
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

AIM302

Deploy ML models for inference at high performance & low cost

Venkatesh Krishnan
Principal Product Manager
Amazon Web Services

Rama Thamman
Sr. Manager
Amazon Web Services

Antoine Diffloth
Director Data Insights,
AT&T Chief Data Office

Matthew Schneid
Chief Architect, USM
AT&T Cybersecurity

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

By 2026, one in every 3 applications will have ML infused in it

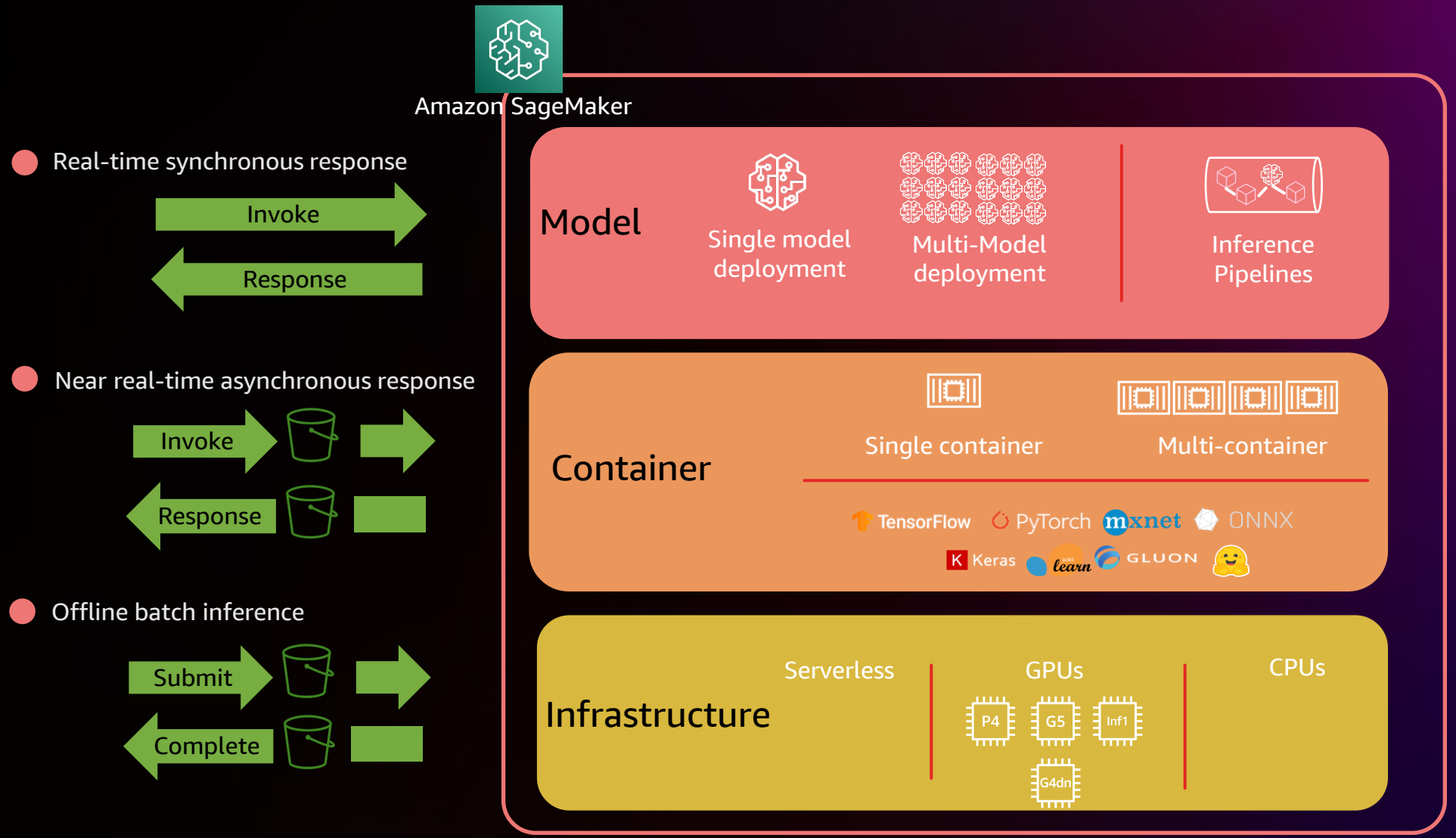
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Amazon SageMaker

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Deploy model to serve inference



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Model deployment – How do you strike the right balance?



Complexity
- Model size
- Payload size
- Inference workflows

Performance
- Latency
- Throughput
- Availability

Cost
- Model hosting cost
- Operational overhead
- Number of models to deploy and manage

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



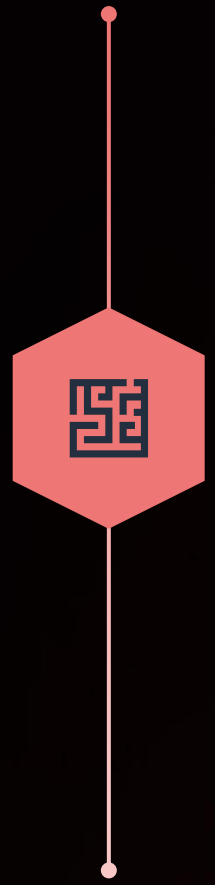
Complexity

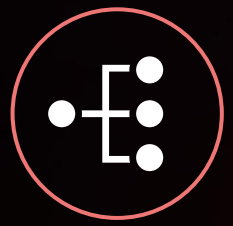
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Deployment complexity

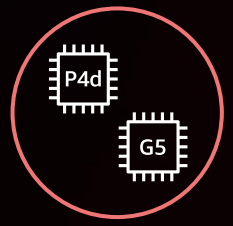
Large ML models

with 100 billion + parameters





Easily parallelize models across multiple GPUs to fit models into the instance and achieve low latency



Deploy models on the most performant and cost-effective GPU-based instances or on AWS Inferentia



