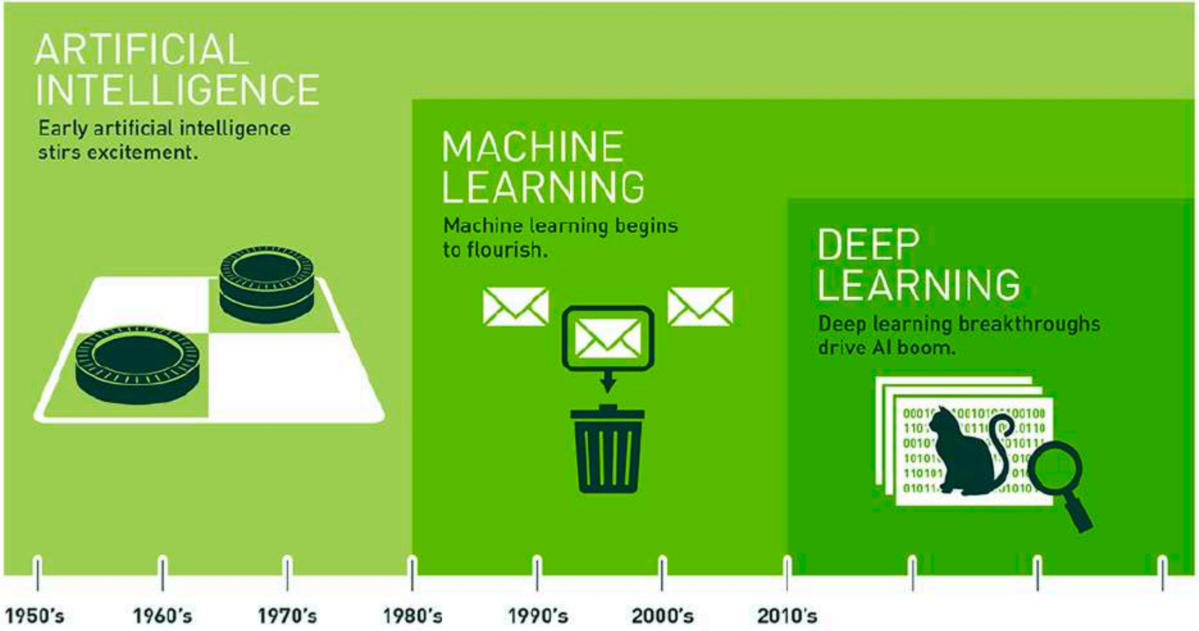
AWS | Artificial Intelligence

Centerpiece for digital transformation

Dylan Tong, AI/Machine Learning Partner Solutions Architect

What is AI?



ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.

MACHINE LEARNING

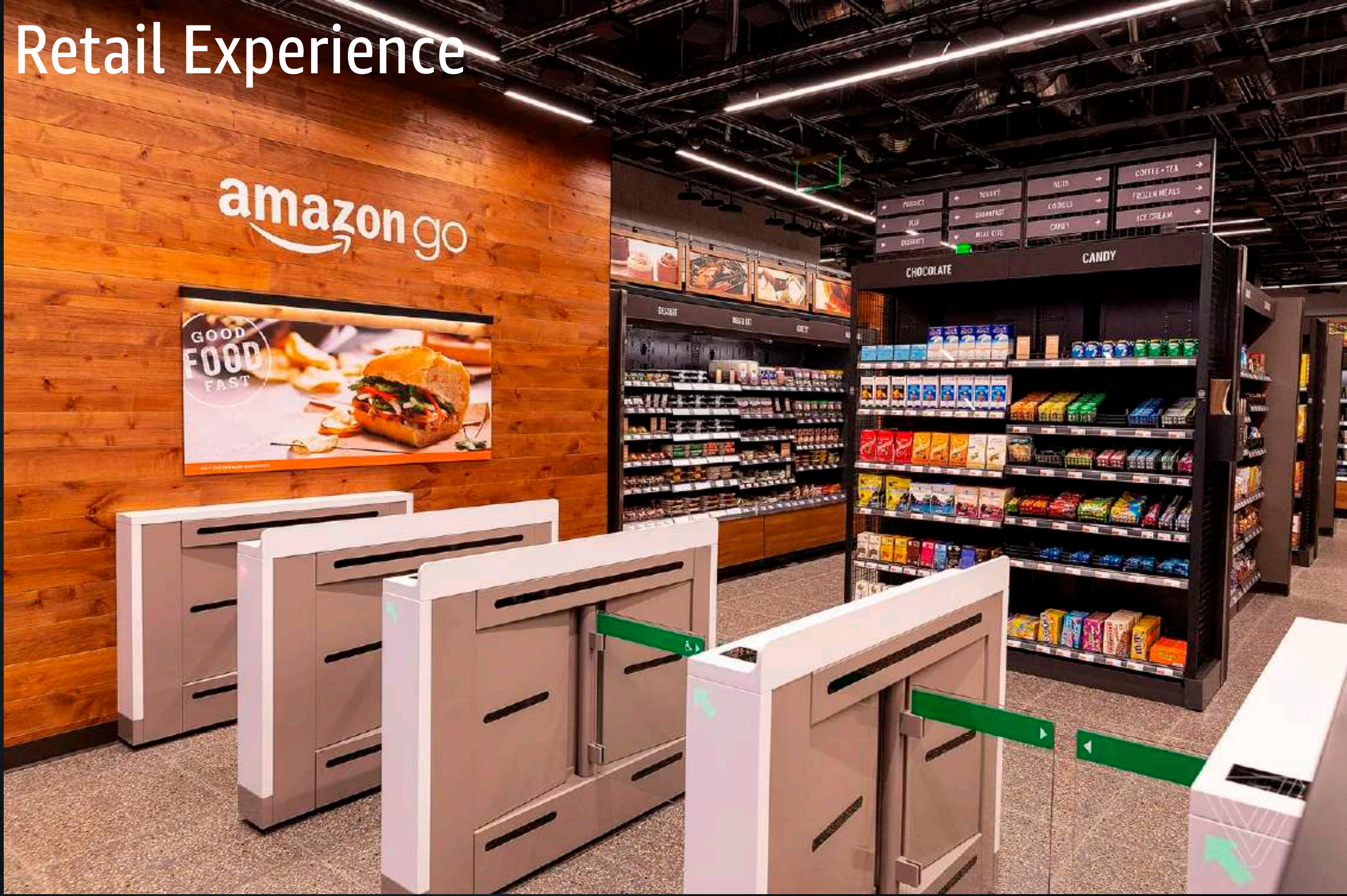
Machine learning begins to flourish.

