

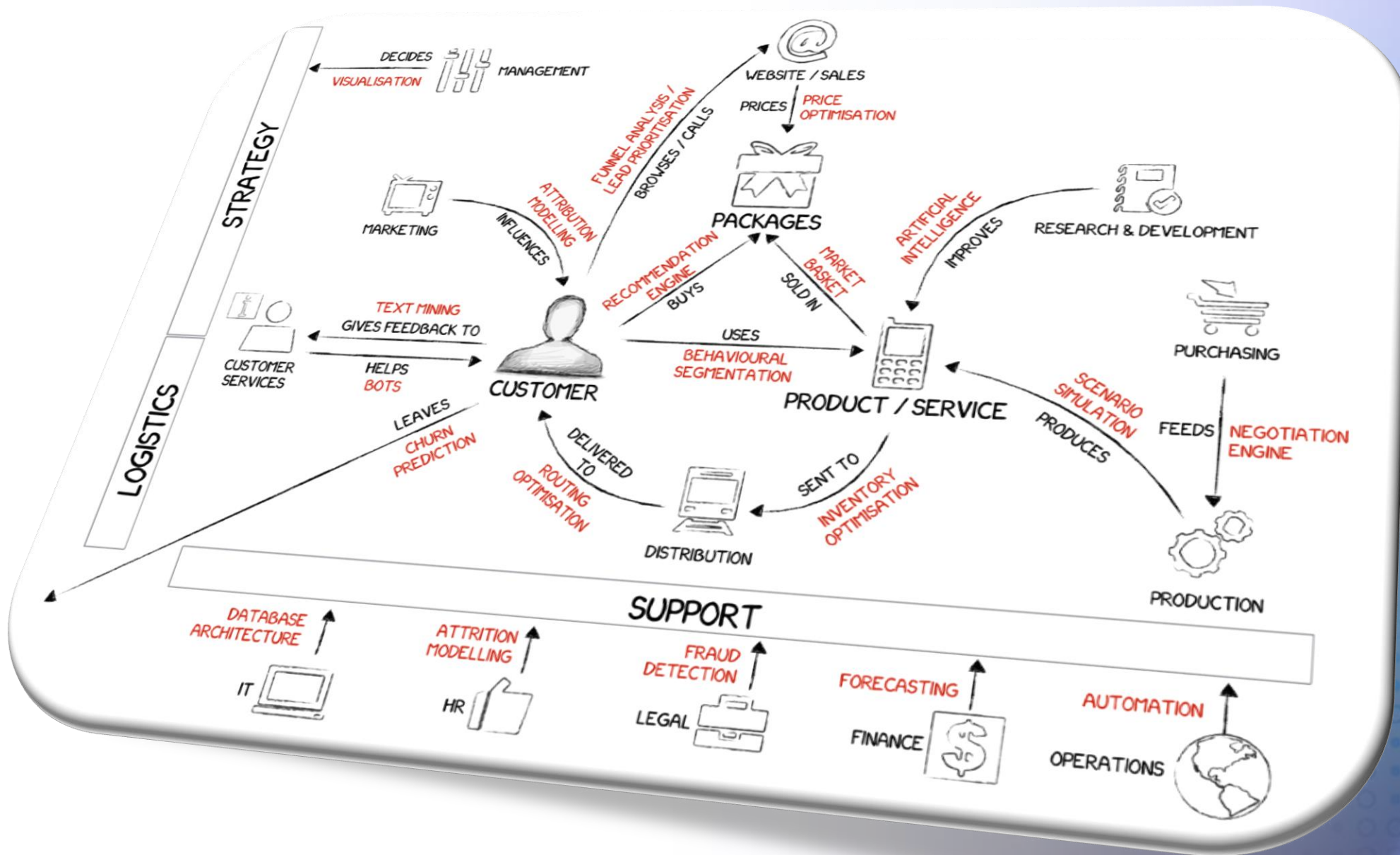
# Managing ML Workflows In Production with Amazon SageMaker

