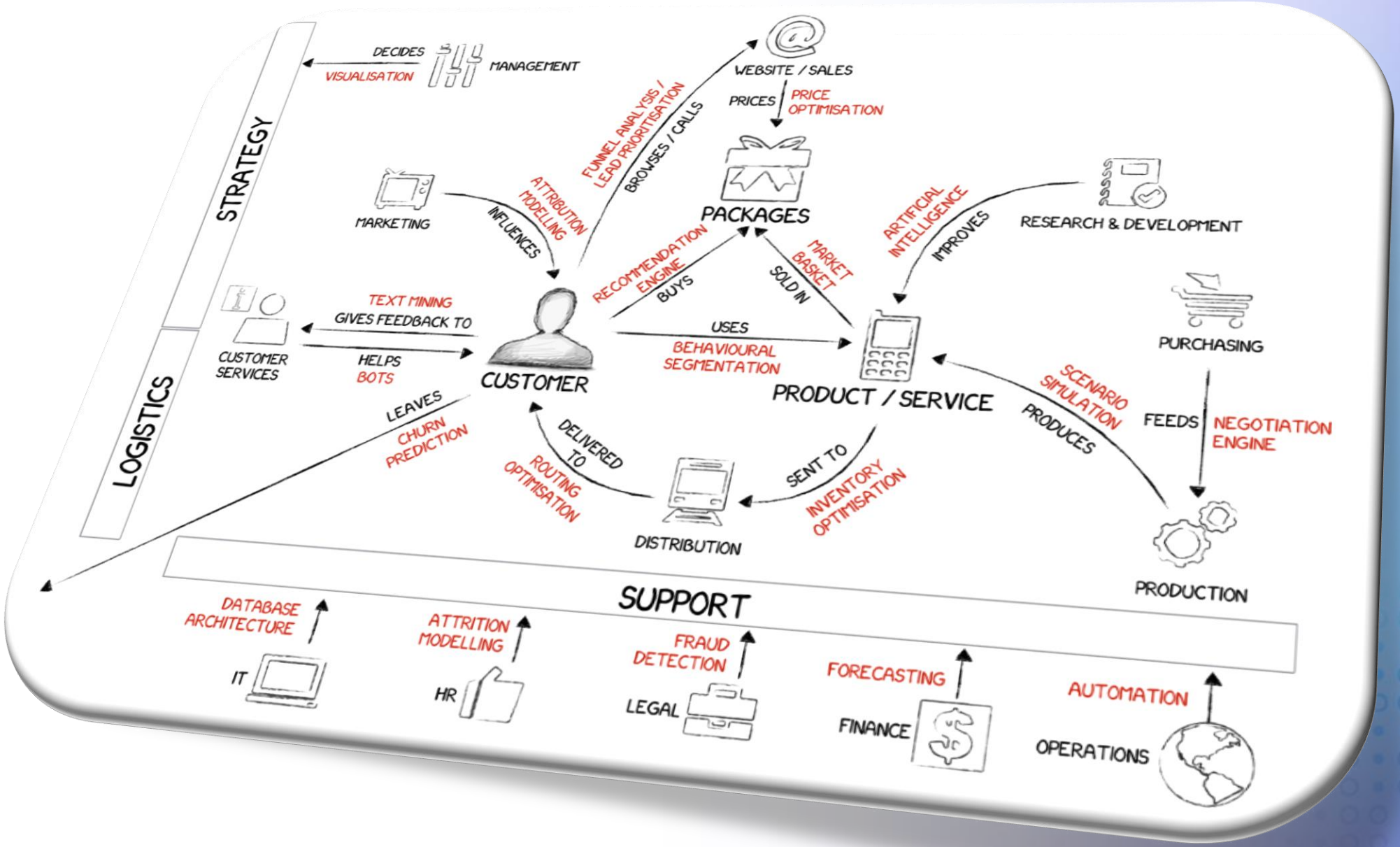
Managing ML Workflows In Production with Amazon SageMaker

Vinicius Caridá

AWS Machine Learning Hero

Head of Digital Customer Service Platforms, PCP, Data and AI at Itaú Unibanco



Fonte: https://applied-data.science

What is the biggest difficulty of machine learning (ML)?

