



Computer Vision on AWS

Partner Enablement

Dylan Tong | dylatong@amazon.com

ML Architect | Global Tech Lead, AI Augmented Analytics

Agenda

| | |
|--|---------------|
| 1. Introductions | |
| 2. Use Cases | |
| 3. AWS Computer Vision Overview: <ul style="list-style-type: none">• Amazon Rekognition Custom Labels• Amazon SageMaker• ... | 30-45 minutes |
| 4. Workshops <ul style="list-style-type: none">• AWS Computer Vision Jump Starter Kit<ul style="list-style-type: none">• Amazon Rekognition Custom Labels Lab• Object Detection on Amazon SageMaker Series: Annotation, Built-in Algorithms, BYOS• BYOS Pose Estimator• CV@Edge Online Series• Challenge | 60-75 minutes |
| 5. Additional Production Guidance | 30-45 minutes |
| 6. Conclusion and Open Q&A | |

Use Cases

Media Analytics



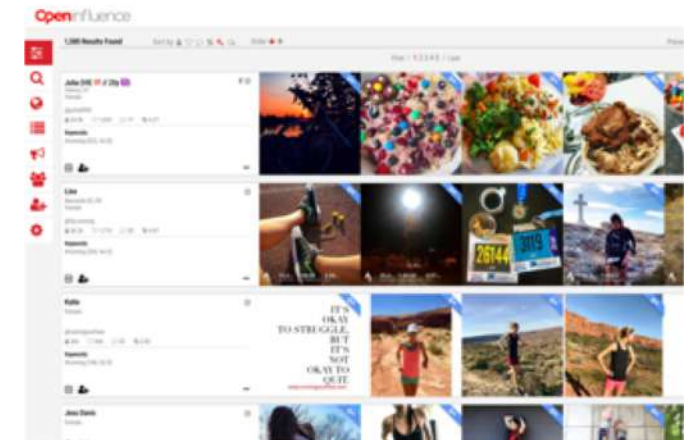
Live
events



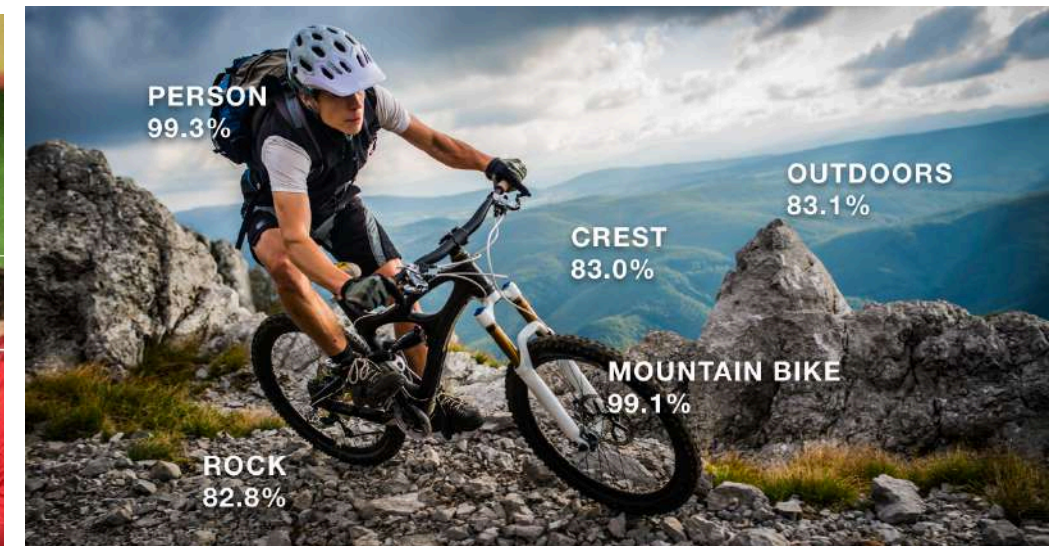
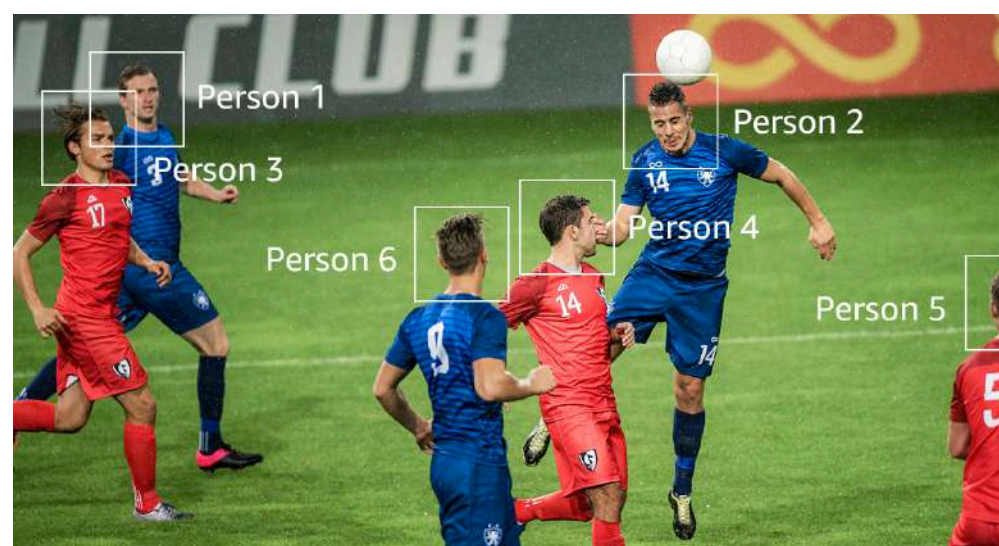
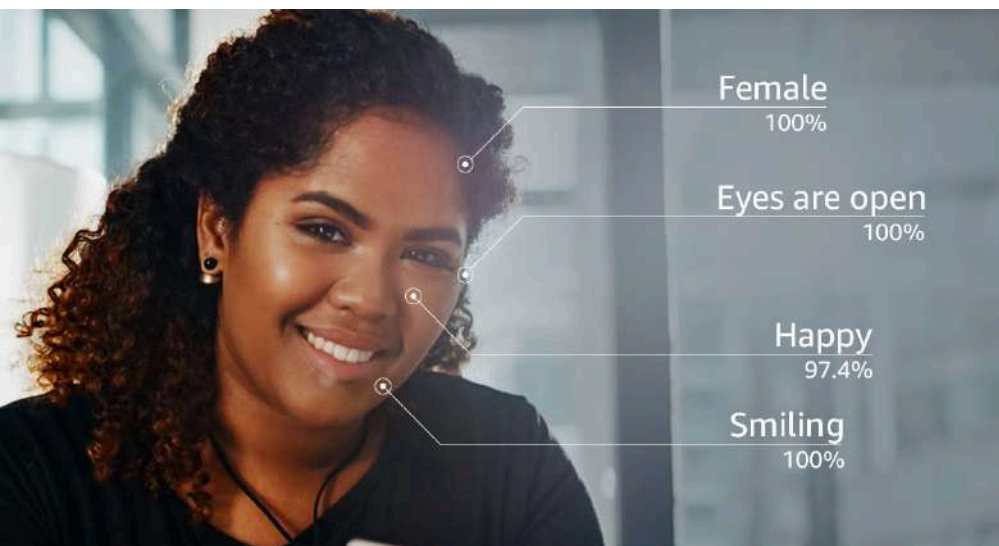
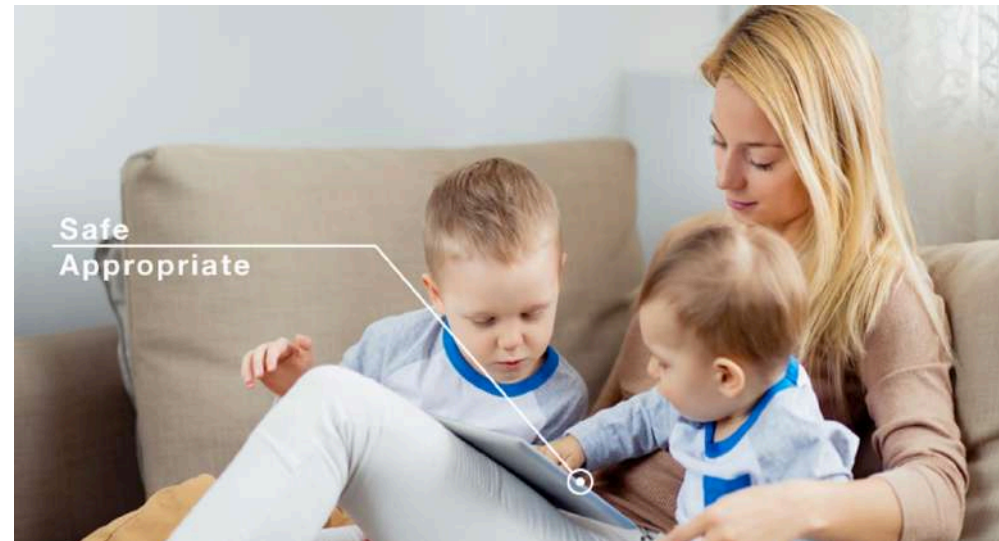
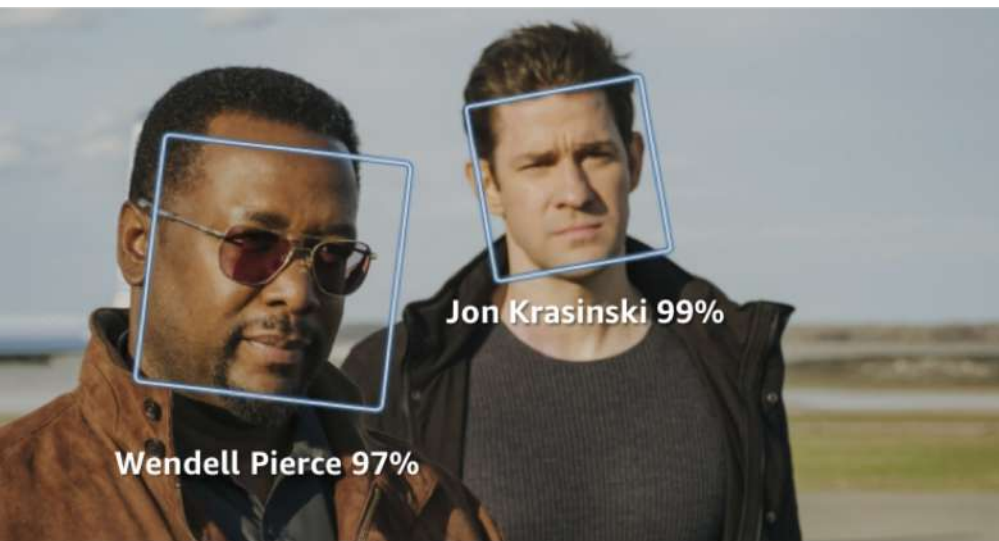
Media
libraries