**Vinicius Caridá**

AWS Machine Learning Hero

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





# 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 ”**

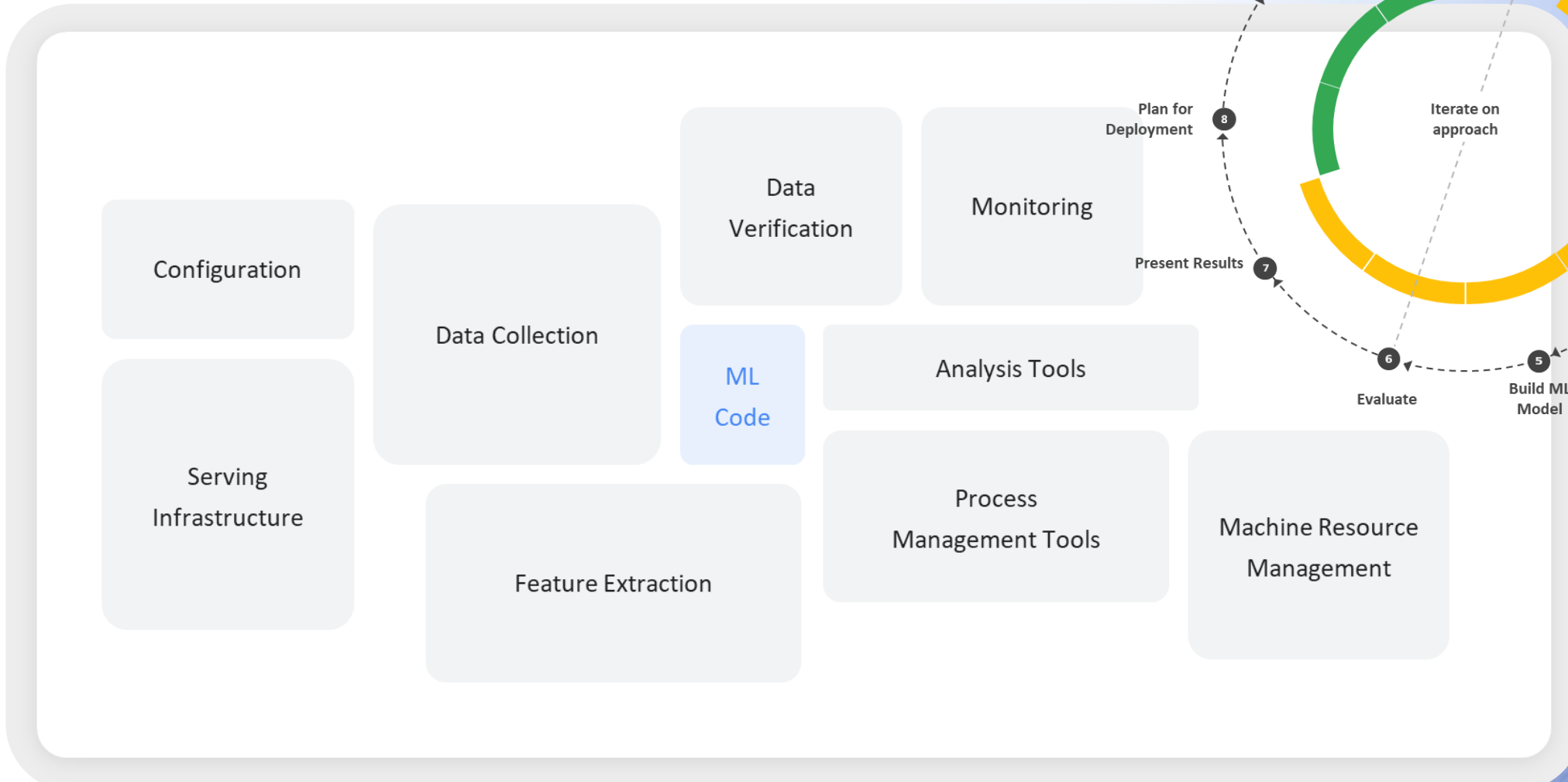
**Anthony Penta**

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”**



**“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

## Store & Query

DATA SOURCES

IOT / DEVICES

APP / LOGS

3RD PARTY DATA

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

 Amazon HealthLake  
 Amazon Transcribe Medical  
 Amazon Comprehend Medical

## Health AI

## 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  
 DeepGraphLibrary

AWS Deep Learning AMIs (DLAMI) and AWS Deep Learning Containers

GPUs and CPUs

Amazon Elastic Inference

AWS Trainium

AWS Inferentia

FPGA

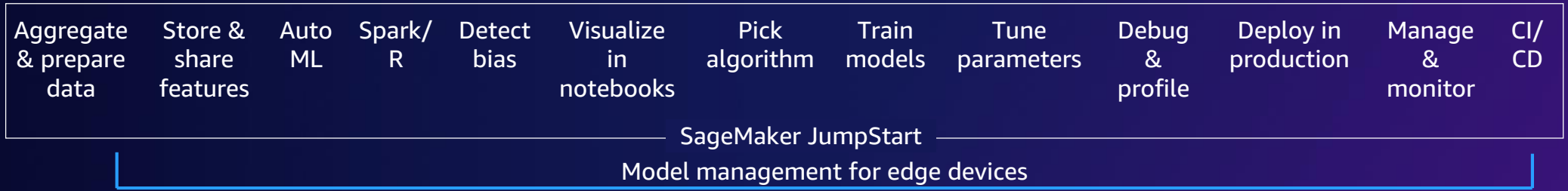




# Amazon SageMaker

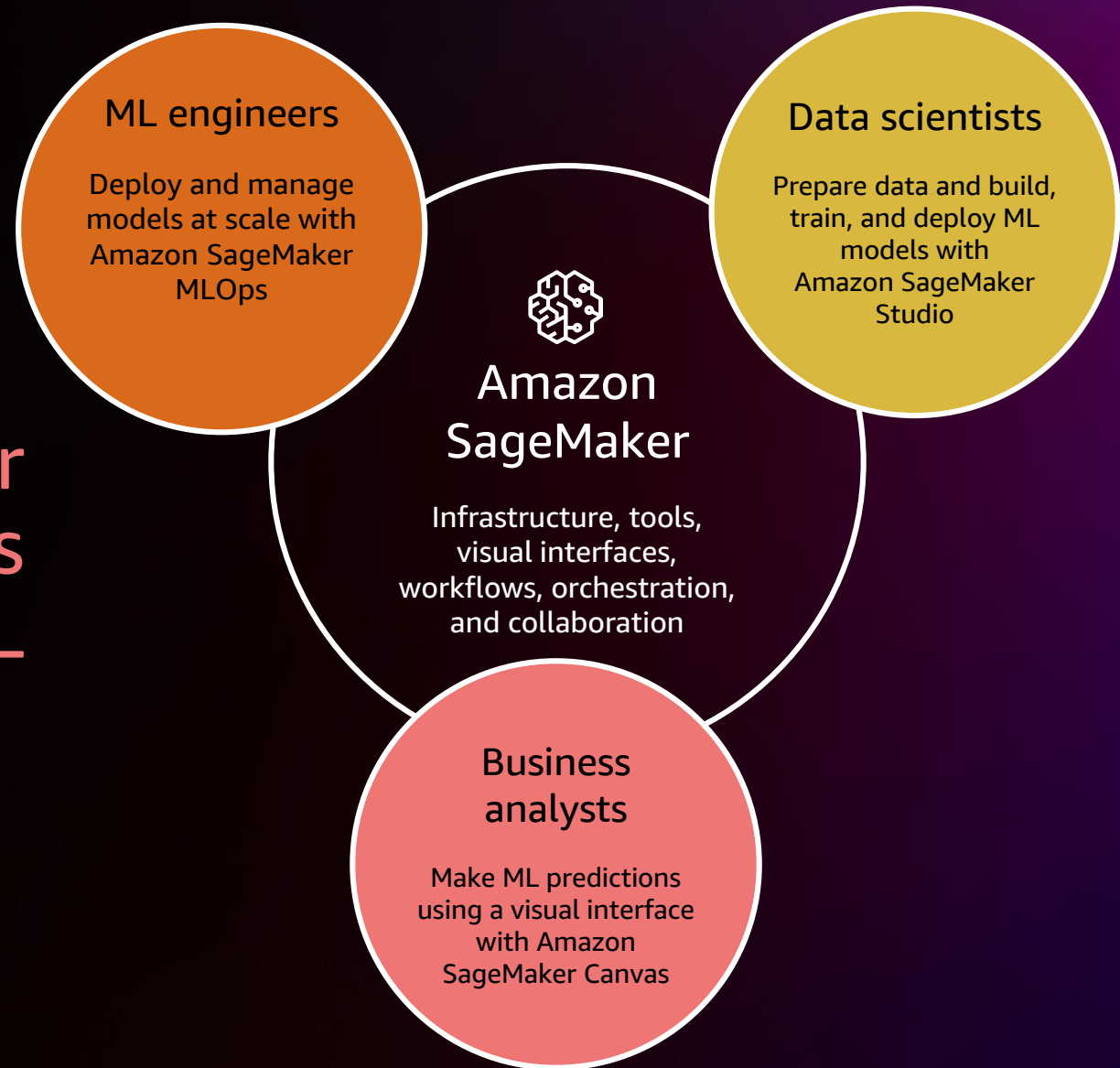


## ML services





# Amazon SageMaker helps organizations harness ML



# Automate ML training workflows

Compose, manage, and reuse ML workflows

