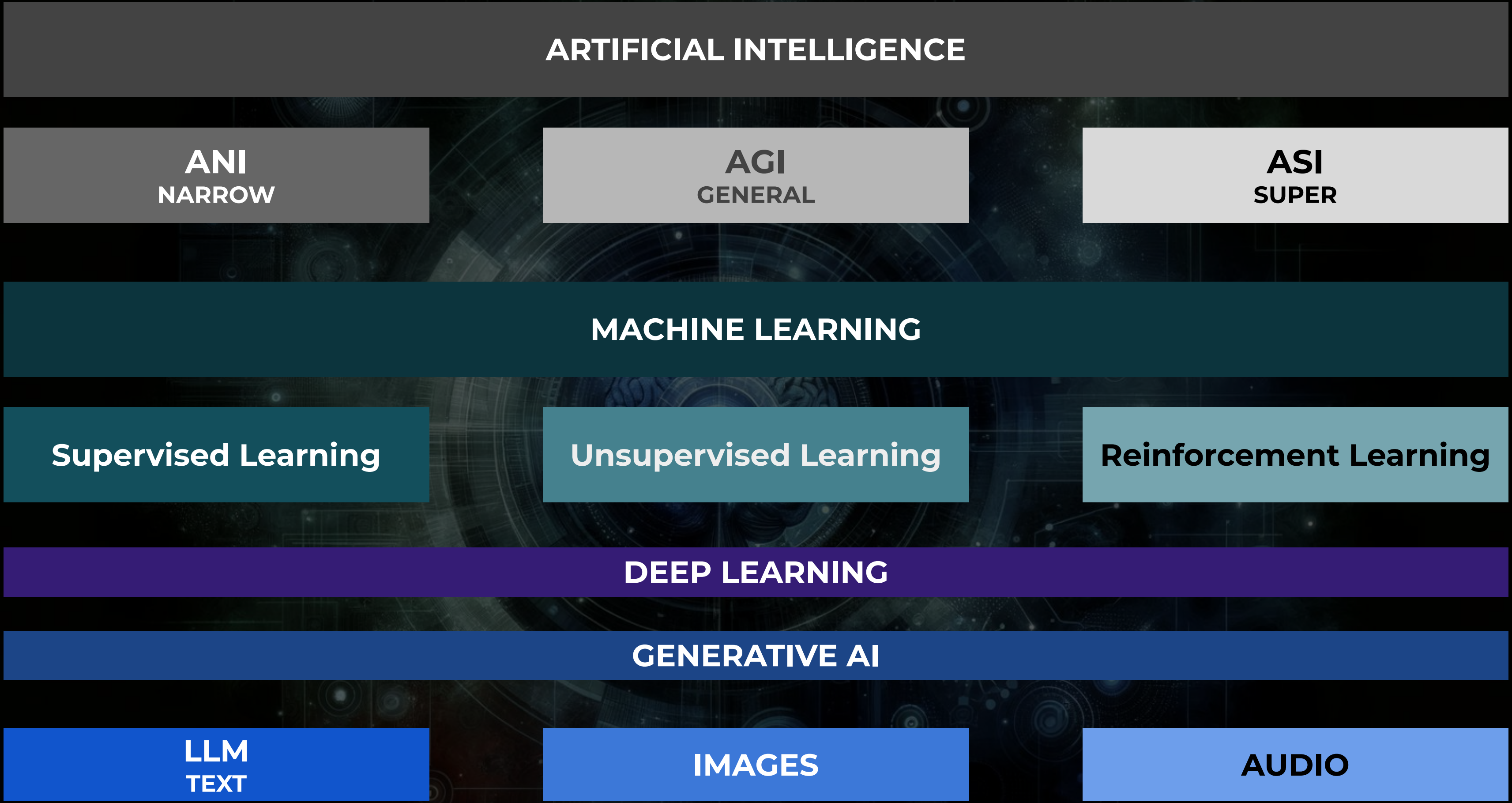
Machine Learning vs Artificial Intelligence