Social
media



Influencer
marketing



Who

Celebrities
Employees
Customers

What

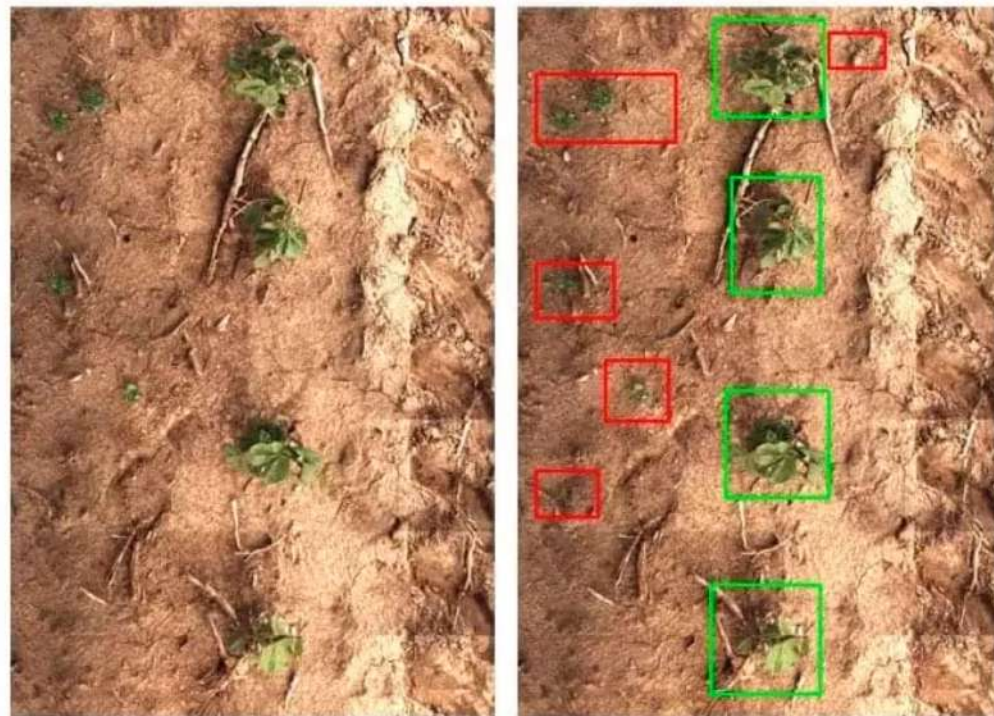
Labels and Activities

Where

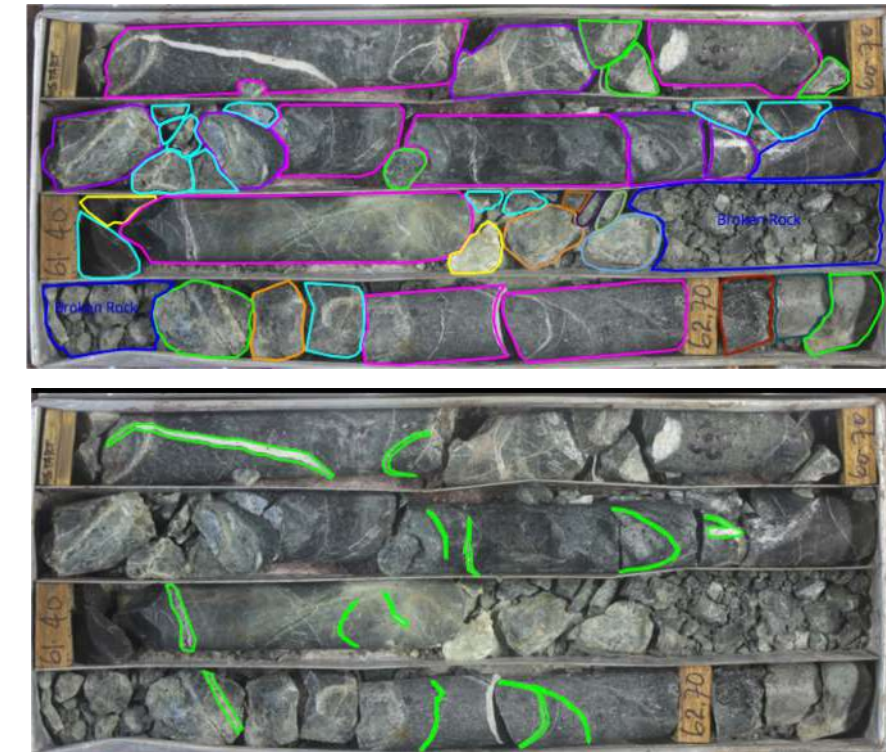
Scenes and Text

Computer Vision

Cloud to Edge | Offline to Online | Cross Industry



Object Detection and Tracking



Segmentation

AI Enhanced Workforce



Opt-in Customer Programs

Personalization: Automate information retrieval, recommendations

Employee Guidance

Customer Insights to improve experience



Human Enhanced Vision

Use cases such as defect detection and parts recognition
Improve service quality and efficiency

New Customer Experiences



TECHNICAL DETAILS

echo look

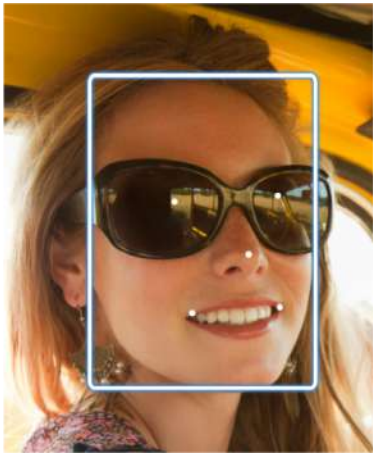
- Built-in LED lighting**
Take photos with four built-in LEDs
- Hands-free camera**
Make your head-to-toe photos pop with intelligent background blur
- Microphone array**
Ask Alexa to take a photo or video of your daily look
- 1.6-watt speaker**
Listen to the weather forecast, news, and more
- Mic & camera off button**
- Standard tripod socket**
Adjustable base mounts on your wall with included kit

intel **REALSENSE**

64%

36%

Details of the outfit are on trend.



티켓 문화의 패러다임을 바꾸다 !!
국내 최초 콘서트에 페이스티켓 도입 성공 (인식률 87%)

FACE TICKET

원본 인증만으로 입장권 및 접근 권한, 메타데이터, 티켓 유효성, 유효기간, 유효지역, 티켓번호, 분실 등 총 10여종의 보안을 한 번에 해결
시행된 한 실재고객의 얼굴이 인증시스템에 등록된 서비스, 유효기간(유효기간) 및 유효지역(유효지역)을 확인

KSTAR

A photograph showing a woman interacting with a facial recognition kiosk at a KSTAR event. The kiosk displays a large screen with a face being scanned. In the background, other people and event staff are visible.

New Wave of Solutions and Strategic Initiatives



Industry 4.0: Smart Factory

Core Toolkit

The AWS ML Stack

Broadest and most complete set of Machine Learning capabilities

AI SERVICES

VISION



Amazon
Rekognition

SPEECH



Amazon
Polly



Amazon
Transcribe

+ Medical
NEW

TEXT



Amazon
Comprehend

+ Medical



Amazon
Translate



Amazon
Textract

SEARCH



Amazon
Kendra

CHATBOTS



Amazon
Lex

PERSONALIZATION



Amazon
Personalize

FORECASTING



Amazon
Forecast

NEW!

FRAUD



Amazon
Fraud Detector

NEW!

DEVELOPMENT



Amazon
CodeGuru

NEW!

CONTACT CENTERS



Contact Lens
For Amazon Connect

ML SERVICES



Amazon SageMaker

Ground
Truth

Augmented
AI

ML
Marketplace

SageMaker Studio IDE

Built-in
algorithms

Notebooks

Experiments

Model
training &
tuning

Debugger

Autopilot

Model
hosting

Model Monitor

Neo

ML FRAMEWORKS & INFRASTRUCTURE



NEW!
TensorFlow



NEW!

NEW!
PYTORCH



GLUON



K Keras

Deep Learning
AMIs & Containers

GPUs &
CPUs

Elastic
Inference

Inferentia

FPGA

Amazon Rekognition

Turn-key Computer Vision Capabilities



Labels (object, scenes, and activities)



Unsafe image and video detection



Text in image



Pathing



Face search



Face detection and analysis



Celebrity recognition



Real-time video analysis

Amazon Rekognition Custom Labels

AutoML for Image Classification and Object Detection



- Image processing and augmentation automation
- State-of-the-art deep learning models for feature extraction and transfer
- Few-shot learning capable
- Model and data versioning
- Fully-managed: annotation, model training and serving

Amazon SageMaker

Prepare

Build

Train & Tune

Deploy & Manage

| | | | |
|--|---|--|---|
| <div>Amazon SageMaker Studio</div> <div>Integrated Development environment(IDE) for Machine Learning</div> | | | |
| <div>Amazon SageMaker Autopilot</div> <div>Automatically build and train models</div> | | | <div>One Click Deployment</div> <div>Supports real-time, batch & multi-model</div> |
| <div>Amazon SageMaker GroundTruth</div> <div>Build and manage training dataset</div> | <div>Amazon SageMaker Notebooks</div> <div>One-click notebooks with elastic compute</div> | <div>One Click Training</div> <div>Supports supervised, unsupervised & RL</div> | <div>Amazon SageMaker Model Monitor</div> <div>Automatically detect concept drift</div> |
| <div>Processing Job</div> <div>Supports Python or Spark</div> | <div>AWS Marketplace</div> <div>Pre-built algorithms, models, and data</div> | <div>Automatic Model Tuning</div> <div>One-click hyperparameter optimization</div> | <div>Amazon SageMaker Neo</div> <div>Train once, deploy anywhere</div> |
| | | <div>Amazon SageMaker Experiments</div> <div>Capture, organize, and compare every step</div> | <div>Amazon Elastic Inference</div> <div>Auto scaling for 75% less</div> |
| | | <div>Amazon SageMaker Debugger</div> <div>Debug and profile training runs</div> | <div>Amazon Augmented AI</div> <div>Add human review of model predictions</div> |



Amazon SageMaker GroundTruth

Image Classification (Single Label)

Get workers to categorize images into individual classes. [Info](#)

☒

Basketball

☐

Soccer



Image Classification (Multi-label)

Get workers to categorize images into one or more classes. [Info](#)

☒


Human

☒

Vehicle

☐

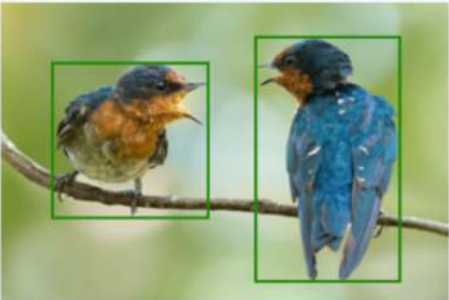
Animal



☒


Bounding box

Get workers to draw bounding boxes around specified objects in your images. [Info](#)



Semantic segmentation

Get workers to draw pixel level labels around specific objects and segments in your images. [Info](#)



Label verification

Get workers to verify existing labels in your dataset. [Info](#)

☒

Correct label

☐

Incorrect label

Car



Automatic Labeling using
active learning

Private workforce
management

© 2019, Amazon Web Services, Inc. or its Affiliates.

