# AWS re:Invent

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

AIM410

# Machine learning at the edge with Amazon SageMaker

Sean Morgan
Senior ML Services Solutions Architect
AWS

Sireesha Muppala
Principal Enterprise Solutions Architect, AI/ML
AWS

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

# Agenda

ML@Edge

Challenges

Amazon SageMaker Edge Manager

Demo

Q&A

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

# ML@Edge use cases



Predictive maintenance



Video security



Personalized computing



Connected buildings and city systems



Industrial IoT and robotics



Health and wellness services

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

# ML@Edge use cases



Predictive maintenance



Video security



Personalized computing



Connected buildings and city systems



Industrial IoT and robotics



Health and wellness services

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

# ML@Edge use cases



Predictive maintenance



Video security



Personalized computing



Connected buildings and city systems



Industrial IoT and robotics



Health and wellness services

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

# ML@Edge use cases



Predictive maintenance



Video security



Personalized computing



Connected buildings and city systems



Industrial IoT and robotics



Health and wellness services

aws

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

# ML@Edge use cases



Predictive maintenance



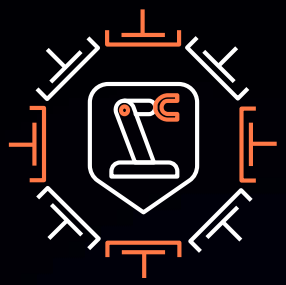
Video security



Personalized computing



Connected buildings and city systems



Industrial IoT and robotics



Health and wellness services

aws

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

# ML@Edge: Computer vision



Need for **low latency** for real-time insights



**Bandwidth constraints** limit what video is sent to the cloud



**On-premises** data processing requirements

aws

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

# Challenges

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

# Edge device challenges

- **Deploy** a trained model to the edge

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

# Edge device challenges

- **Deploy** a trained model to the edge
- **Secure** my model while **deploying/running** it at the edge

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

# Edge device challenges

- **Deploy** a trained model to the edge
- **Secure** my model while **deploying/running** it at the edge
- **Eliminate** the need for installing a **big framework** like TensorFlow/PyTorch on my device

aws

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

# Edge device challenges

- **Deploy** a trained model to the edge
- **Secure** my model while **deploying/running** it at the edge
- **Eliminate** the need for installing a **big framework** like TensorFlow/PyTorch on my device
- **Monitor** my model's **performance** and **retrain it, if needed**

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

# Edge ML scalability challenges

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

# Edge ML scalability challenges

- **Operate** ML models on a **fleet of devices** at the edge

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

# Edge ML scalability challenges

- **Operate** ML models on a **fleet of devices** at the edge
- **Prepare** and **deploy** an ML model to **multiple edge devices**

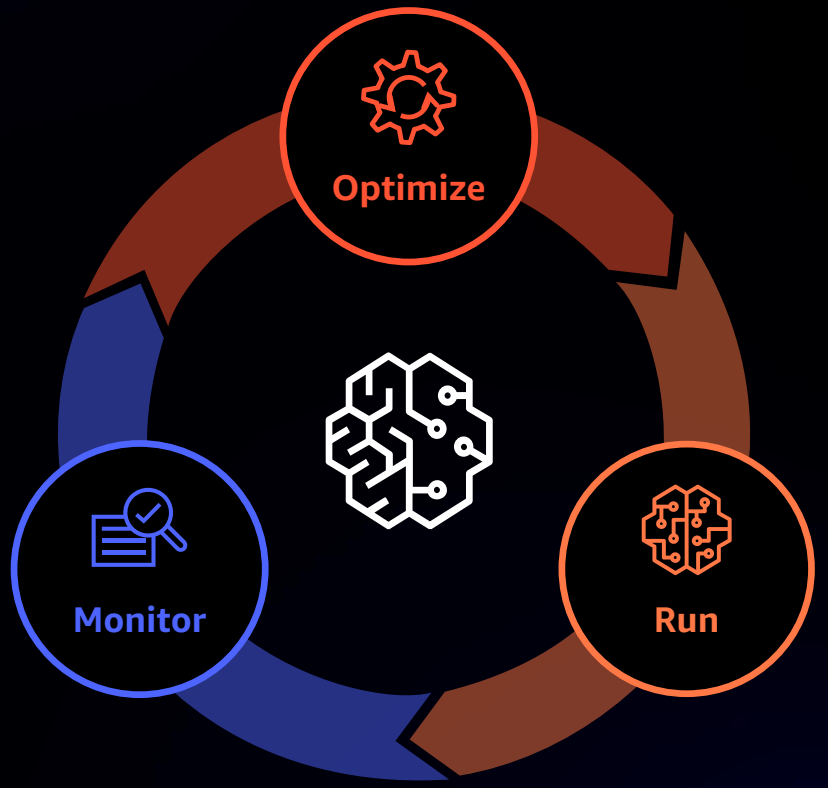
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon SageMaker Edge Manager

