AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

AIM414

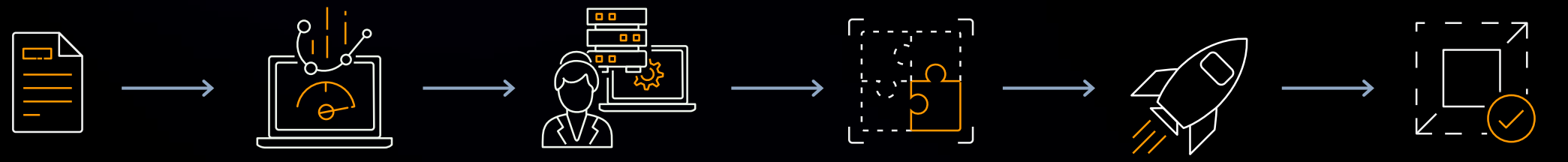
End-to-end 3D machine learning on Amazon SageMaker

Isaac Privitera (he/him)
Senior Data Scientist at ML Solutions Lab
Amazon Web Services

Vidya Sagar Ravipati (he/him)
Senior Deep Learning Architect
Amazon Web Services

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Typical machine learning process



Collect, prepare, and label training data

Choose and optimize ML algorithm

Set up and manage environments for training

Train and tune ML models

Deploy models in production

Scale and manage the production environment

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

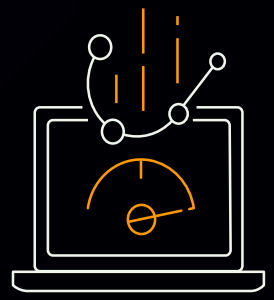
The problem

- 3D ML is complex!



Collect, prepare, and label training data

Point cloud data labeling is challenging and expensive



Choose and optimize ML algorithm

Distributed training is necessary to iterate quickly, but is resource intensive, difficult to setup, and difficult to optimize



Set up and manage environments for training/inference

3D models have complex dependencies and can be difficult to deploy

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Workshop overview

- Notebook 1:
  - Point cloud visualization
  - Sensor fusion
  - Amazon SageMaker Ground Truth point cloud labeling
- Notebook 2:
  - Launch SageMaker distributed training with MMDetection3D
  - Monitor your training job with SageMaker Debugger Profiler
- Notebook 3:
  - Deploy an asynchronous SageMaker endpoint
  - Visualize results

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Workshop overview

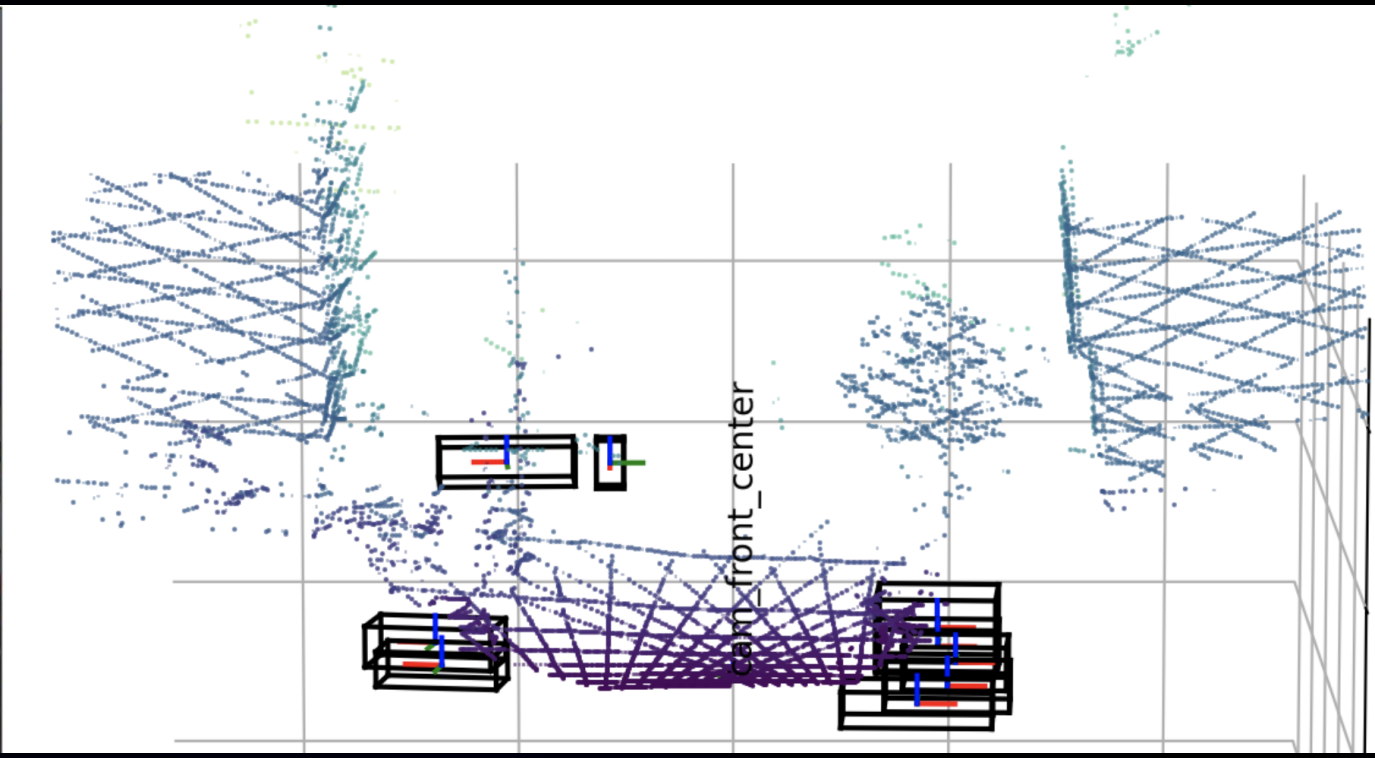
- Notebook 1:
  - Point cloud visualization
  - Sensor fusion
  - Amazon SageMaker Ground Truth point cloud labeling
- Notebook 2:
  - Launch SageMaker distributed training with MMDetection3D
  - Monitor your training job with SageMaker Debugger Profiler
- Notebook 3:
  - Deploy an asynchronous SageMaker endpoint
  - Visualize results