less than 20 seconds ago

execution-1654644962339

Status: ● 6/7/2022, 4:36 PM 16m3s

Started time Elapsed time

Graph Parameters Settings

Search for step...

AbaloneProcess

AbaloneTrain

AbaloneEval

AbaloneMSECond

AbaloneRegisterModel-RegisterModel

AbaloneTrain

Input	Output	Logs	Information
Metrics			Value
train:rmse			1.57085
validation:rmse			2.27563
Files			Source
model.tar.gz			s3://sagemaker-us-east...



# Automatic tracking of models

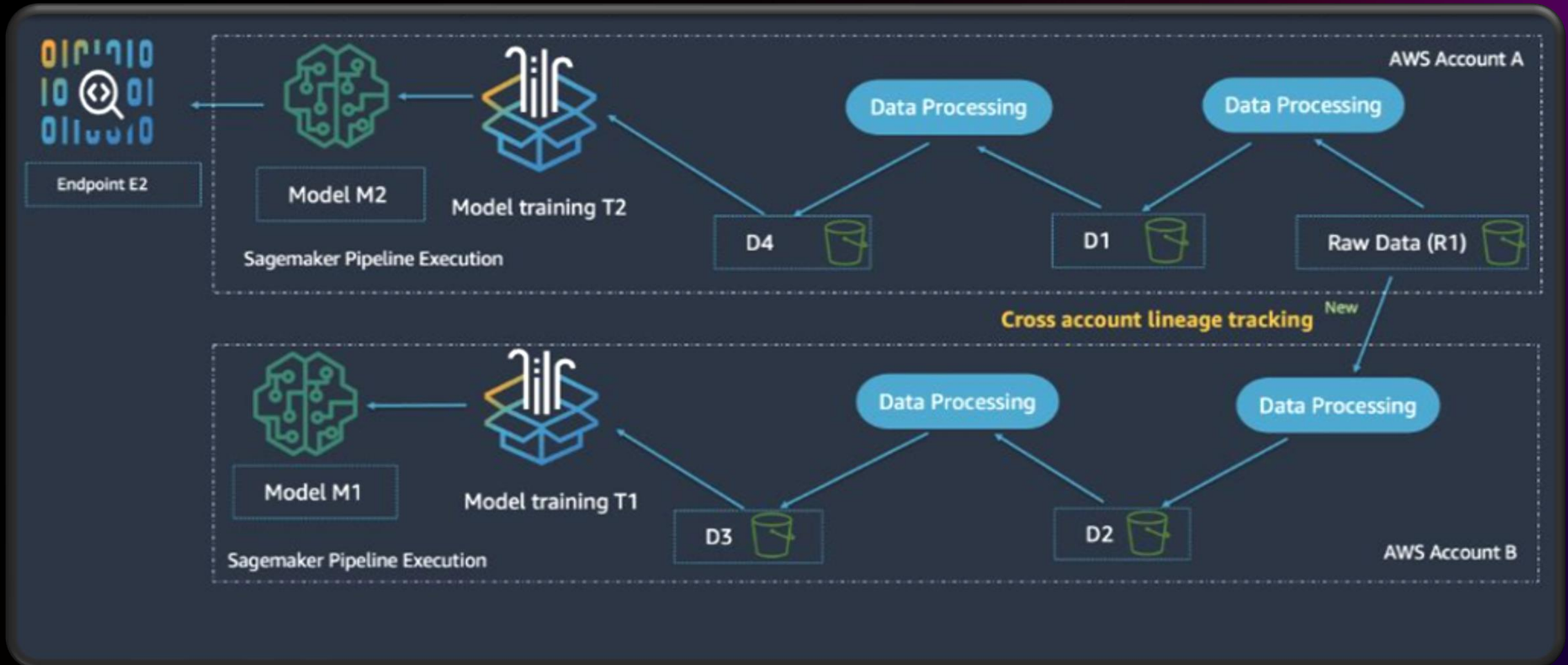
The screenshot displays the Amazon SageMaker Studio interface. On the left, the 'Components and registries' sidebar shows the 'Model registry' section with a table of model packages. The main area shows the 'Version 3' settings for a specific model. The status is 'Approved' and the pipeline is 'UCNov23E2EBTD-p-wg7uomn1kp'. The execution ID is 'execution-160623656...'. The model group is 'UCNov23E2EBTD-p-wg7uomn1kp'. The settings table below provides detailed information about the model, including its project, pipeline, execution, and metadata.