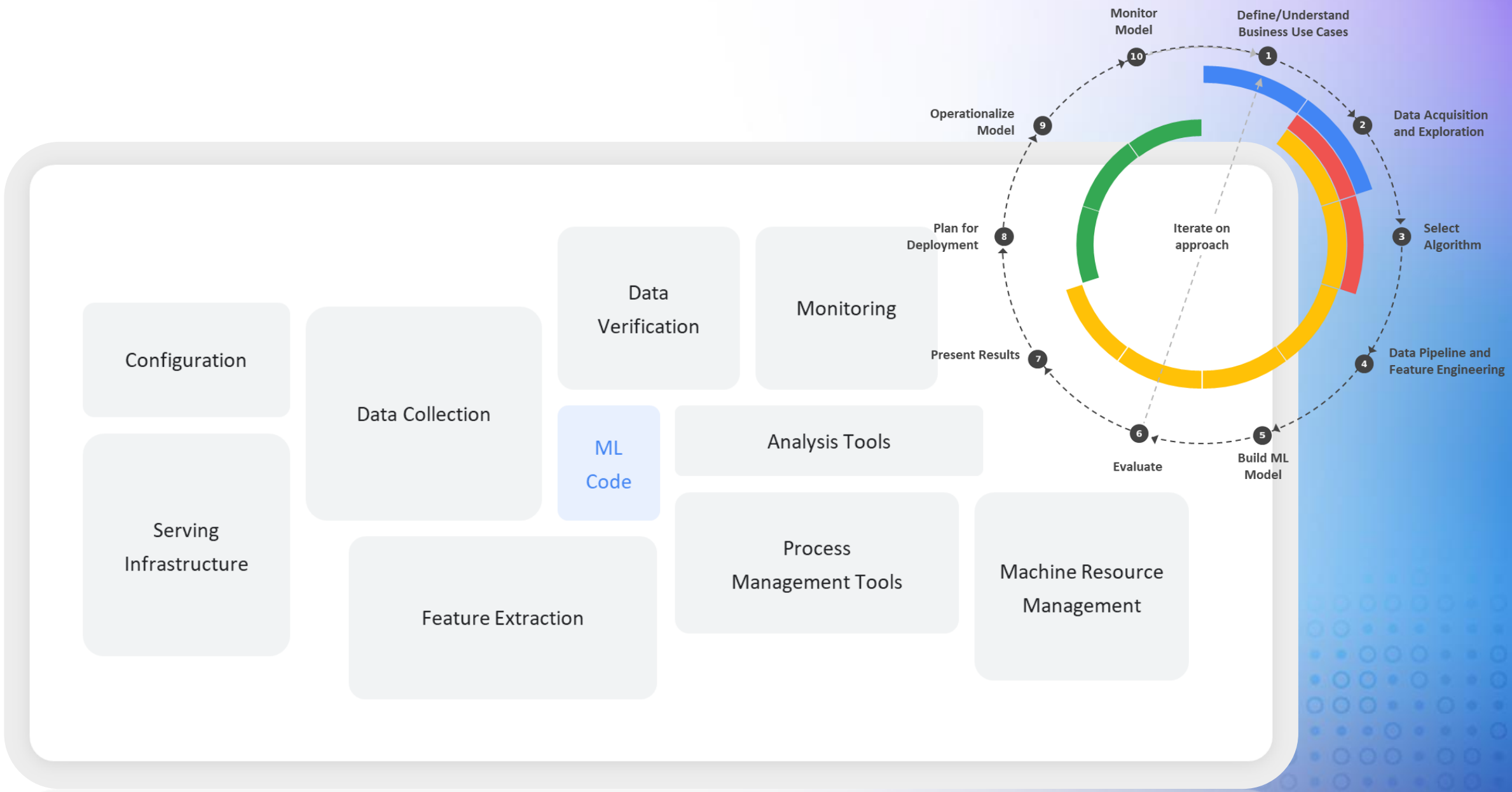
> “The hardest part of ML is not the ML, but the massive amount of effort put into maintaining ML systems. It's easy to become dependent and difficult to support”

Anthony Penta

Sr. Manager & Principal Scientist, Amazon Consumer Payments

Launching is easy, Operating is hard.

“The real problems with a ML system will be found while you are continuously operating it for the long term”



Hidden Technical Debt in Machine Learning Systems
https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

“More than 87% of data science projects never make it into production“

- Multiple studies and surveys -

Fonte: https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/

BUILDING AN END-TO-END DATA STRATEGY

DATA SOURCES

IOT / DEVICES

APP / LOGS

3RD PARTY DATA

Store & Query

FOR APPLICATIONS

- Amazon Aurora
- Amazon DynamoDB
- Amazon Kinesis
- Amazon MSK

FOR ANALYTICS & ML

Data Warehouse

- Amazon Redshift

Data Lake

- Amazon S3

Act

ANALYTICS

- Redshift Query Engine
- Amazon EMR
- Amazon Athena
- Amazon OpenSearch

MACHINE LEARNING

- Amazon SageMaker

BUSINESS INTELLIGENCE

- Amazon QuickSight

Catalog & Govern

AWS Lake Formation

Amazon DataZone

PEOPLE

APPS

DEVICES

The AWS ML Stack

BROADEST AND MOST COMPLETE SET OF MACHINE LEARNING CAPABILITIES

AI services

Health AI
- Amazon HealthLake
- Amazon Transcribe Medical
- Amazon Comprehend Medical

Industrial AI
- AWS Panorama
- AWS Panorama Appliance
- Amazon Monitron
- Amazon Lookout for Equipment
- Amazon Lookout for Vision

Anomaly detection
- Amazon Lookout for Metrics

Code and DevOps
- Amazon DevOps Guru
- Amazon CodeGuru

Vision
- Amazon Rekognition

Speech
- Amazon Polly

Text
- Amazon Transcribe
- Amazon Comprehend
- Amazon Translate
- Amazon Textract

Search
- Amazon Kendra

Chatbots
- Amazon Lex

Personalization
- Amazon Personalize

Forecasting
- Amazon Forecast

Fraud
- Amazon Fraud Detector

Contact centers
- Contact Lens for Amazon Connect
- Voice ID

ML services

- Amazon SageMaker
- Label data

SageMaker Studio IDE

Aggregate & prepare data | Store & share features | Auto ML | Spark/ R | Detect bias | Visualize in notebooks | Pick algorithm | Train models | Tune parameters | Debug & profile | Deploy in production | Manage & monitor | CI/ CD

- Human review

SageMaker JumpStart

Model management for edge devices

Frameworks and infrastructure

- TensorFlow
- mxnet
- PyTorch
- GLUON
- Keras
- scikit learn
- Horovod
- DeepGraphLibrary
- AWS Deep Learning AMIs (DLAMI) and AWS Deep Learning Containers
- GPUs and CPUs
- Amazon Elastic Inference
- AWS Trainium
- AWS Inferentia
- FPGA