DEEP LEARNING

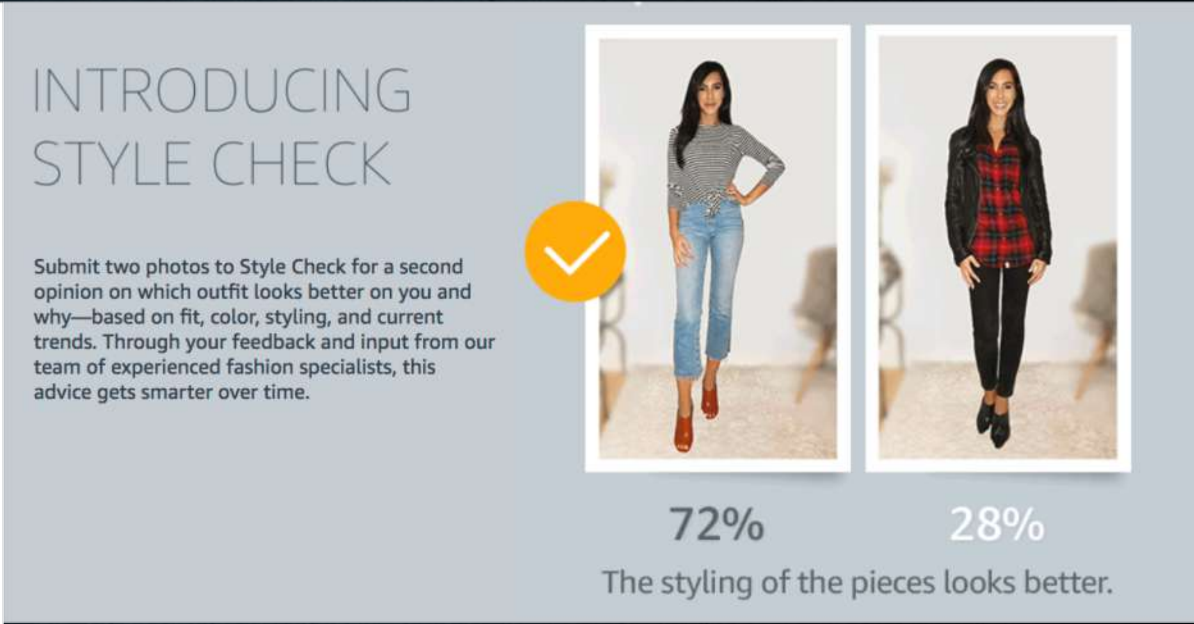
Deep learning breakthroughs drive AI boom.

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Reinventing the Retail Experience



AI Driven Stylist



INTRODUCING
STYLE CHECK

Submit two photos to Style Check for a second opinion on which outfit looks better on you and why—based on fit, color, styling, and current trends. Through your feedback and input from our team of experienced fashion specialists, this advice gets smarter over time.

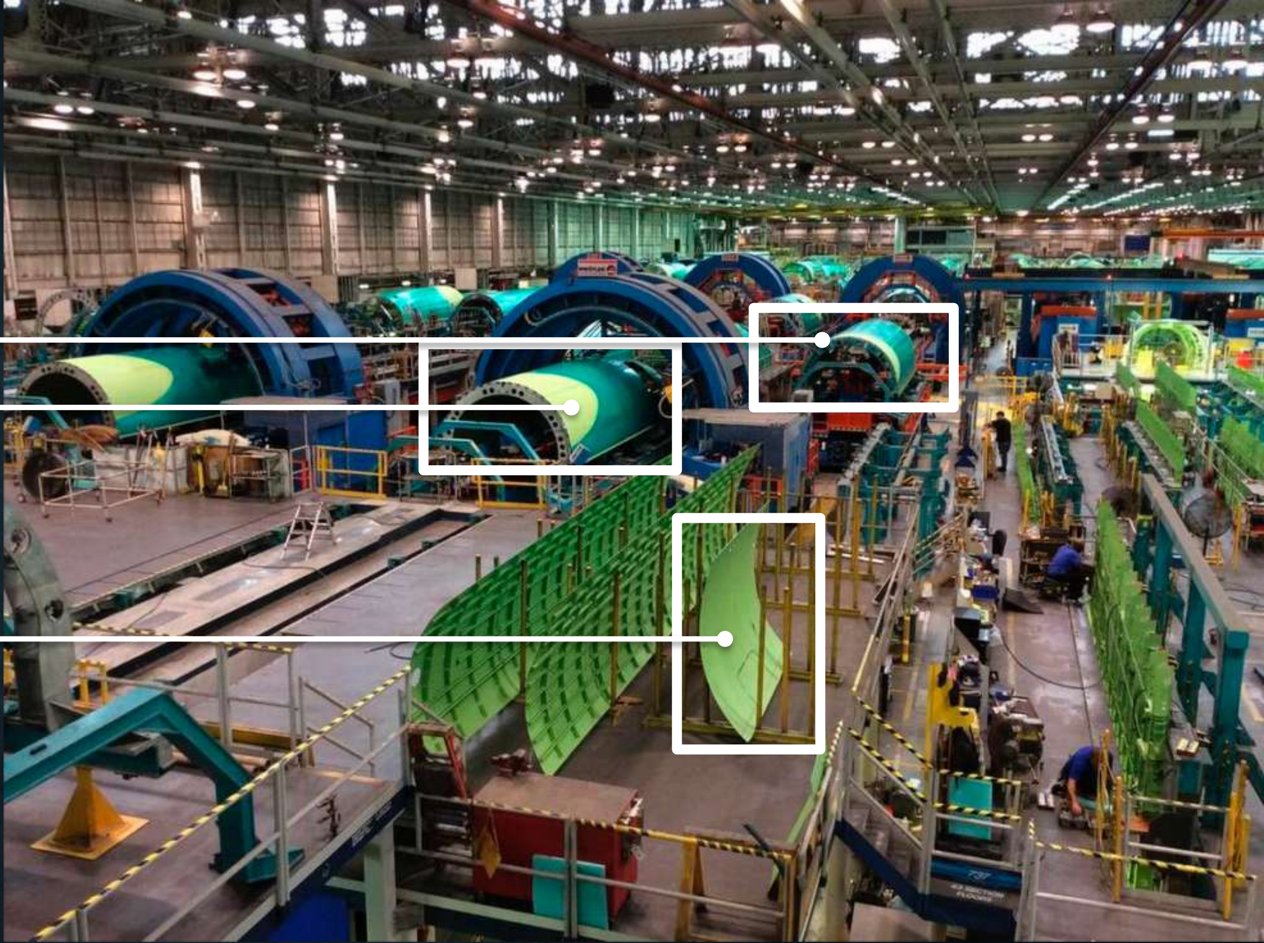
72% 28%

The styling of the pieces looks better.