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The data

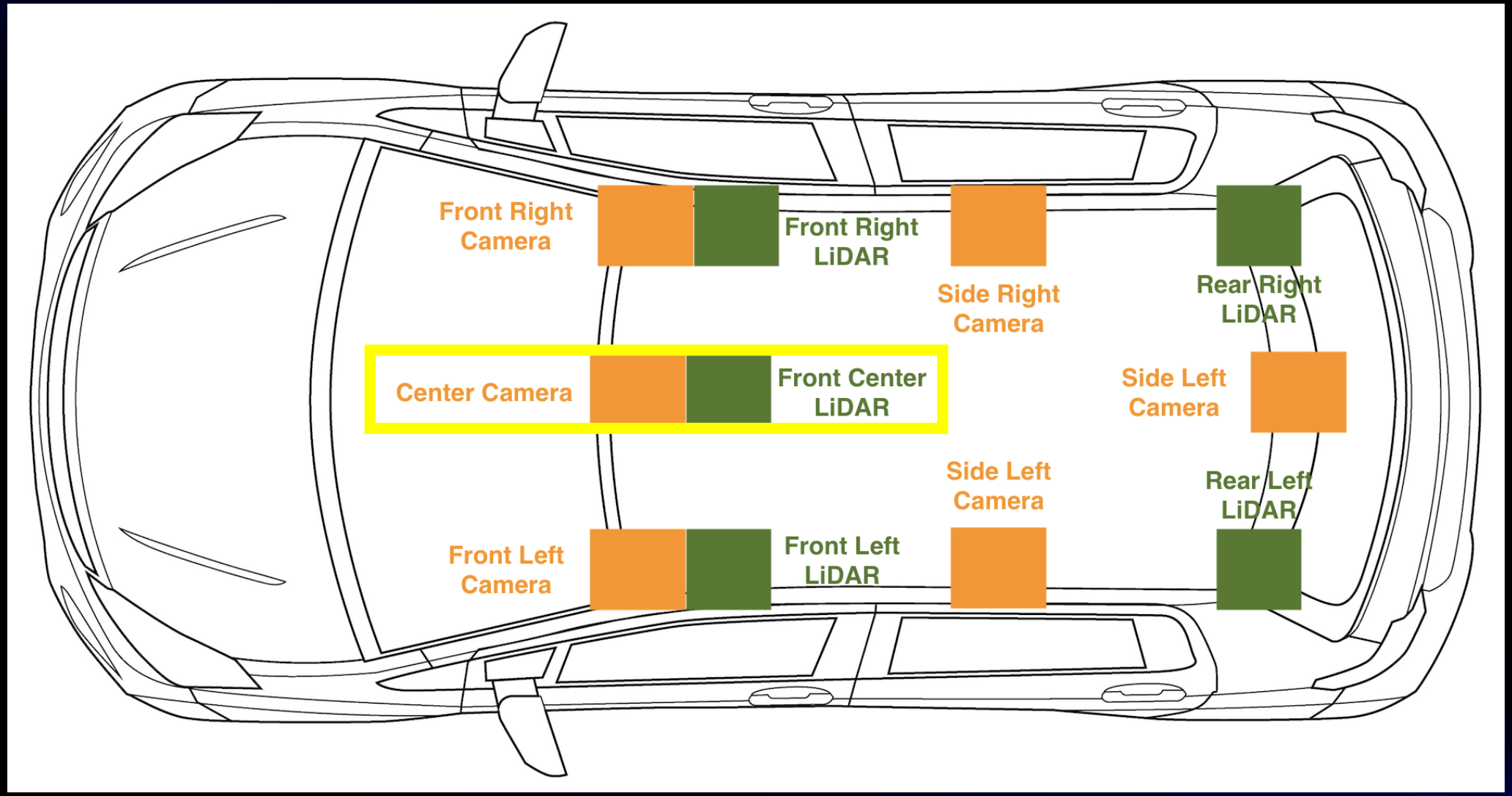
AUDI AUTONOMOUS DRIVING DATASET (A2D2)





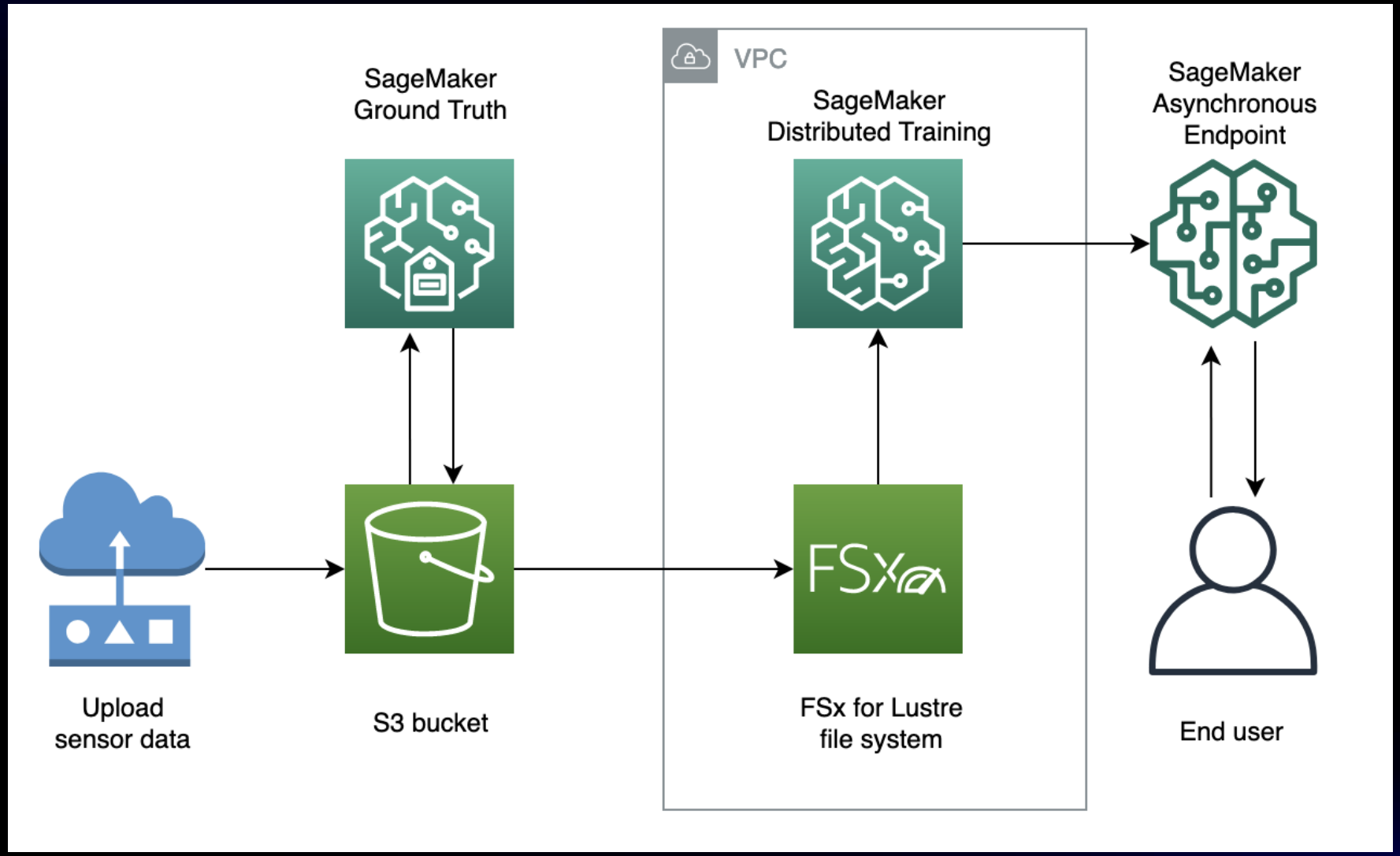
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The sensors