Leverage 500GB of Amazon EBS volume per endpoint

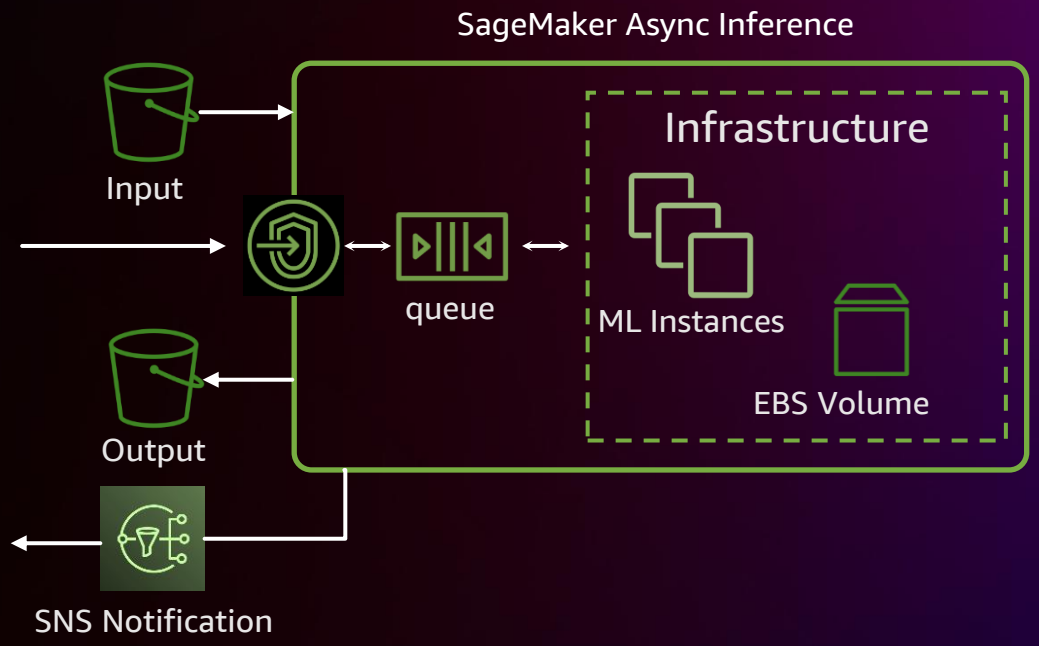
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Deployment complexity

Large payloads or long running inference





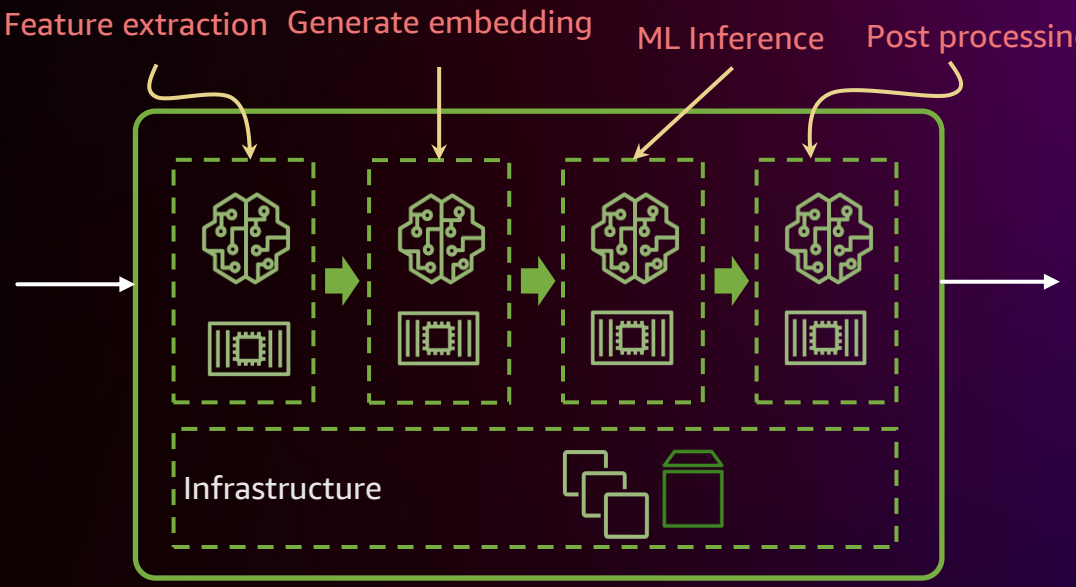
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Deployment complexity

Inference workflows with pre- and post processing steps



SageMaker serial inference pipelines



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

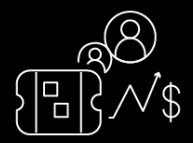


Performance

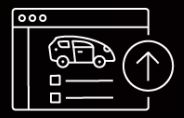
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Inference performance

ML-BASED APPLICATIONS HAVE DIVERSE AND STRINGENT PERFORMANCE REQUIREMENTS



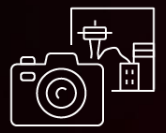
Ad serving



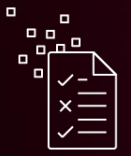
Personalized recommendations



Fraud detection



Object detection



Document analysis

• • •

Latency

Ultra-low latency for real-time interactions

Availability

Minimize downtime and improve reliability

Throughput

Serve millions of transactions per second

Scale

Thousands of models – one per person

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

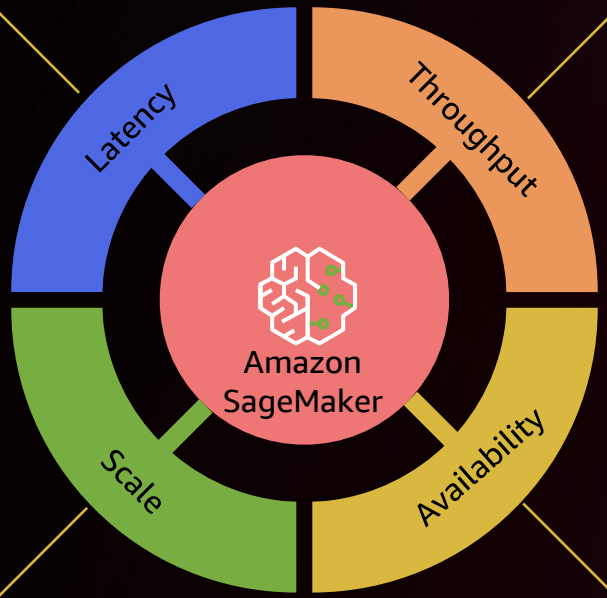
SageMaker for best inference performance