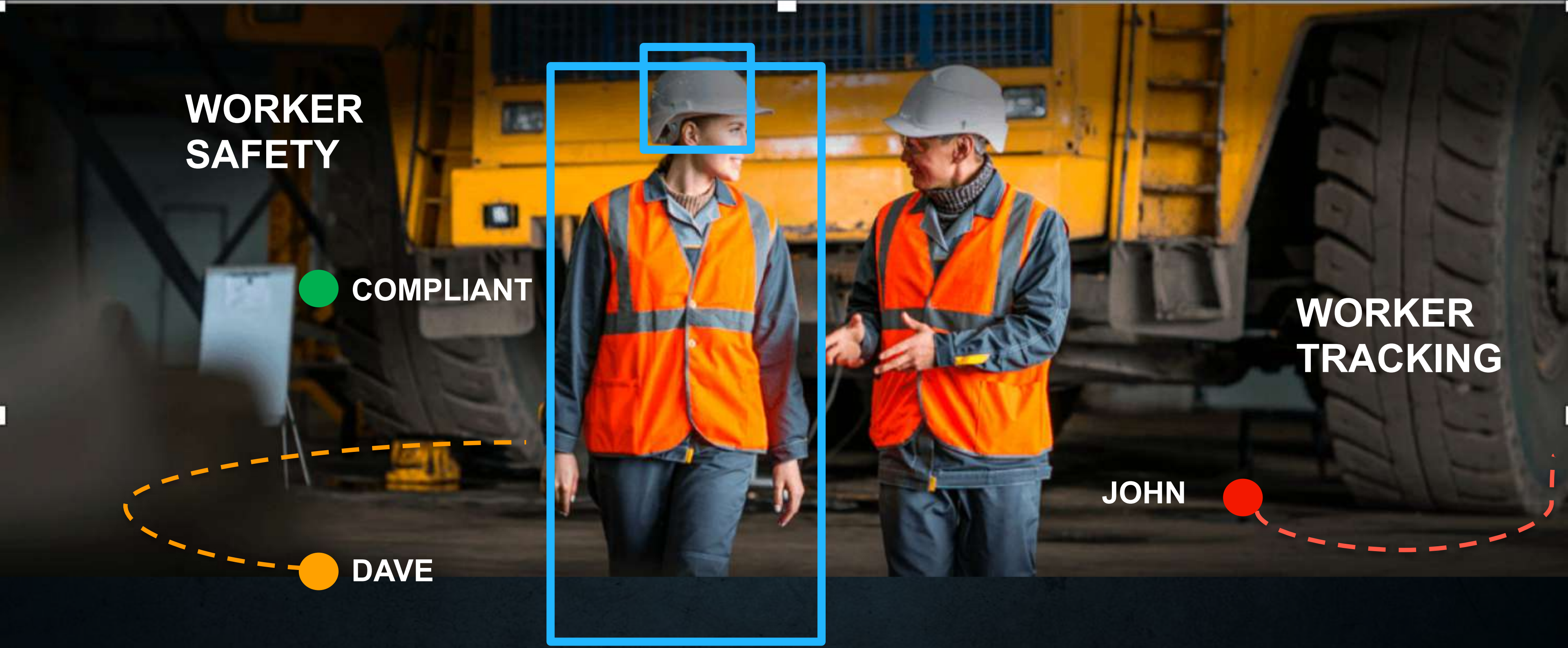
Smart Factory

Inspection:
Granular Quality and Progress Tracking

Asset Tracing (SKU)



The Connected Worker



The Connected Worker

Improve worker productivity through new digital experiences:

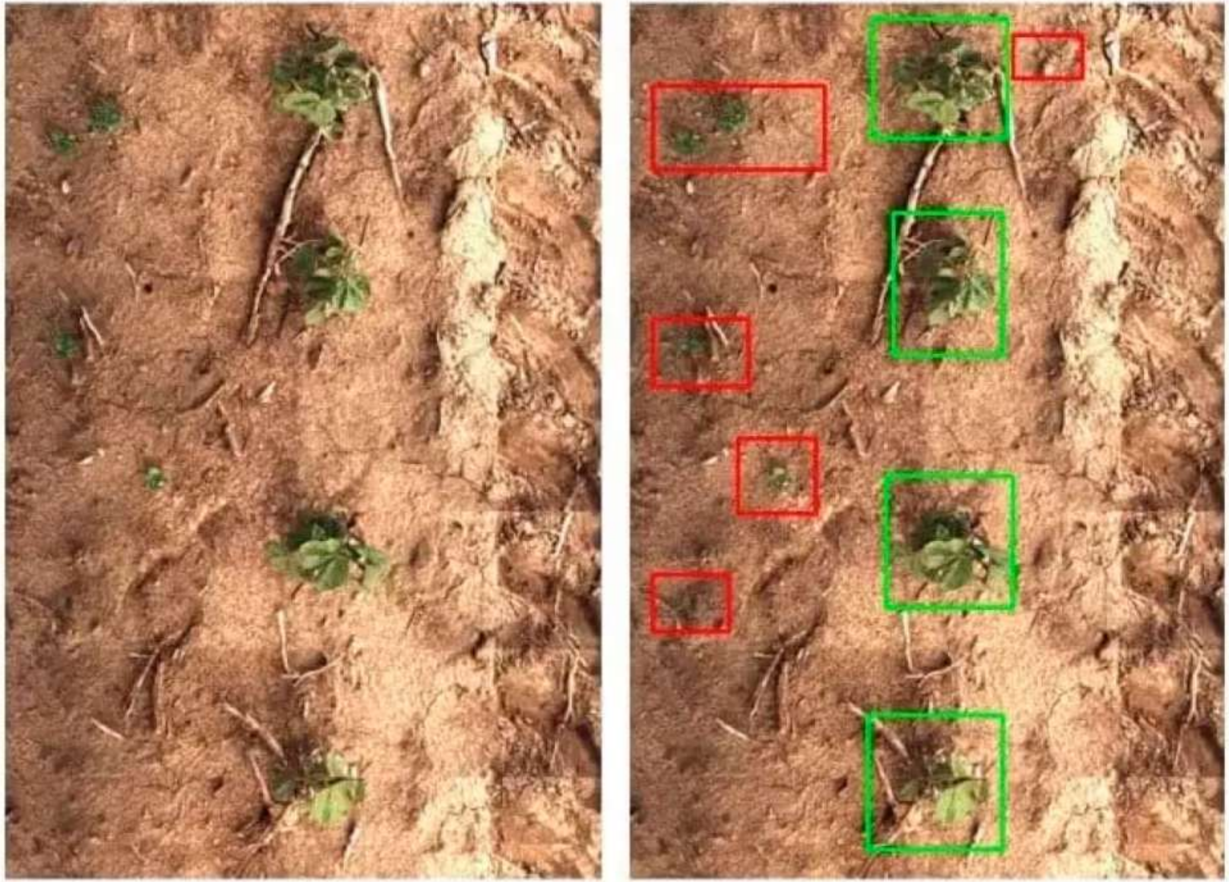
- Voice-enabled Interfaces
- Augmented Reality

Reduce key-strokes and lower the learning curve of complex machinery.