Info	Value
Name	—
Description	—
Tags	—
Metadata	Value
Project	UCNov23E2EBTD
Pipeline	UCNov23E2EBTD-p-wg7uomn1kp
Execution	execution-1606236560678
Model group	UCNov23E2EBTD-p-wg7uomn1kp
ECR image URI	257758044811.dkr.ecr.us-east-2.amazonaws.com/sagemaker-registry-1.0-1-cpu-py3
Model location (S3)	s3://sagemaker-project-p-wg7uomn1kp/UCNov23E2EBTD-p-wg7uomn1kp/AbaloneTrain/pipelines-w7eBm6qpmh-ProjectAbaloneTrain-bFn50Yin0C/output/model.tar.gz
Modified on	—
Modified by	—
Created on	5 days ago
Created by	—
ARN	arn:aws:sagemaker:us-east-2:101263914973:model-package/ucnov23e2ebtd-p-wg7uomn1kp/3



# 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

### Comparing model versions

Model group: modelGroup-5

Model metrics (20)	version 3	version 4	Actions	
Confusion matrix			🔍	
Receiver operating characteristic curve			🔍	
PRC			🔍	
Metric	Value	SD	Value	SD
Recall	0.25	0.25	0.25	0.25
Precision	0.25	0.25	0.25	0.25
Accuracy	0.625	0.625	0.625	0.625
Balanced accuracy	0.0	0.0	0.0	0.0
Precision best constant classifier	0.0	0.0	0.0	0.0
Accuracy best constant classifier	0.0	0.0	0.0	0.0
True positive rate	0.25	0.25	0.25	0.25
True negative rate	0.25	0.25	0.25	0.25
False negative rate	0.25	0.25	0.25	0.25
False negative rate	0.25	0.25	0.25	0.25

Amazon SageMaker Studio

### Components and registries

Select the component or registry to view.

#### Model registry

MODEL REGISTRY

1 row selected 0/20 filters

Search column name to start

Name	Created
11272020	5 hours ago
AbaloneModelPackageGroup...	7 hours ago
AbaloneModelPackageGroup...	7 hours ago
shellbee-std-forecast-p-lcfe...	7 hours ago
abn4-p-iScymmgkdaq	7 hours ago
forecastMG	7 hours ago
abn3-p-ks0vcpc2on9	1 day ago
AbaloneModelPackageGroup...	7 days ago
abn2-p-kamjgrzpf0e	2 days ago
UC-EZE-STD-Nov27-p-t1hbq...	2 days ago

### UCNov23E2EBTD-p-wg7uommnr1kp

Versions Settings

Search column name to start

Version	Stage	Status	Short description	Modified by	Last modified
6	None	Pending			
5	None	Approved		urvatlic-407	4 days ago
4	staging	Approved		urvatlic-407	5 days ago
3	prod	Approved			
2	None	Approved			
1	prod	Approved			