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

# Amazon SageMaker Edge Manager

Prepare models for deployment at the edge (SageMaker Neo)




Run one or more models on each device in a fleet (Edge Manager agent)

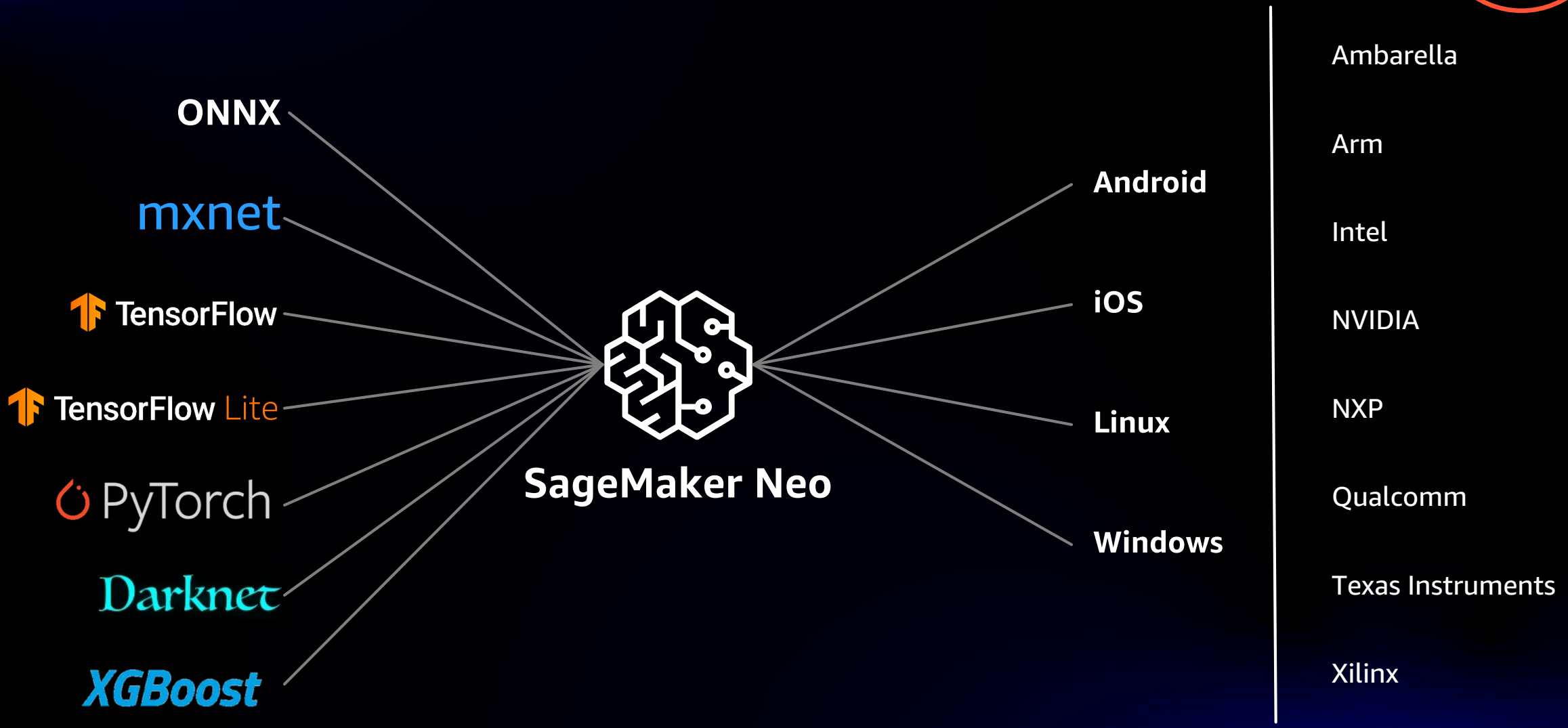
Detect model decay and improve model quality (SageMaker and other AWS services)