Amazon SageMaker Training

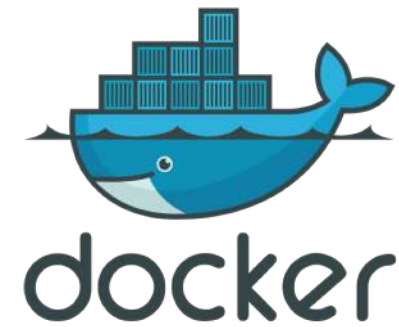
BUILT-IN ALGORITHMS

- 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

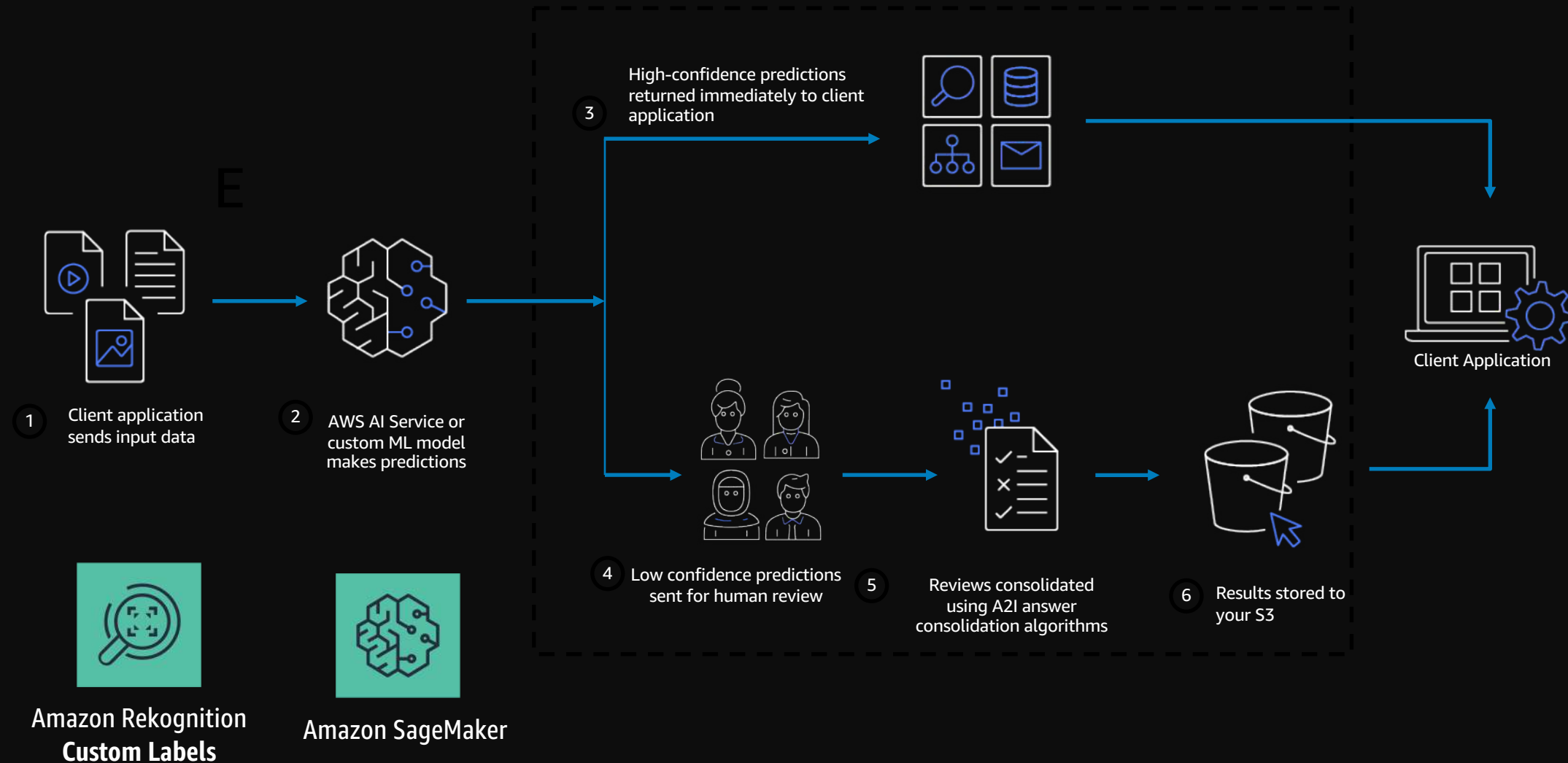
BRING YOUR OWN SCRIPT



BRING YOUR OWN ALGORITHM



Amazon SageMaker Augmented AI (A2I) for Human Review



[Aws-Sample](#) for Rekognition

Requirements for CV@Edge

BANDWIDTH



1 billion cameras WW (2020)
10's of petabytes per day

LATENCY



30 images per second
200ms latency

PRIVACY



Confidentiality
Private cloud or on-premises storage

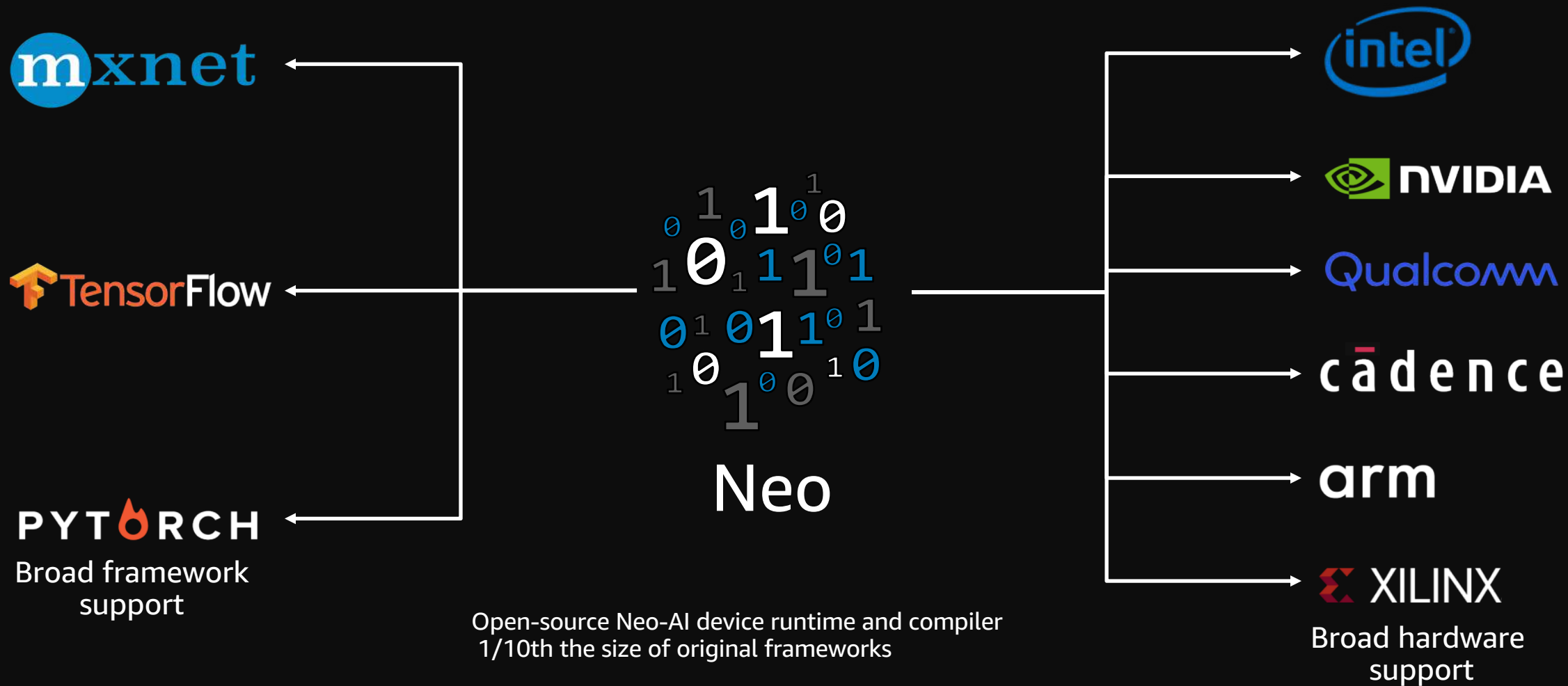
AVAILABILITY



50% of populated world < 8mbps
Bulk of uninhabited world no 3G+

Amazon SageMaker Neo

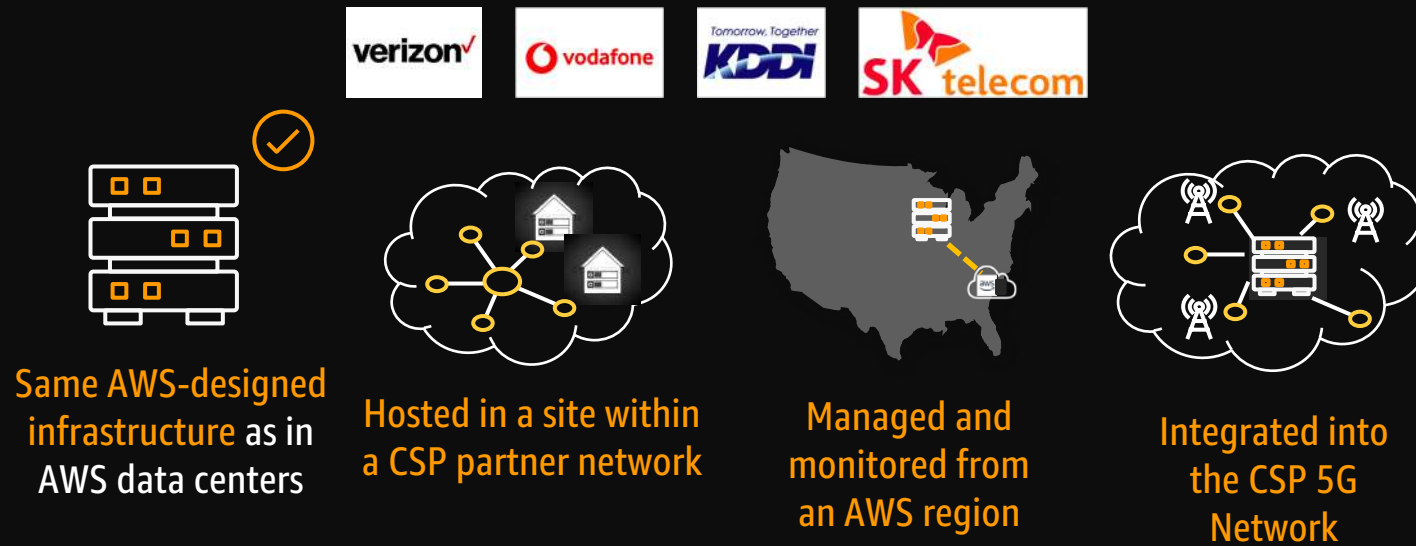
Train once and run anywhere with improved performance



CV Near the Edge

Wavelength

Run latency-sensitive portions of applications in “Wavelength Zones,”



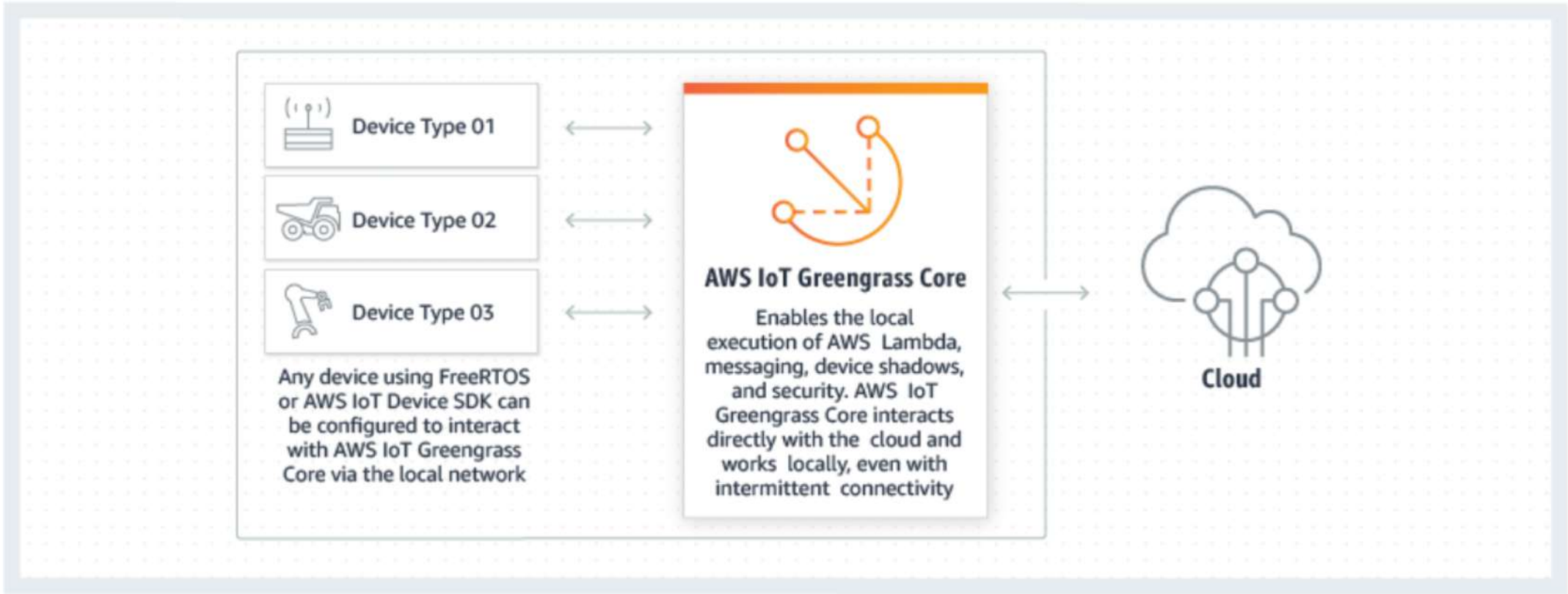
Local Zones





AWS Greengrass

MANAGING EDGE TO CLOUD DEVICE MECHANICS

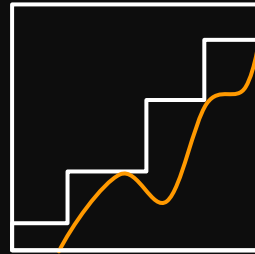


Amazon Elastic Inference

Reduce Deep Learning Inference costs up to 75%



Lower inference costs



Match capacity
to demand



Available between 1 to 32 TFLOPS per
accelerator

KEY FEATURES

Integrated with
Amazon EC2 and
Amazon SageMaker

Support for TensorFlow, Apache MXNet,
and ONNX
with PyTorch coming soon

Single and
mixed-precision
operations

AWS Inferentia

Low cost Machine Learning Inference Optimized Hardware

- 4 Neuron Cores
- Up to 128 TOPS
- 2-stage memory hierarchy
 - Large on-chip cache and commodity DRAM
- Supports FP16, BF16, INT8 data types
- Fast chip-to-chip interconnect

| Instance size | vCPUs | Memory (GiB) | Storage | Inferentia Chips | Neuron Core Pipeline Mode | Network B/W | EBS B/W |
|---------------|-------|--------------|----------|------------------|---------------------------|---------------|----------------|
| inf1.xlarge | 4 | 8 | EBS only | 1 | N/A | Up to 25 Gbps | Up to 3.5 Gbps |
| inf1.2xlarge | 8 | 16 | EBS only | 1 | N/A | Up to 25 Gbps | Up to 3.5 Gbps |
| inf1.6xlarge | 24 | 48 | EBS only | 4 | Yes | 25 Gbps | 3.5 Gbps |
| inf1.24xlarge | 96 | 192 | EBS only | 16 | Yes | 100 Gbps | 14 Gbps |