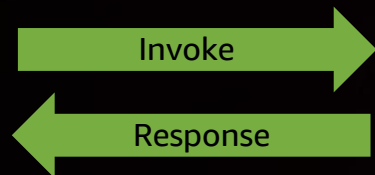


# Deploy model to serve inference

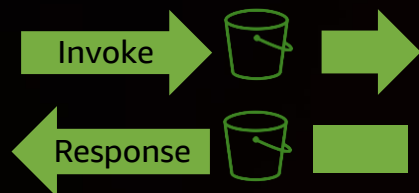


Amazon SageMaker

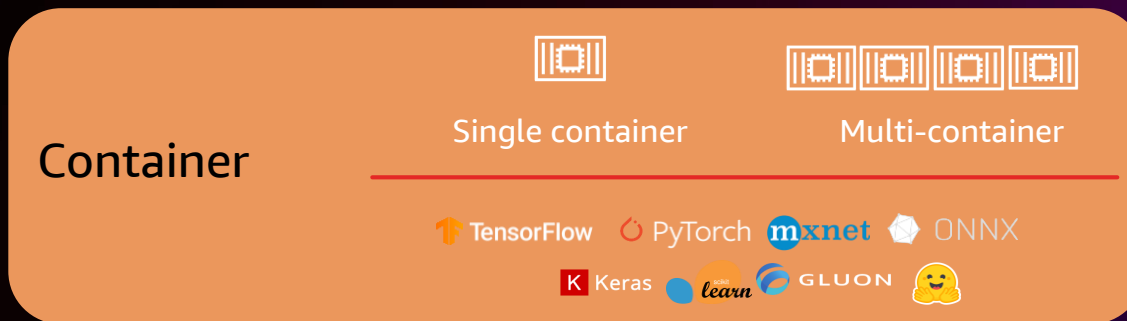
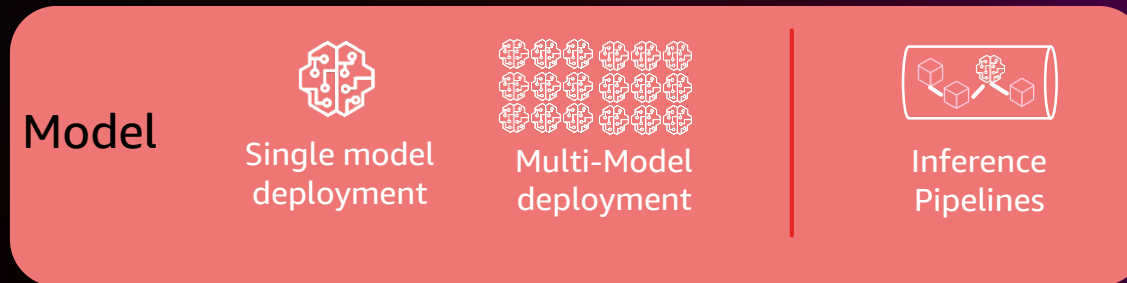
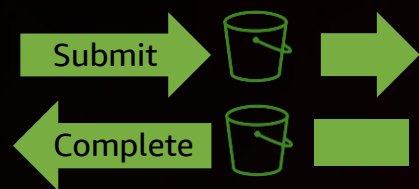
- Real-time synchronous response