Under 5 ms for small payloads

70+ instances to fit performance requirements
Ability to co-locate endpoint with application

1M+ transactions per second

Built-in routing layer to distribute your traffic and prevent overloading



Deploy thousands of models

Use multi-model endpoints

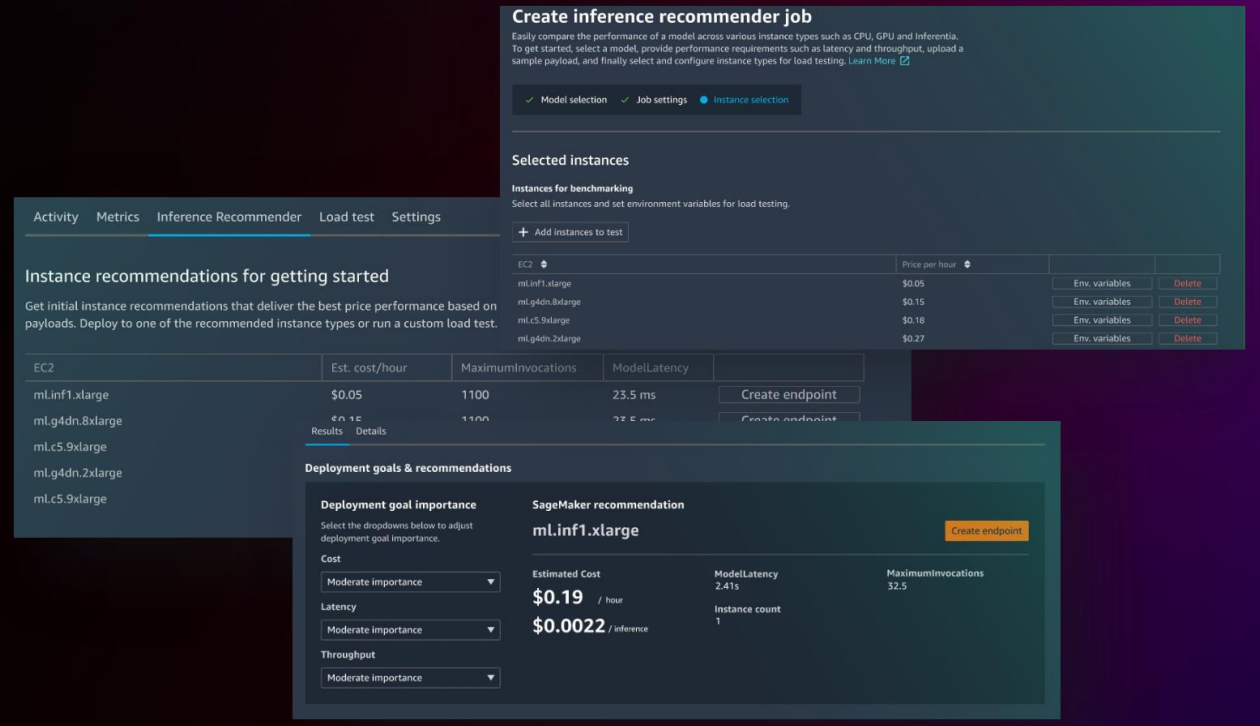
99.95% availability SLA

- Automatically deploy across multiple AZs
- Routine health checks and replacement of unhealthy instances

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker Inference Recommender

- Get instance recommendation in just a few clicks
- Run extensive load tests to optimize cost, latency, and throughput
- Deploy models with confidence



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



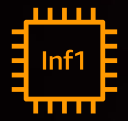
Cost

- Infrastructure costs
- Operations and maintenance costs

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Infrastructure cost

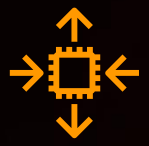
Choose the right instances



Inferentia: 2.3x throughput with 70% lower cost compared to GPU instances



Graviton 2 and 3: 40% better price performance over comparable current generation x86-based instances



Autoscaling: Provision instances dynamically to meet traffic pattern; auto scales in a few minutes

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Deploy on SageMaker Serverless Inference