aws

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

# Neo: Optimize the models

SAGEMAKER ALGORITHMS

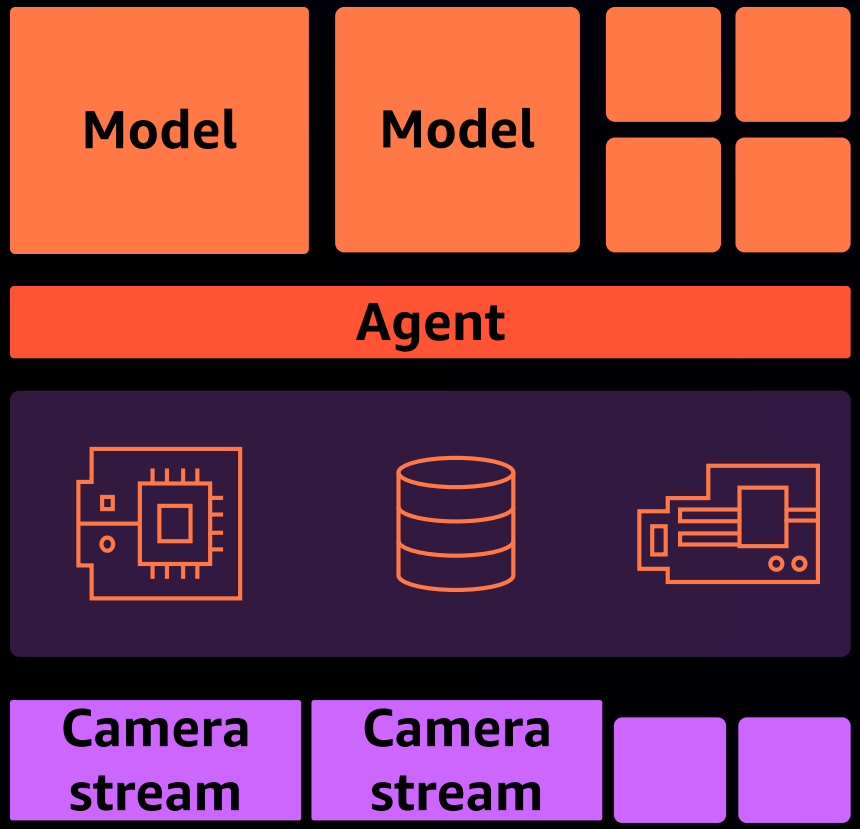



TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.
PyTorch, the PyTorch logo and any related marks are trademarks of Facebook, Inc.