- Available in 4 sizes
- Single and multi chip instances
- AWS 2nd Gen Intel Xeon Scalable Processors
- Up to 100Gbps networking bandwidth
- Will support managed services such as Amazon SageMaker, EKS and ECS



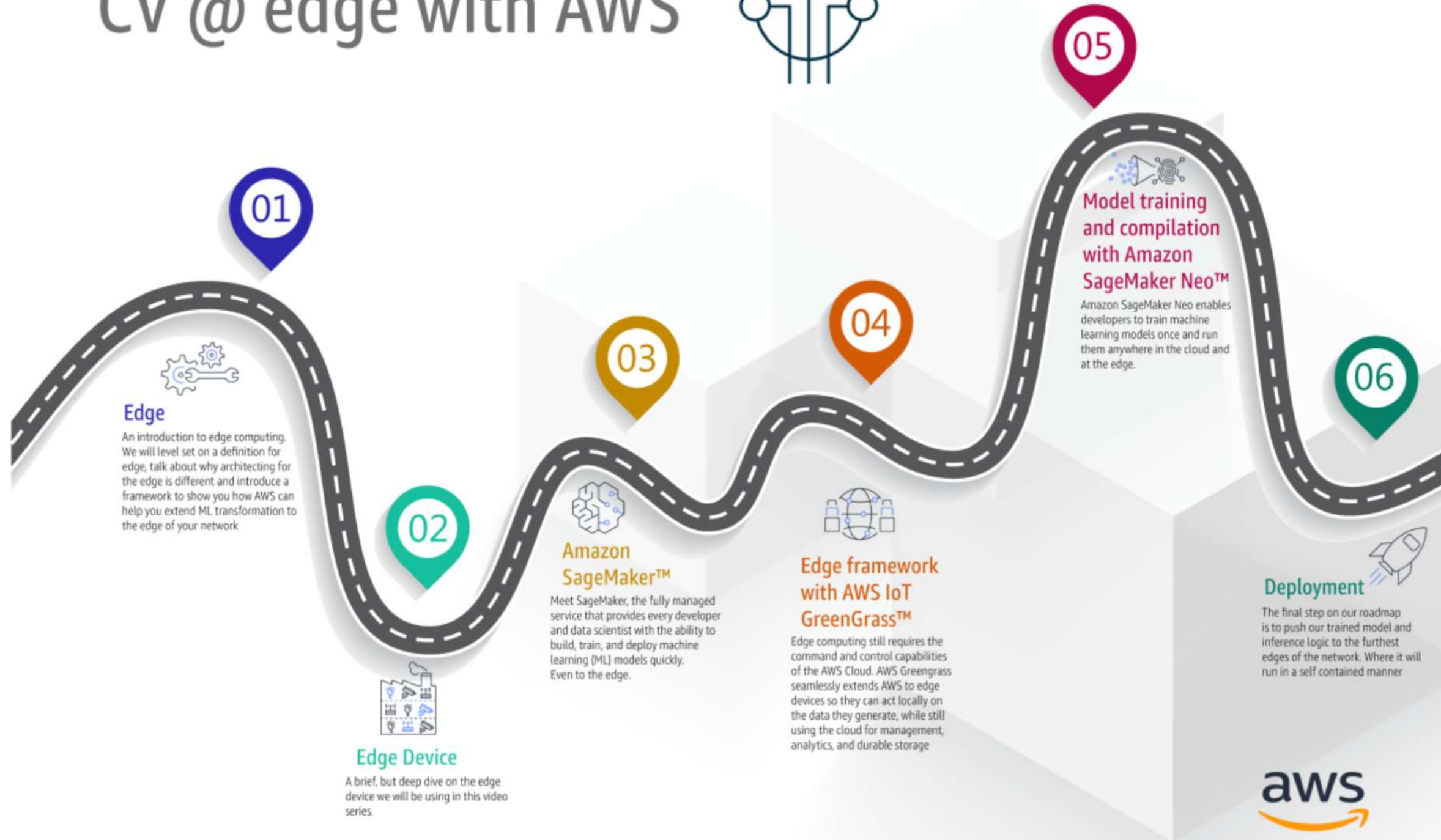
Workshops

AWS Computer Vision Jump Starter Kit

[Repository](https://github.com/dylan-tong-aws/aws-cv-jumpstarter.git): **git clone** <https://github.com/dylan-tong-aws/aws-cv-jumpstarter.git>

- **Lab:** Amazon Recognition **Custom Labels**
- Object Detection Series:
 - **Lab1:** Amazon **GroundTruth**
 - **Lab2:** Amazon SageMaker **Object Detection Algorithm**
 - **Lab3:** Amazon SageMaker + GluonCV **YOLOv3** Object Detection (BYOS)
- **Lab4:** Amazon SageMaker + GluonCV **Simple Pose** Estimation

CV @ edge with AWS



Challenge: Jetson Nano Smart Cam



NVIDIA Jetson Nano Developer Kit
Quad-core Arm A57 processor @ 1.43 GHz
System Memory: 4GB
128-core Maxwell GPU: 0.5 TFLOPs