aws

Amazon SageMaker

ML services

SageMaker Studio IDE

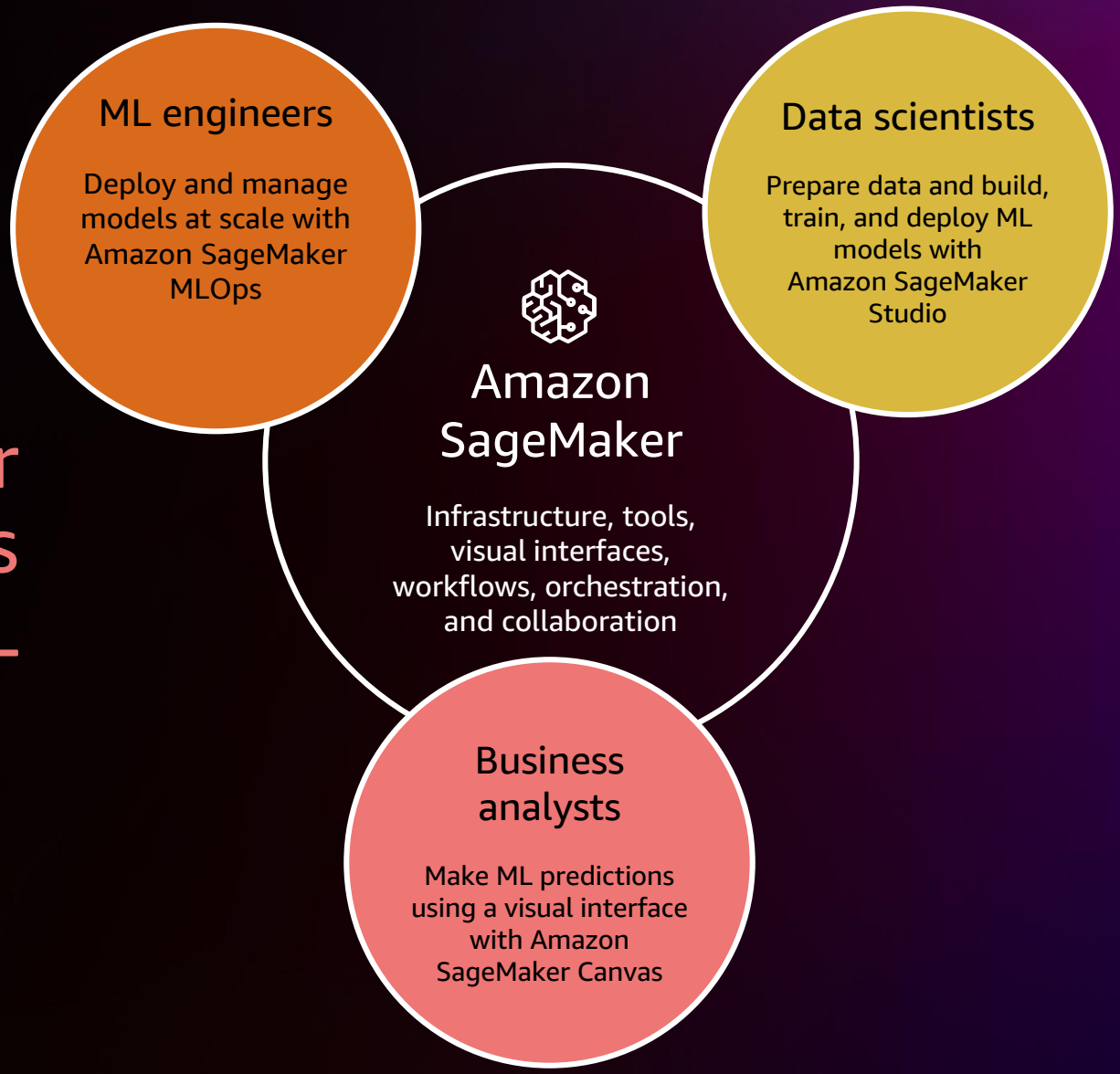
Amazon SageMaker

Label data | Aggregate & prepare data | Store & share features | Auto ML | Spark/ R | Detect bias | Visualize in notebooks | Pick algorithm | Train models | Tune parameters | Debug & profile | Deploy in production | Manage & monitor | CI/ CD | Human review