1 
ECR image location for inference code

2 
S3 location for model artifacts

3 
Choose a memory size

Serverless Inference endpoint



Automatically spins up and manages compute resources

Automatic scaling based on demand

Managed logging and monitoring

Sends inference requests

Inference results

Trigger from client application or other AWS services



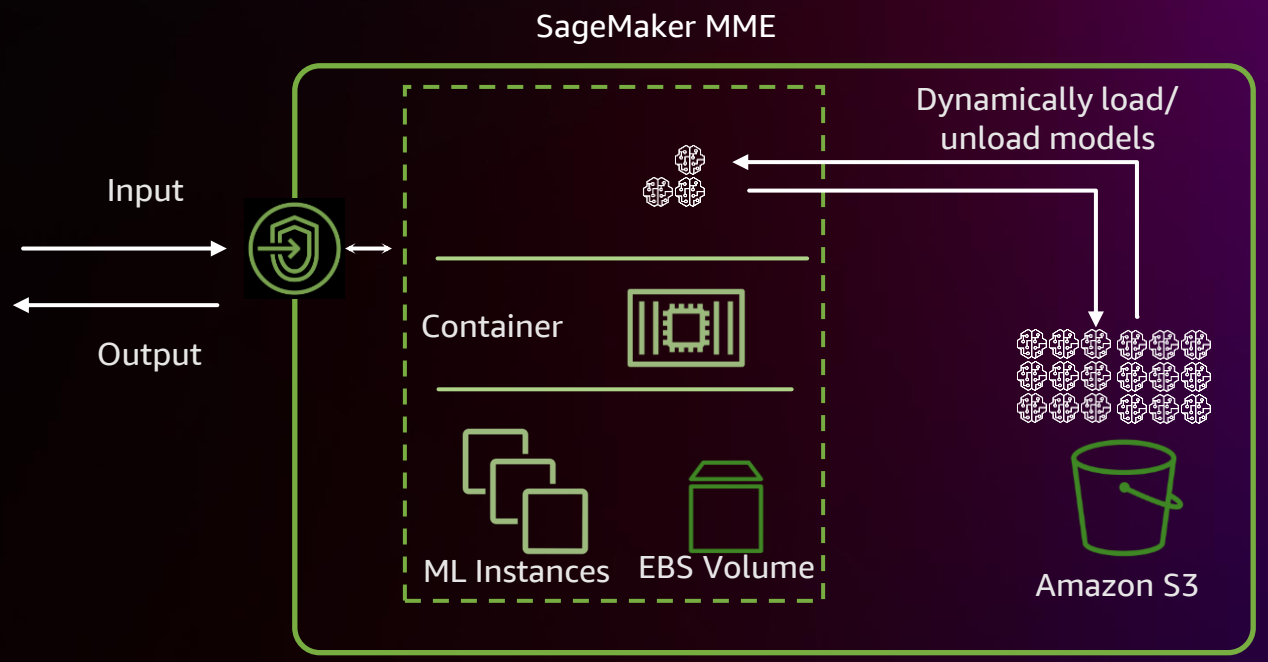
End user

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker Multi-model endpoints

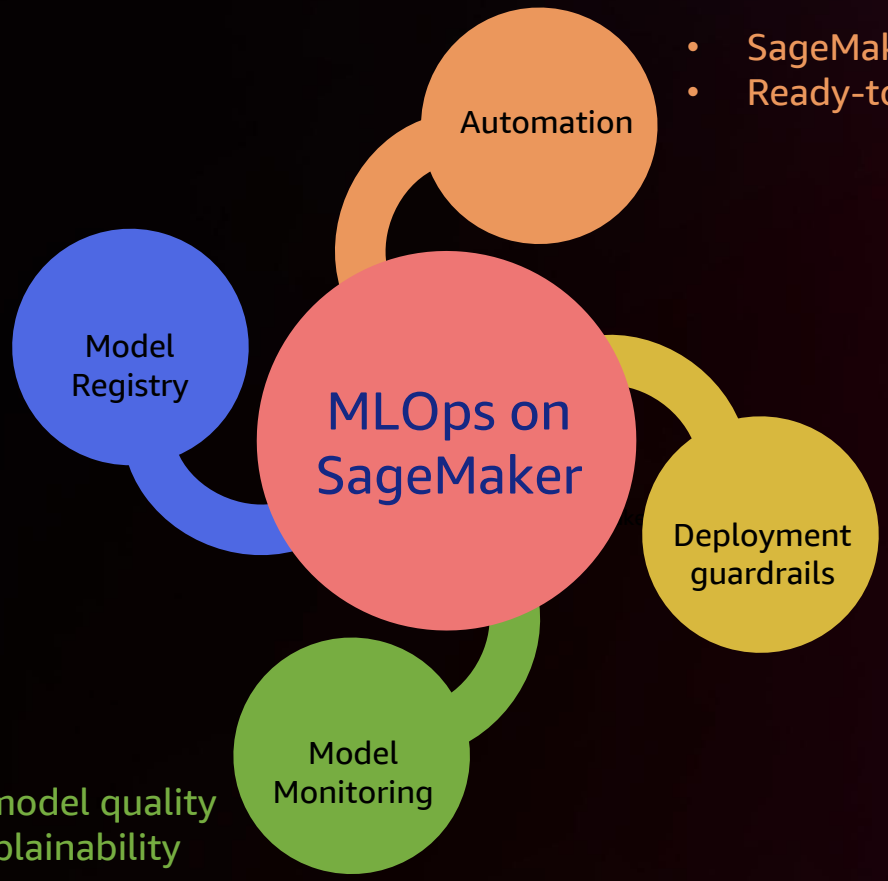
- Deploy thousands of models to a single endpoint
- Save 90% of cost deploying personalization models



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cost of operations

SAGEMAKER HELPS ACHIEVE MLOPS EXCELLENCE



- SageMaker Pipelines to automate model workflow
- Ready-to-use CI/CD templates via SageMaker projects

- Catalog models and track metrics
- End-to-end lineage tracking

- No loss of availability during updates
- Blue/green deployment
- Canary and linear traffic shifting

- Measure model quality
- Online explainability

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Should I use Amazon SageMaker or build my own deployment platform?

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Inference on Amazon SageMaker