- Near real-time asynchronous response

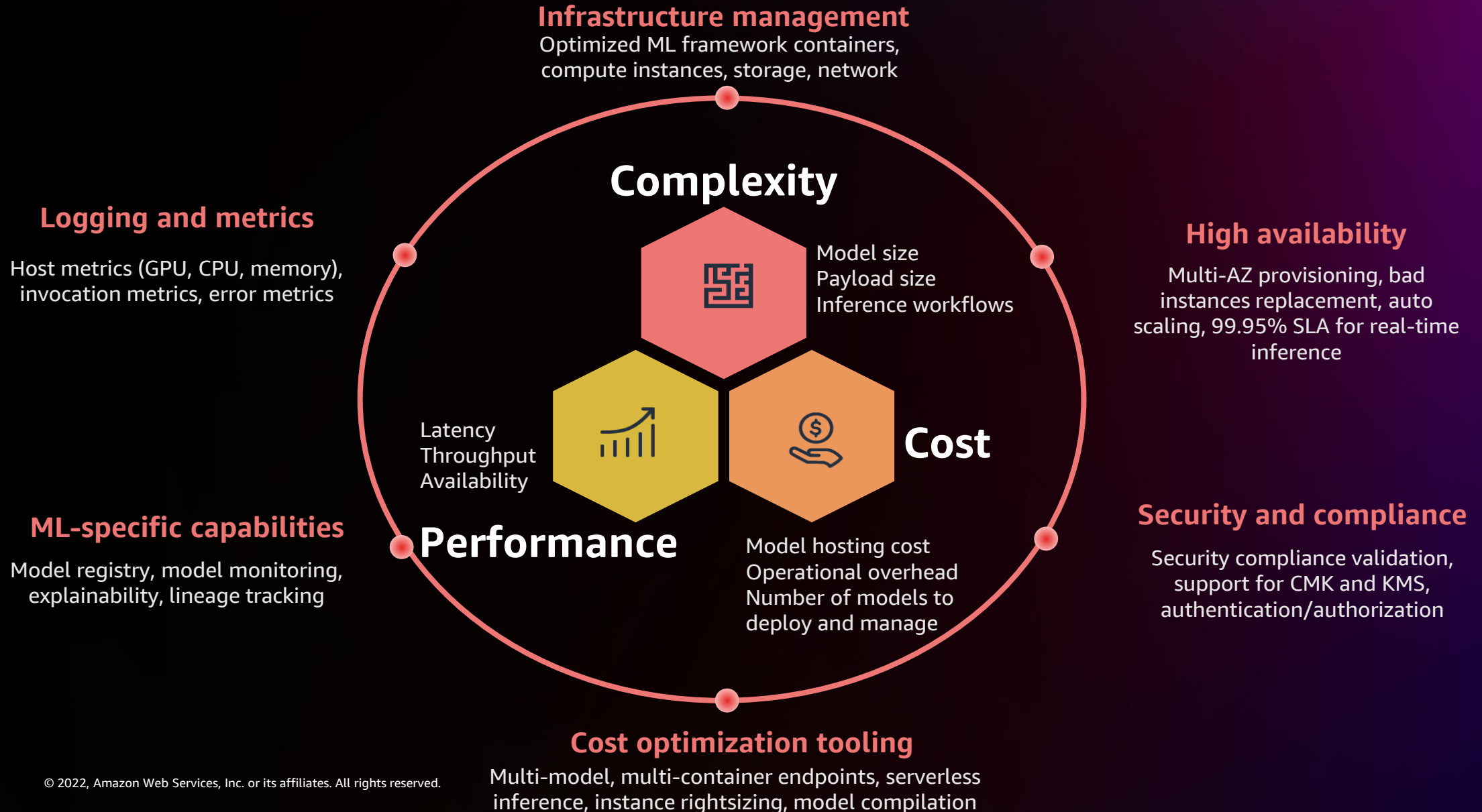


- Offline batch inference

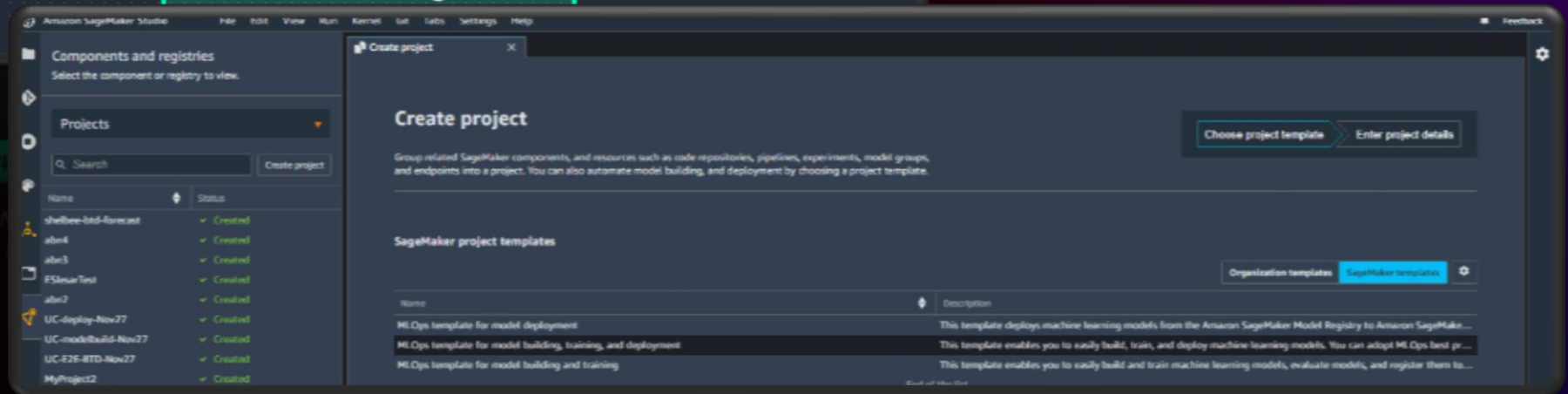
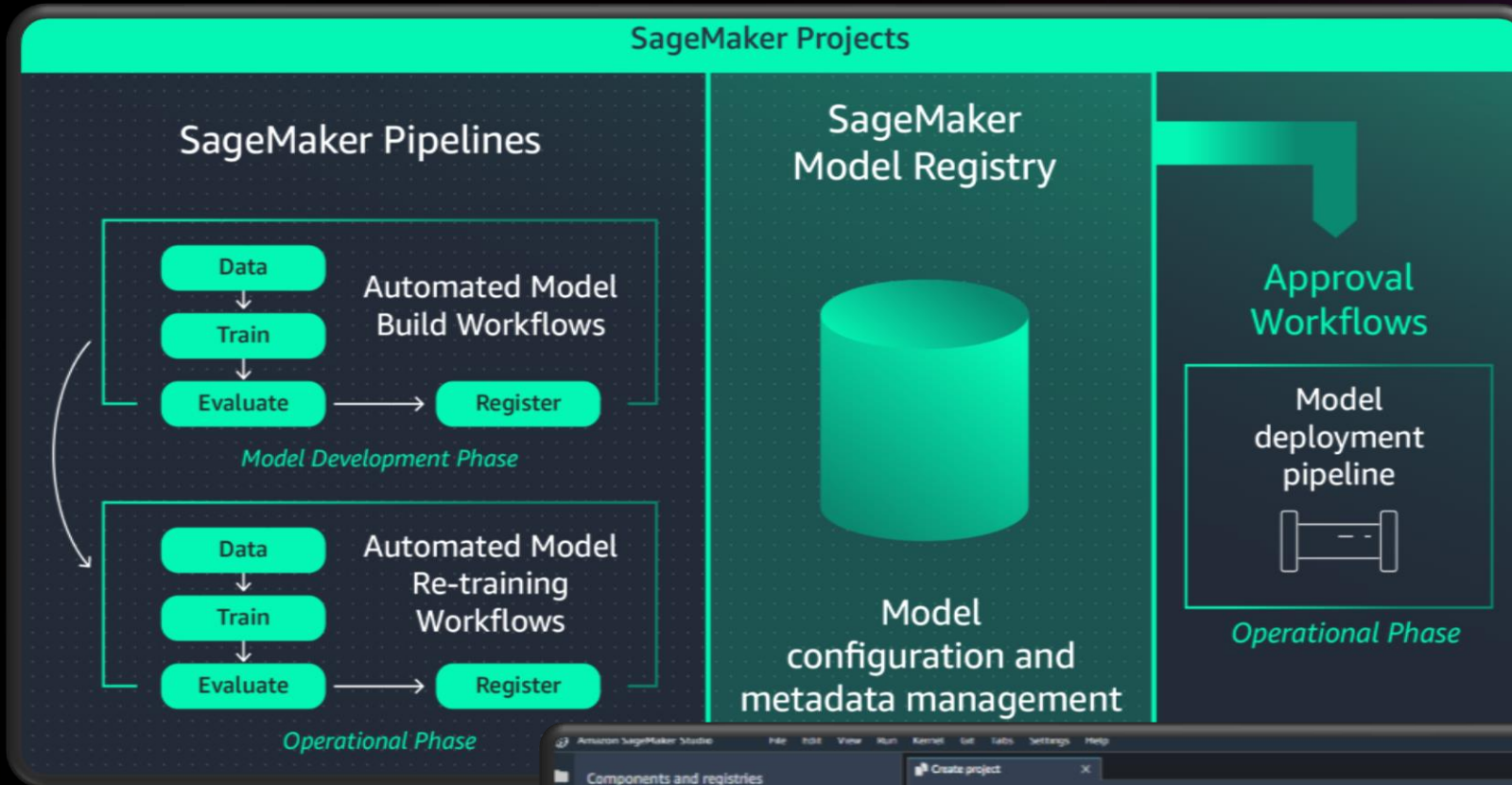


# Inference on Amazon SageMaker

How do you strike the right balance?