Arducam 8MP Autofocus
Replacement for Raspberry
Pi Camera Module V2,
IMX219

FAST MULTI-OBJECT TRACKING



INFERENCE TIME: ~0.01-0.1s
CAM MAX FPS: 21

PREDICTIVE AUTO FOCUS



SRGAN: SUPER RESOLUTION ZOOM

SUPER RES ZOOM



LOW RES ZOOM



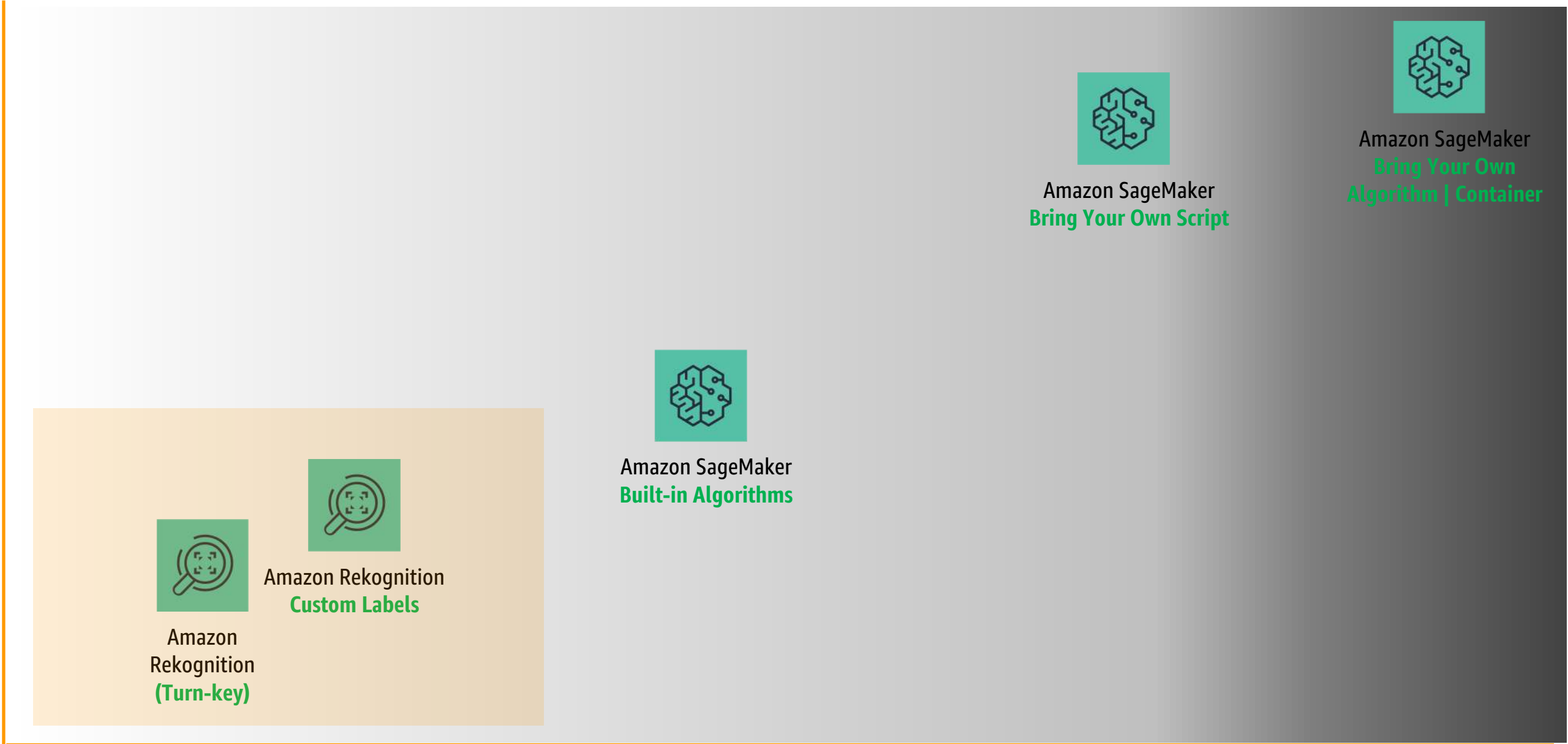
Additional Production Guidance

Strategy for Selecting Your Tools

LOW SKILL AND RESOURCES

HIGH SKILL AND RESOURCES

USE CASE
COVERAGE

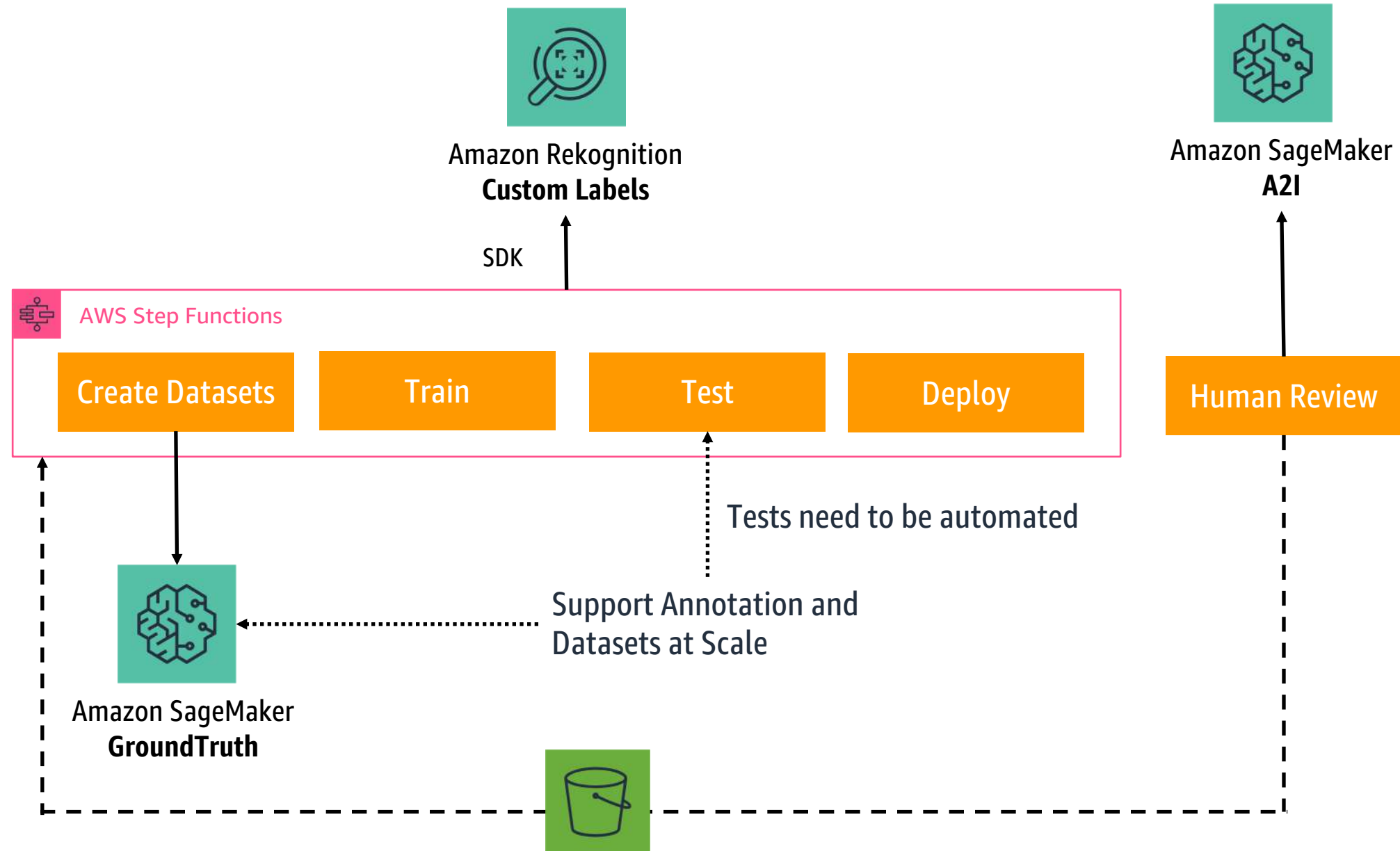


* *Conceptual Illustration*

TIME TO VALUE



MLOps for Rekognition Custom Labels

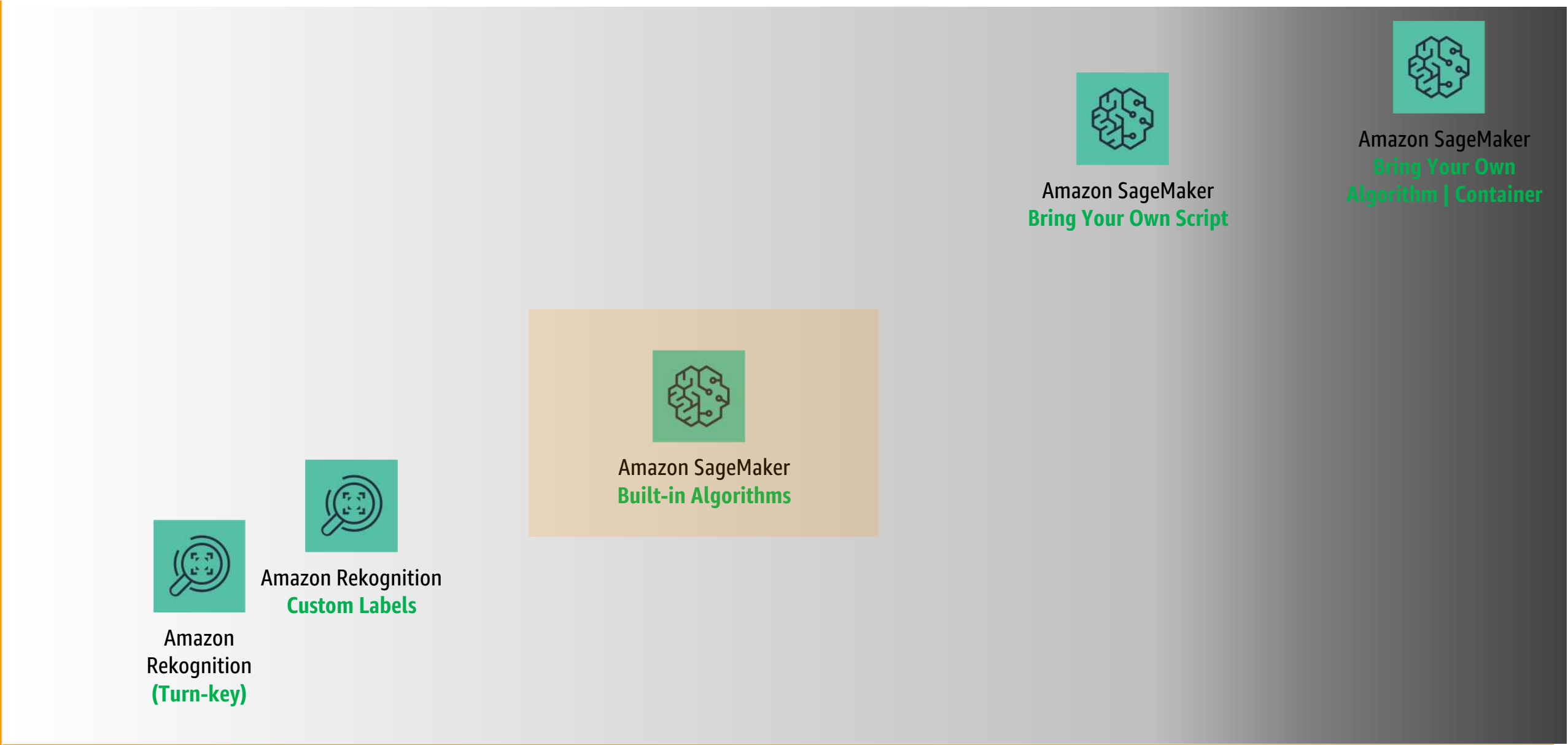


Strategy for Selecting Your Tools

LOW SKILL AND RESOURCES

HIGH SKILL AND RESOURCES

USE CASE
COVERAGE



** Conceptual Illustration*

TIME TO VALUE

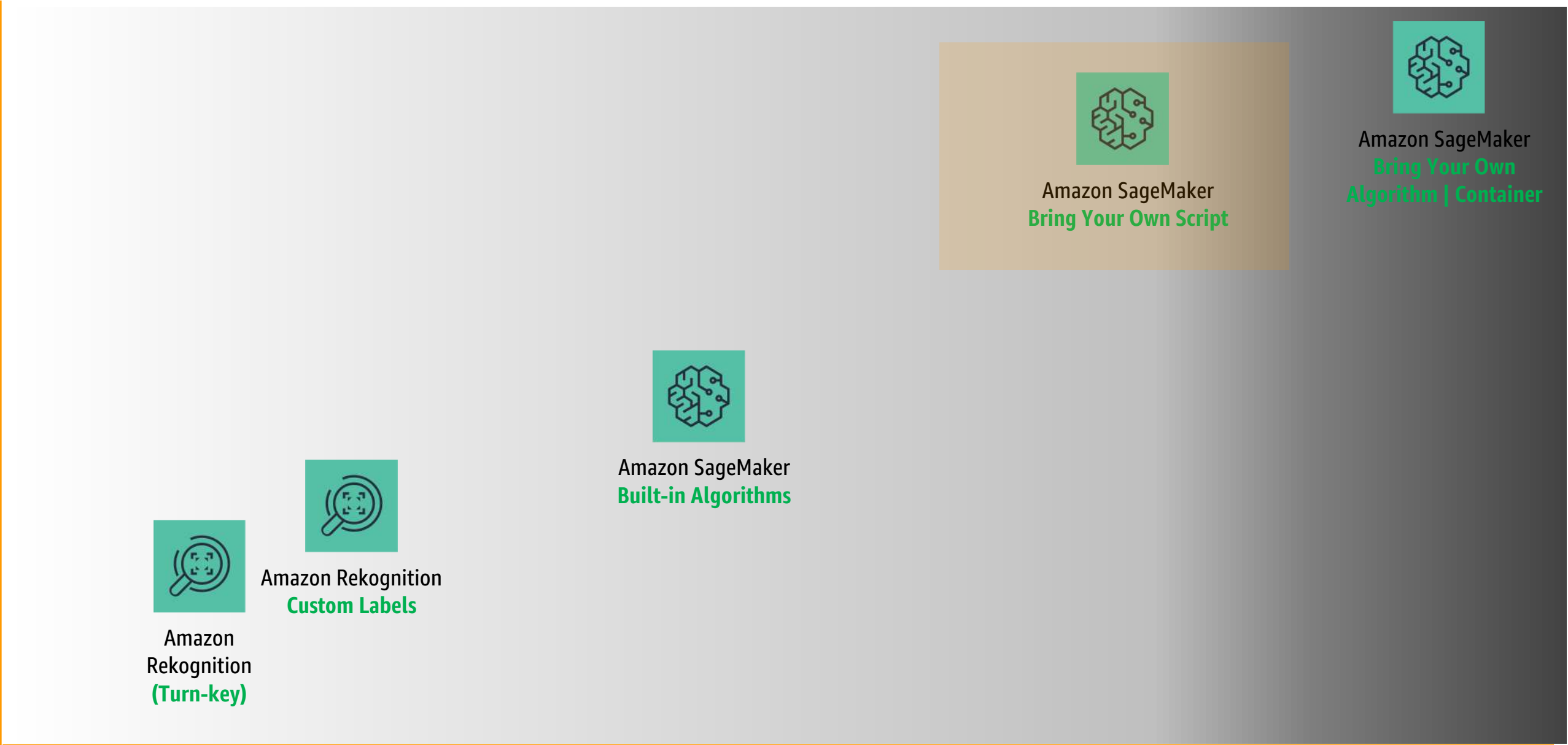


Strategy for Selecting Your Tools

LOW SKILL AND RESOURCES

HIGH SKILL AND RESOURCES

USE CASE
COVERAGE



* Conceptual Illustration


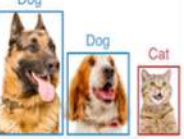
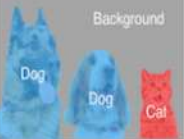