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

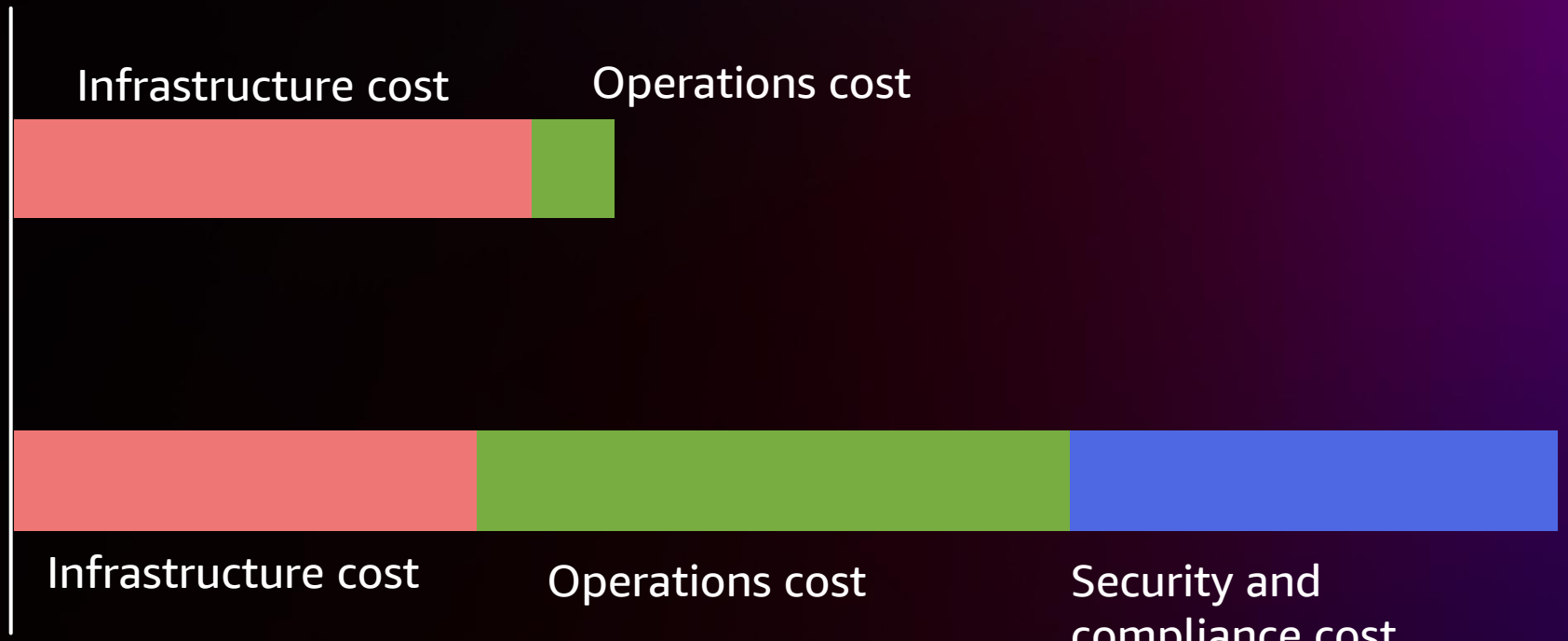
Cost optimization tooling

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

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Save costs by deploying on SageMaker



- Compute instances
- Storage
- Network

Operating, managing, and maintaining infrastructure

Security and compliance for ML features, encrypt data and models, access policies, track and trace

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Model deployment

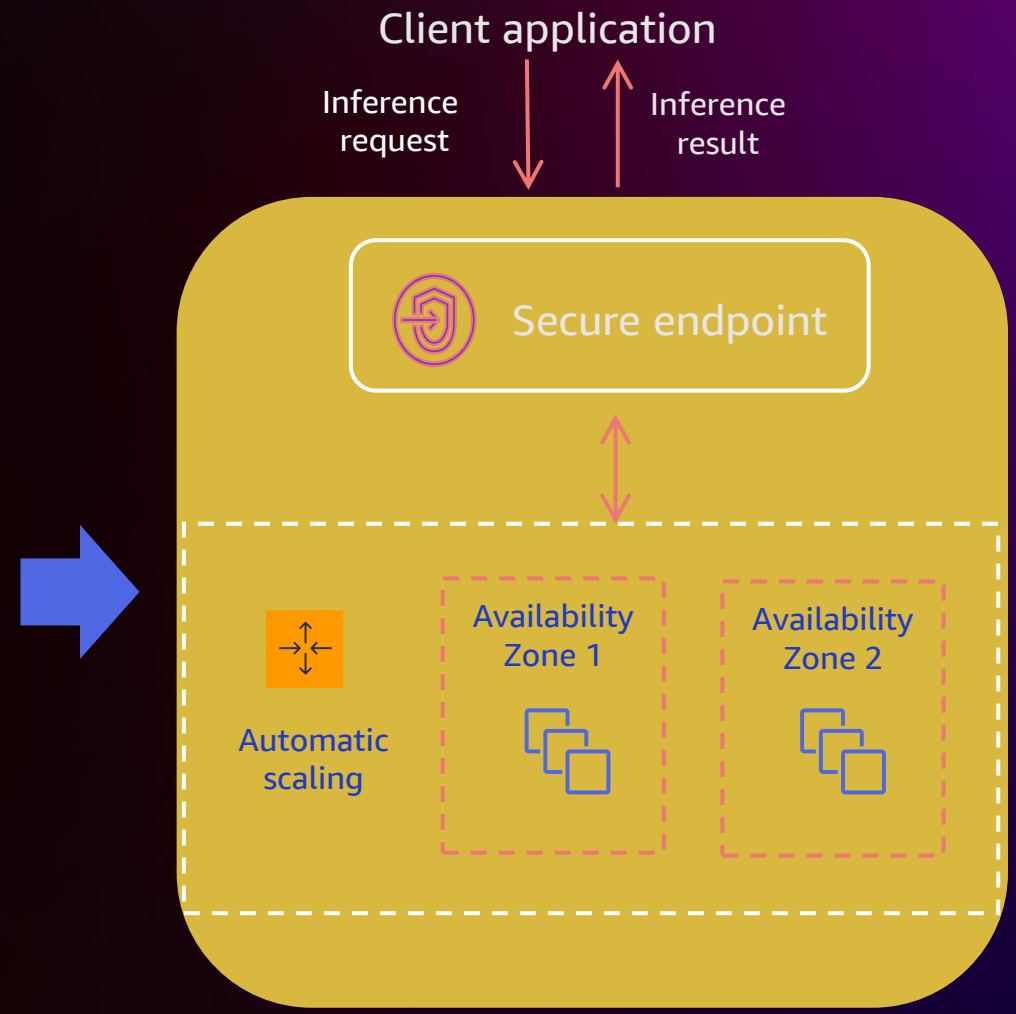
```
import sagemaker

from sagemaker.tensorflow.model import
TensorFlowModel

model_data_file = ”<location>"

sagemaker_model =
TensorFlowModel(model_data=model_data_file,
entry_point=None,….))

predictor =
sagemaker_model.deploy(initial_instance_count=
6, instance_type='ml.m4.xlarge')
```



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Demo

- Train models to predict house prices
- One model per city
- Deploy the models behind a multi-model endpoint

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi-model endpoint demo

1

Amazon S3
model storage



s3://bucket/your-endpoint-models/

Chicago_IL.tar.gz

Dallas_TX.tar.gz

Houston_TX.tar.gz

SanDiego_CA.tar.gz

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi-model endpoint demo

1

Amazon S3
model storage



s3://bucket/your-endpoint-models/

Chicago_IL.tar.gz

Dallas_TX.tar.gz

Houston_TX.tar.gz

SanDiego_CA.tar.gz

2



Amazon SageMaker
Multi-model endpoint

Mode: MultiModel

Artifact location:
s3://bucket/your-endpoint-models

Chicago_IL.tar.gz	Houston_TX.tar.gz
Dallas_TX.tar.gz	SanDiego_CA.tar.gz

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi-model endpoint demo

1

Amazon S3
model storage



s3://bucket/your-endpoint-models/

Chicago_IL.tar.gz

Dallas_TX.tar.gz

Houston_TX.tar.gz

SanDiego_CA.tar.gz

2



Amazon SageMaker
Multi-model endpoint

Mode: MultiModel

Artifact location:
s3://bucket/your-endpoint-models

Chicago_IL.tar.gz	Houston_TX.tar.gz
Dallas_TX.tar.gz	SanDiego_CA.tar.gz

3



Predict

```
predict('nevada.tar.gz',
features)
```

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

amazon-sagemaker-examples | AWS Management Console | CloudWatch Management Con

github.com/aws/amazon-sagemaker-examples/blob/main/advanced_functionality/multi_model_xgboost_home_value/xgboost_multi_model_endpoint_home_value.ipynb

Update

Reading List | AWS Services | Account | TFC | Hiring | Self | Tools | SM SA | AI/ML | To Do | Quip - SA | Imported From Fir... | RTM Deck | (37) AI/ML Org Pr... | AWS Wisdom Sag...