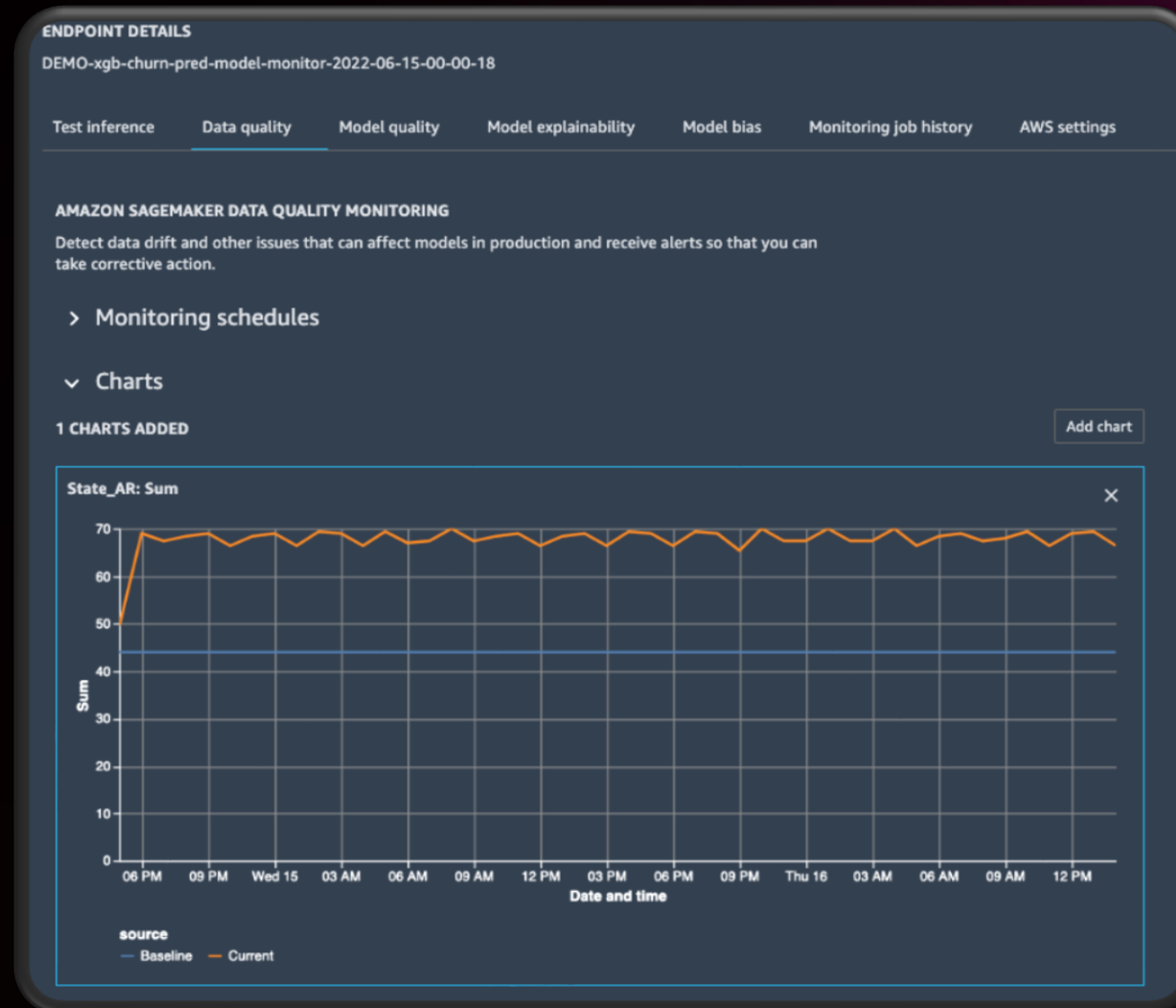


# Automate integration and deployment (CI/CD) workflows





# Continuously retrain models to maintain prediction quality





Contact Us Support English My Account

Sign In to the Console

Products Solutions Pricing Documentation Learn Partner Network AWS Marketplace Customer Enablement Events Explore More

Getting Started Resource Center Getting Started Community Learning AWS re:Post Libraries More Resources

Getting Started / Hands-on / ...