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The architecture



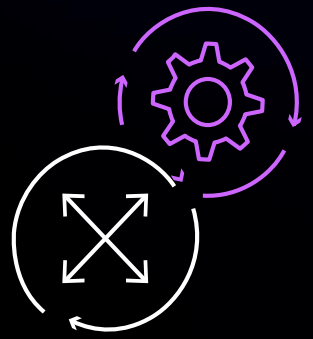
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data labeling



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data labeling challenges



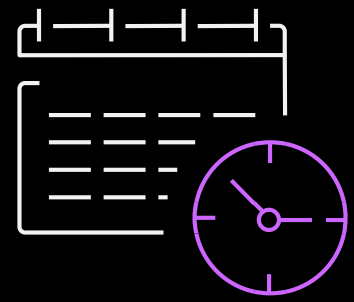
Massive scale

ML models need large, labeled datasets



High accuracy

ML models depend on accurately labeled data



Time consuming

Building training datasets takes up to 80% of a data scientist's time

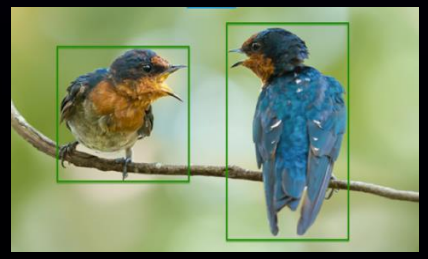
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker Ground Truth