Search or jump to... / Pull requests Issues Marketplace Explore

Branch master was renamed to main.

aws / amazon-sagemaker-examples Public

Watch 251 | Fork 5.6k | Star 7.4k

Code | Issues 563 | Pull requests 126 | Discussions | Actions | Projects | Security | Insights

main | amazon-sagemaker-examples / advanced_functionality / multi_model_xgboost_home_value / xgboost_multi_model_endpoint_home_value.ipynb | Go to file

haixiw update xgboost version to 1.5-1 in notebooks (#3439) ... | Latest commit 93163a8 on Jun 16 | History

8 contributors

1597 lines (1597 sloc) | 52.8 KB | Raw | Blame

Amazon SageMaker Multi-Model Endpoints using XGBoost

With Amazon SageMaker multi-model endpoints, customers can create an endpoint that seamlessly hosts up to thousands of models. These endpoints are well suited to use cases where any one of a large number of models, which can be served from a common inference container to save inference costs, needs to be invokable on-demand and where it is acceptable for infrequently invoked models to incur some additional latency. For applications which require consistently low inference latency, an endpoint deploying a single model is still the best choice

AT&T Cybersecurity

Unified Security Management

USM Anywhere centralizes security monitoring of networks and devices in the cloud, on premises, and in remote locations, helping you to detect threats virtually *anywhere*

A modern software-defined platform, USM Anywhere automates threat detection, powered by AT&T Alien Labs threat intelligence

Plus, it orchestrates incident response with other leading security technologies – helping you to achieve security without the seams

Data analysis

Threat detection

Automation and response



aws

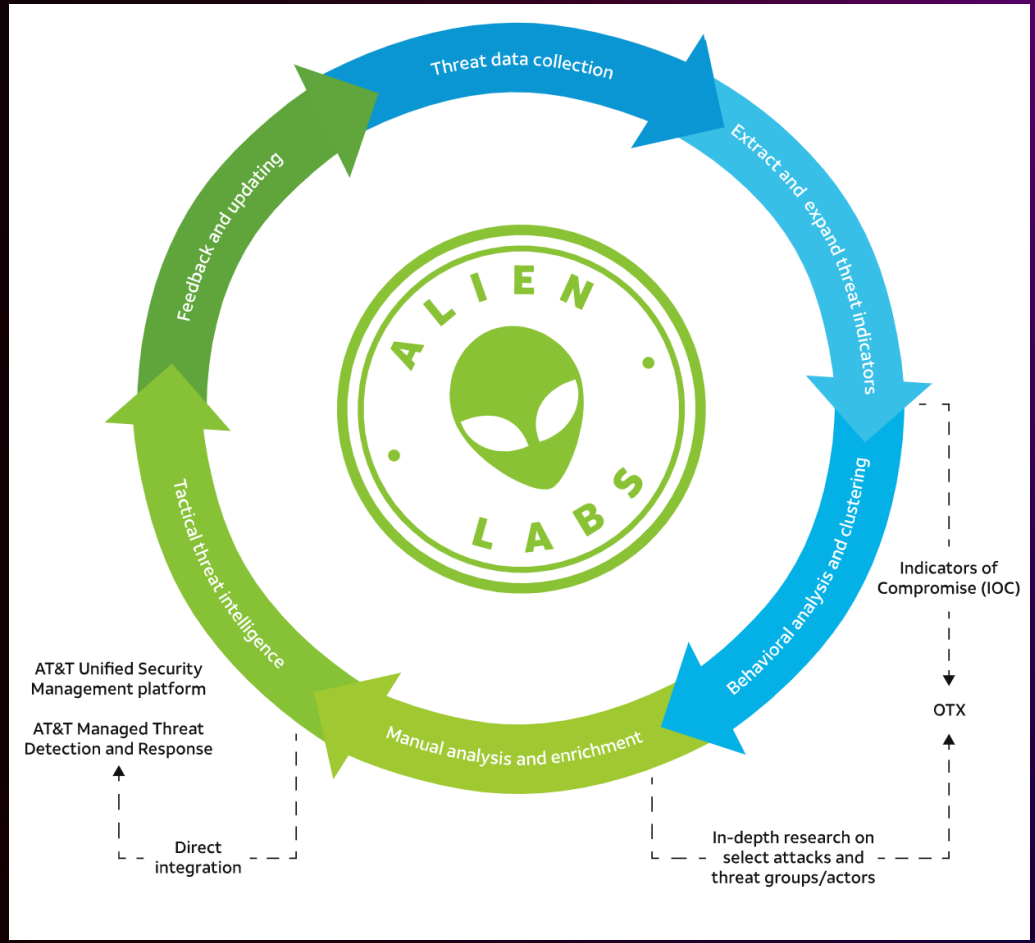
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AT&T Business

AT&T Cybersecurity ML use case

- Data from USM sensor network, Open Threat Exchange, Alien Labs threat intel
- Hand-built correlation rules developed by Alien Labs threat researchers
- Objectives
  - Produce higher quality detections
  - Reduce alarm fatigue
  - Create personalized detections at scale