JOHN DEERE

Smart Machine Era:

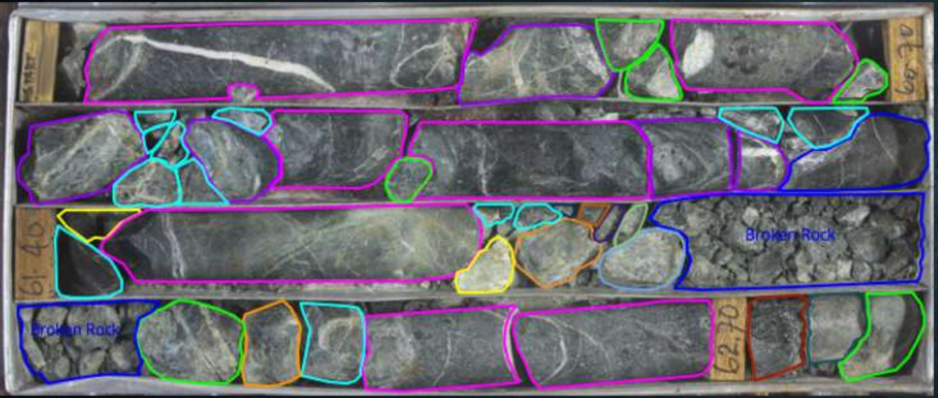
Reduces 90% use of herbicide through Blue River's smart sprayer. Computer Vision technology built on AWS and NVIDA GPU.

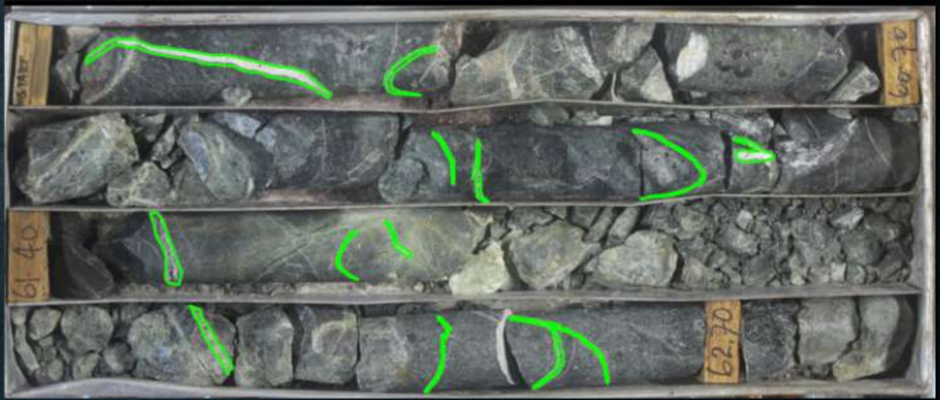


aws

DIUS

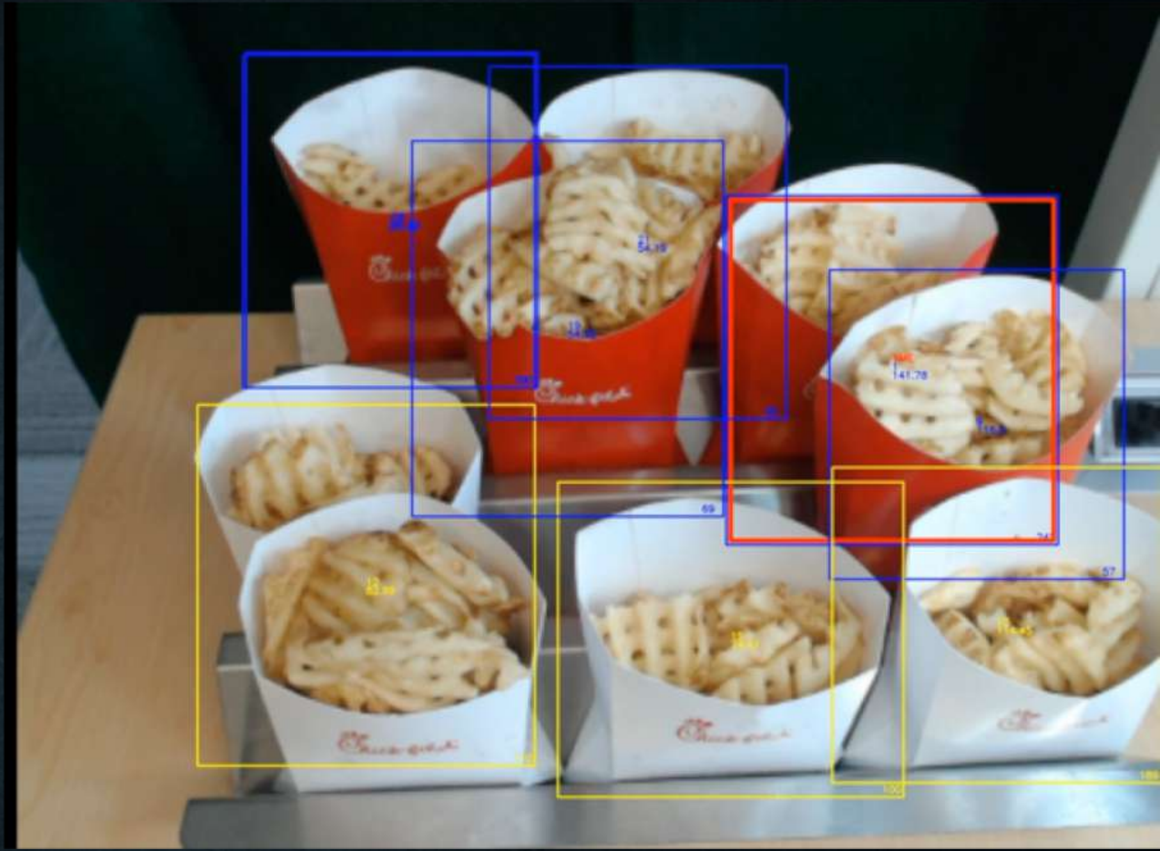
SOLVE GEOSOLUTIONS





aws

Chick-fil-A



Technology

Mission: to put machine learning in the hands of every developer.

The Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

Vision

- REKOGNITION IMAGE
- REKOGNITION VIDEO
- TEXTRACT

Speech

- POLLY
- TRANSCRIBE

Language

- TRANSLATE
- COMPREHEND

Chatbots

- LEX

Forecasting

- FORECAST

Recommendations

- PERSONALIZE

ML SERVICES

AMAZON SAGEMAKER

BUILD

Pre-built algorithms & notebooks

Data labeling (GROUND TRUTH)

Algorithms & models (AWS MARKETPLACE FOR MACHINE LEARNING)

TRAIN

One-click model training & tuning

Optimization (NEO)

Reinforcement learning

DEPLOY

One-click deployment & hosting

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

- TensorFlow
- mxnet
- PYTORCH

Interfaces

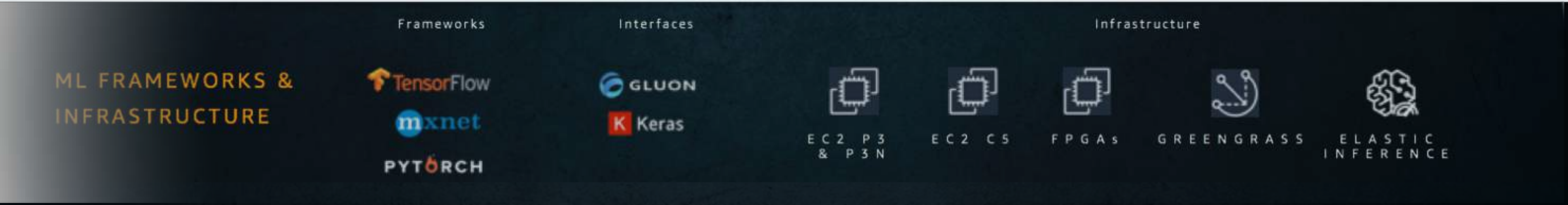
- GLUON
- Keras

Infrastructure

- EC2 P3 & P3N
- EC2 C5
- FPGAs
- GREENGRASS
- ELASTIC INFERENCE

aws

The Amazon ML stack: Frameworks and Infrastructure



ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

- TensorFlow
- mxnet
- PYTORCH

Interfaces

- GLUON
- Keras

Infrastructure

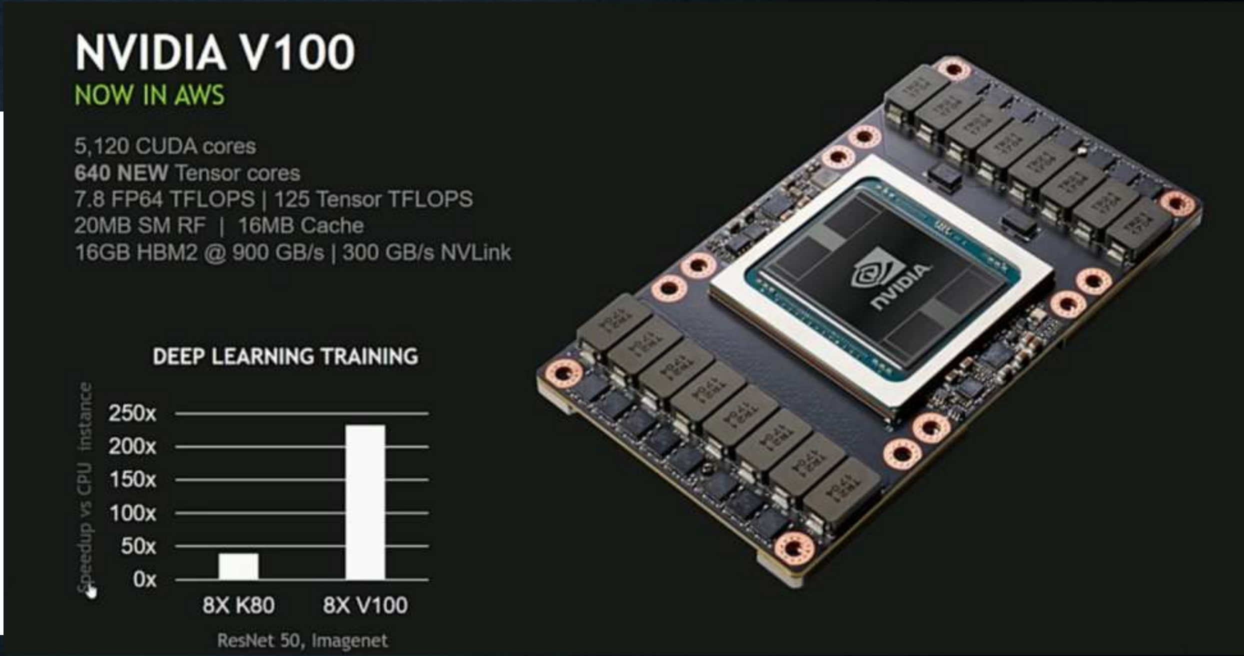
- EC2 P3 & P3N
- EC2 C5
- FPGAs
- GREENGRASS
- ELASTIC INFERENCE

ML Optimized Environments and Infrastructure

More data science, less setup

Deep Learning AMI

TensorFlow
PyTorch
mxnet
Chainer
theano
K
GLUON
HOROVOD
pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
scikit learn
CONDA
NVIDIA. CUDA.
cuDNN