Image classification



Bounding boxes

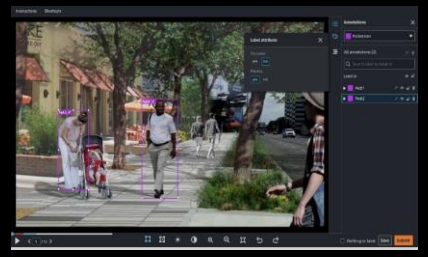


Semantic segmentation

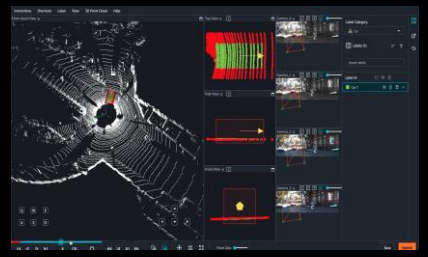


Label verification

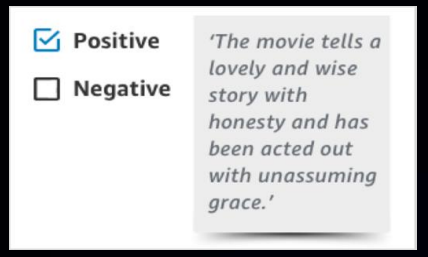




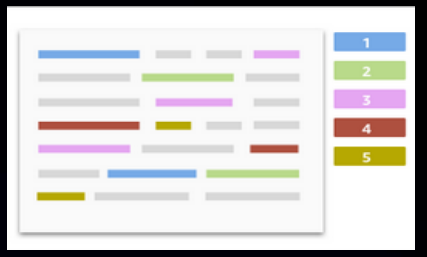
Video



LIDAR 3D point cloud



Text classification



Named entity recognition

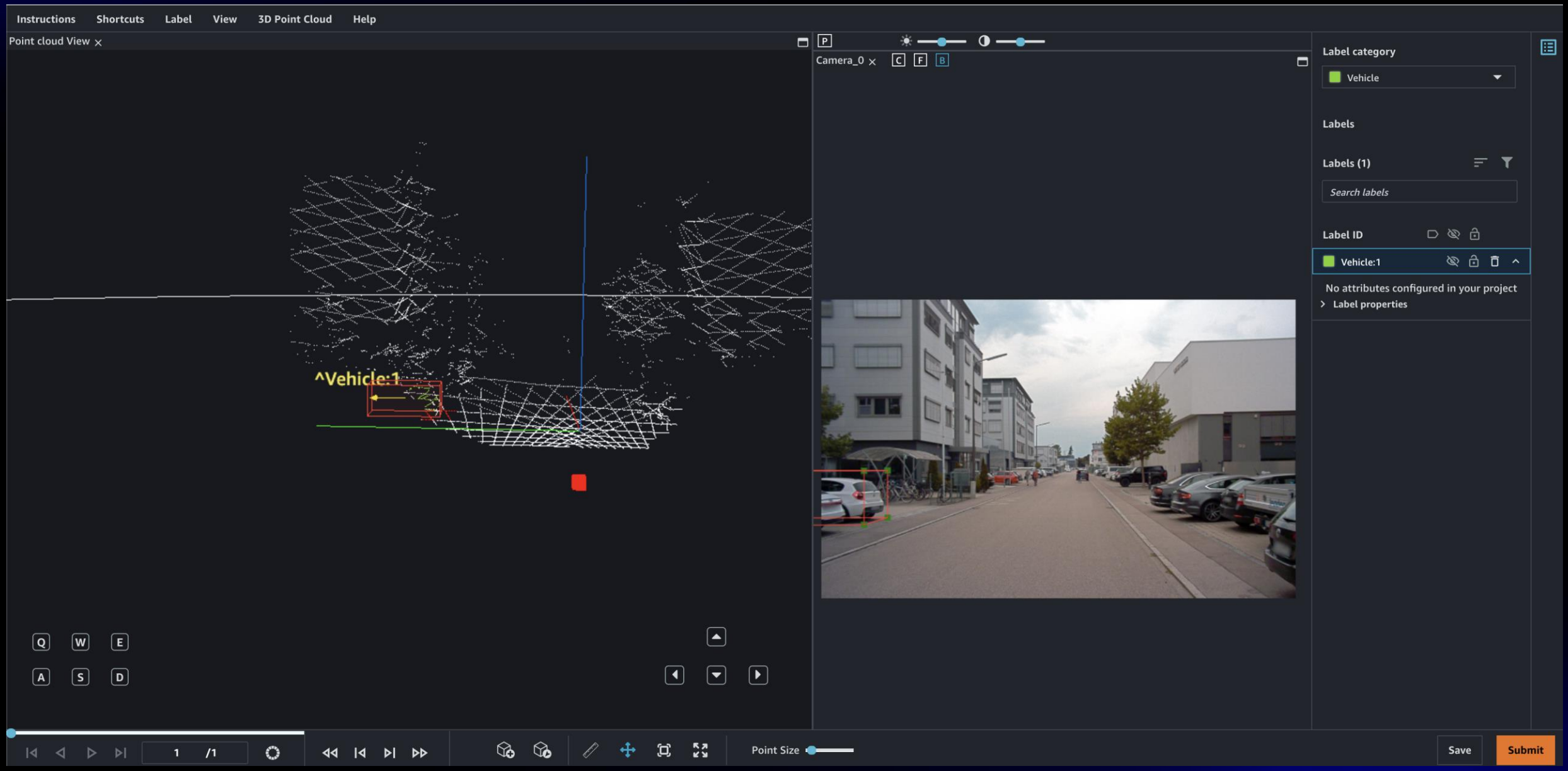


Custom (25 templates)

- Key point
- Line
- Instance segmentation

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Point cloud object detection labeling



Notebook 1

- Should take ~30 minutes

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Workshop overview

- Notebook 1:
  - Point cloud visualization
  - Sensor fusion
  - Amazon SageMaker Ground Truth point cloud labeling
- Notebook 2:
  - Launch SageMaker distributed training with MMDetection3D
  - Monitor your training job with SageMaker Debugger Profiler
- Notebook 3:
  - Deploy an asynchronous SageMaker endpoint
  - Visualize results

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Distributed training

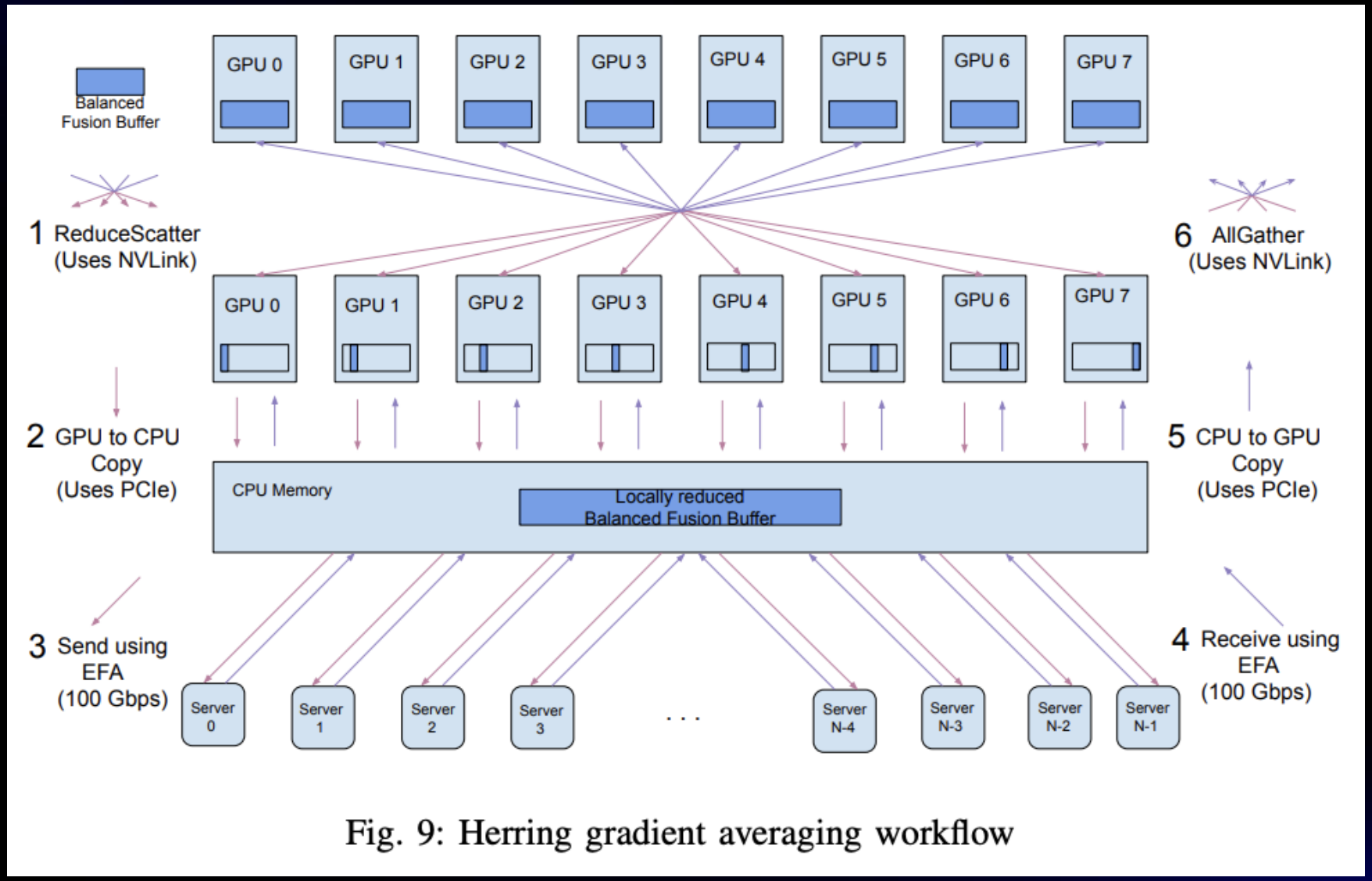


Fig. 9: Herring gradient averaging workflow

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon SageMaker distributed training libraries