Resilient threat detection



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.
AT&T Business

Engineering challenges



Complexity

Performance

Cost

Self-managed endpoints per customer

Customer: XXX

Endpoint: https://sample.com/XXX/O365

Model1.tar.gz

Customer: YYY

Endpoint: https://sample.com/YYY/O365

Model2.tar.gz

Customer: ZZZ

Endpoint: https://sample.com/ZZZ/O365

Model3.tar.gz

Versus

Amazon SageMaker MME

Mode: MultiModel

Endpoint: https://sample.com/O365

Artifact location:
s3://bucket/your-endpoint-models

Model1.tar.gz

Model2.tar.gz

Model3.tar.gz

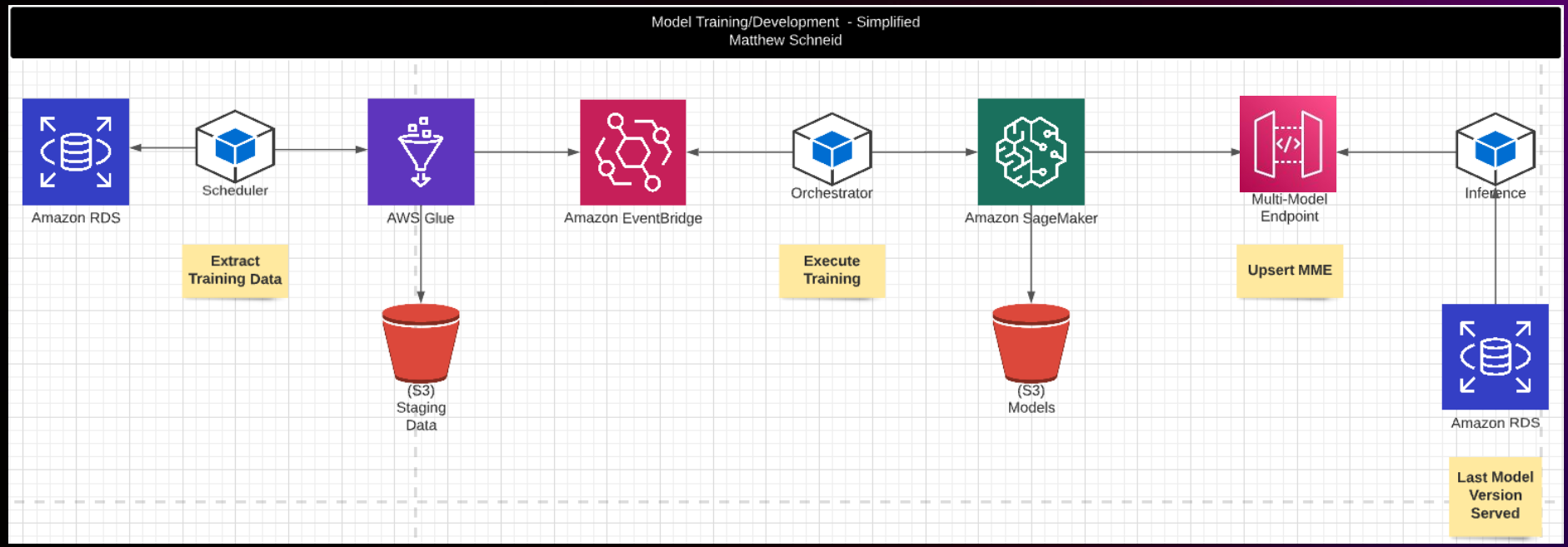
Modeln.tar.gz

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Architecture (Training / MME)

- Extract
- Train
- Produce model
- Upsert MME
- Infer



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Engineering key learnings

1  SageMaker MME – Most cost effective

2  MME performance optimization

3  EventBridge is your friend

4  Use your AWS resources!

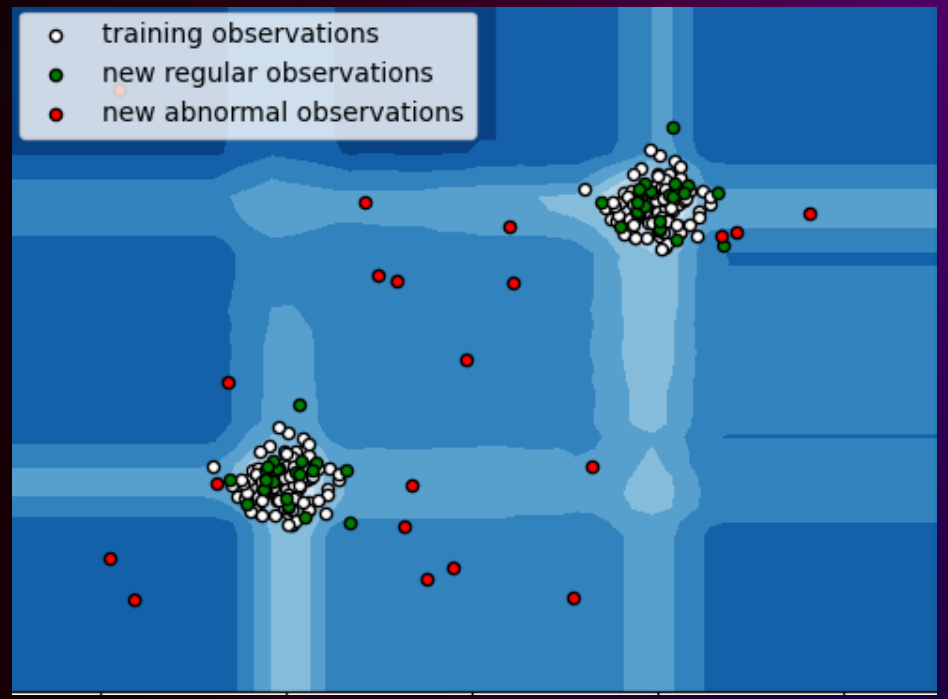
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AT&T Business

Data science: Development

- Anomaly detection on streaming events
- Isolation forest, histogram-based outlier score, autoencoders, embeddings
- Components used
  - SageMaker notebooks
  - Processing
  - Clarify
- 10 ML models delivered in 2022
- Benefits to data scientists