Machine Learning $\subset$ Artificial Intelligence



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

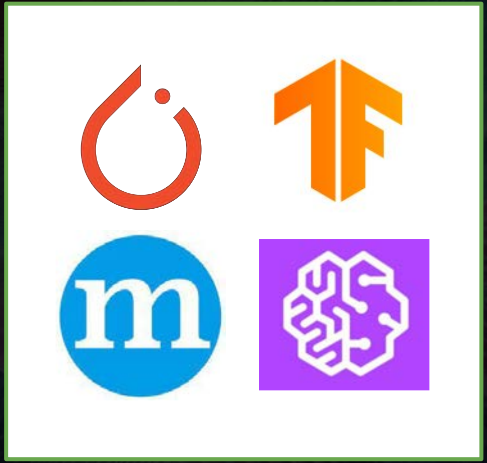
AUDIO

SELF-HOSTED OPEN SOURCE LLMS

YEARS AGO...

BUILD EVERYTHING FROM SCRATCH

VS



PRESENT

HAVE A SELF-HOSTED OPEN-SOURCE LLM SETUP

VS

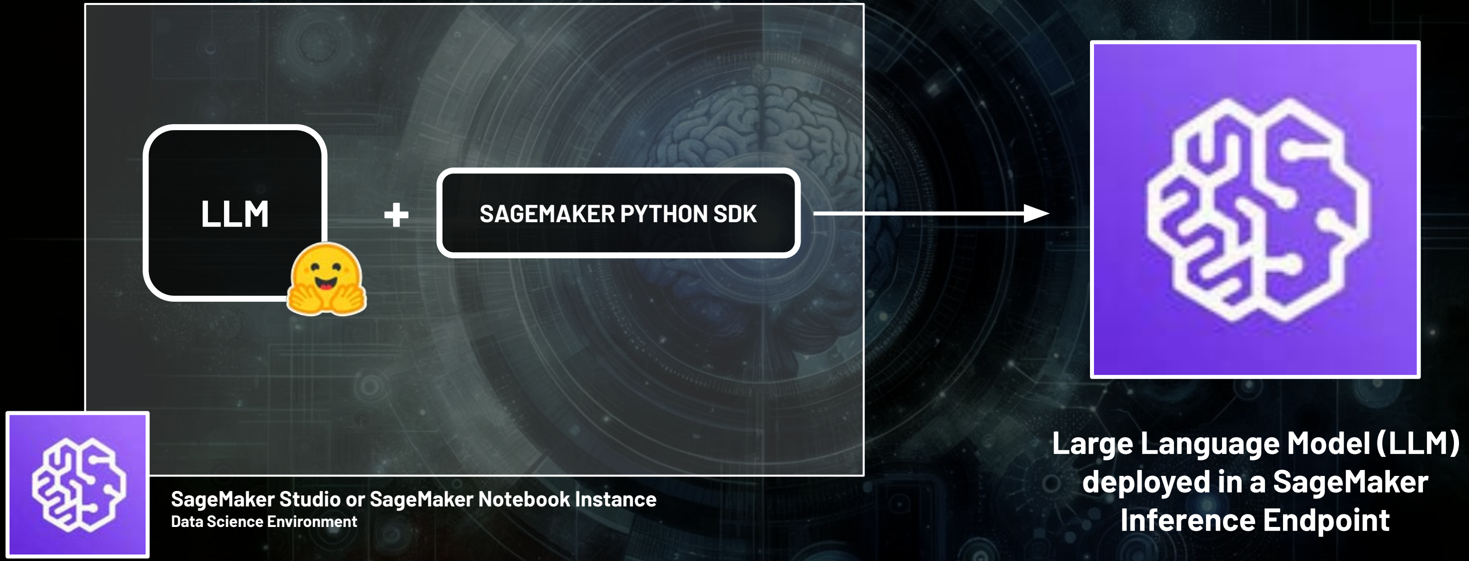