The fastest and easiest way to train large deep learning models



Reduced training time

Reduce training time by 25% with synchronization across GPUs

Optimized for AWS

Achieve near-linear scaling efficiency with data parallelism designed for AWS

Support for popular ML framework APIs

Re-use existing APIs such as Horovod without custom training code

Automatic and efficient model partitioning

Avoid experimentation with automated model profiling and partitioning

Minimal code change

Implement model parallelism with fewer than 10 lines of code change

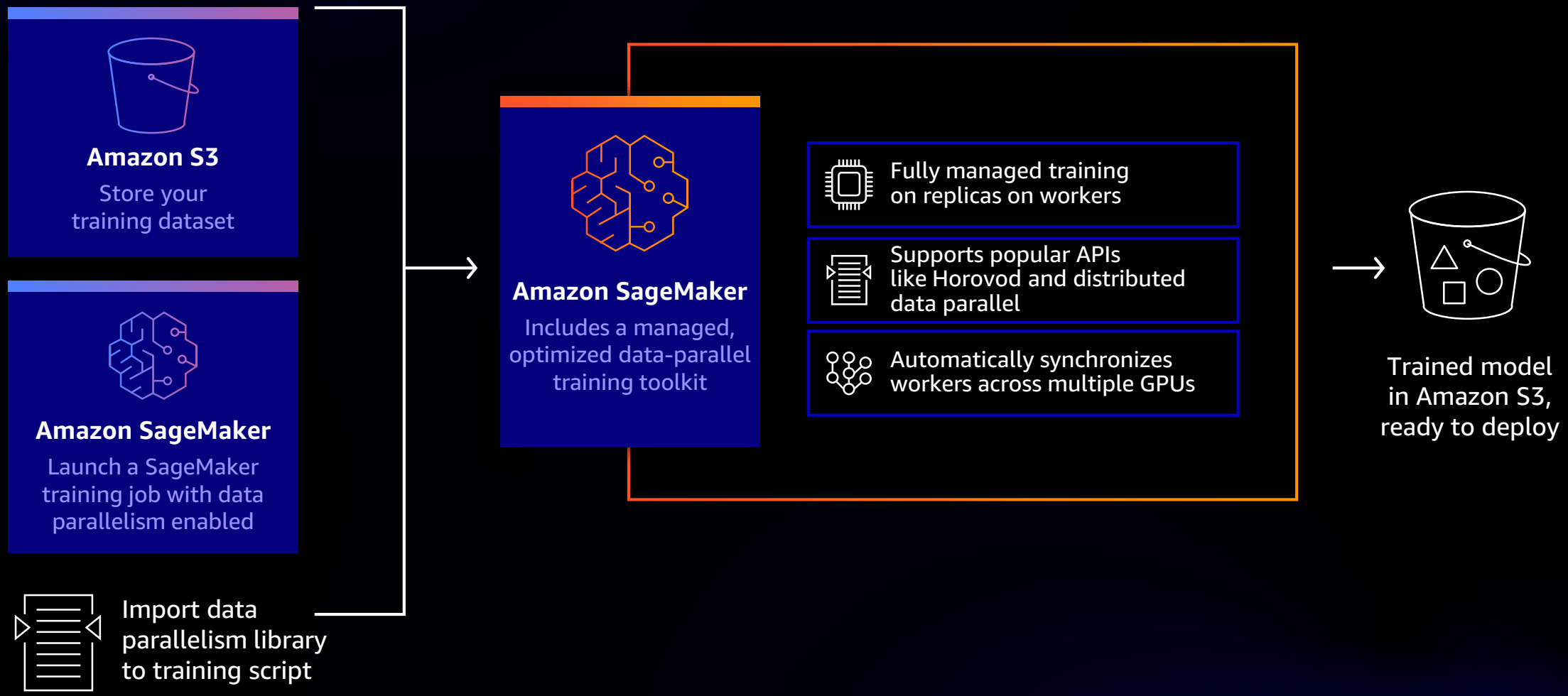
Efficient pipelining

Maximize resource usage with pipelining of micro-batches that keeps all GPUs active