NVIDIA V100

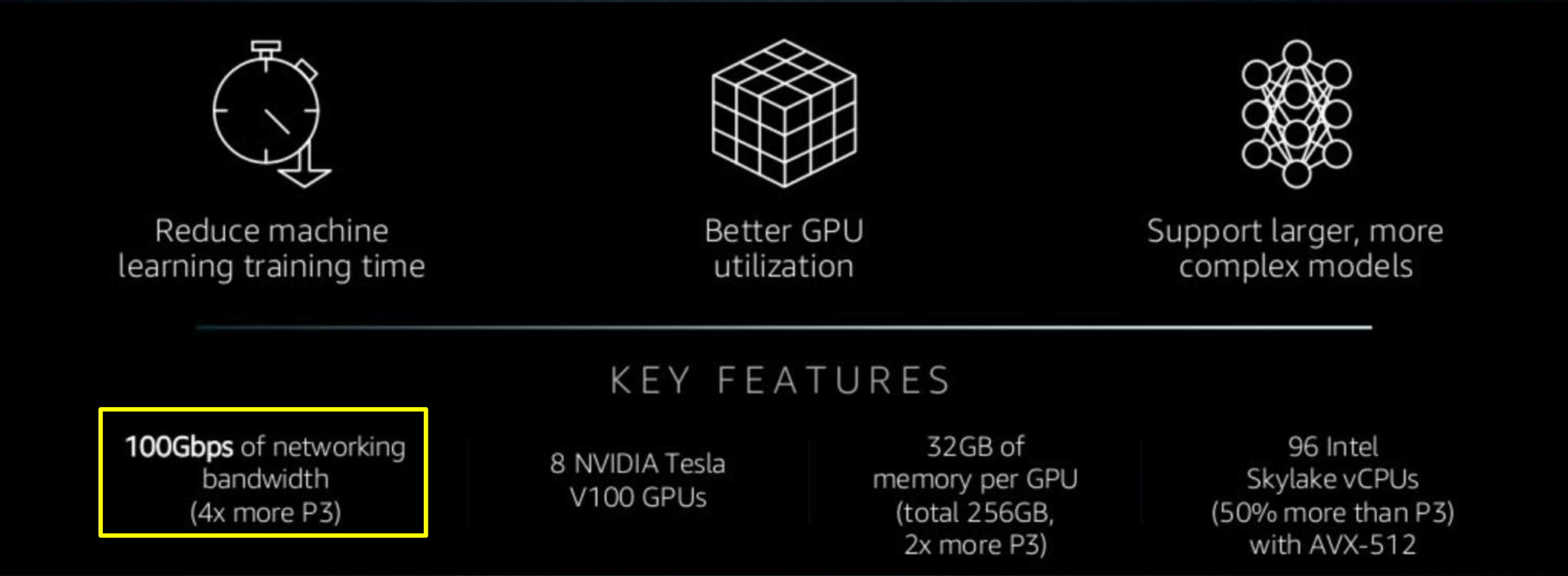
NOW IN AWS

5,120 CUDA cores
640 NEW Tensor cores
7.8 FP64 TFLOPS | 125 Tensor TFLOPS
20MB SM RF | 16MB Cache
16GB HBM2 @ 900 GB/s | 300 GB/s NVLink

DEEP LEARNING TRAINING

ResNet 50, Imagenet

P3dn Instances: Optimized for Training at Scale



Reduce machine learning training time

Better GPU utilization

Support larger, more complex models

KEY FEATURES

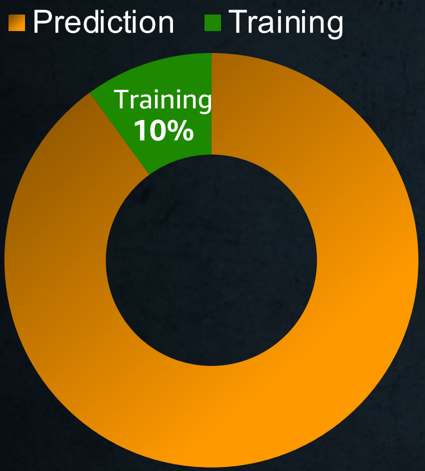
100Gbps of networking bandwidth (4x more P3)

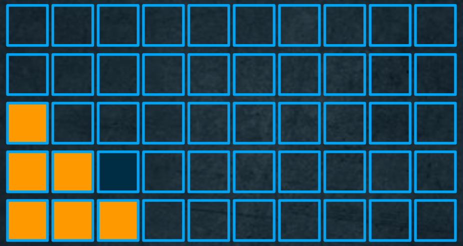
8 NVIDIA Tesla V100 GPUs

32GB of memory per GPU (total 256GB, 2x more P3)

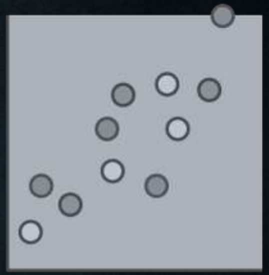
96 Intel Skylake vCPUs (50% more than P3) with AVX-512

Challenges with Inference in Production





Low utilization and high costs



One size does not fit all

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

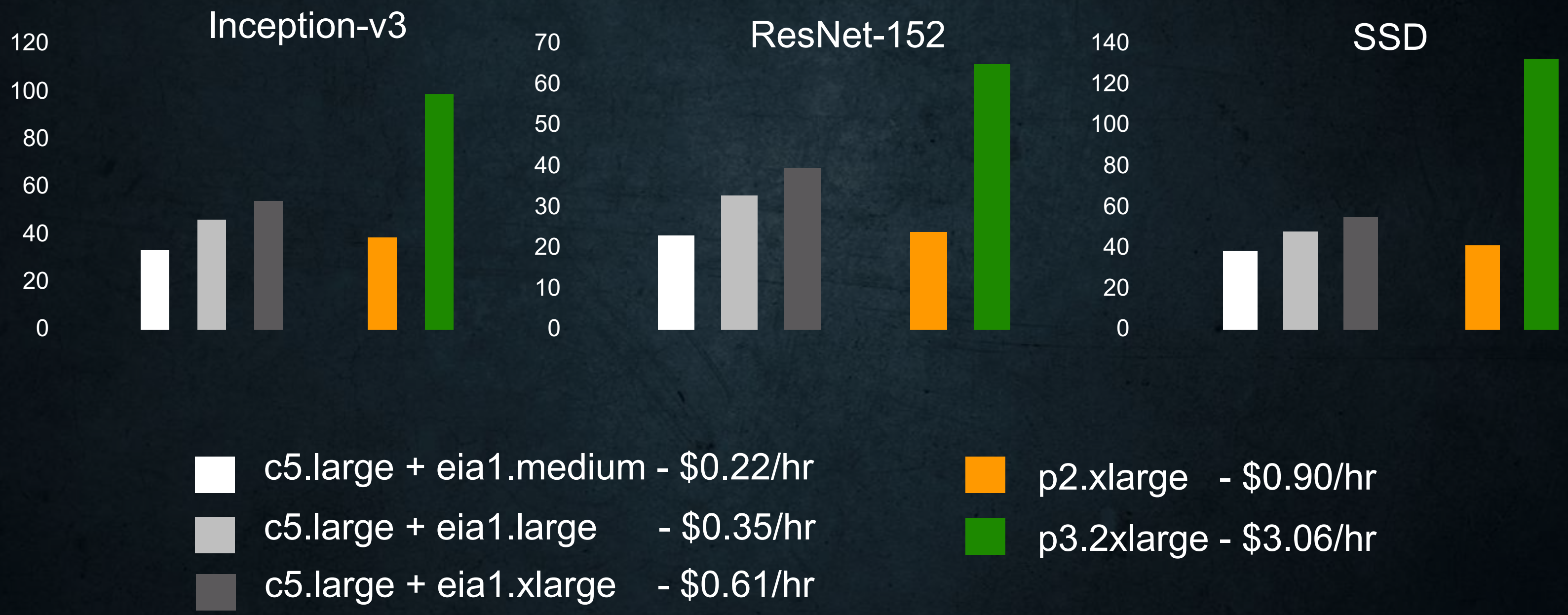
Integrated with
Amazon EC2 and
Amazon SageMaker