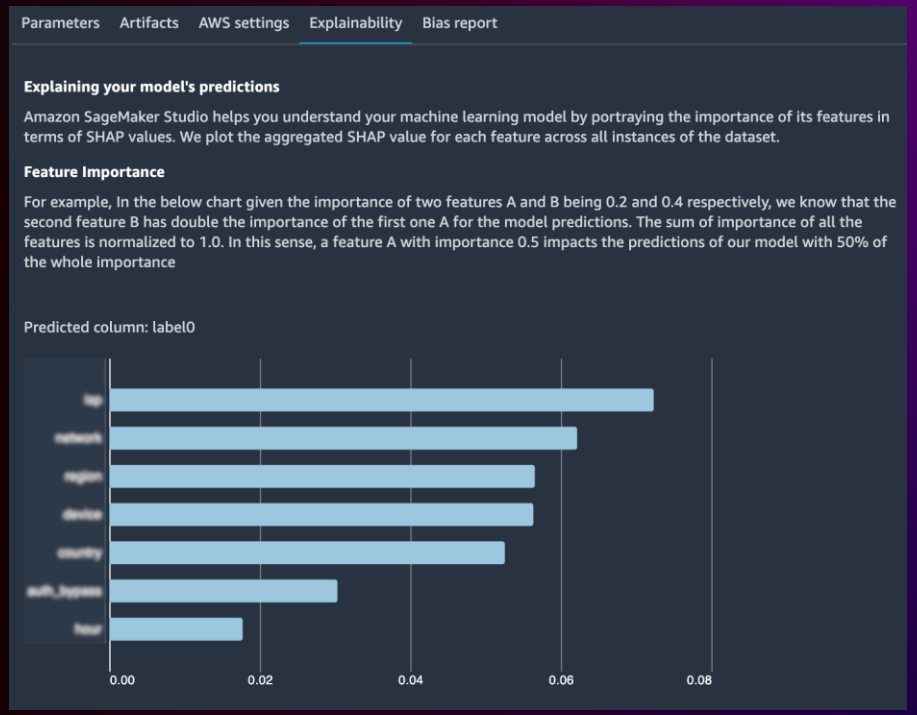


Isolation forest anomaly detection algorithm

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.
AT&T Business

Data science: Operational improvements

- What did we learn about 1:1 models?
- SageMaker Clarify SHAP analysis as an MLOps tool
  - Feature importance monitoring
  - Validation of auto-retrained models
  - Discovery of model clusters



Model explainability in SageMaker Clarify

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

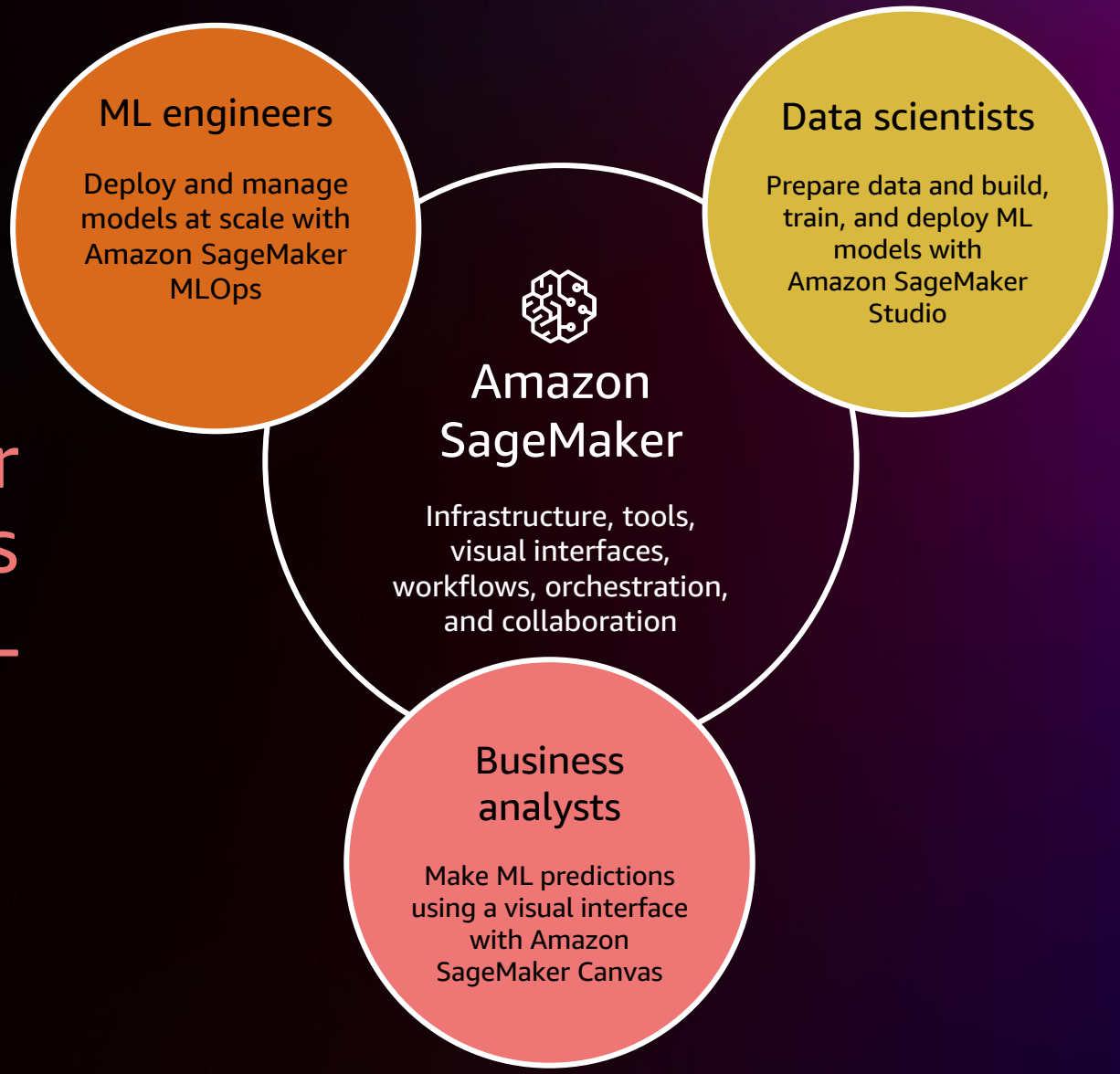
Striking the right balance as projects evolve



- Several thousand models reduced to 12
- Benefits
  - Reduces complexity with minimal impact to model performance
  - Reduces ML cold start problem with a selection of pretrained models

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.
AT&T Business

Amazon SageMaker helps organizations harness ML



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.