aws

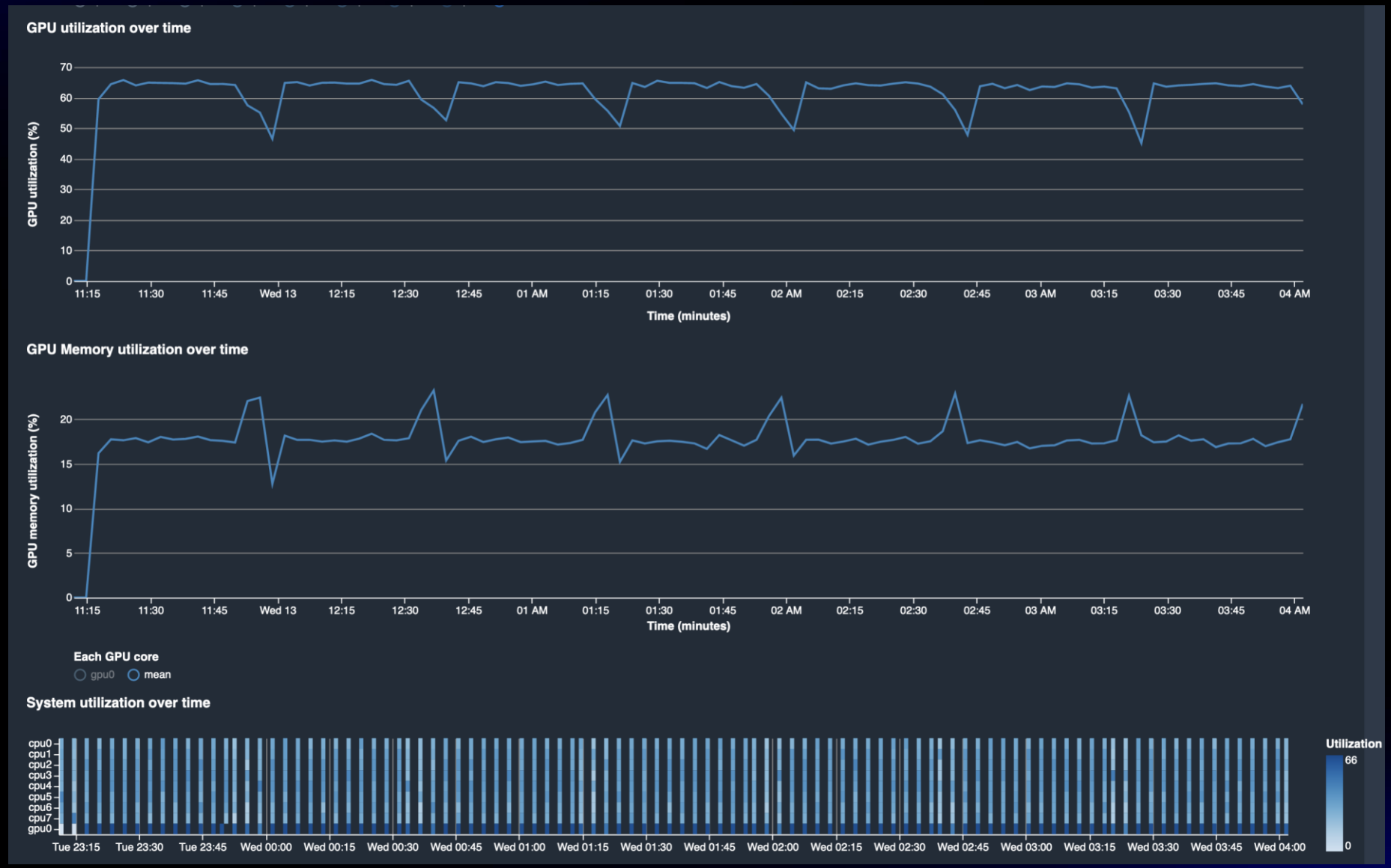
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How it works: Data parallelism library



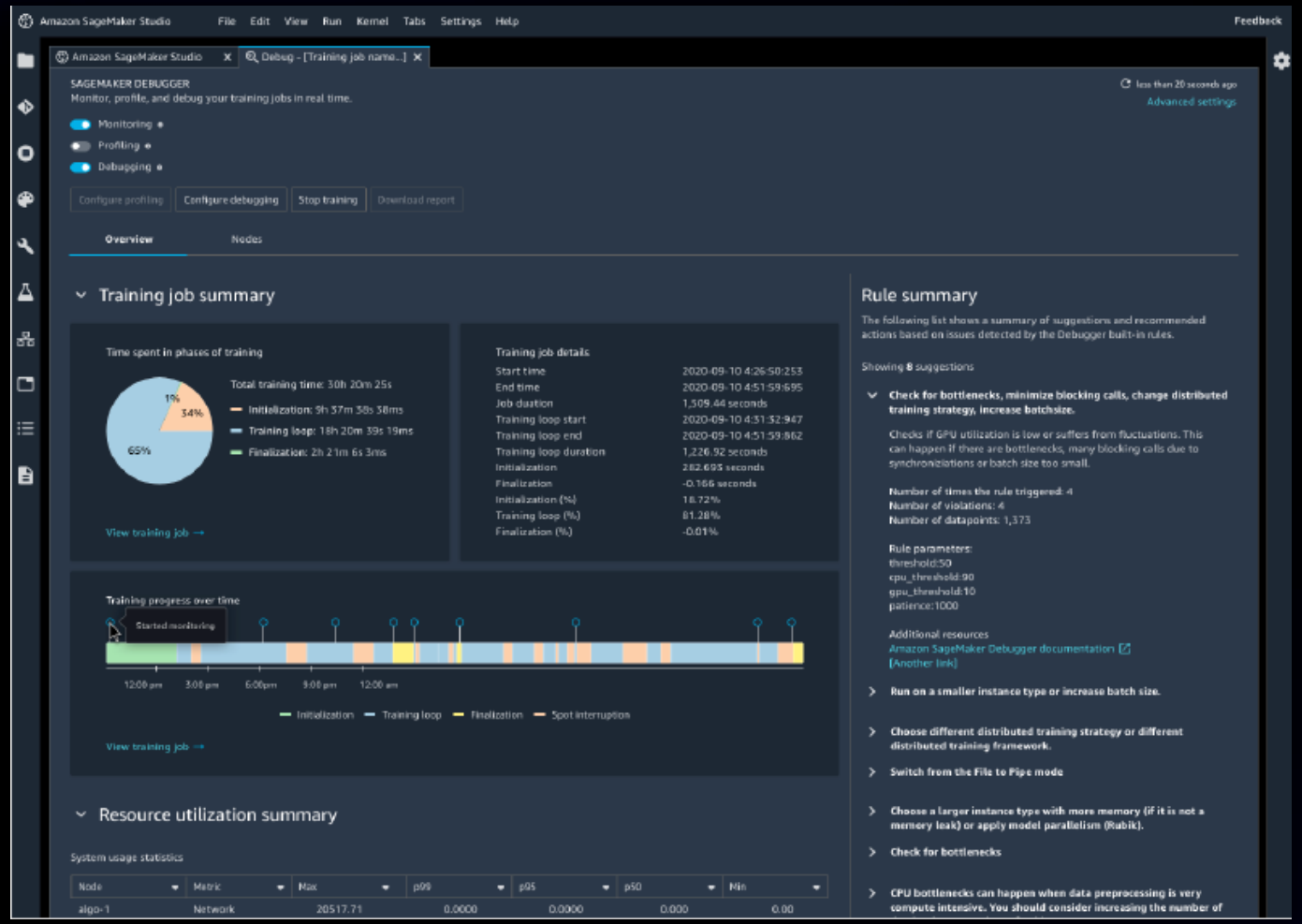
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker Debugger Profiler