SageMaker JumpStart

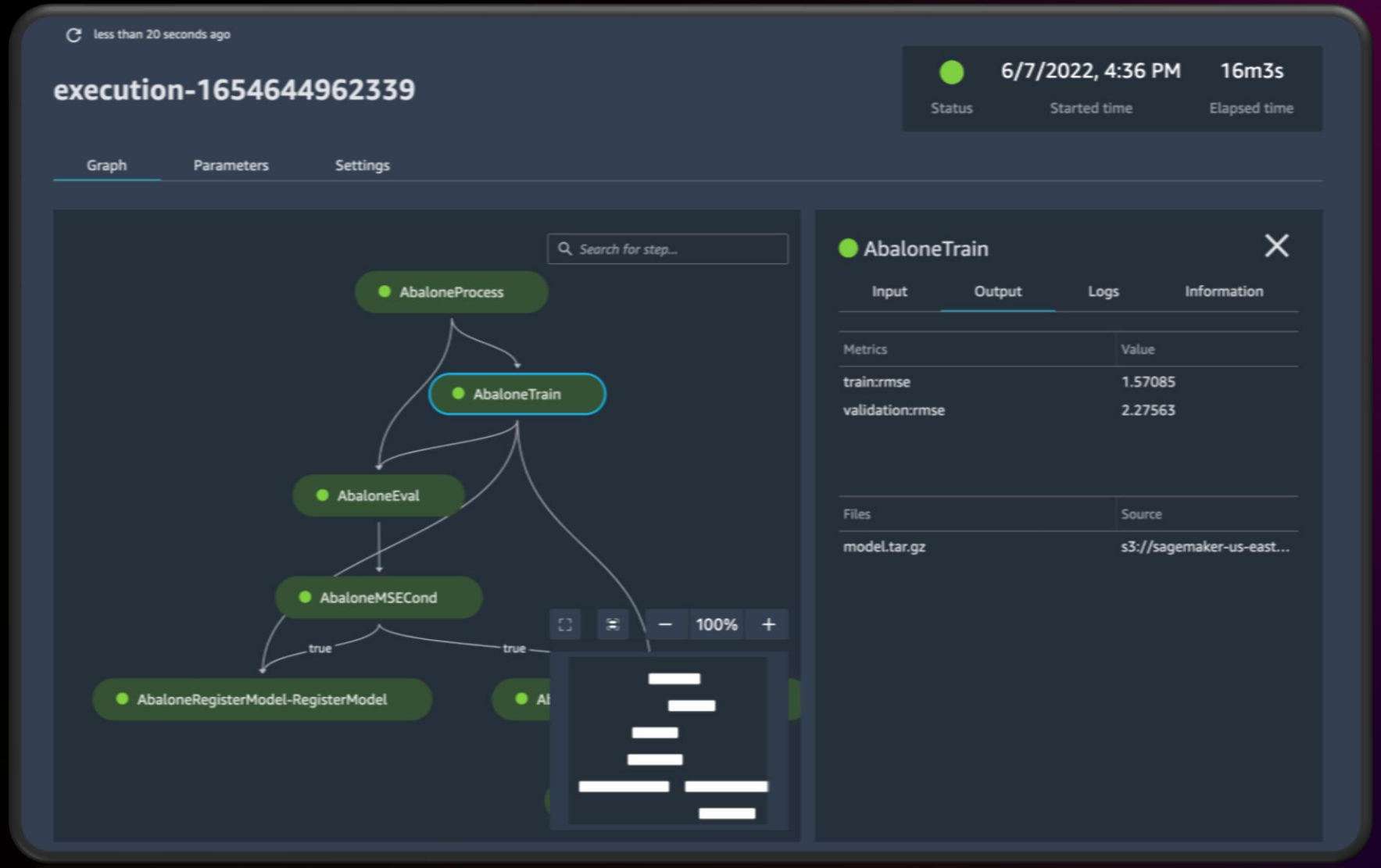
Model management for edge devices

Amazon SageMaker helps organizations harness ML



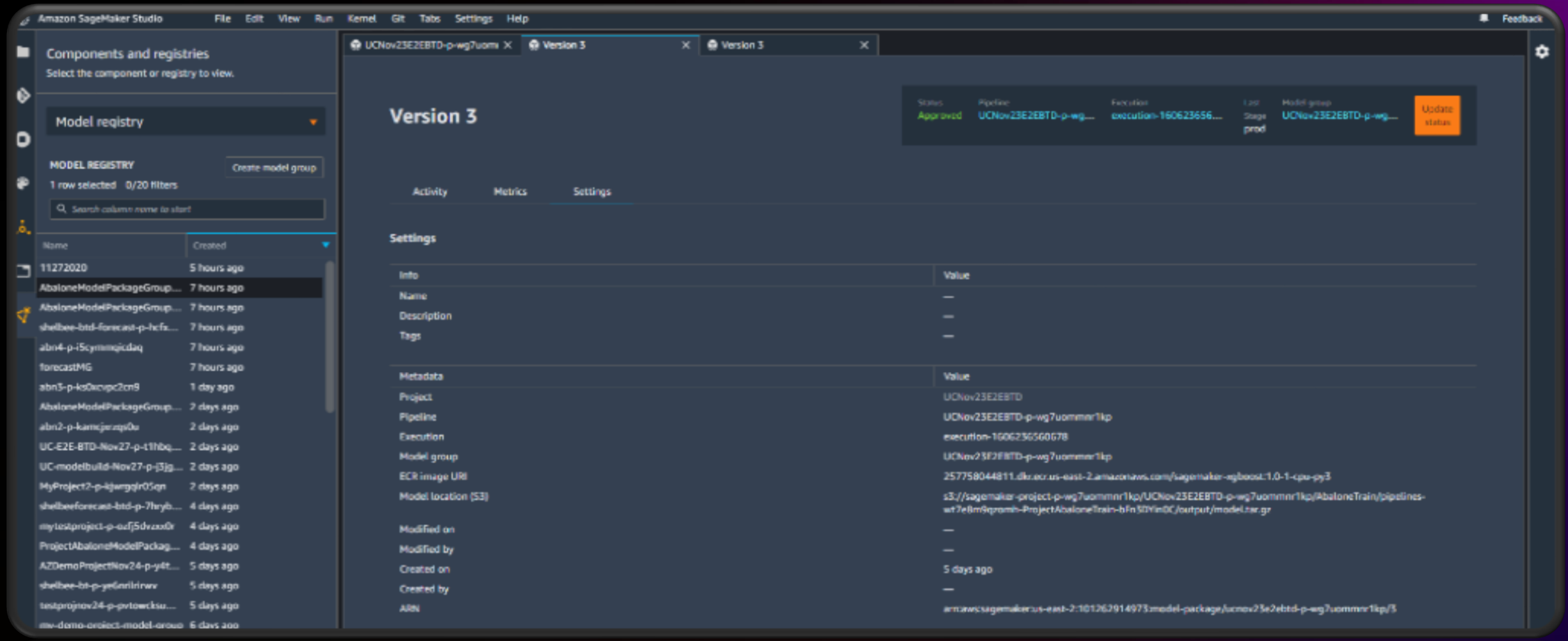
Automate ML training workflows

Compose, manage, and reuse ML workflows



