AWS re:Invent

ARC340

Amazon.com automating machine learning deployments at scale

Fei Yuan
Senior Software Engineer
Amazon Consumer Payments

Kieran Kavanagh
Senior Solutions Architect
Amazon Web Services

Kunal Batra
Senior Technical Evangelist
Amazon Web Services

Agenda

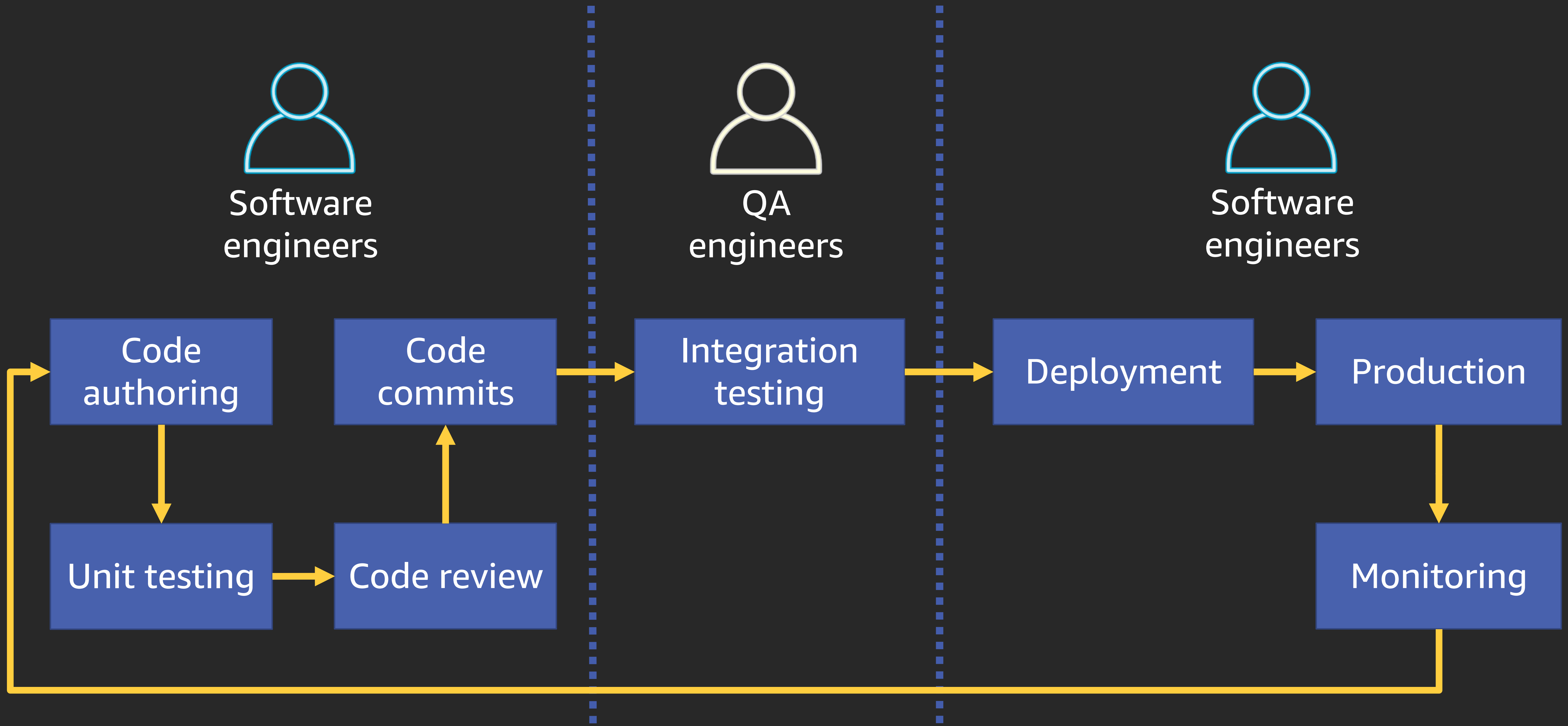
- Overview of machine learning Model Development Life Cycle (MDLC)
- Overview of AWS services used in today's discussion
- Deep dive into MDLC
- Assess common challenges and options
- Describe how Amazon Consumer Payments used AWS services to solve these common challenges

What is the hard part of machine learning (ML)?

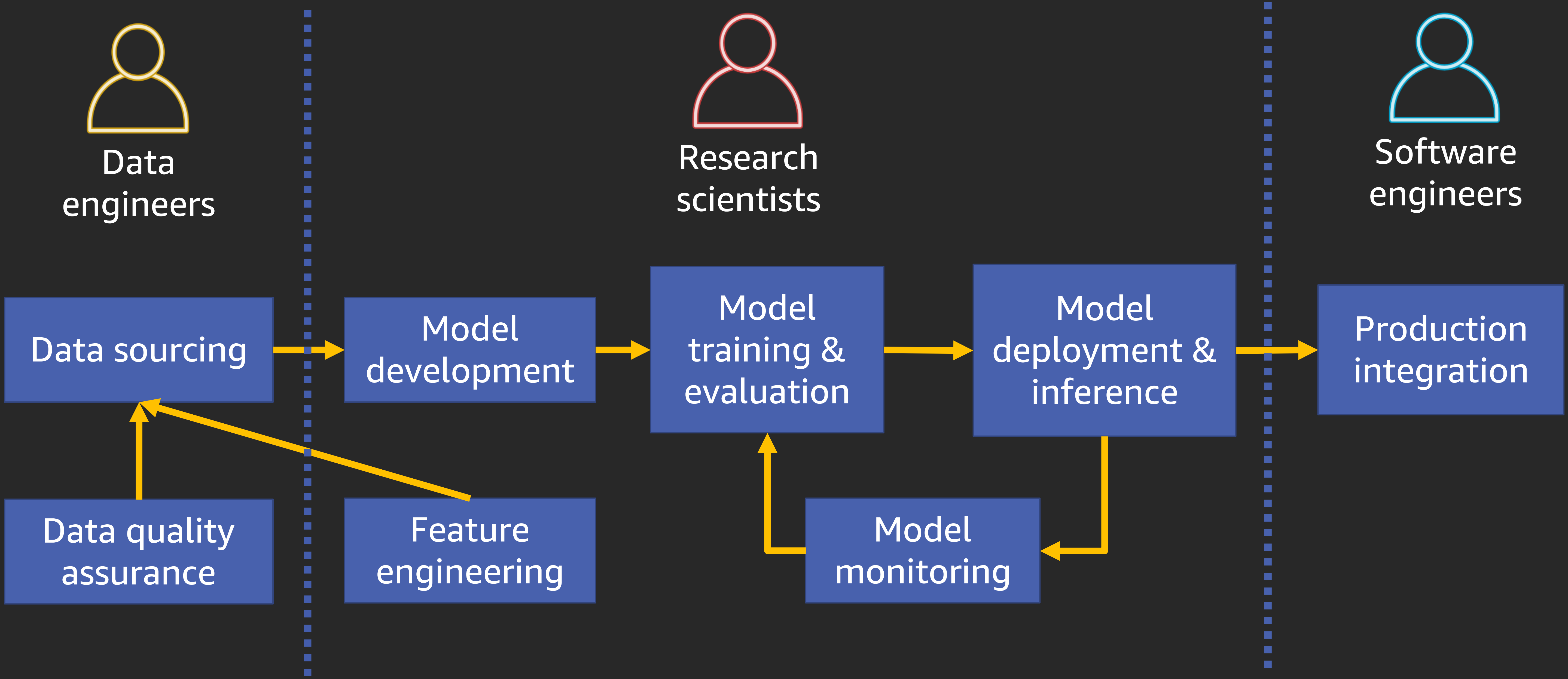
“The hard part of ML is not ML, it is the massive ongoing effort to maintain ML systems. Easy to incur but expensive to sustain.”

- Anthony Penta, Sr. Manager & Principal Scientist, Amazon Consumer Payments

Software Development Life Cycle (SDLC)



Model Development Life Cycle (MDLC)



AWS service overview

AWS services

Amazon Simple Storage Service

Object storage service that offers industry-leading scalability, data availability, security, and performance

AWS IAM

Manage access to AWS services and resources securely

AWS CloudFormation

Model your entire infrastructure with either a text file or programming languages

Amazon SageMaker

Fully-managed service that covers the entire machine learning workflow

AWS Step Functions

Coordinate multiple AWS services into serverless workflows so you can build and update apps quickly

AWS Lambda

Run code without provisioning or managing servers

AWS services

AWS CodeCommit

Fully-managed source control service that hosts secure Git-based repositories

AWS CodePipeline

Fully managed continuous delivery service that helps you automate your release pipelines for fast and reliable application and infrastructure updates

Amazon ECR

Fully-managed Docker container registry

AWS CodeBuild

Fully managed continuous integration service that compiles source code, runs tests, and produces software packages that are ready to deploy.

Amazon Kinesis

Collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.