USE A PAID GENERATIVE AI API

BARRIER TO ENTRY

INFRASTRUCTURE COST

FLEXIBILITY AND LEVEL OF CONTROL

SELF-HOSTED LLM SETUP



LLM

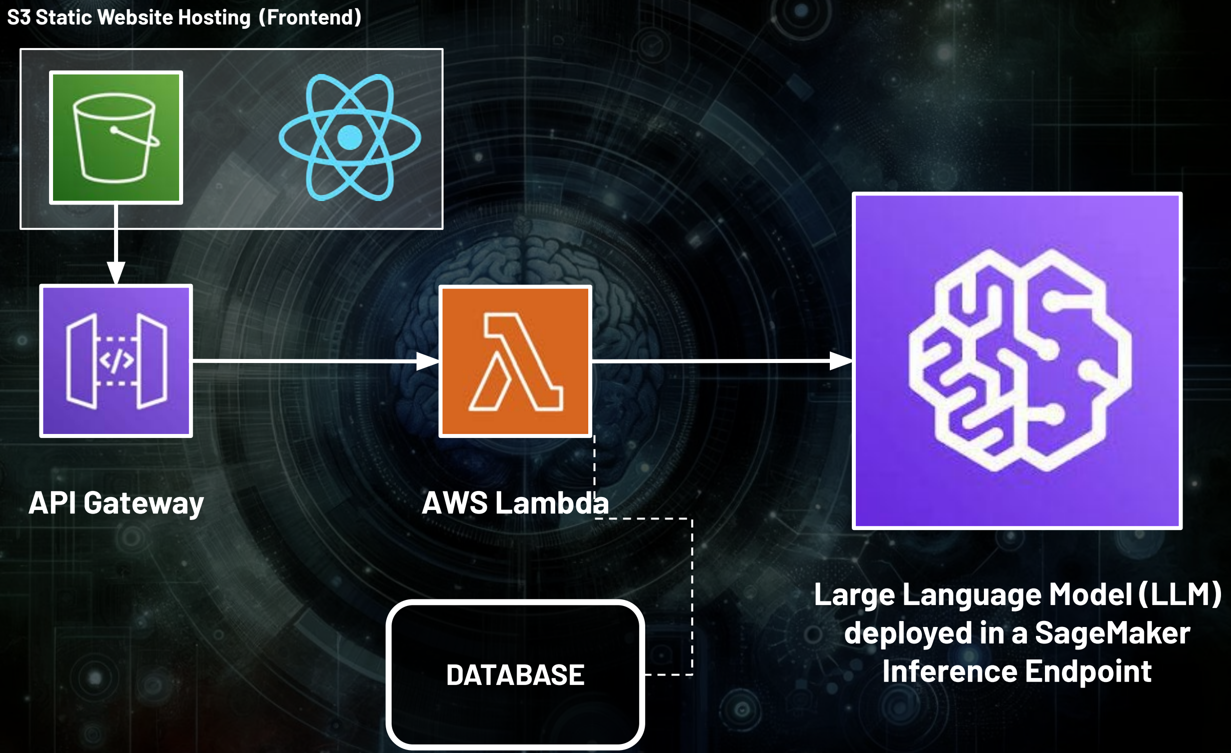
+

SAGEMAKER PYTHON SDK

SageMaker Studio or SageMaker Notebook Instance
Data Science Environment

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