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

# Edge Manager agent: Run the models




One agent serves multiple models

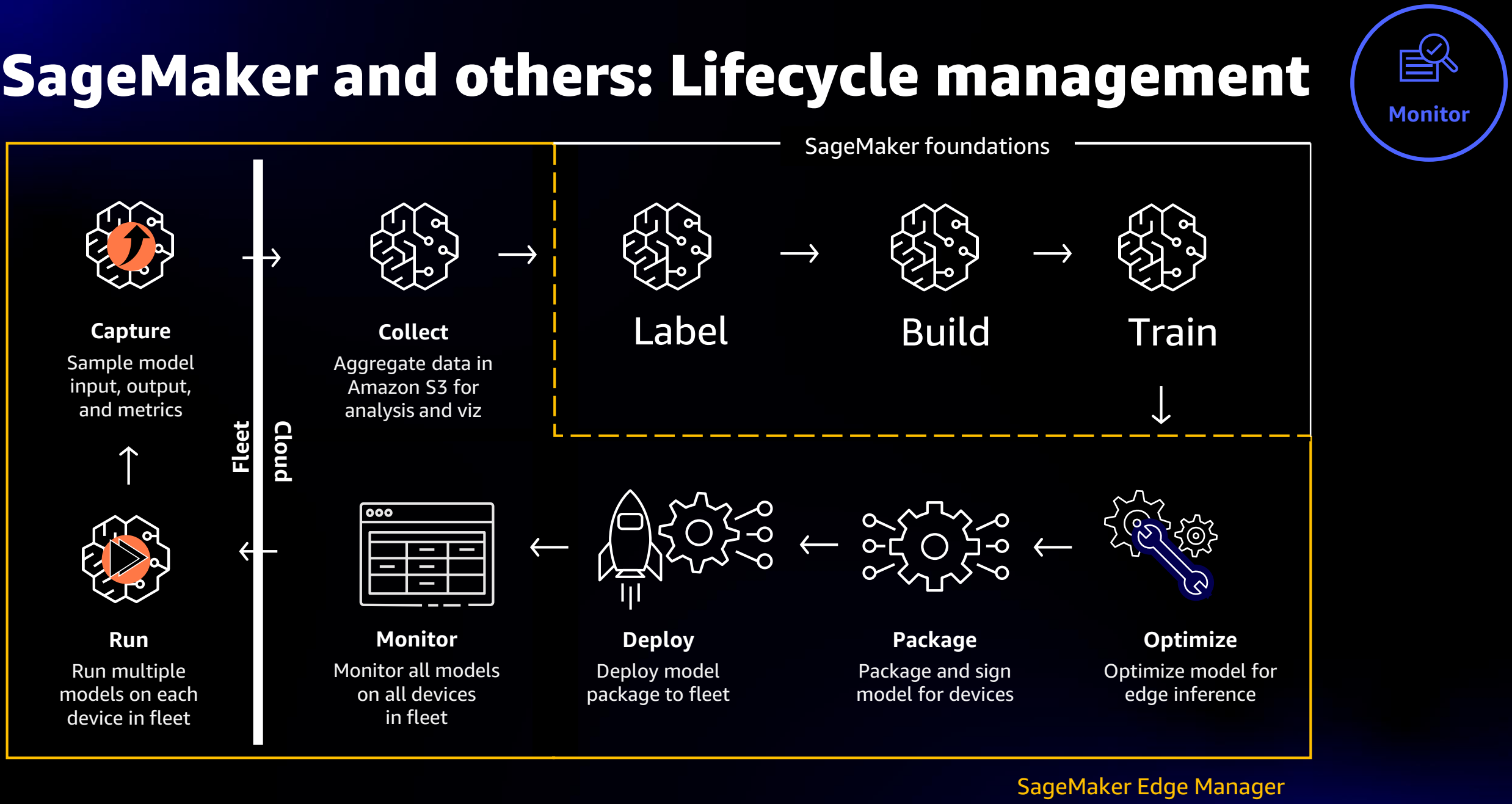
The agent consumes <2 MB of memory, freeing up resources for multiple models