TIME TO VALUE



Bring Your Own Script Scenario



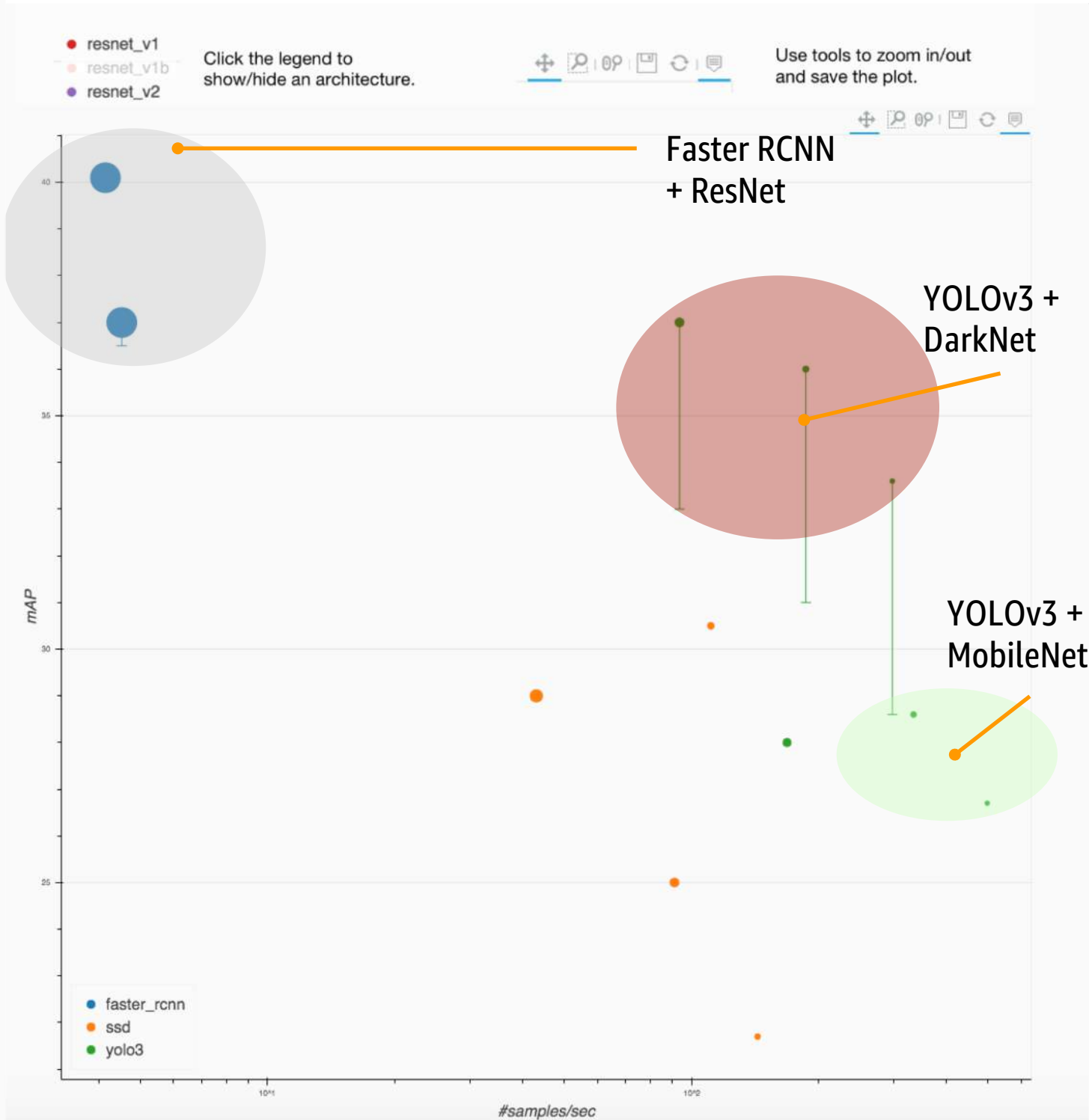
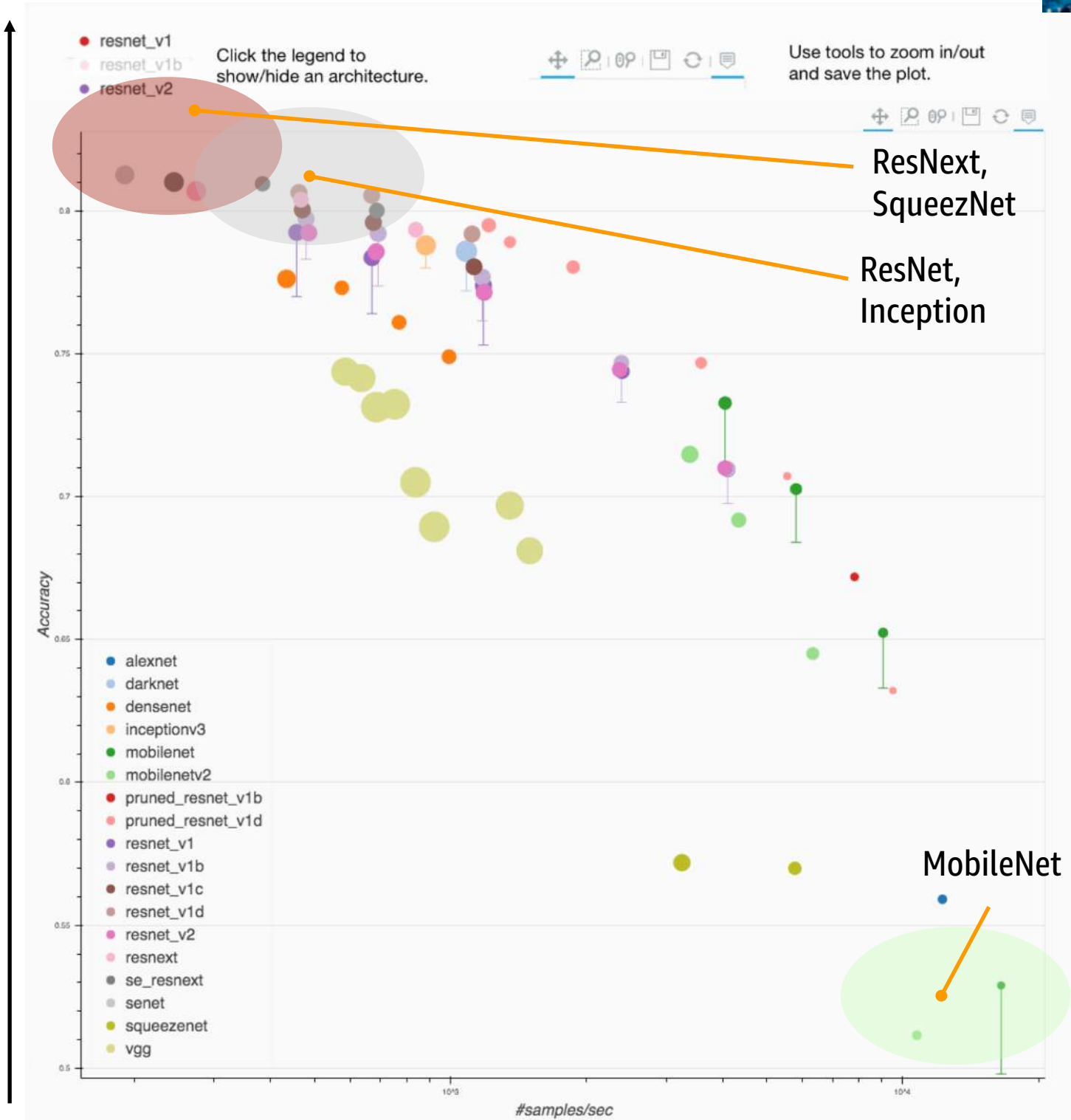
- Image Classification
- Object Detection
- Semantic Segmentation
- Instance Segmentation
- Pose Estimation
- Action Recognition
- Depth Prediction
- GAN (e.g. SRGAN)
- Person Re-Id

| Application | Illustration | Available Models |
|--|--|---|
| Image Classification: recognize an object in an image. |  Dog | 50+ models, including ResNet, MobileNet, DenseNet, VGG, ... |
| Object Detection: detect multiple objects with their bounding boxes in an image. |  Dog Dog Cat | Faster RCNN, SSD, Yolo-v3 |
| Semantic Segmentation: associate each pixel of an image with a categorical label. |  Dog Dog Cat | FCN, PSP, ICNet, DeepLab-v3, DeepLab-v3+, DANet, FastSCNN |
| Instance Segmentation: detect objects and associate each pixel inside object area with an instance label. |  Dog 1 Dog 2 Cat 1 | Mask RCNN |
| Pose Estimation: detect human pose from images. |  | Simple Pose |
| Video Action Recognition: recognize human actions in a video. |  Biking | TSN, C3D, i3D, P3D, R3D, R2+1D, Non- local, SlowFast |
| Depth Prediction: predict depth map from images. |  | Monodepth2 |
| GAN: generate visually deceptive images |  | WGAN, CycleGAN, StyleGAN |
| Person Re-ID: re-identify pedestrians across scenes |  | Market1501 baseline |

HIGHER PREDICTIVE
PERFORMANCE (ACCURACY)



HIGHER THROUGHPUT
AND LOWER LATENCY

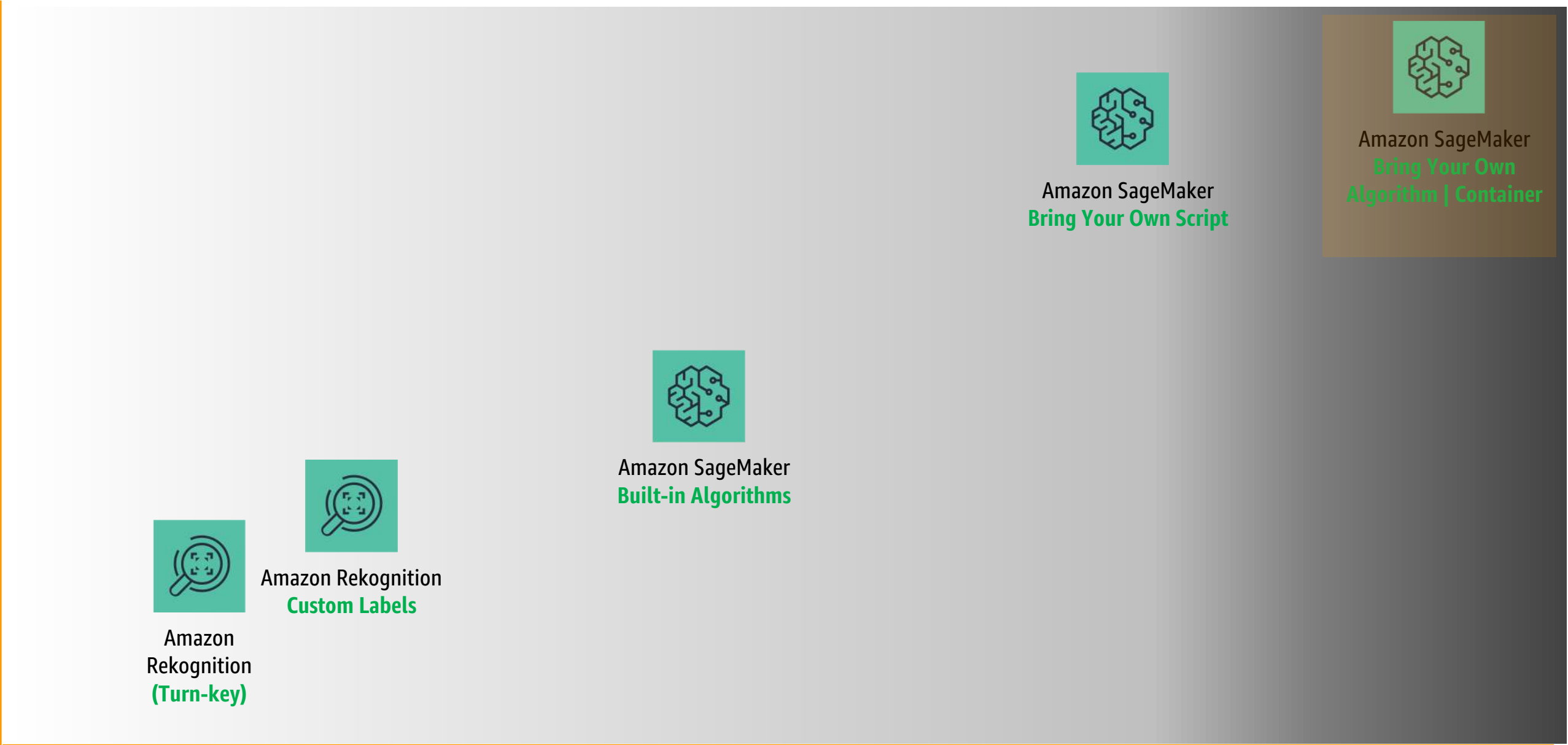


Strategy for Selecting Your Tools

LOW SKILL AND RESOURCES

HIGH SKILL AND RESOURCES

USE CASE
COVERAGE

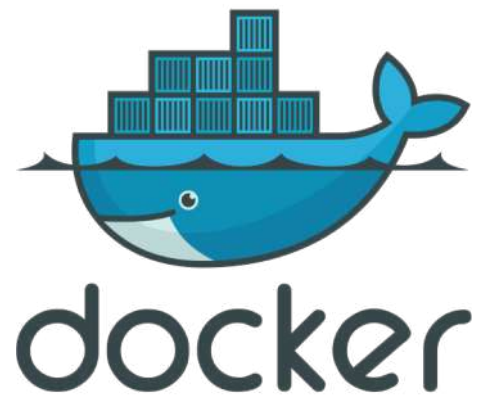


** Conceptual Illustration*

TIME TO VALUE



Bring Your Own Algorithm Scenarios



- Custom implementation such as on C/C++ for lower latency
- DevOps guidelines and custom libraries required.

Training Optimizations: Data Ingestion

- Use framework optimized formats like **RecordIO**. Leverage the [Im2rec tool](#) to pre-process your images.
- 2nd best option. Create an [Augmented Manifest](#) file to enable **Pipe Mode** (Note: SageMaker GroundTruth will package annotations into a compatible manifest file).

If you're bringing your own script or algorithm, you'll need to implement pipe-mode streaming ingest logic. [This notebook](#) shows you how.

Training Optimizations: Time vs. Cost

- P3 instances are equipped with NVIDIA V100s. Currently the fastest.
- G4 instances are equipped with T4 GPUs. They're slower than V100s but reportedly are better in terms of performance/\$.
- Consider SageMaker Managed Spot Training. Offers a good trade-off between GPU training costs and training time.

Up to 90% discount, but interruptions can slow down training. SageMaker handles the interruptions automatically. [This notebook](#) will show you how to use Managed Spot Training.