SELF-HOSTED LLM SETUP



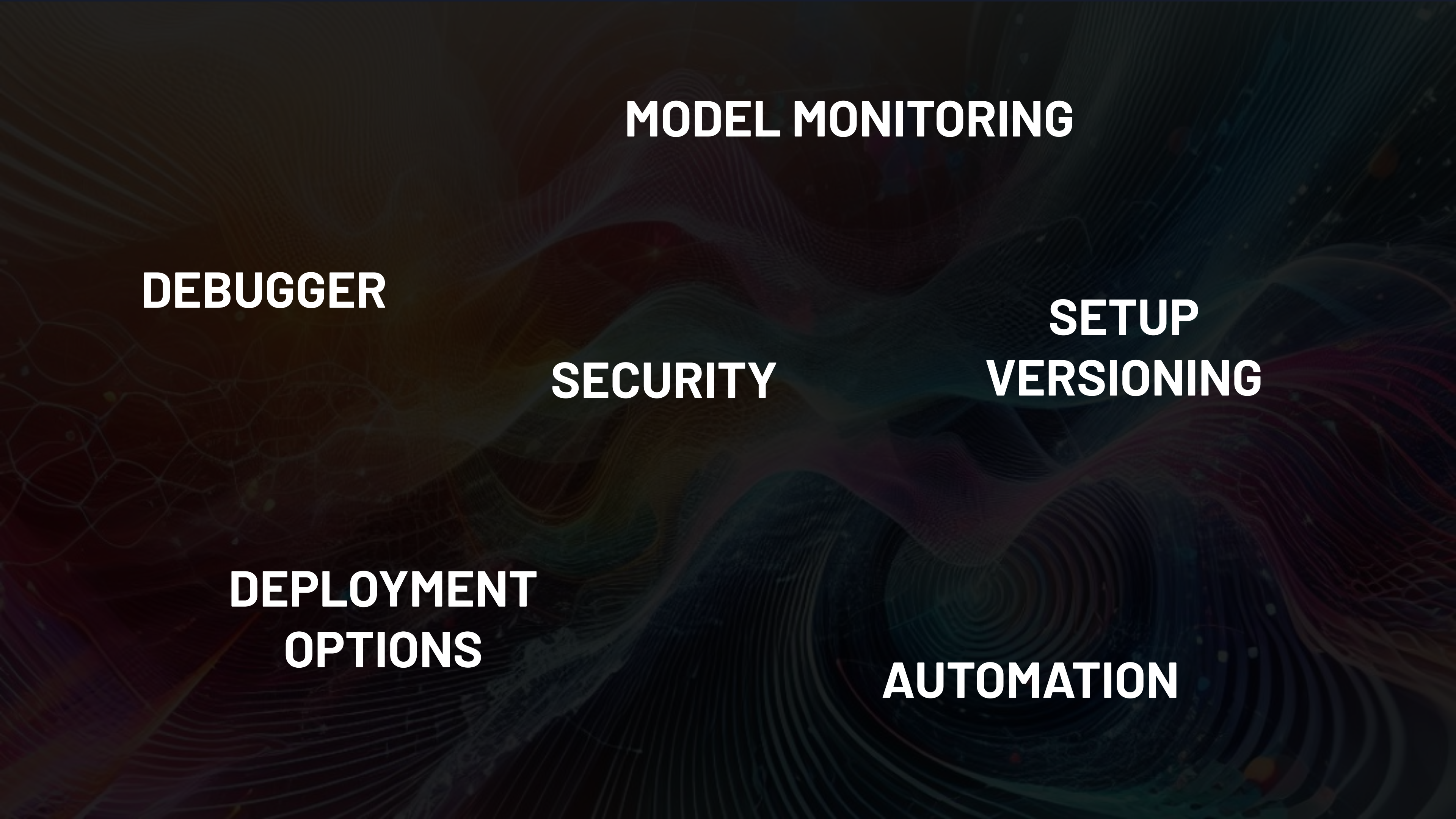
S3 Static Website Hosting (Frontend)

API Gateway

AWS Lambda

DATABASE

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



MODEL MONITORING

DEBUGGER

SETUP
VERSIONING

SECURITY

DEPLOYMENT
OPTIONS

AUTOMATION

THE END