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Monitor and profile system resource utilization



- Automatically monitor system resource utilization
- Profile training jobs to collect ML framework metrics
- Visualize system resource utilization for GPU, CPU, network, memory within Amazon SageMaker Studio

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Notebook 2

- Should take ~45 minutes

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

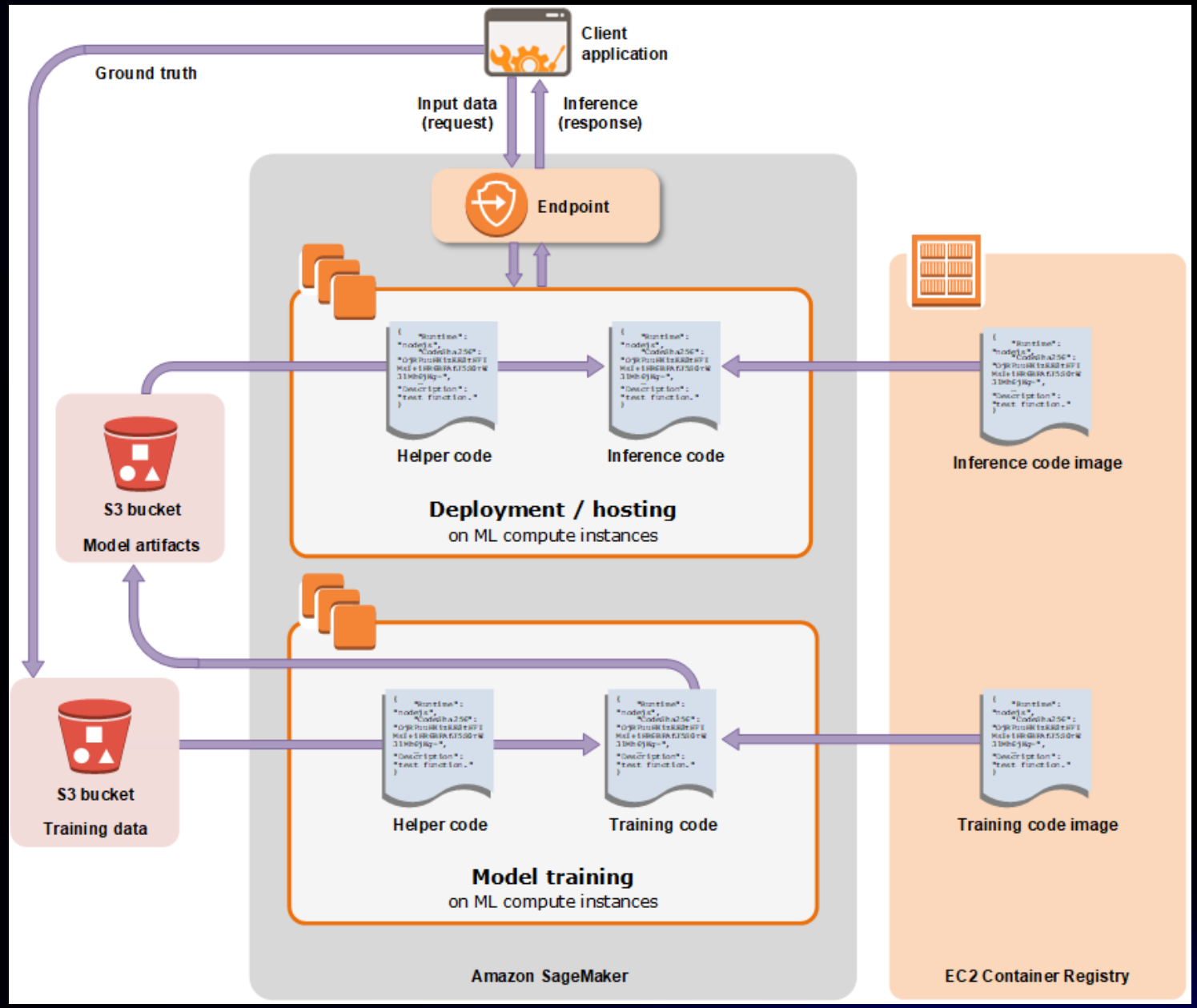
Workshop overview

- Notebook 1:
  - Point cloud visualization
  - Sensor fusion
  - Amazon SageMaker Ground Truth point cloud labeling
- Notebook 2:
  - Launch SageMaker distributed training with MMDetection3D
  - Monitor your training job with SageMaker Debugger Profiler
- Notebook 3:
  - Deploy an asynchronous SageMaker endpoint
  - Visualize results

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker deployment



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker model deployment options

Real-time inference

- RESTful HTTP requests
- Elastic environment

- Low latency
- Ultra-high throughput
- Multi-model endpoints
- A/B testing

Example use cases:
Ad serving, personalized recommendations, fraud detection

Asynchronous inference

new

- RESTful HTTP requests
- Queue-based system

- Large payloads (1 GB)
- Long timeouts (15 min)
- Scale down to zero

Example use cases:
Computer vision, object detection, NLP

Batch transform

- Job-based system
- Non-persistent environments

- Process large datasets
- Associate results with records

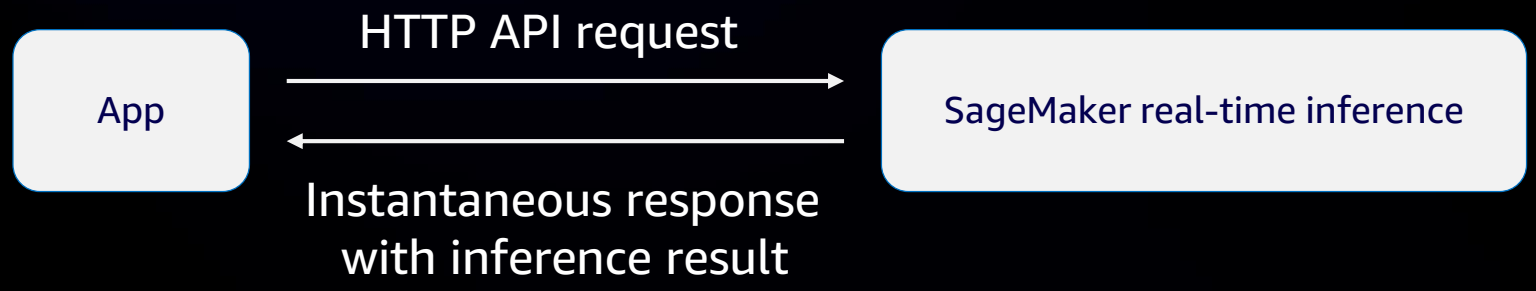
Example use cases:
Churn prediction, predictive maintenance, demand forecasting