Inference Optimizations: Infrastructure Selection



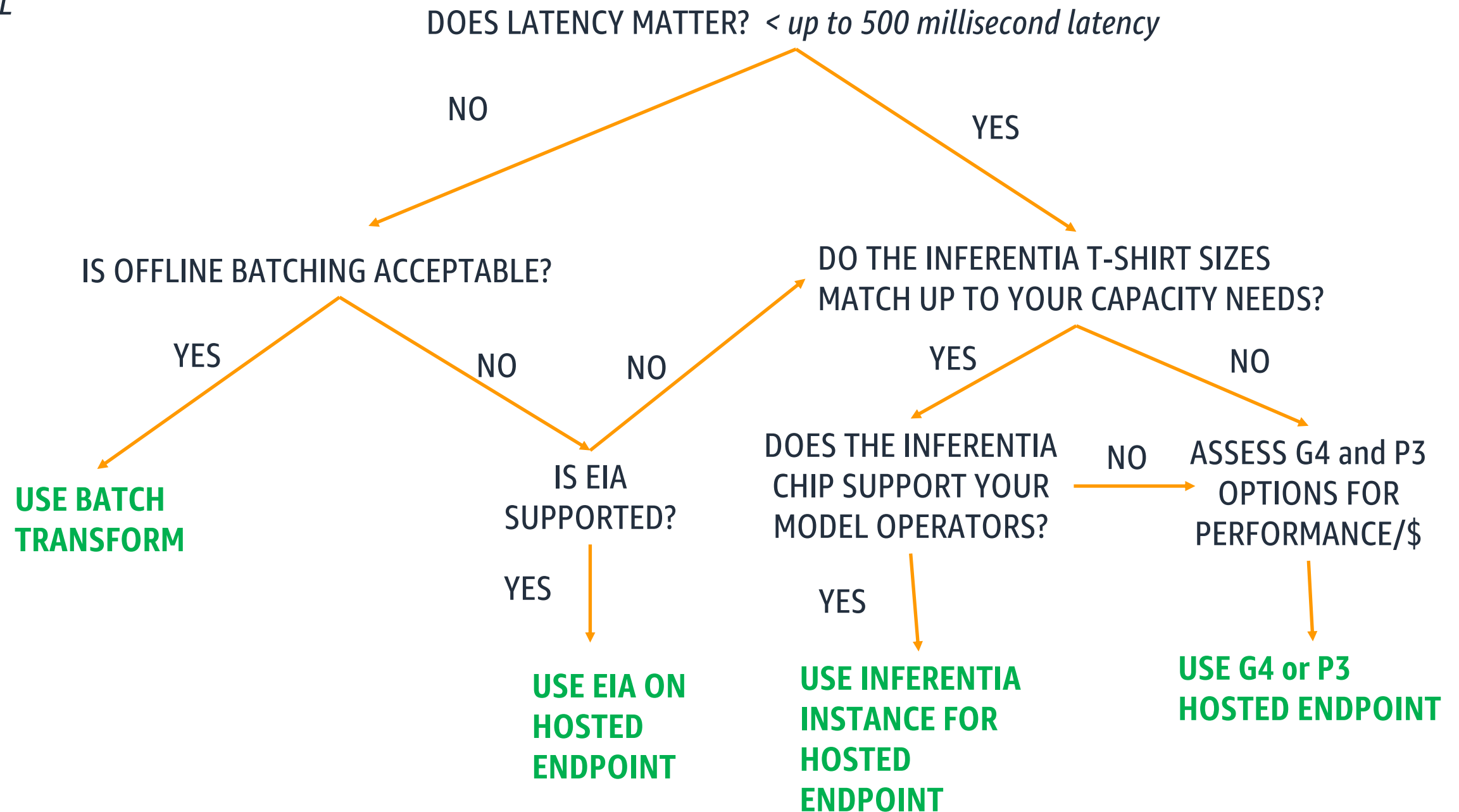
Model complexity: mobilenetV2 vs. resnet-152
Inference batch size
Input size

Resource availability and competing workloads for CPU and RAM
Framework: CUDA and [MKL-DNN](#) Support

Inference Optimizations: GPU Inference Cost Optimization

PERFORMANCE BENCHMARK
DRIVEN DECISIONS ARE STILL
PREFERRED. REMEMBER:

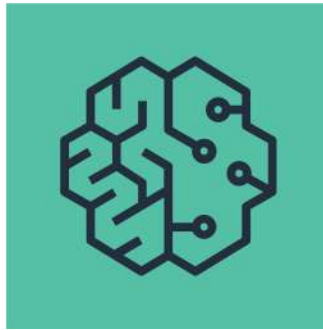
- Adjust for performance inefficiencies in production i.e.. Low GPU utilization.
- Use latency and TPS/\$ as your metrics.



Inference Optimizations: General Cost Optimizations

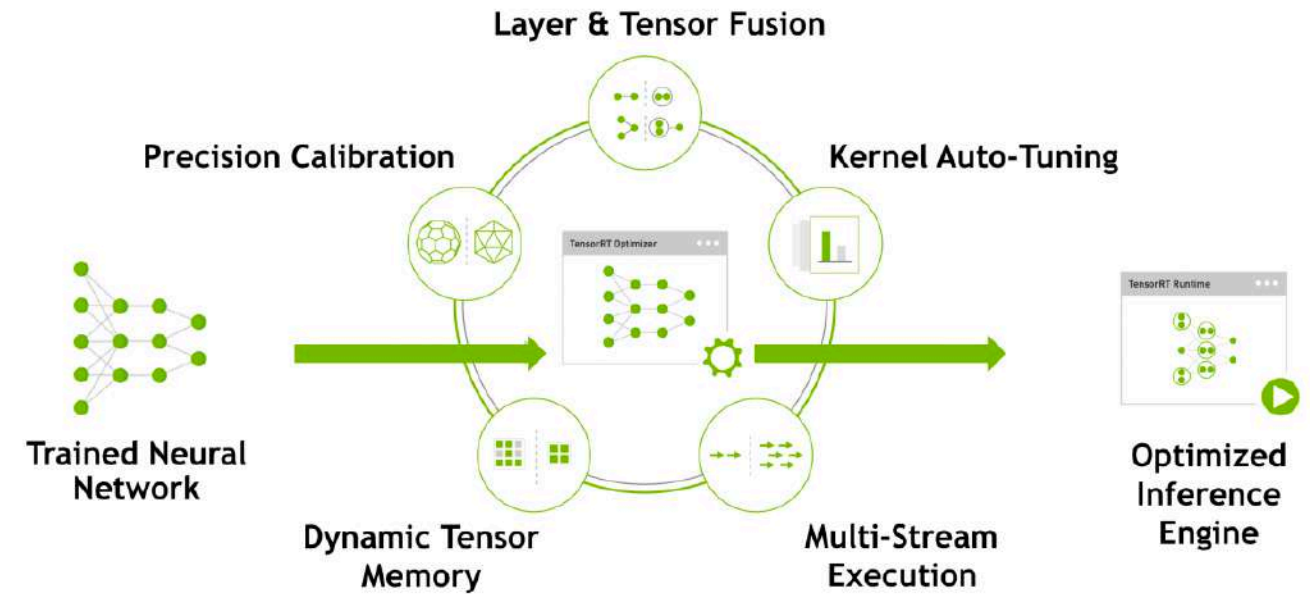
- Configure auto-scaling
- Create [multi-model endpoints](#) whenever possible.

Inference Optimizations: Model Optimization



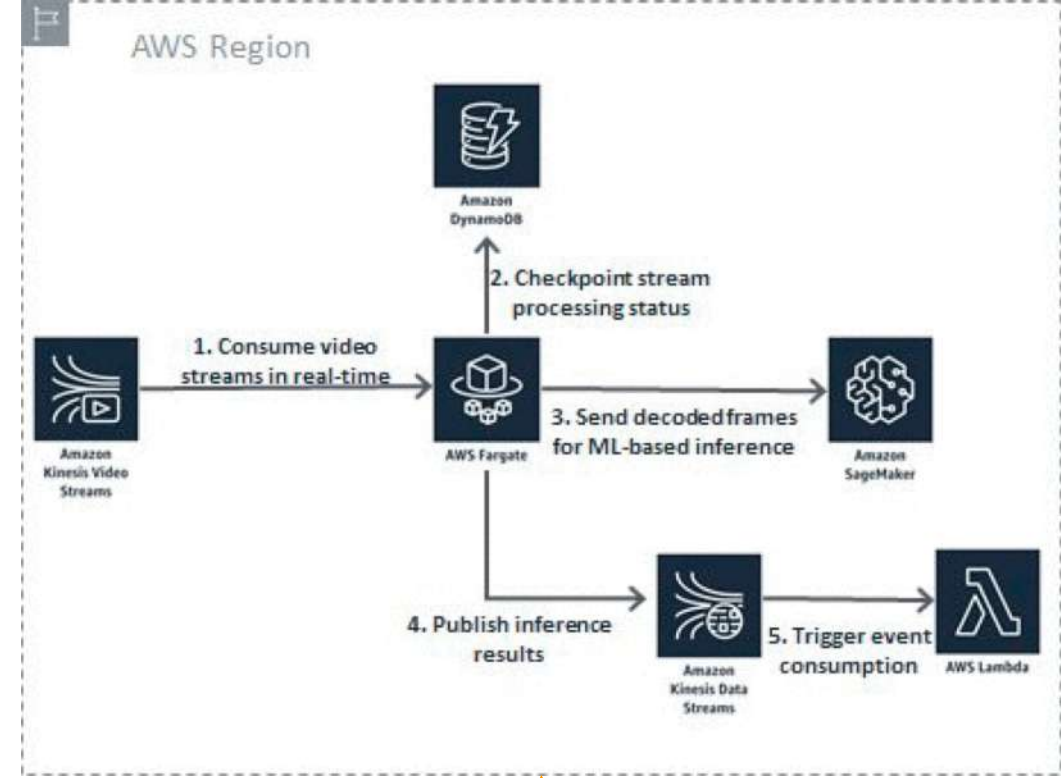
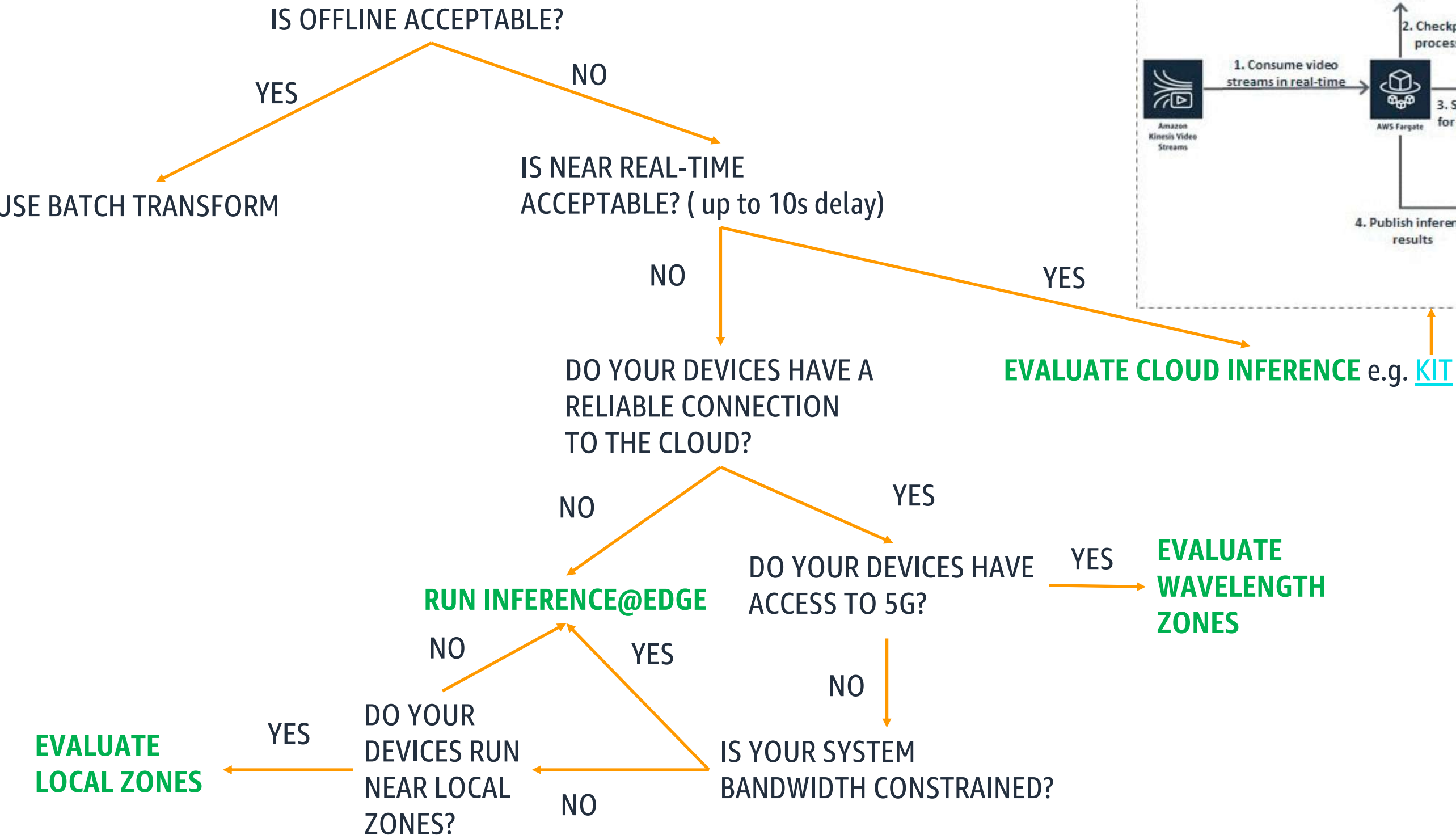
Amazon
SageMaker **Neo**