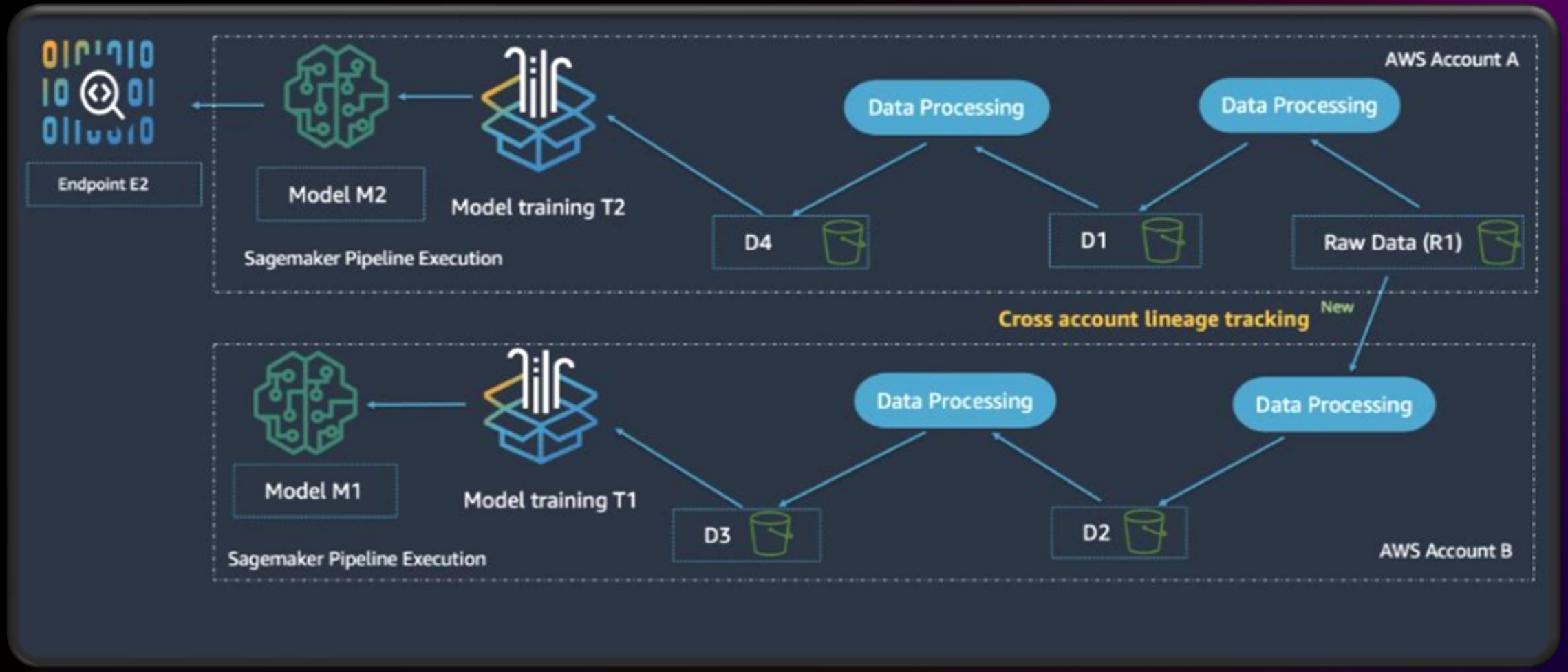
https://docs.aws.amazon.com/sagemaker/latest/dg/workflows.html

Automatic tracking of models



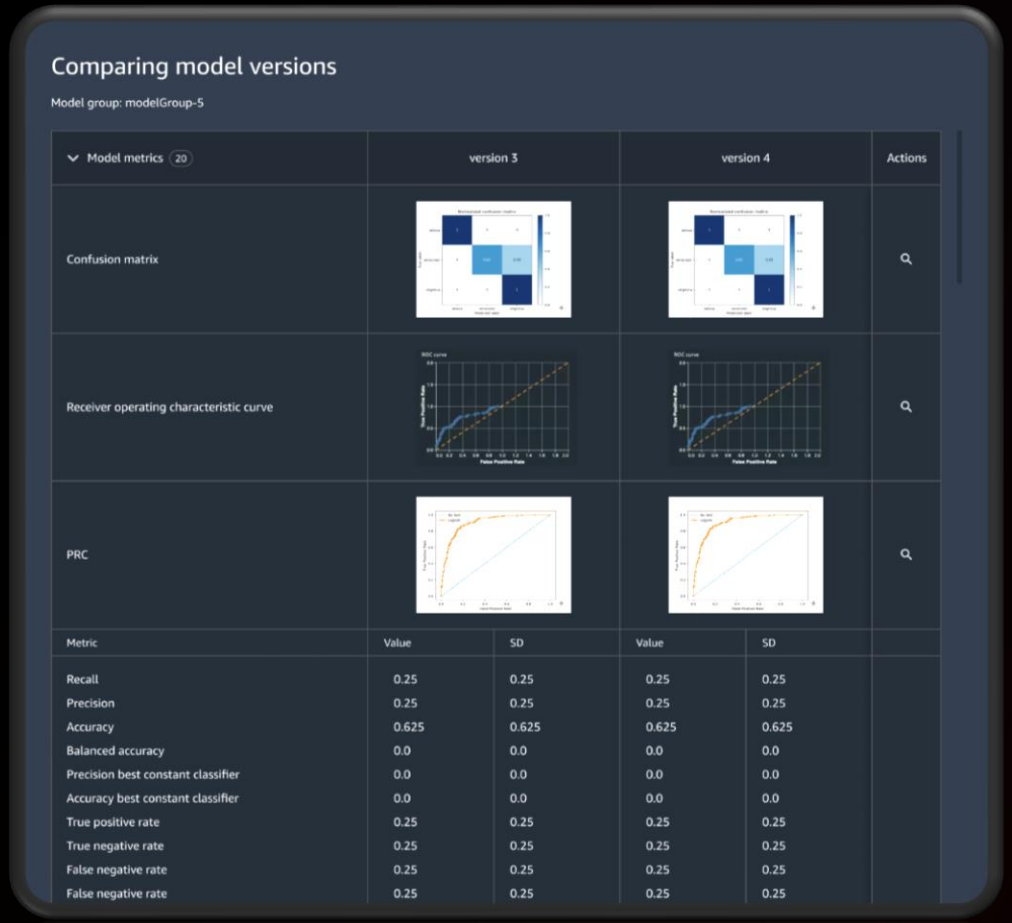
Easily deploy and manage models in production