Amazon CloudWatch Events

Near real-time stream of system events

MDLC: Data preparation

MDLC Stage 1: Data sourcing

What is data sourcing?

- To discover and retrieve data for both training and inference

	Dataset 9/23	Dataset 9/24	Dataset 9/25	Dataset Today (9/26)
Order Amount	50	25	75	25
Account Tenure	364	365	366	367
Fraud Status	No	No	Yes	?

MDLC Stage 2: Data quality monitoring

What is data quality monitoring?

- A model is only as accurate as its data quality
- To monitor data completeness, consistency, and accuracy

	Dataset 9/23	Dataset 9/24	Dataset 9/25	Dataset Today (9/26)
Order Amount	50	25	75	25
Account Tenure	364	54	54	54
Fraud Status	No	No	Yes	?

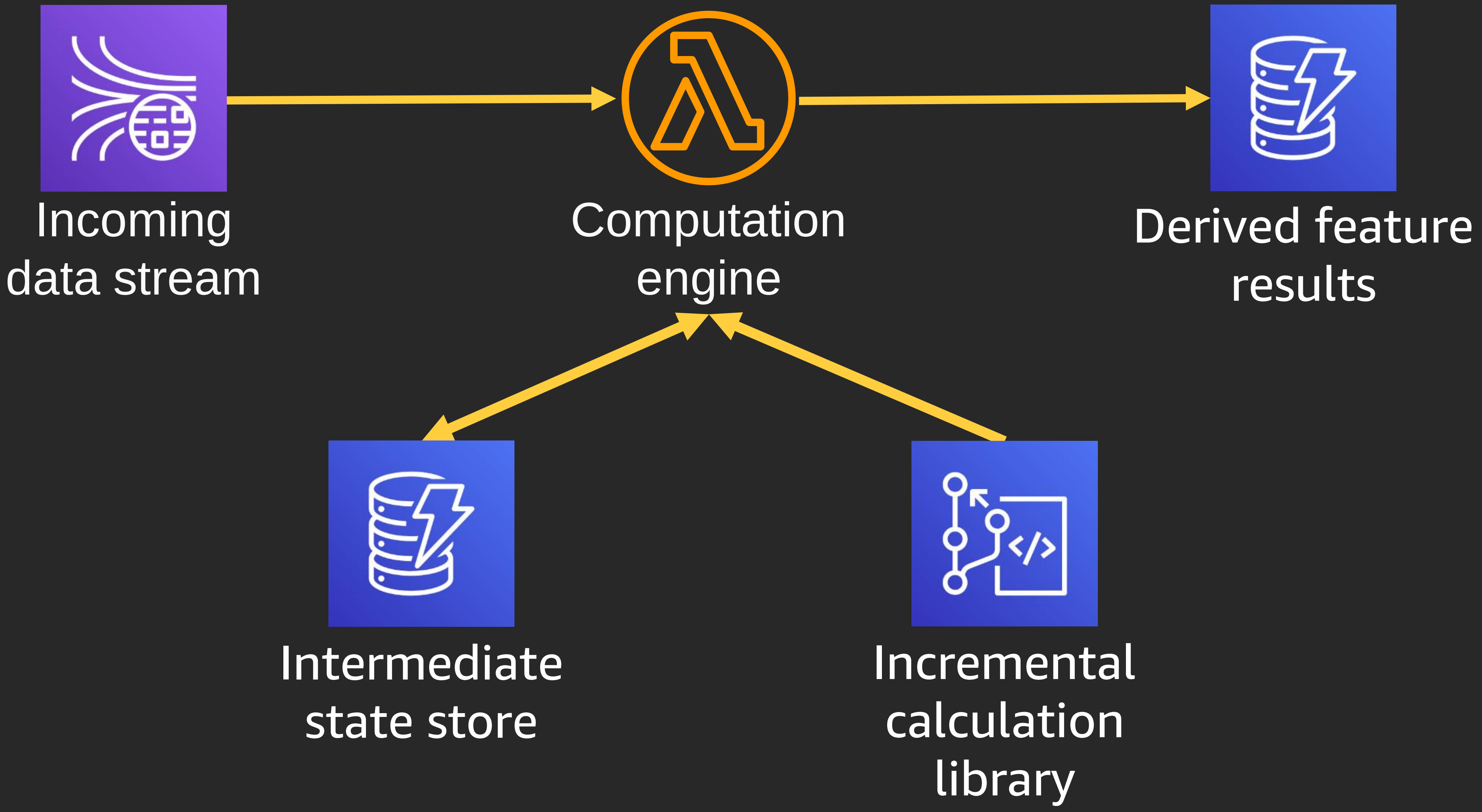
MDLC Stage 3: Feature engineering

What is feature engineering?

- To calculate derived features from raw features
- E.g., total amount of orders a customer made in the past 30 days

	Dataset 9/23	Dataset 9/24	Dataset 9/25	Dataset Today (9/26)
Order Amount	50	25	75	25
Order Total	50	75	150	175
Account Tenure	364	365	366	367
Fraud Status	No	No	Yes	?

MDLC Stage 3: Feature engineering (optimized)