Applications can run multiple models, increasing efficiency

aws

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

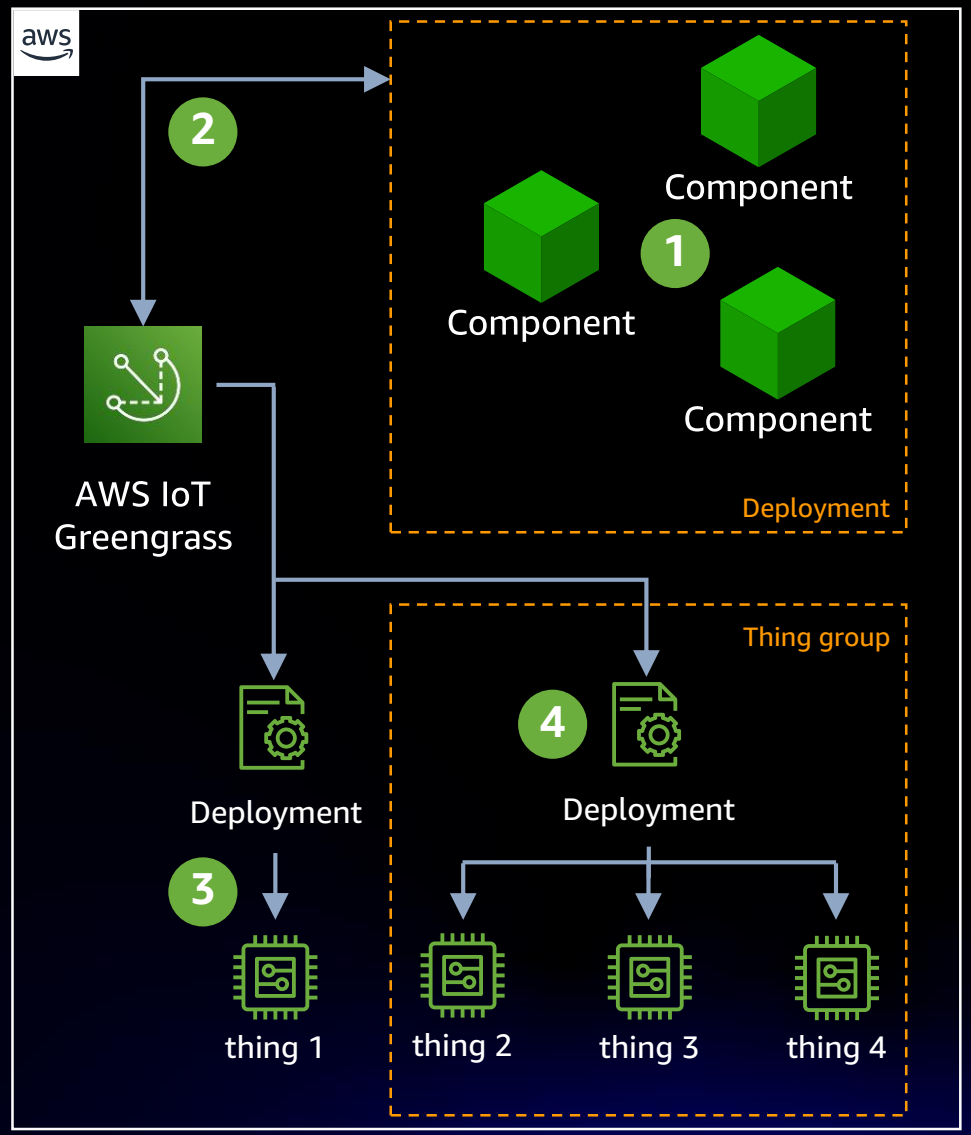
# SageMaker and others: Lifecycle management




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

# AWS IoT Greengrass V2 deployment

- AWS-provided **Edge Manager component**
- Build **custom components** for application code




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

# Service choices

- SageMaker MLOps workflow management
  - SageMaker Pipelines, Apache Airflow, AWS Step Functions, KubeFlow
- Inference servers
  - SageMaker Edge Manager, NVIDIA Triton
- Runtime environments
  - Containers, system installs
- Deployment
  - AWS IoT GreenGrass, MQTT/IoT jobs, SCPs

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

# Next steps

Learn more about Amazon SageMaker Edge Manager

https://aws.amazon.com/sagemaker/edge-manager

Practice with the Amazon SageMaker Edge Manager Workshop

https://github.com/aws-samples/amazon-sagemaker-edge-manager-workshop

Self-contained defect detection sample

https://github.com/aws-samples/amazon-sagemaker-edge-defect-detection-computer-vision

aws

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

# Thank you!

Sean Morgan

seanmrgn@amazon.com
Twitter: @spmorgan_

Sireesha Muppala

smuppala@amazon.com

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