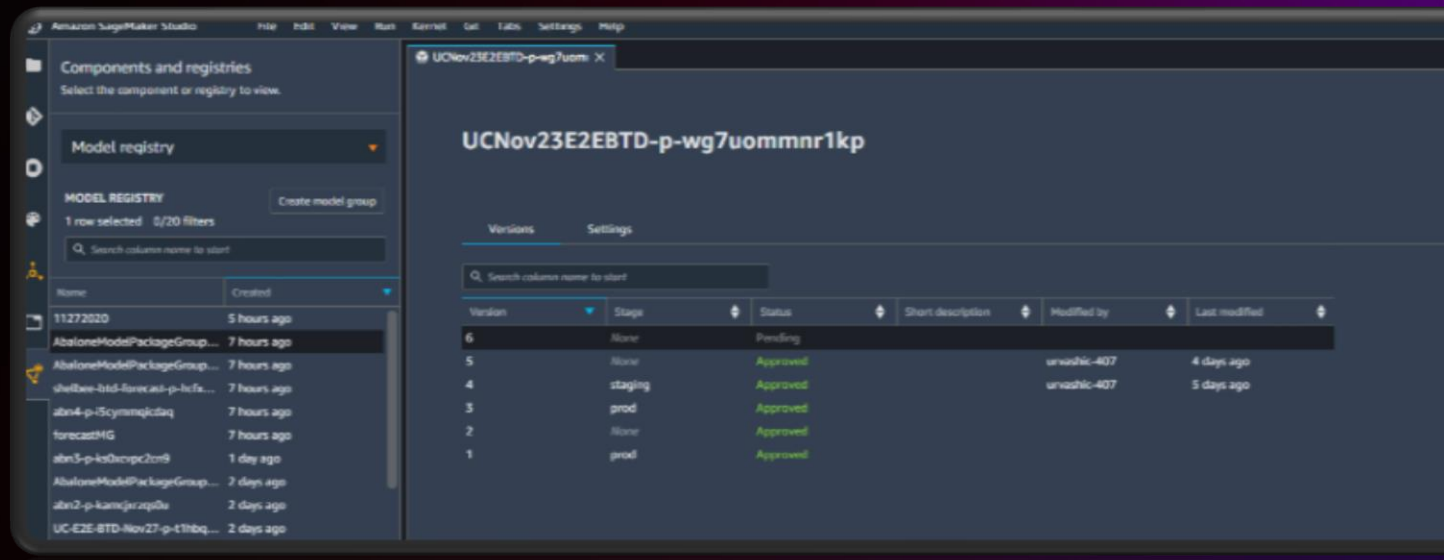
Quickly reproduce your models for troubleshooting



Centrally track and manage model versions

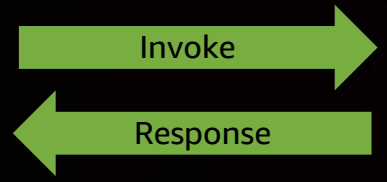
Choose the best models for deploying into production



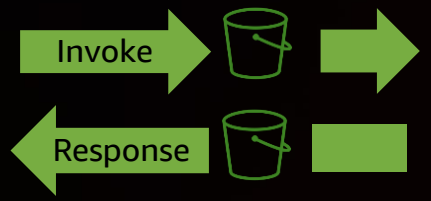


Deploy model to serve inference

- Real-time synchronous response



- Near real-time asynchronous response



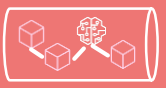
- Offline batch inference

Amazon SageMaker

Model

Single model deployment

Multi-Model deployment



Inference Pipelines

Container

Single container

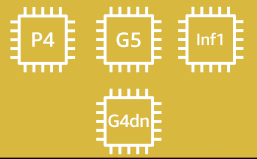
Multi-container

TensorFlow PyTorch mxnet ONNX Keras learn GLUON

Infrastructure

Serverless

GPUs



CPUs

Inference on Amazon SageMaker

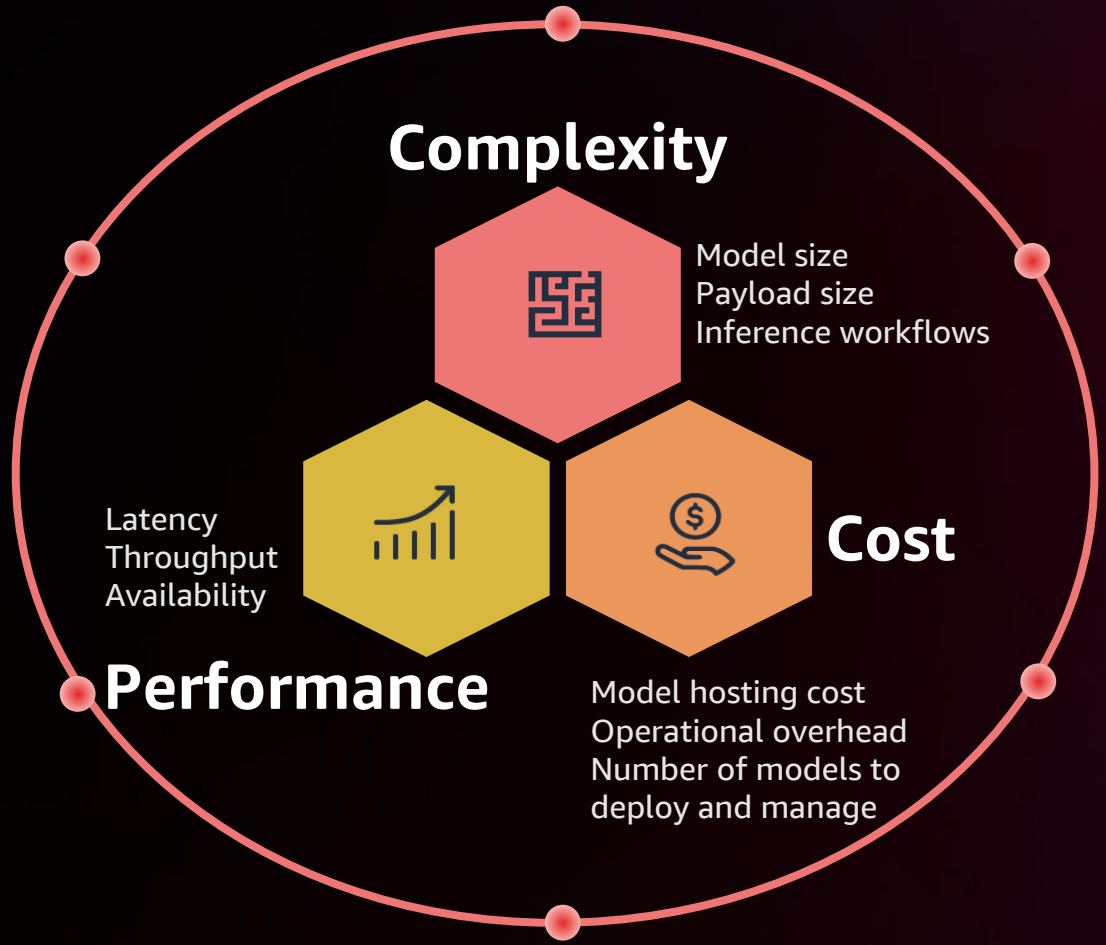
How do you strike the right balance?