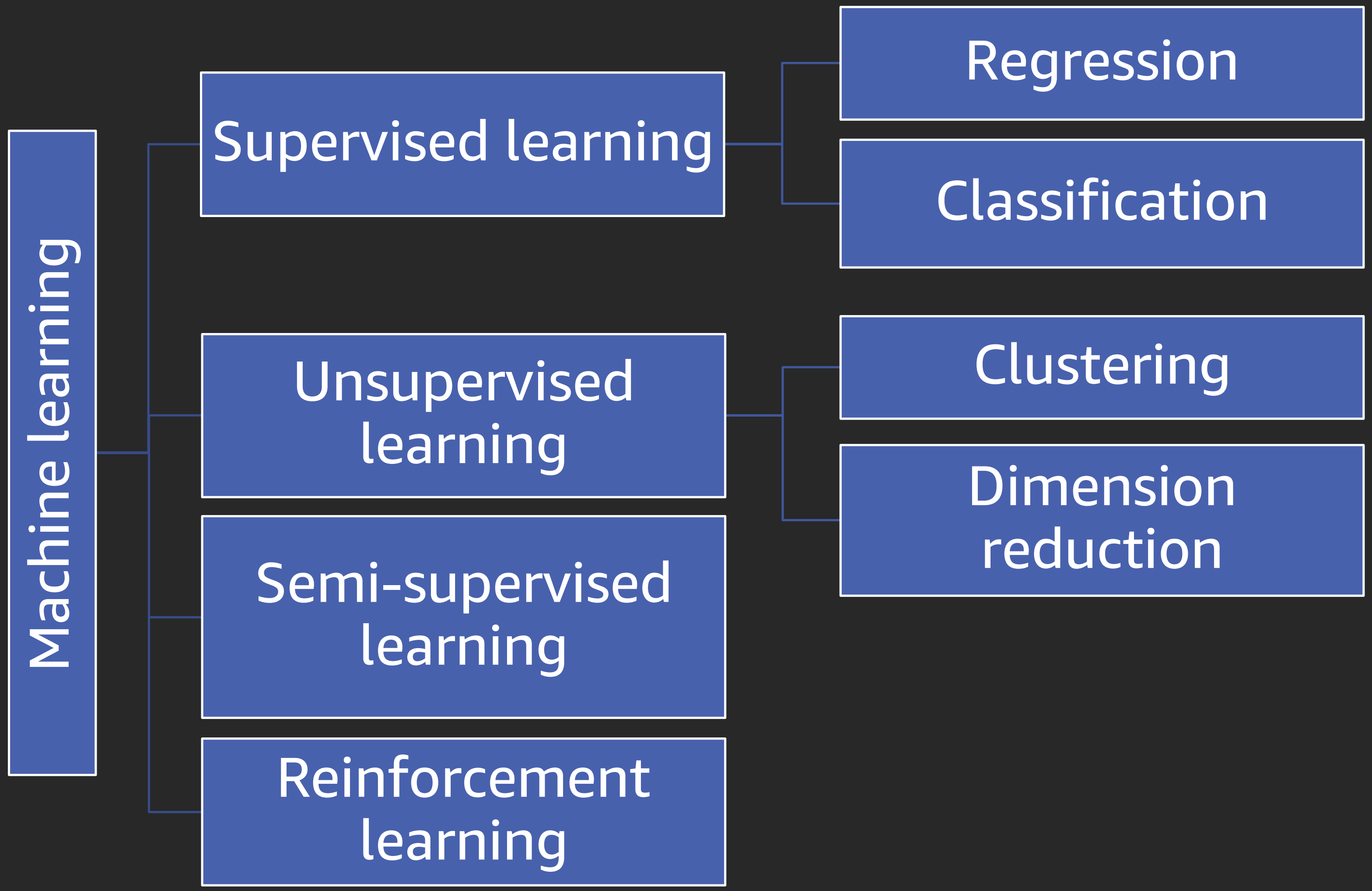


MDLC: Model development

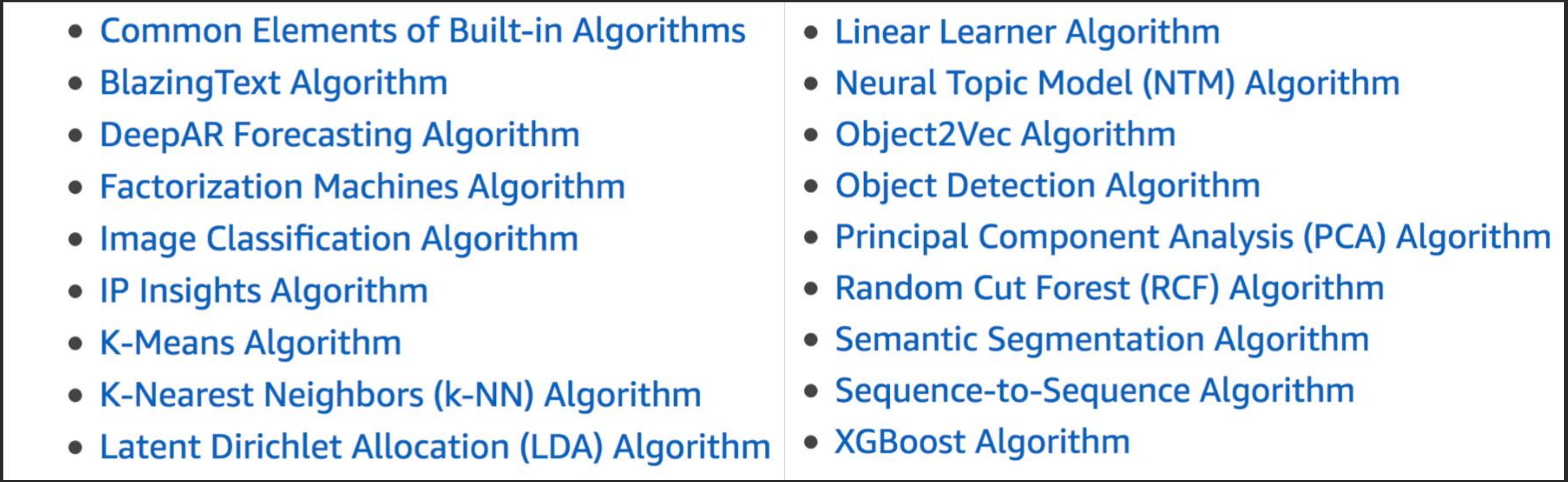
MDLC Stage 4: Model development

What is model development?

- To choose and import an existing ML algorithm, or
- To develop a custom ML algorithm

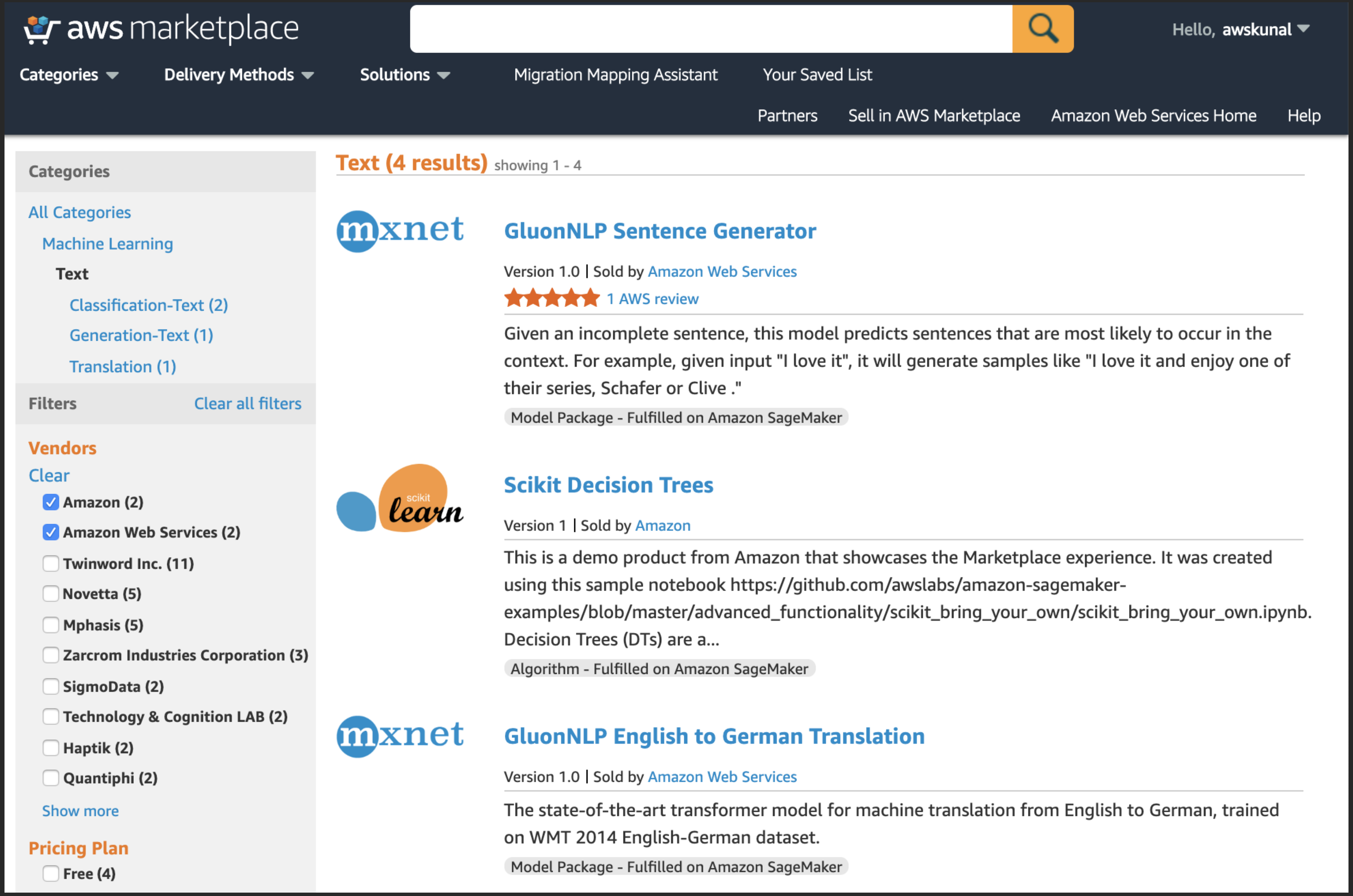


MDLC Stage 4: Model development (built-in)



- Common Elements of Built-in Algorithms
- BlazingText Algorithm
- DeepAR Forecasting Algorithm
- Factorization Machines Algorithm
- Image Classification Algorithm
- IP Insights Algorithm
- K-Means Algorithm
- K-Nearest Neighbors (k-NN) Algorithm
- Latent Dirichlet Allocation (LDA) Algorithm
- Linear Learner Algorithm
- Neural Topic Model (NTM) Algorithm
- Object2Vec Algorithm
- Object Detection Algorithm
- Principal Component Analysis (PCA) Algorithm
- Random Cut Forest (RCF) Algorithm
- Semantic Segmentation Algorithm
- Sequence-to-Sequence Algorithm
- XGBoost Algorithm

MDLC Stage 4: Model development (AWS Marketplace)



Categories

All Categories

Machine Learning

Text

Classification-Text (2)

Generation-Text (1)

Translation (1)

Filters Clear all filters

Vendors

Clear

- ☑ Amazon (2)
- ☑ Amazon Web Services (2)
- □ Twinword Inc. (11)
- □ Novetta (5)
- □ Mphasis (5)
- □ Zarcrom Industries Corporation (3)
- □ SigmoData (2)
- □ Technology & Cognition LAB (2)
- □ Haptik (2)
- □ Quantiphi (2)

Show more

Pricing Plan

- □ Free (4)

Text (4 results) showing 1 - 4

GluonNLP Sentence Generator

Version 1.0 | Sold by Amazon Web Services

★★★★★ 1 AWS review

Given an incomplete sentence, this model predicts sentences that are most likely to occur in the context. For example, given input "I love it", it will generate samples like "I love it and enjoy one of their series, Schafer or Clive ."

Model Package - Fulfilled on Amazon SageMaker

Scikit Decision Trees

Version 1 | Sold by Amazon

This is a demo product from Amazon that showcases the Marketplace experience. It was created using this sample notebook https://github.com/awslabs/amazon-sagemaker-examples/blob/master/advanced_functionality/scikit_bring_your_own/scikit_bring_your_own.ipynb. Decision Trees (DTs) are a...