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

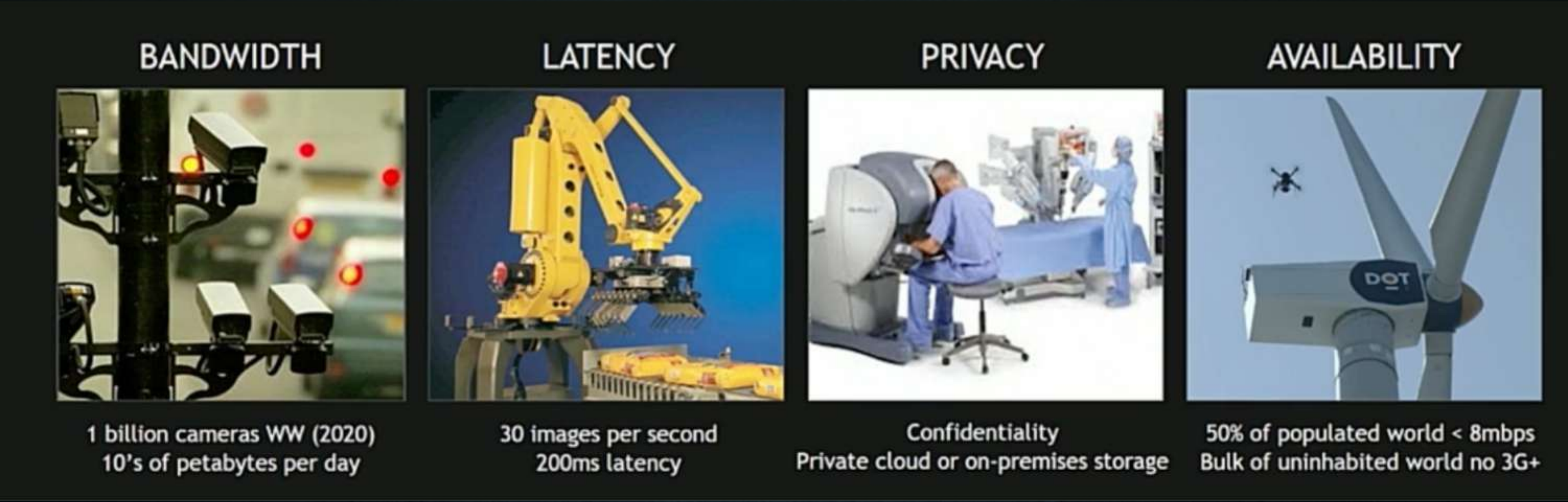
Single and
mixed-precision
operations

Up to 75% Reduction in Inference Costs

Inferences per second



Requirements for Inference at the 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

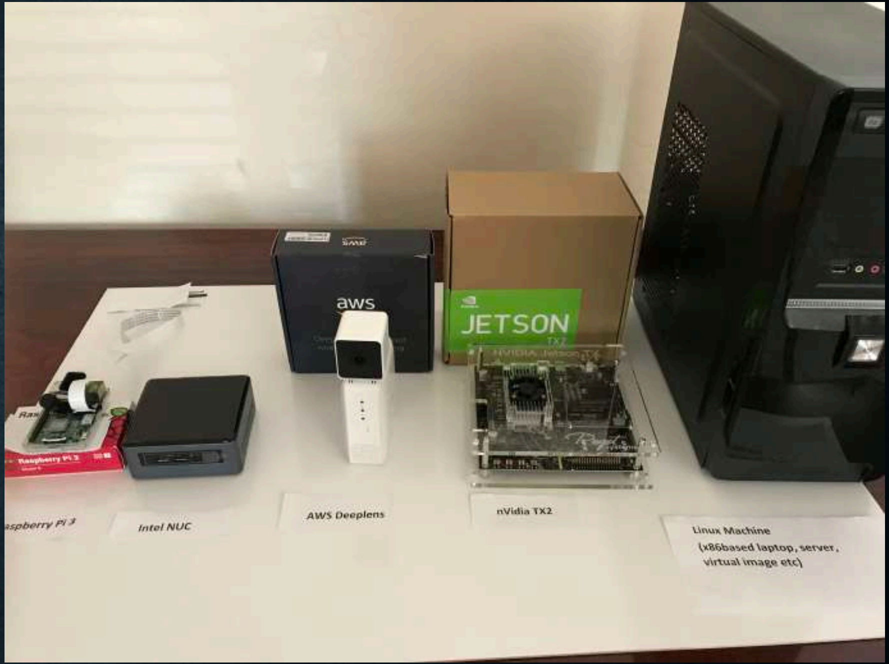
AVAILABILITY

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

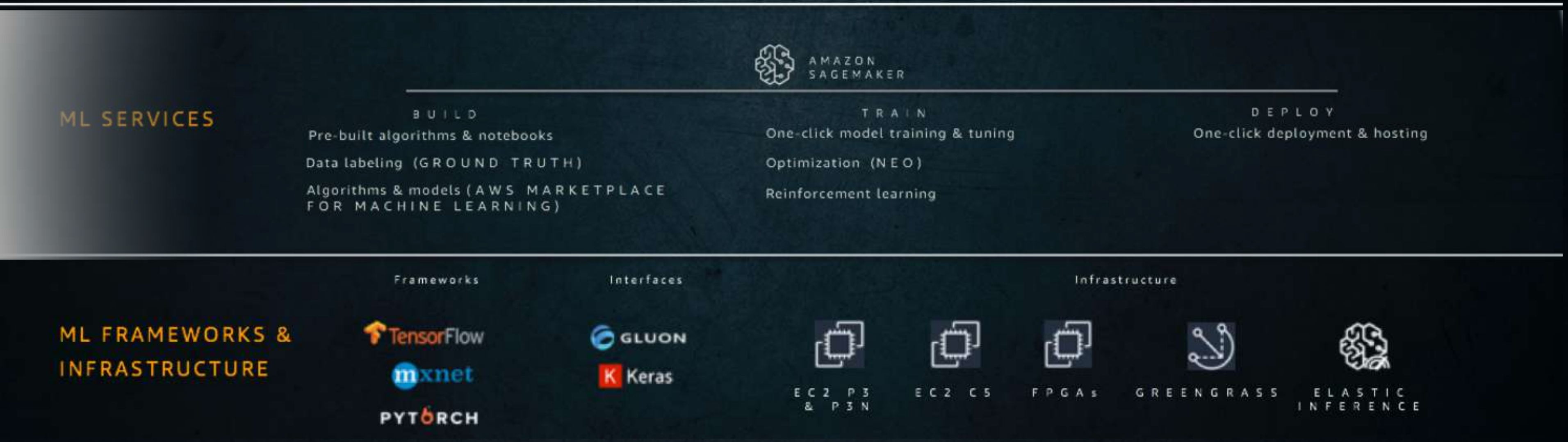
AWS Greengrass

Extend intelligence to the edge

- Local actions
- Local triggers
- Data and state sync
- Security
- Local resource access
- Local ML inference



The Amazon ML stack: ML Services



AMAZON SAGEMAKER

ML SERVICES

BUILD

Pre-built algorithms & notebooks

Data labeling (GROUND TRUTH)

Algorithms & models (AWS MARKETPLACE FOR MACHINE LEARNING)