Infrastructure management

Optimized ML framework containers, compute instances, storage, network

Logging and metrics

Host metrics (GPU, CPU, memory), invocation metrics, error metrics



High availability

Multi-AZ provisioning, bad instances replacement, auto scaling, 99.95% SLA for real-time inference

ML-specific capabilities

Model registry, model monitoring, explainability, lineage tracking

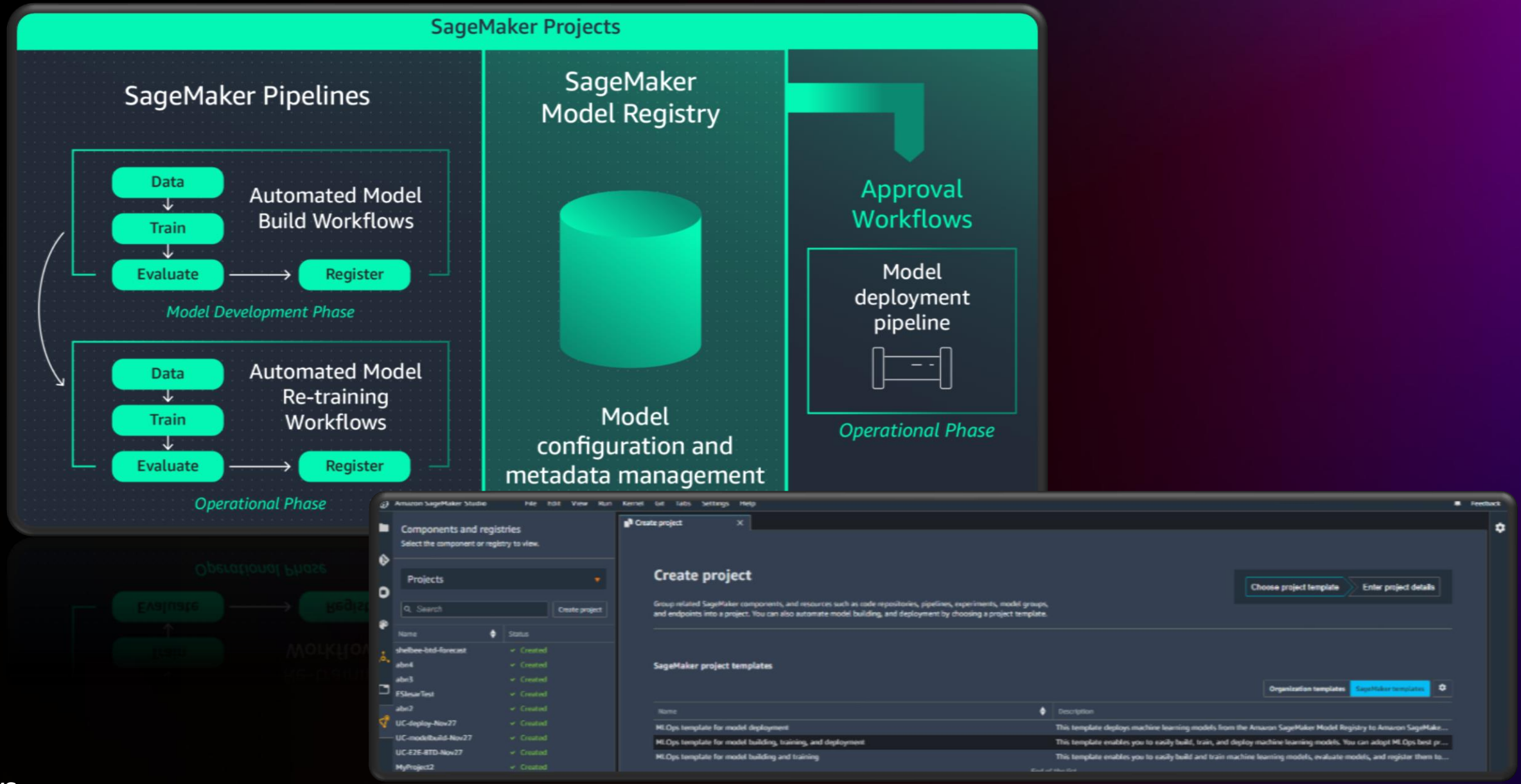
Security and compliance

Security compliance validation, support for CMK and KMS, authentication/authorization

Cost optimization tooling

Multi-model, multi-container endpoints, serverless inference, instance rightsizing, model compilation