Algorithm - Fulfilled on Amazon SageMaker

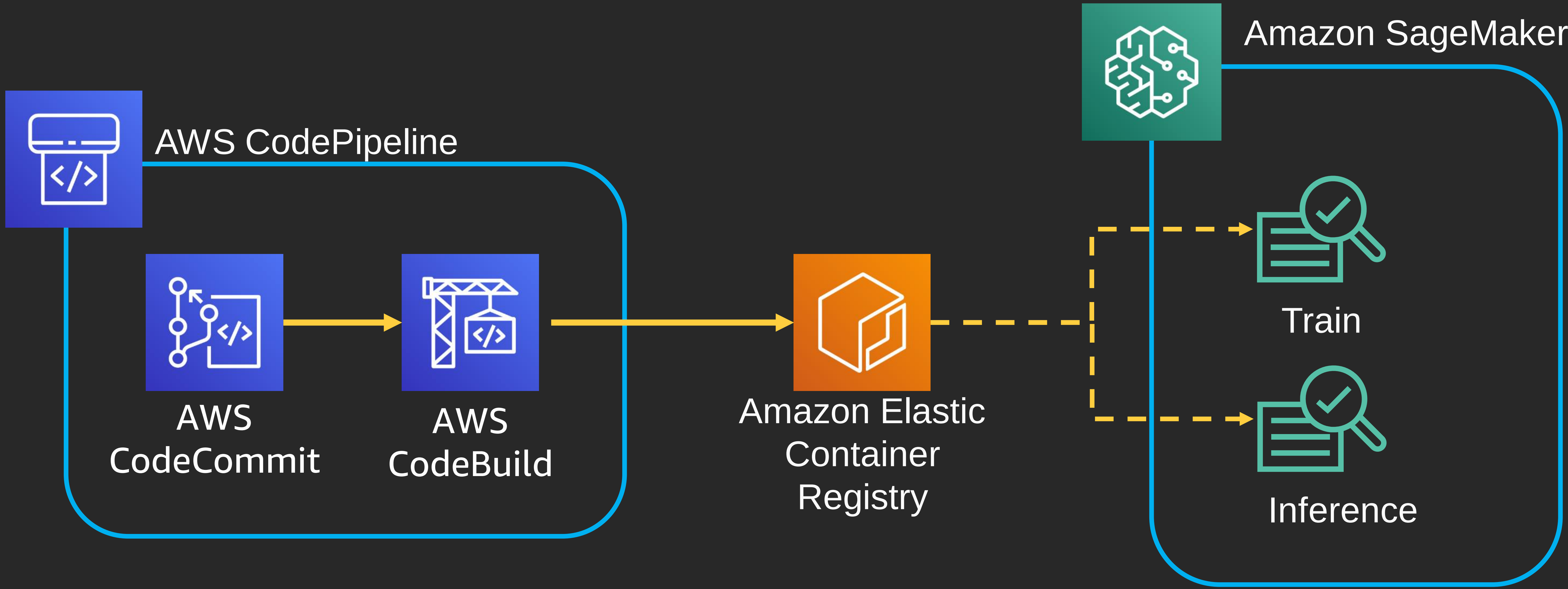
GluonNLP English to German Translation

Version 1.0 | Sold by Amazon Web Services

The state-of-the-art transformer model for machine translation from English to German, trained on WMT 2014 English-German dataset.

Model Package - Fulfilled on Amazon SageMaker

MDLC Stage 4: Model development (custom)

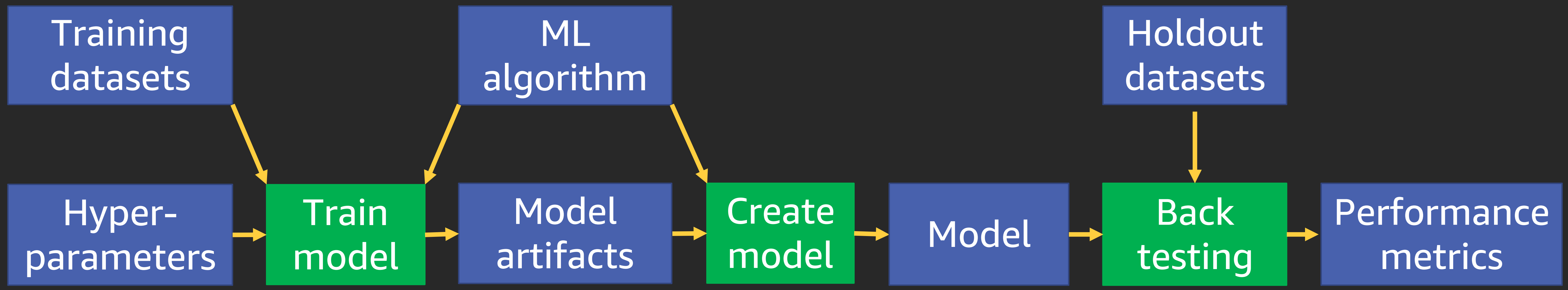


MDLC: Training and deployment

MDLC Stage 5: Model training & evaluation

What is model training & evaluation?

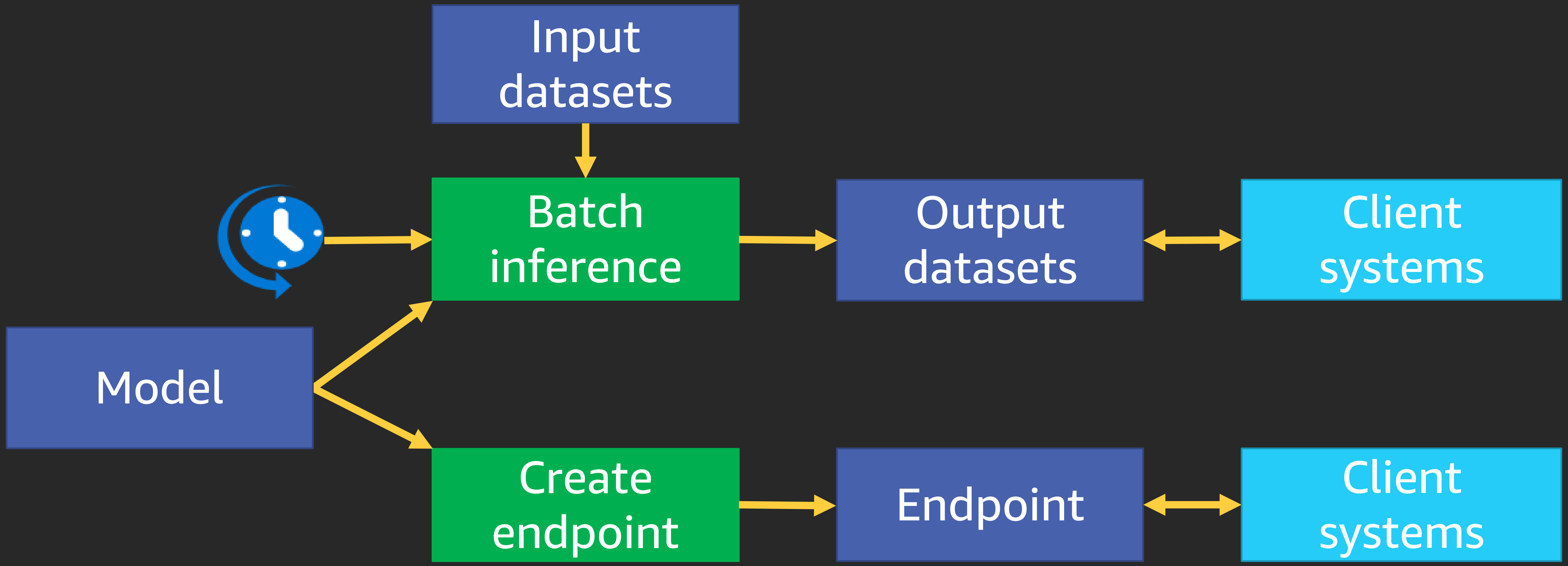
- Derive rules by learning from the past data using an ML algorithm
- Evaluate how effective the rules are before using it in production



MDLC Stage 6: Model deployment & inference

What is model deployment and inference?