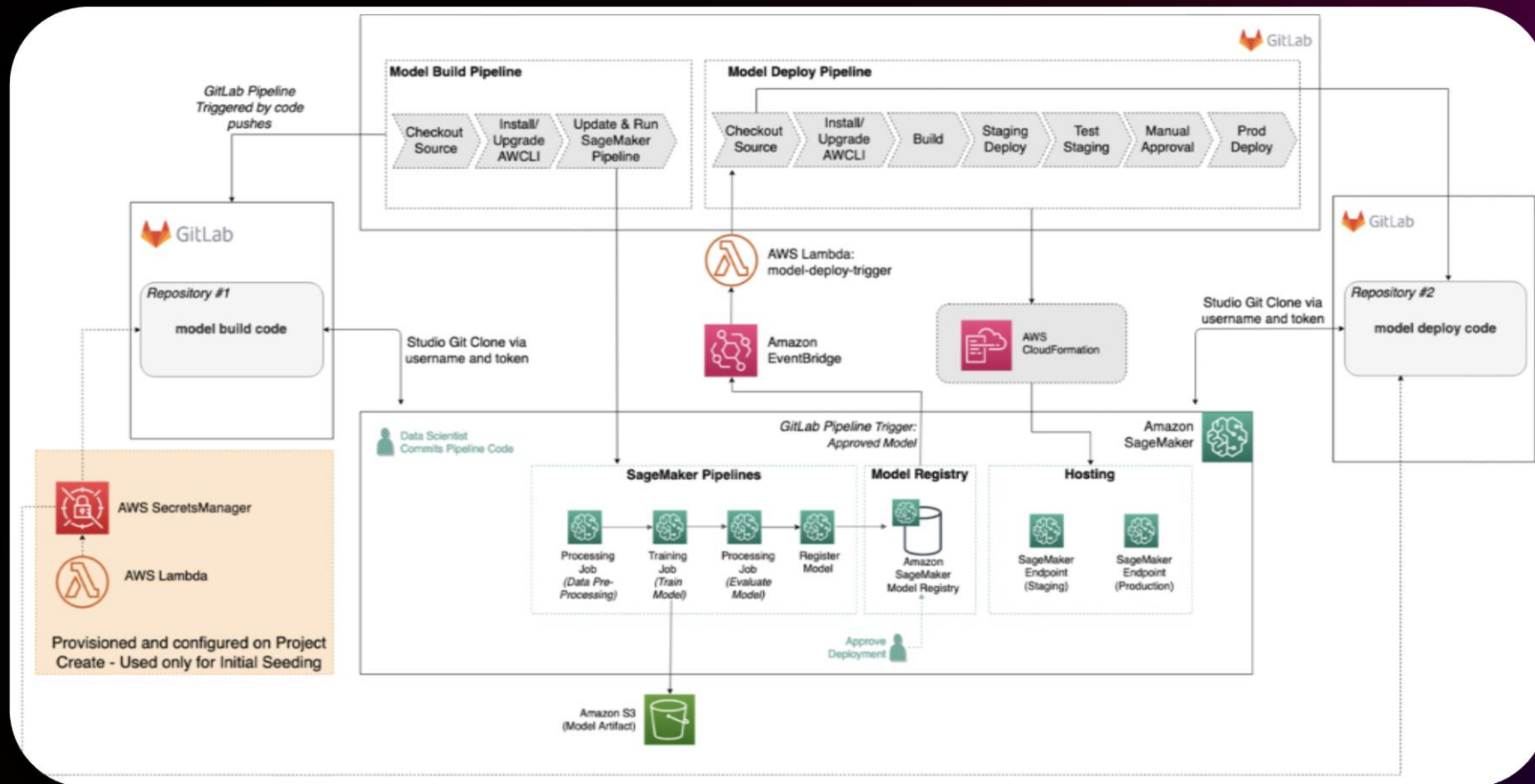
# Automate Machine Learning Workflows

TUTORIAL

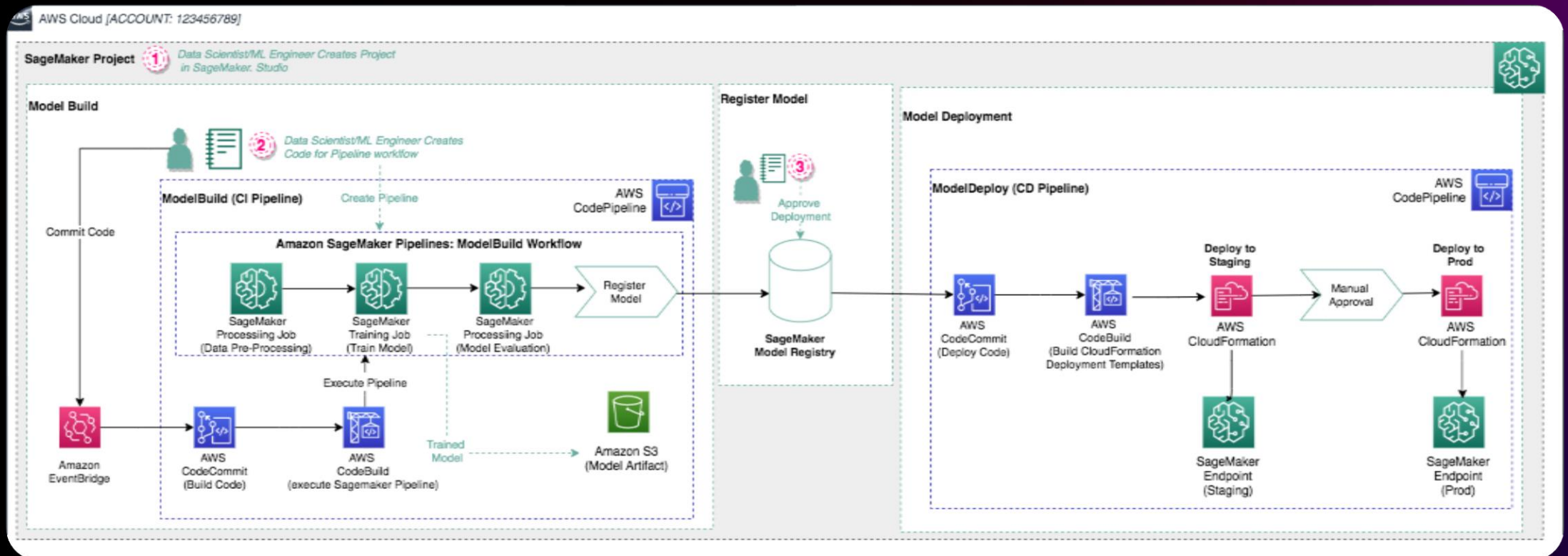
[https://aws.amazon.com/getting-started/hands-on/machine-learning-tutorial-mlops-automate-ml-workflows/?nc1=h\\_ls](https://aws.amazon.com/getting-started/hands-on/machine-learning-tutorial-mlops-automate-ml-workflows/?nc1=h_ls)



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



# 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/>