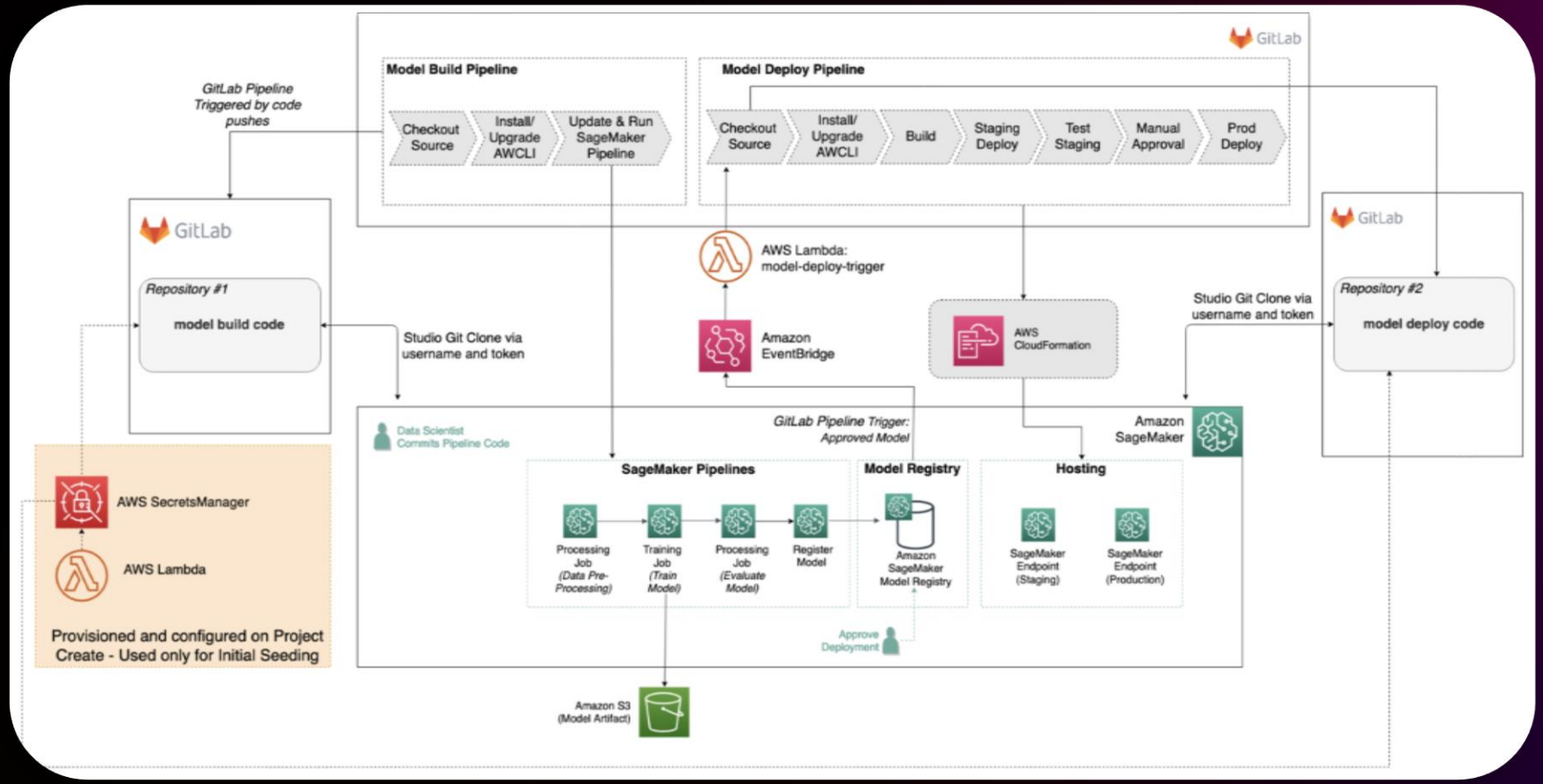
Automate integration and deployment (CI/CD) workflows



Continuously retrain models to maintain prediction quality

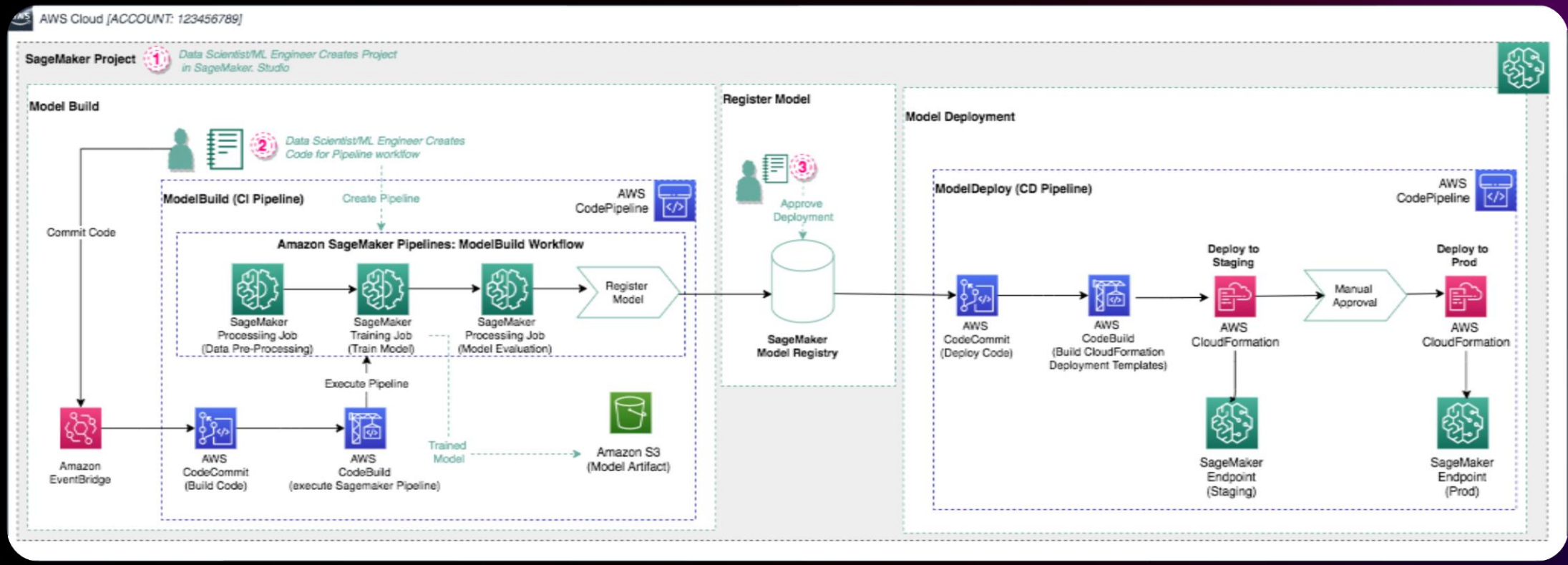
https://aws.amazon.com/getting-started/hands-on/machine-learning-tutorial-mlops-automate-ml-workflows/?nc1=h_ls

Build MLOps workflows with Amazon SageMaker projects, GitLab, and GitLab pipelines



https://aws.amazon.com/pt/blogs/machine-learning/build-mlops-workflows-with-amazon-sagemaker-projects-gitlab-and-gitlab-pipelines/

Building, automating, managing, and scaling ML workflows using Amazon SageMaker Pipelines



https://aws.amazon.com/pt/blogs/machine-learning/building-automating-managing-and-scaling-ml-workflows-using-amazon-sagemaker-pipelines/

Thank you!

Vinicius Caridá

@vfcarida
vfcarida@gmail.com
https://www.linkedin.com/in/viniciuscarida/