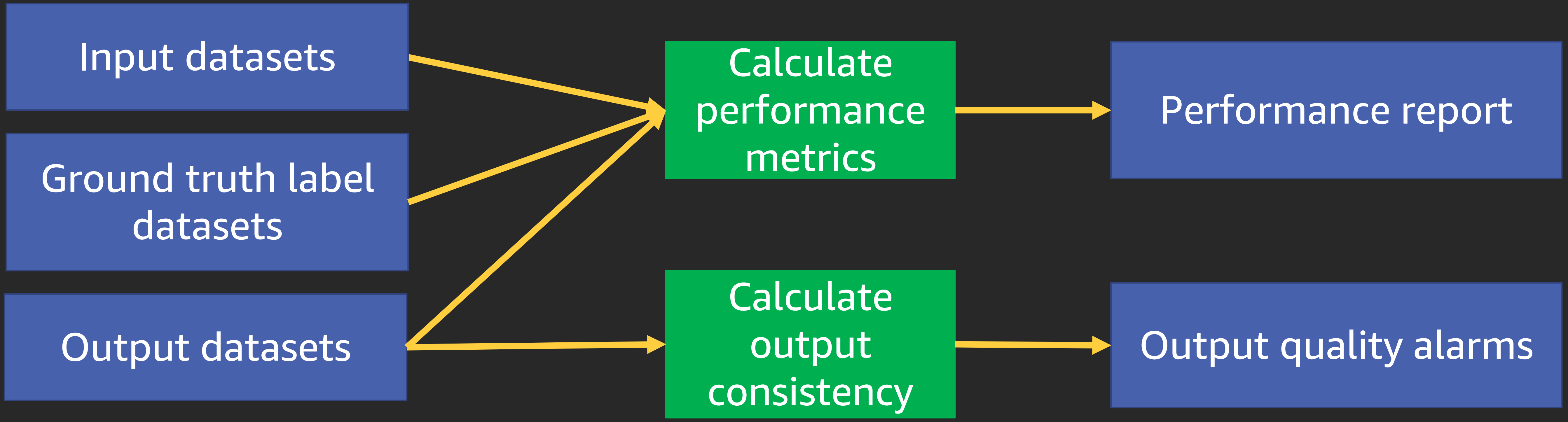
- Deploy a trained model to production to run offline inferences
- Deploy a trained model to production to run real-time inferences



MDLC Stage 7: Model monitoring

What is model monitoring?

- Monitor model output accuracy (closed-loop)
- Monitor model output consistency (open-loop)


Input datasets
Ground truth label datasets
Output datasets
Calculate performance metrics
Calculate output consistency
Performance report
Output quality alarms

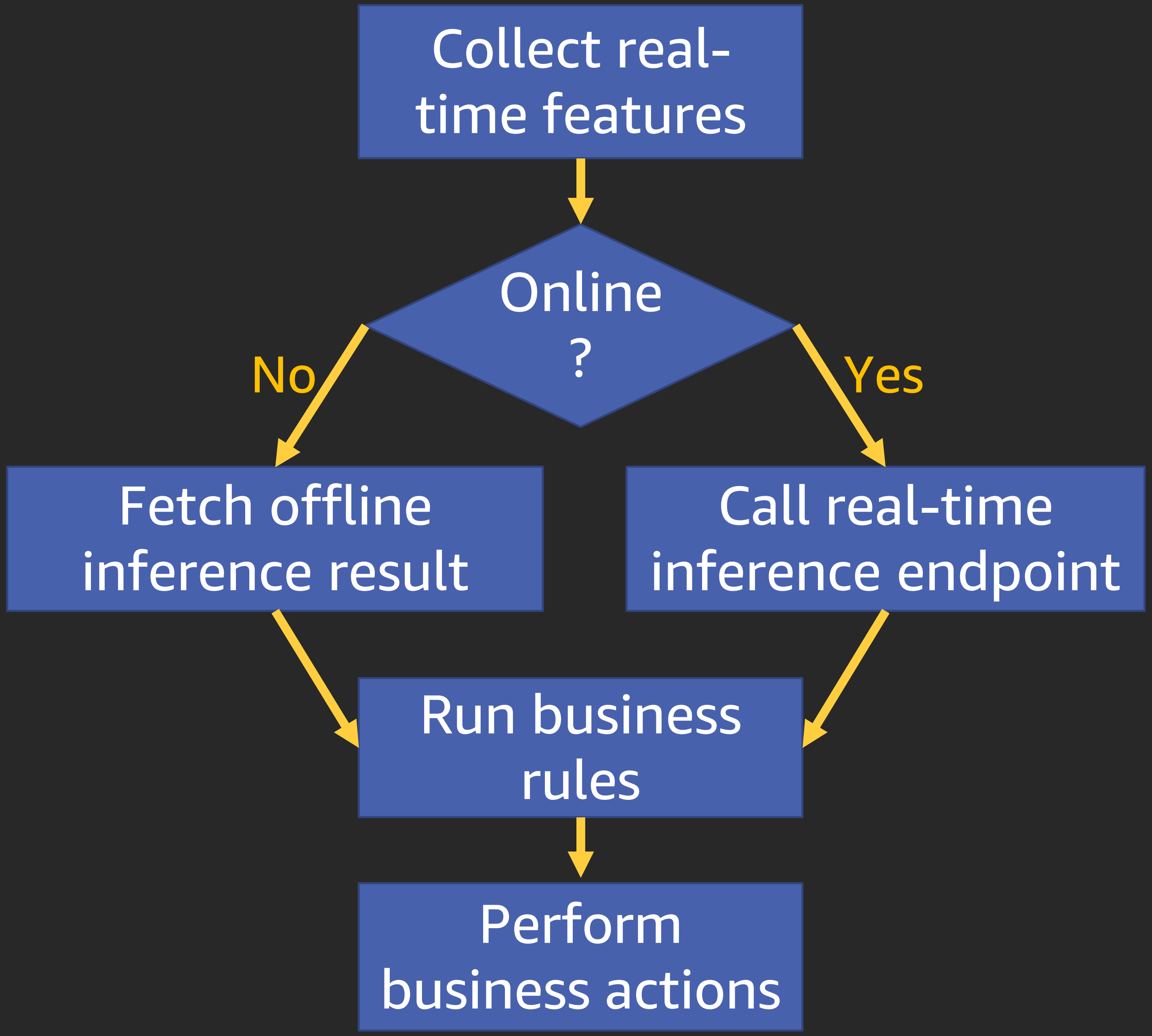
MDLC Stage 8: Client integration

What is client integration

- Consume model results and turn them into actions

What tools/methods did we use?

- Collect Features (Amazon DynamoDB)
- Run Inference (Amazon SageMaker)
- Execute Rules (AWS Lambda)
- Perform Action (AWS Lambda)


Collect real-time features
Online ?
No
Yes
Fetch offline inference result
Call real-time inference endpoint
Run business rules
Perform business actions

MDLC: Bringing it all together

Challenges of ML model life cycle management

What are the challenges?

- Manual processes incur 50-80% inefficiency
- Deciding who owns what in the entire end-to-end ML life cycle
- Versioning, auditing
- Safely deploying models to production
- Scaling to many models with frequent re-training
- Sharing best practices among the data scientist team

Solution: – Ignore it

Keep it manual

- A valid option to be scrappy at first
- Risks that can impact the business
- Ongoing maintenance costs

Image source: PublicDomainPictures.net

Solution 2: MDLC workflows

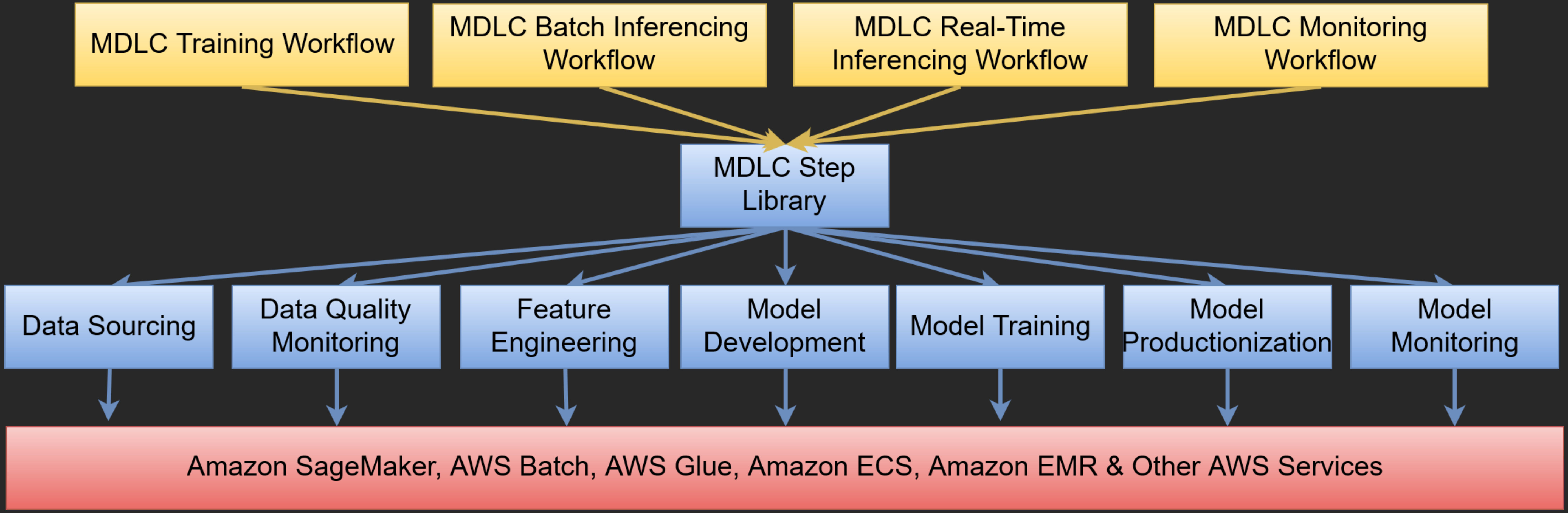
What are MDLC workflows?