Neo Runtime: [DLR](#)



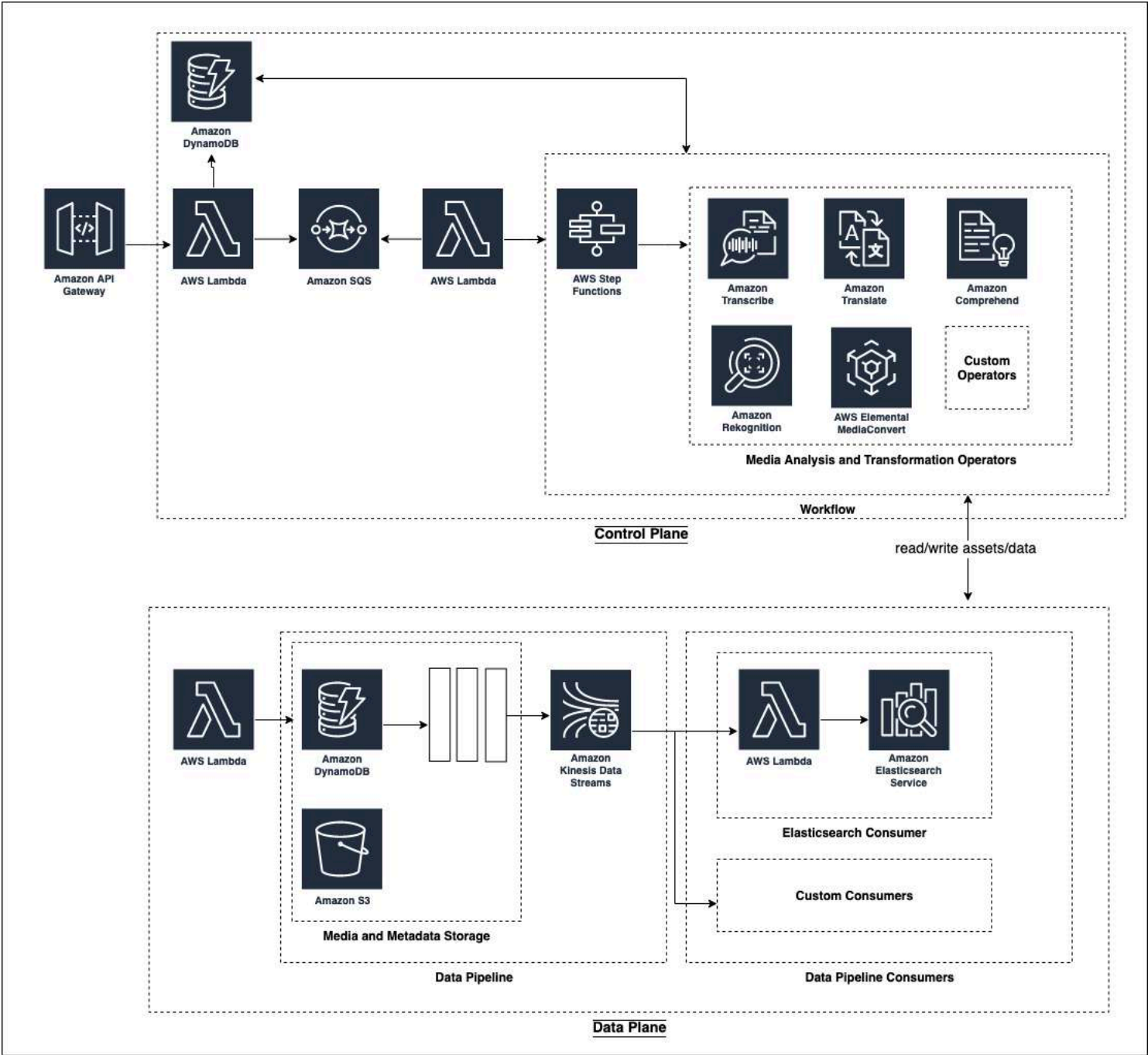
NVIDIA TensorRT
Programmable Inference Accelerator

Inference Optimizations: Video Analysis



AWS CV Solutions

Media Insights Engine



Media Insights Engine - - Execute Workflow Architecture

Media Insights Engine Upload Collection Analytics Help

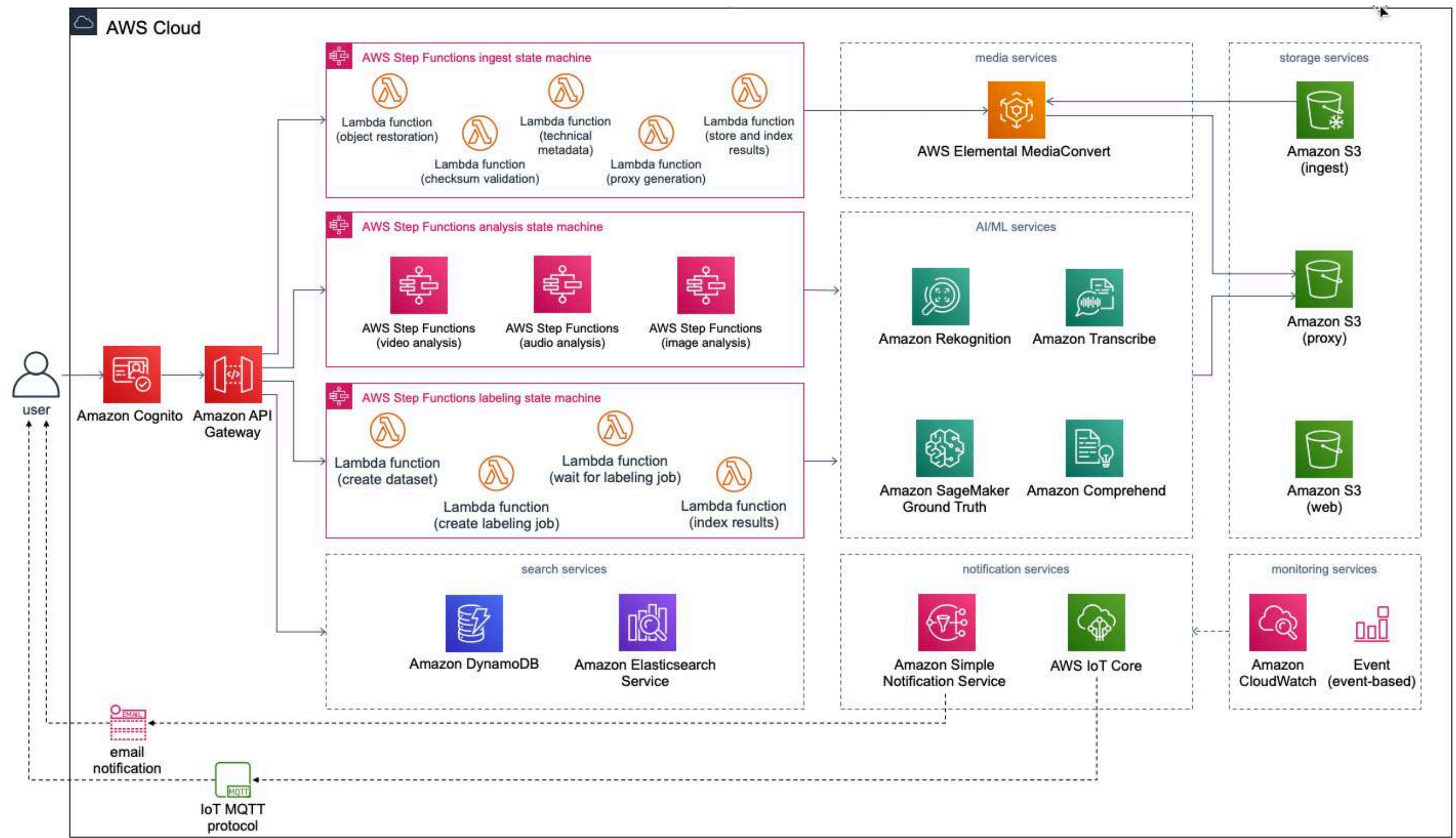
Media Collection

Discover insights in your media by searching for keywords, objects, or even people.

Search Collection...

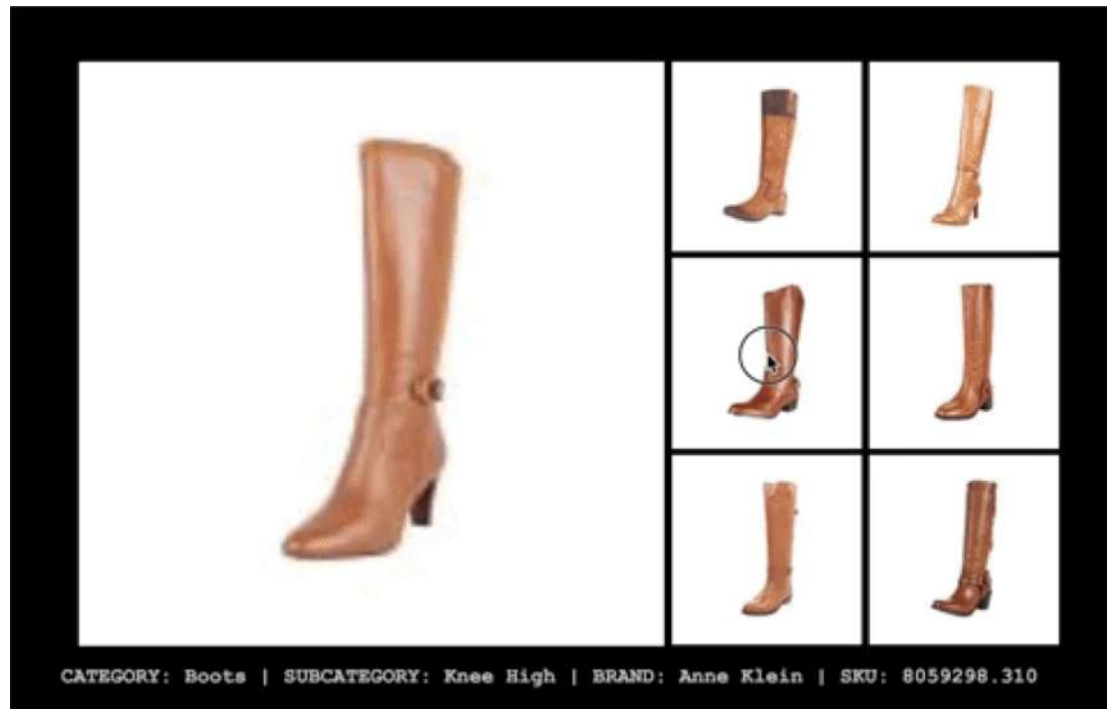
| Thumbnail | File Name | Status | Asset ID | Created | Actions |
|-----------|--------------------------------|----------|--------------------------------------|--------------------------|--|
| | GrandeTour.mp4 | Complete | 7bed493d-d376-491f-899a-465bacdb279c | 2019-09-24T21:32:04.000Z | Analyze Delete |
| | MozartInTheJungle-language.mp4 | Complete | a56229ed-154d-40a9-8f18-e5ef802b61c | 2019-09-24T21:30:39.000Z | Analyze Delete |
| | RoseParadePasadena.mp4 | Complete | 6d13316b-903a-4323-93be-e52ab50d1edd | 2019-09-24T21:28:21.000Z | Analyze Delete |
| | AmazonVideoSizzle2019.mp4 | Complete | fc17ca4b-337b-44e4-a2c9-a076f1ec7b2c | 2019-09-24T21:28:09.000Z | Analyze Delete |
| | boulder_eats-ft-01.mp4 | Complete | 77dc708c-17e0-46c1-a3fd-afe31fafd92a | 2019-09-24T21:16:41.000Z | Analyze Delete |

Media2Cloud Solution



Additional Examples

PyTorch Siamese Network on SageMaker



[Blog](#)

[Repository](#)