aws

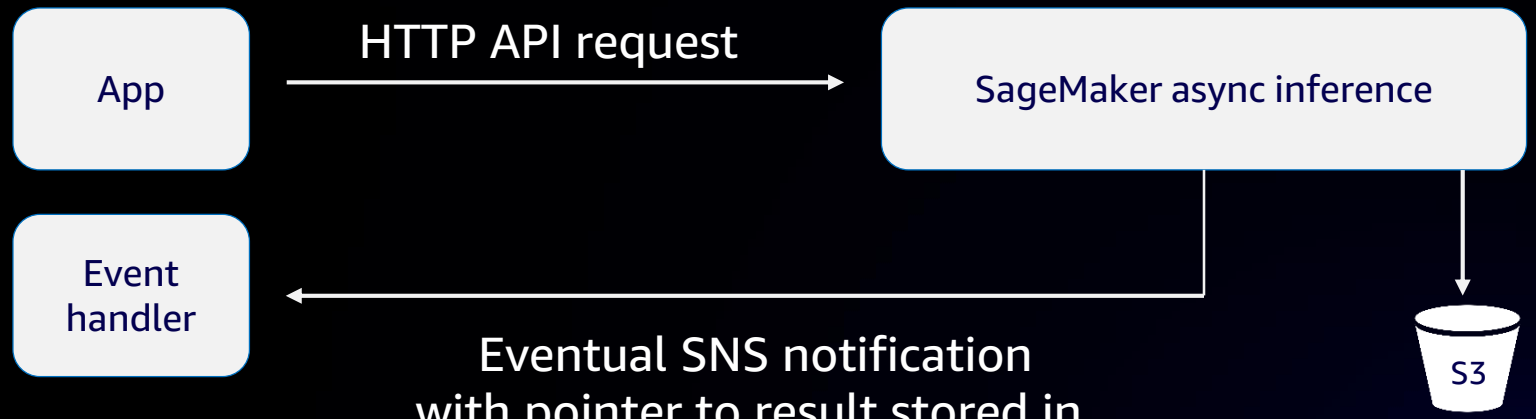
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How is real-time inference different from async inference?

REAL-TIME INFERENCE

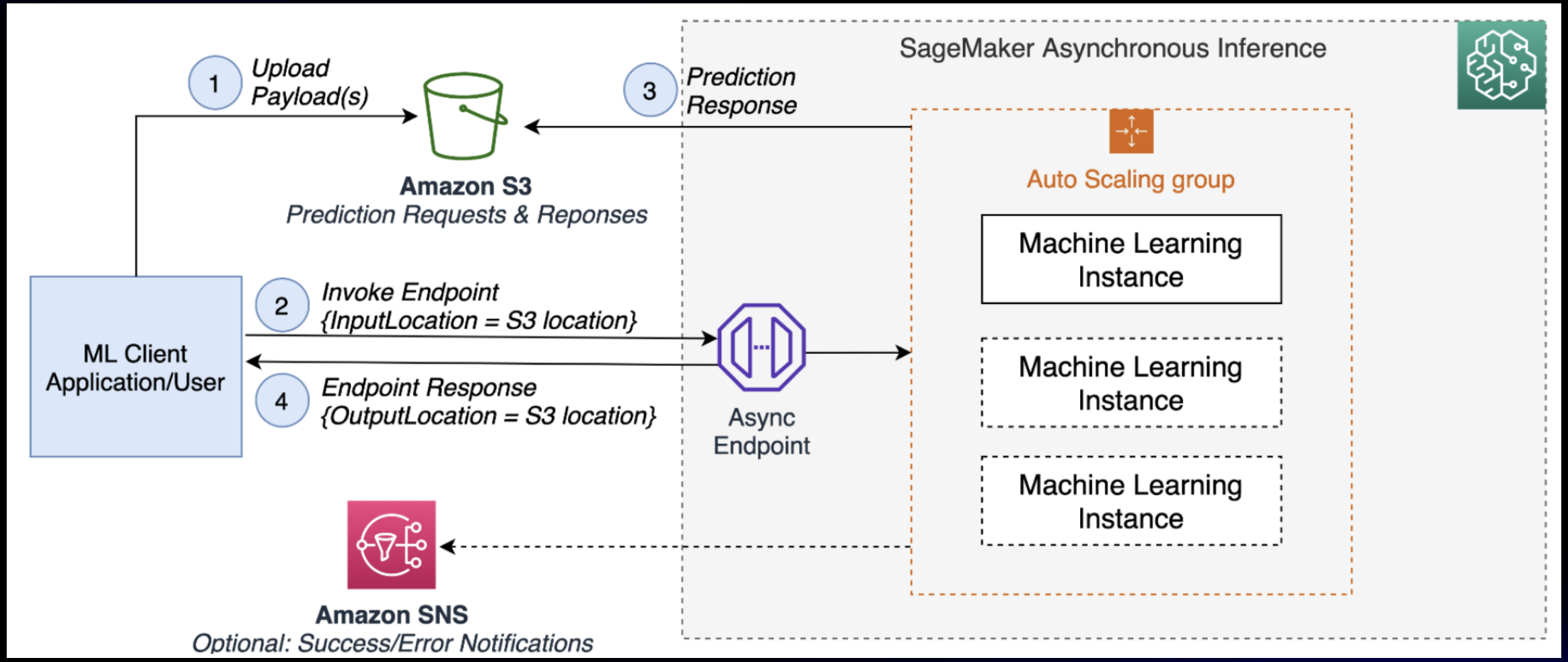


Async inference



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Asynchronous inference



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Inference visualization



its affiliates. All rights reserved.

Notebook 3

- Should take ~20 minutes

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Isaac Privitera

privisaa@amazon.com
https://www.linkedin.com/in/isaac-privitera-b8183a78/

Vidya Sagar Ravipati

ravividy@amazon.com

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.