- Templates for automating the ML model life cycle
- Built by using Amazon SageMaker, AWS Step Functions, and other AWS services

Templates for:

- Model training
- Batch inference
- Real-time inference
- Model monitoring
- Model re-training

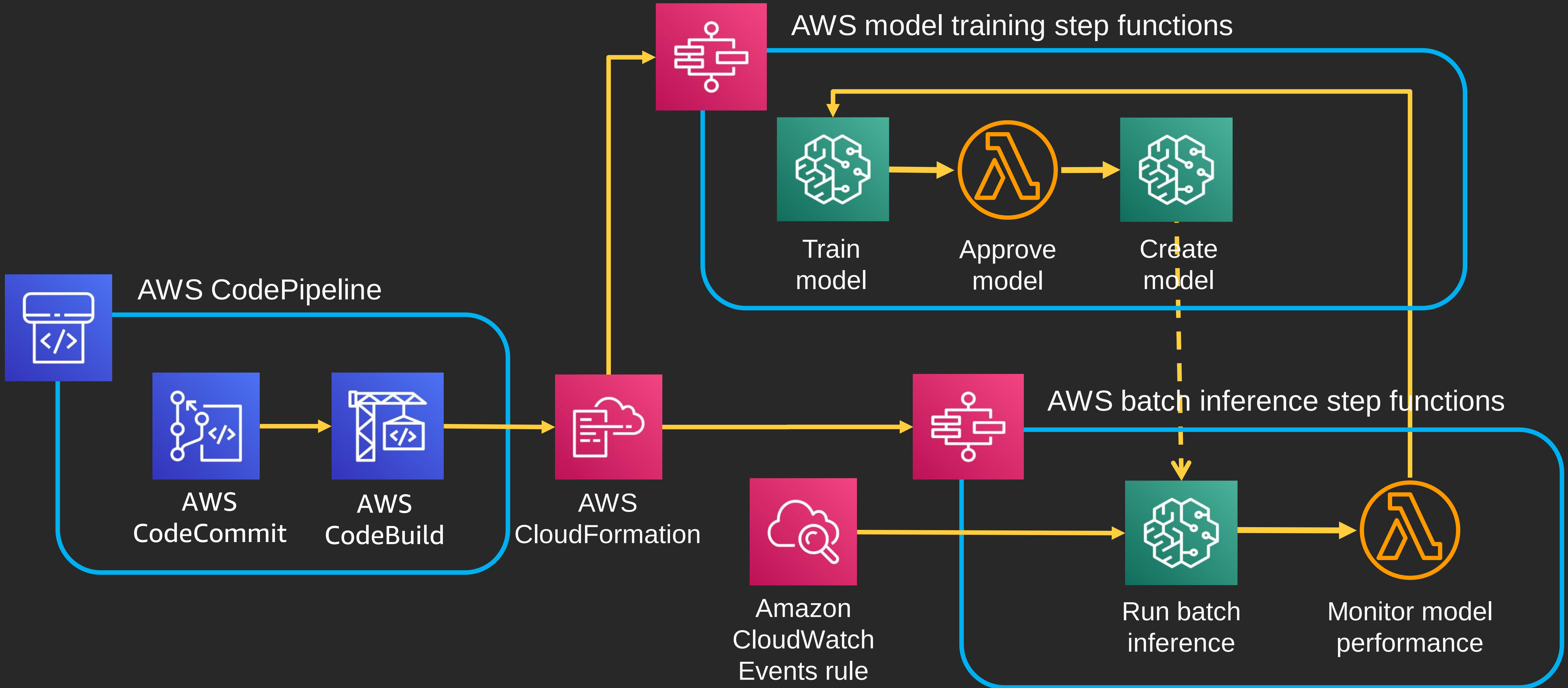
MDLC workflows: High-level components



MDLC workflows: Example 1

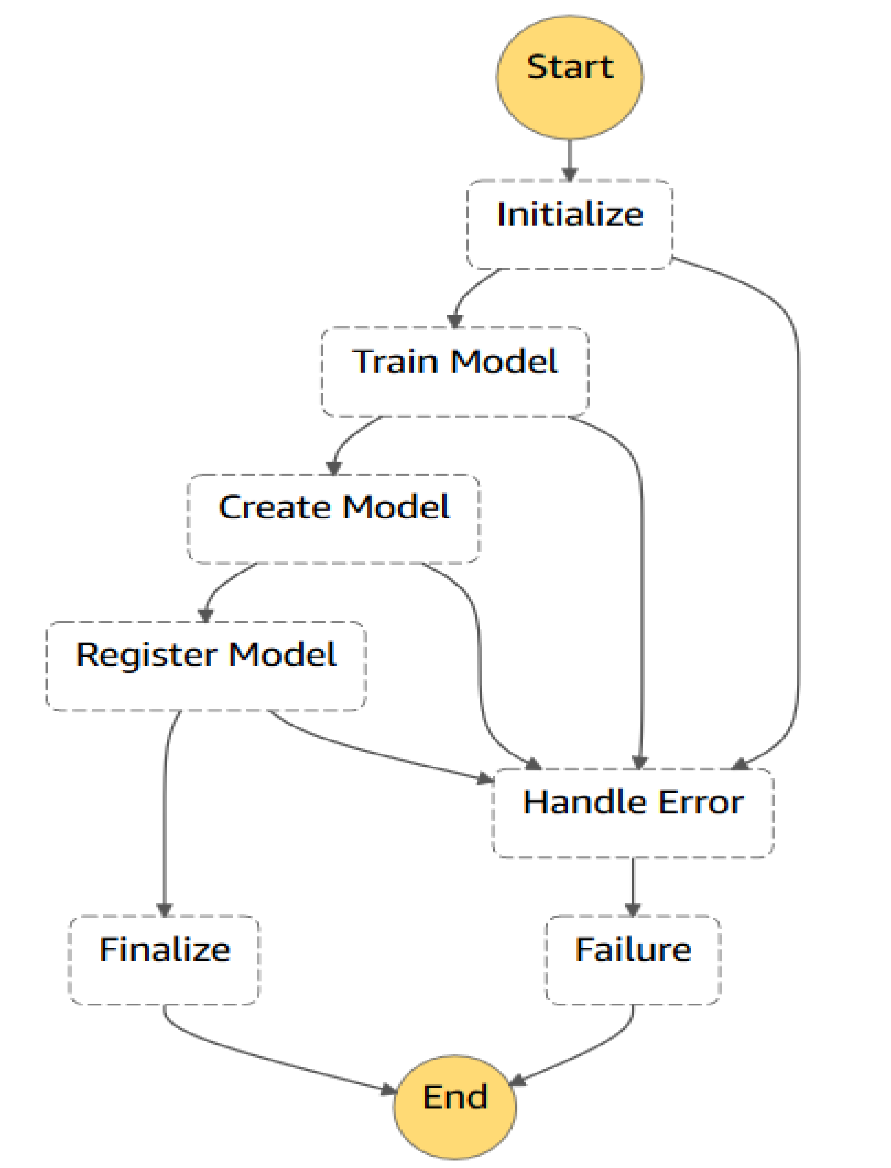
Batch (Offline) Inference Workflow

MDLC batch (offline) inference model workflow

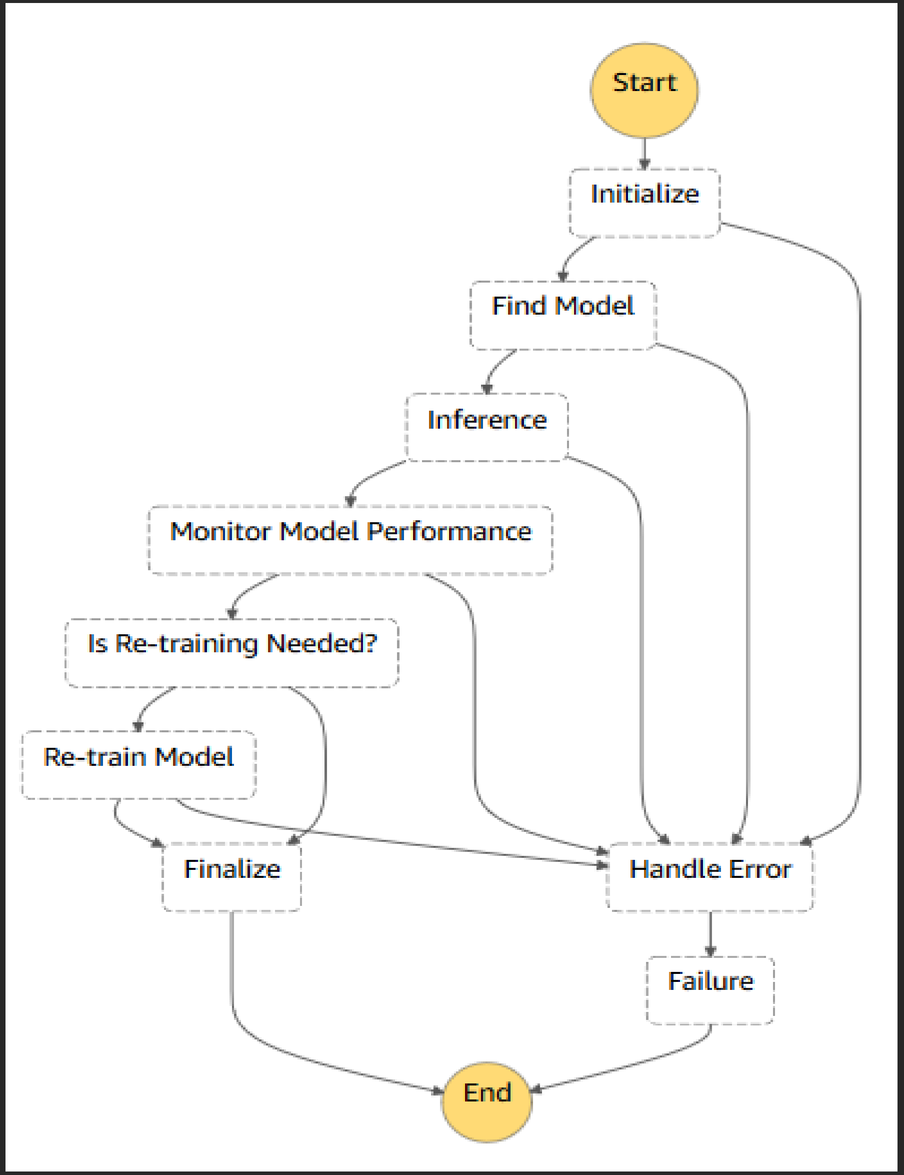

AWS model training step functions
Train model
Approve model
Create model
AWS CodePipeline
AWS CodeCommit
AWS CodeBuild
AWS CloudFormation
AWS batch inference step functions
Amazon CloudWatch Events rule
Run batch inference
Monitor model performance

MDLC workflow: AWS Step Functions

- State machine for training



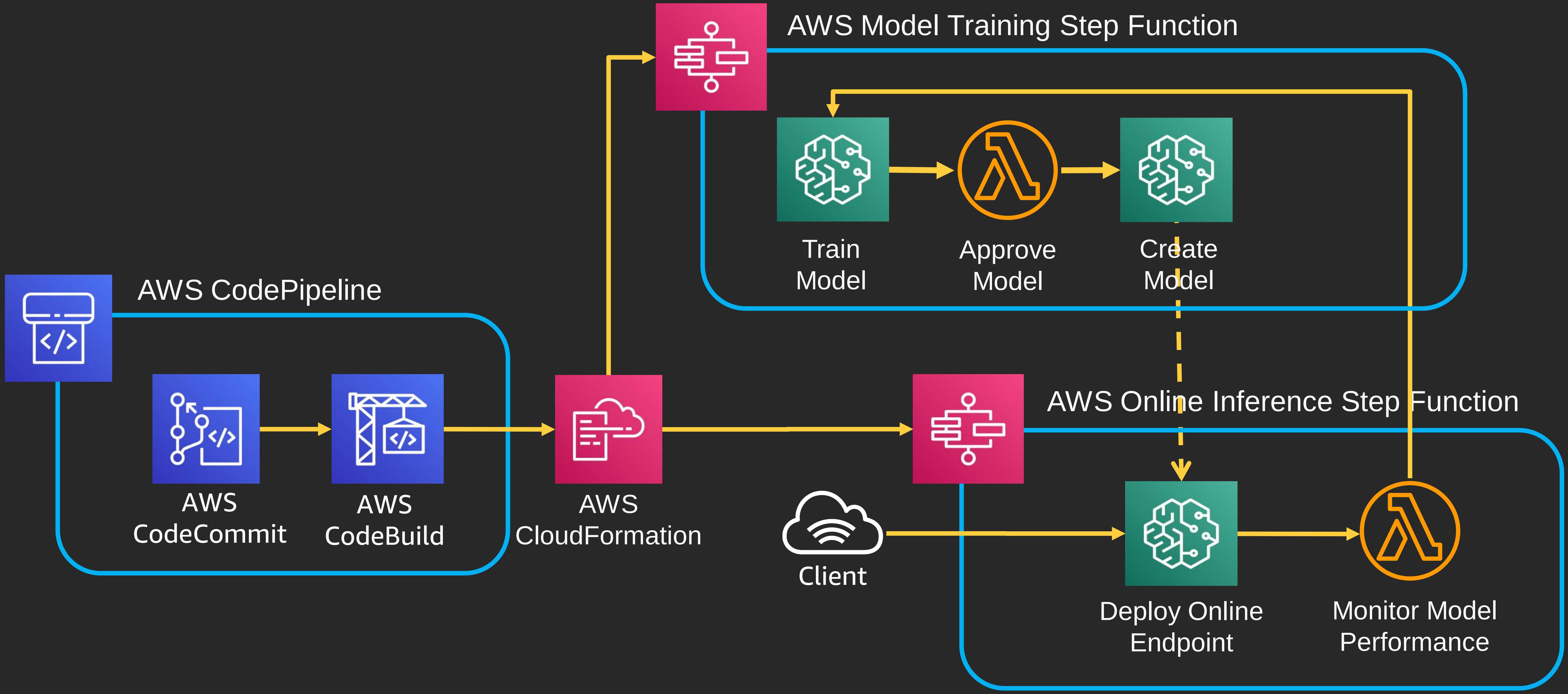
- State machine for inference



MDLC Workflows: Example 2

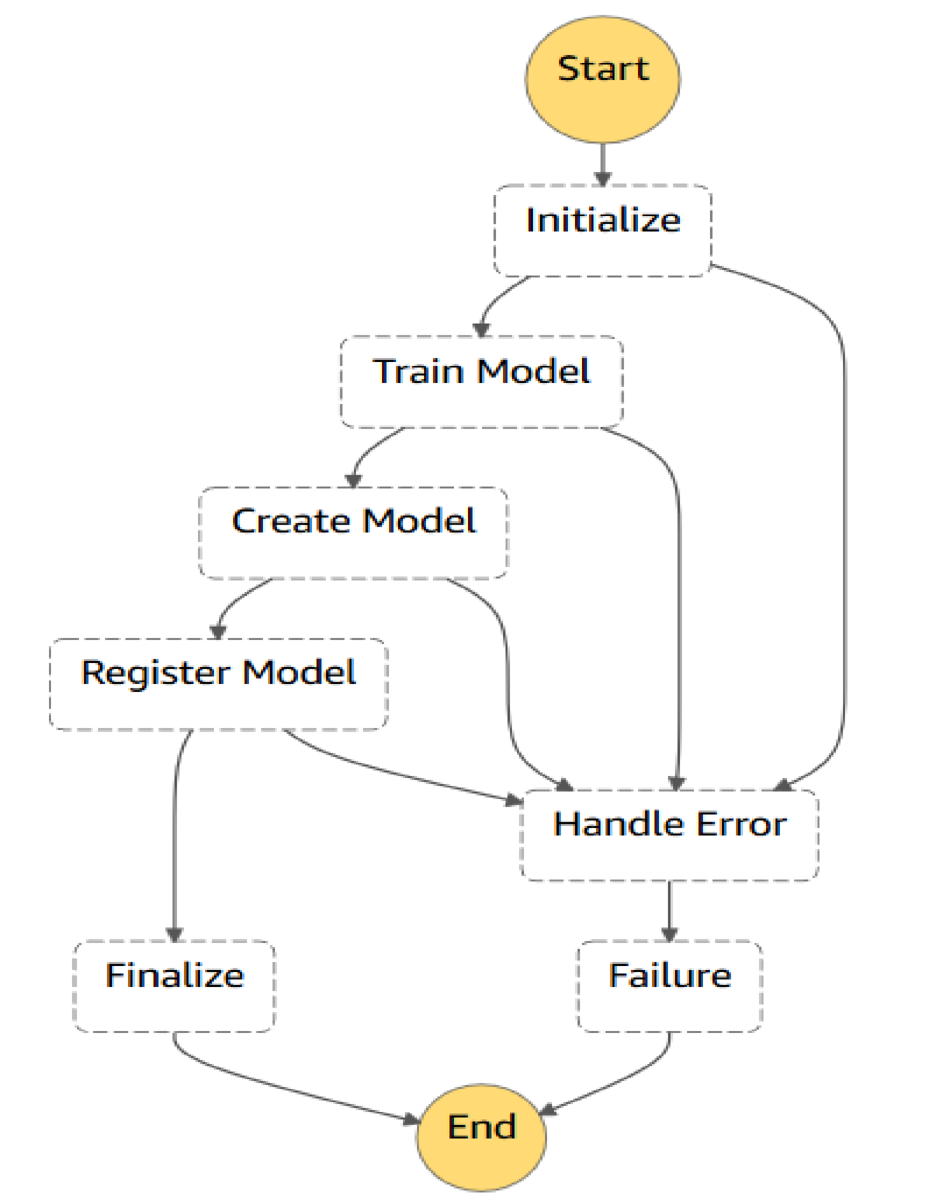
Real-Time (Online) Inference Workflow

Real-time (online) inference workflow

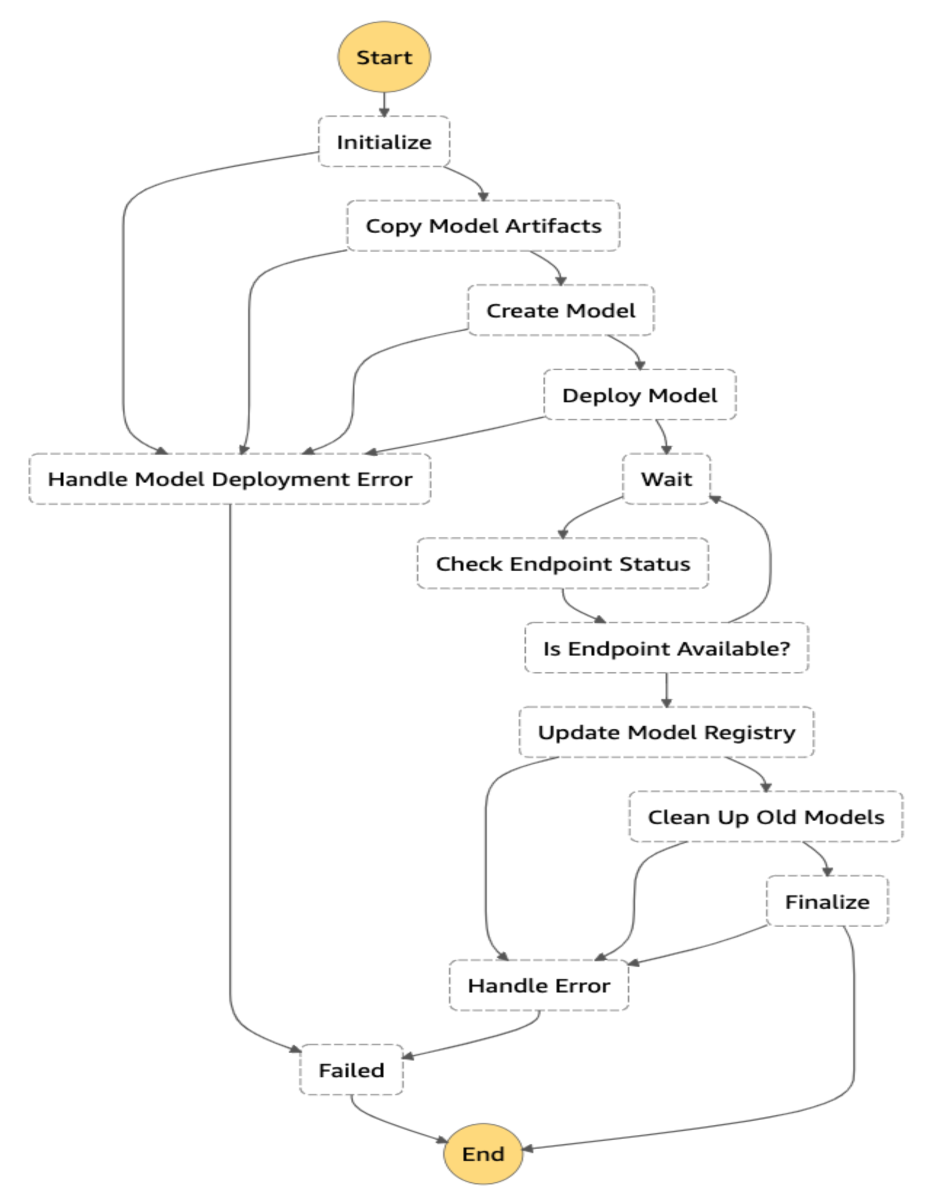


MDLC workflow: AWS Step Functions

- State machine for training

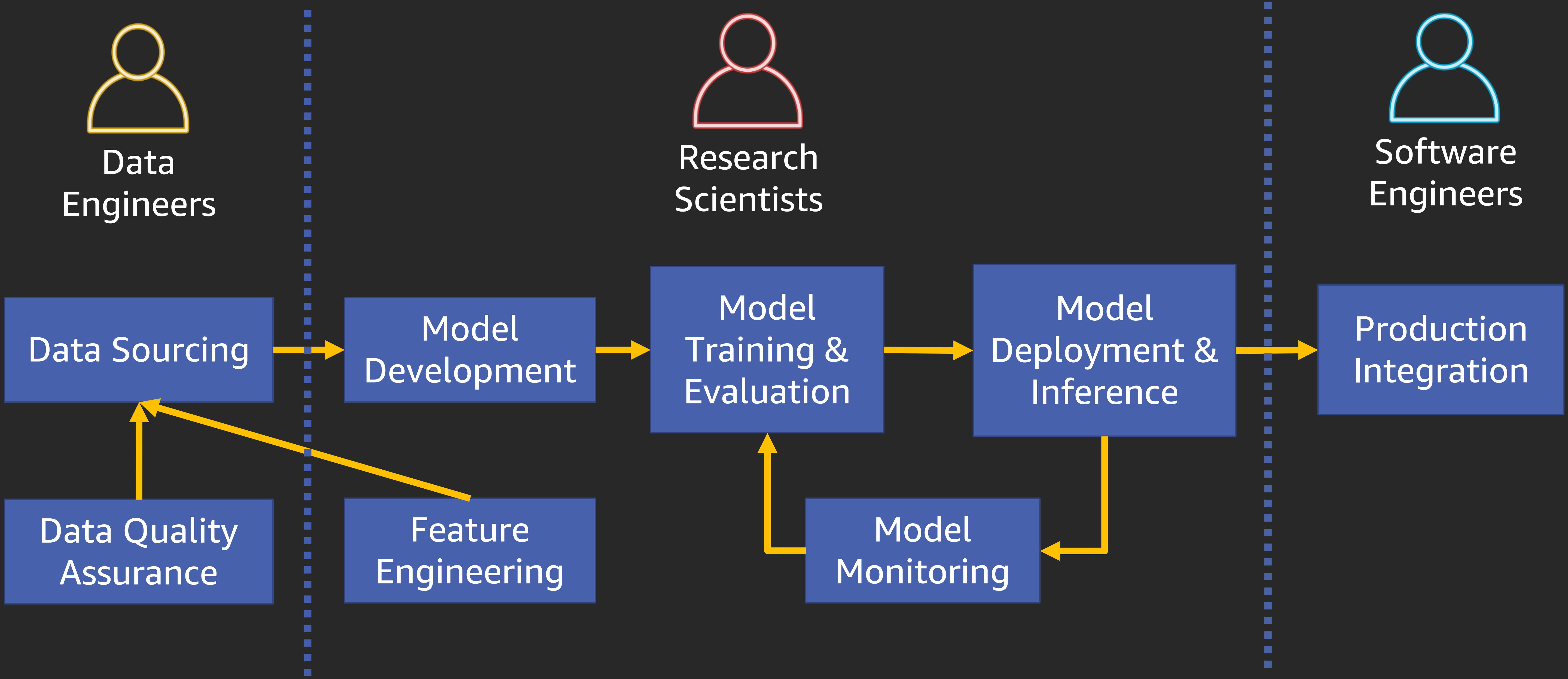


- State machine for inference



Summary

Model Development Life Cycle (MDLC)



MDLC workflows: Benefits

What are the benefits of MDLC workflows?

- Rapid bootstrapping of ML pipelines
- Integrate and leverage existing technologies
- Continuous Deployment/Integration
- Versioning & auditing
- Full ownership for the research scientist team to deploy and iterate their models

Where to learn more

Machine Learning University

https://aws.training/machinelearning

Uses the same materials used to train Amazon developers

Foundational knowledge with real-world application

Structured courses and specialist certification

Learn to architect with AWS Training and Certification

Resources created by the experts at AWS to propel your organization and career forward

Free foundational to advanced digital courses cover AWS services and teach architecting best practices

Classroom offerings, including Architecting on AWS, feature AWS expert instructors and hands-on labs

Validate expertise with the **AWS Certified Solutions Architect - Associate** or **AWS Certification Solutions Architect - Professional** exams

Visit aws.amazon.com/training/path-architecting/

aws training and certification

Questions?

Thank you!

Fei Yuan
Senior Software Engineer
Amazon Consumer Payments

Kieran Kavanagh
Senior Solutions Architect
Amazon Web Services

Kunal Batra
Senior Technical Evangelist
Amazon Web Services

Please complete the session survey in the mobile app.