TRAIN

One-click model training & tuning

Optimization (NEO)

Reinforcement learning

DEPLOY

One-click deployment & hosting

ML FRAMEWORKS & INFRASTRUCTURE

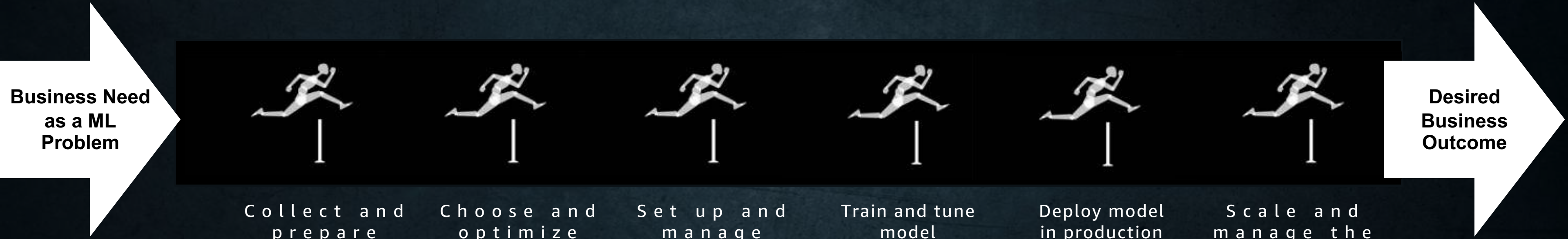
Frameworks

Interfaces

Infrastructure

Amazon SageMaker

BRINGING MACHINE LEARNING TO ALL DEVELOPERS



Business Need as a ML Problem

Collect and prepare training data

Choose and optimize your ML algorithm

Set up and manage environments for training

Train and tune model (trial and error)

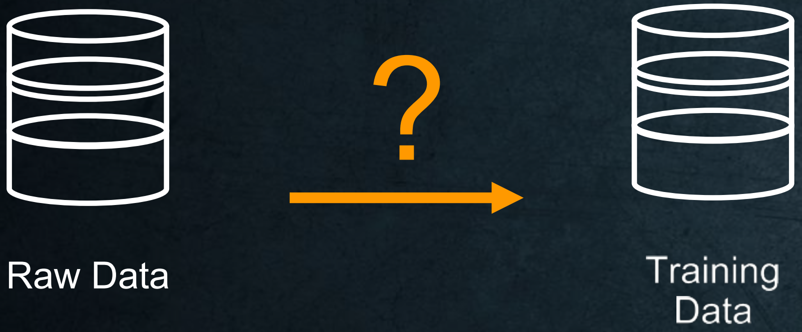
Deploy model in production

Scale and manage the production environment

Desired Business Outcome

SIMPLIFY THE END-TO-END MACHINE LEARNING PROCESS

Training Data for Supervised Learning

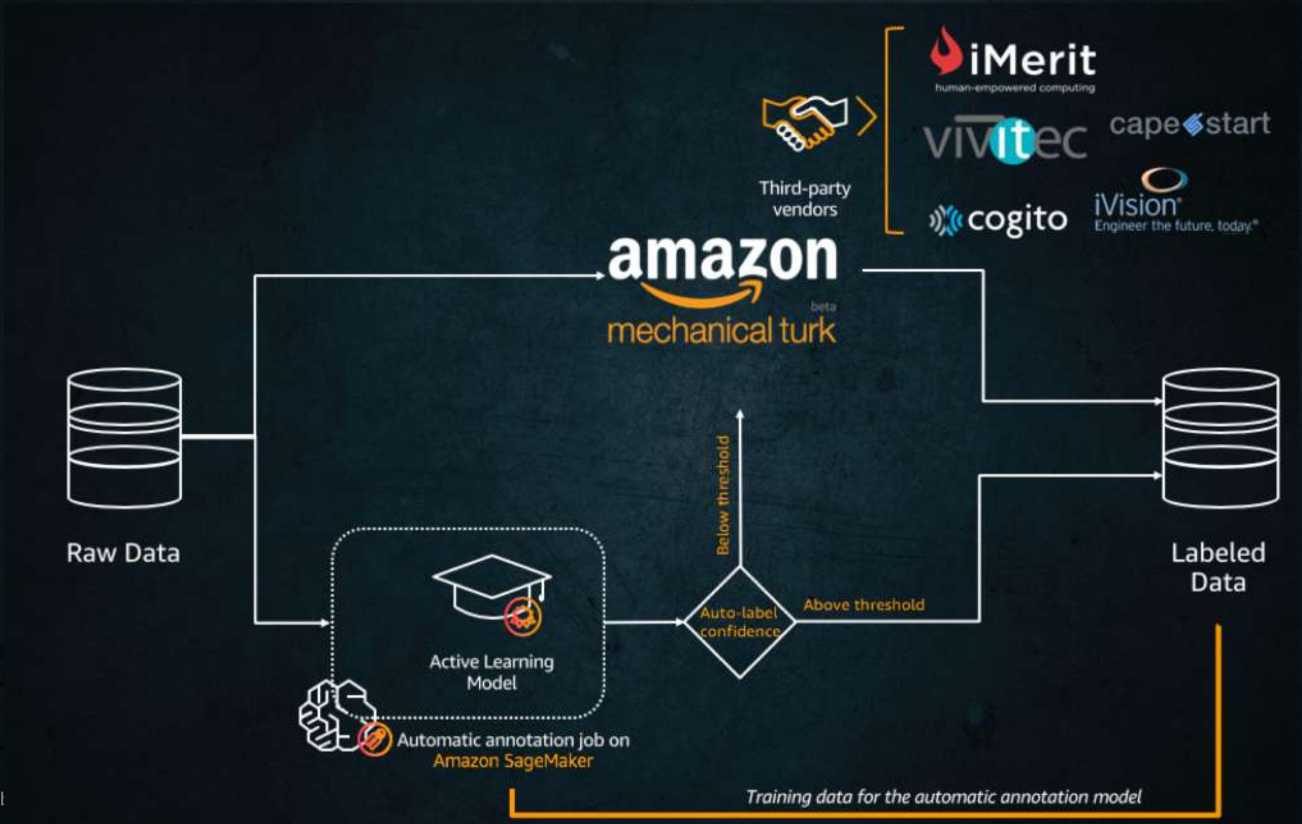


Challenges

- Timely annotation of large data sets
- Managing workforces.
  - Ensuring labeling quality
  - Handling workflows
- Integration with ML development environment

Amazon SageMaker Ground Truth

Label machine learning training data easily and accurately



Third-party vendors

iMerit human-empowered computing

vivitec

capestart

cogito

iVision Engineer the future, today.

amazon mechanical turk beta

Raw Data

Labeled Data

Below threshold

Auto-label confidence

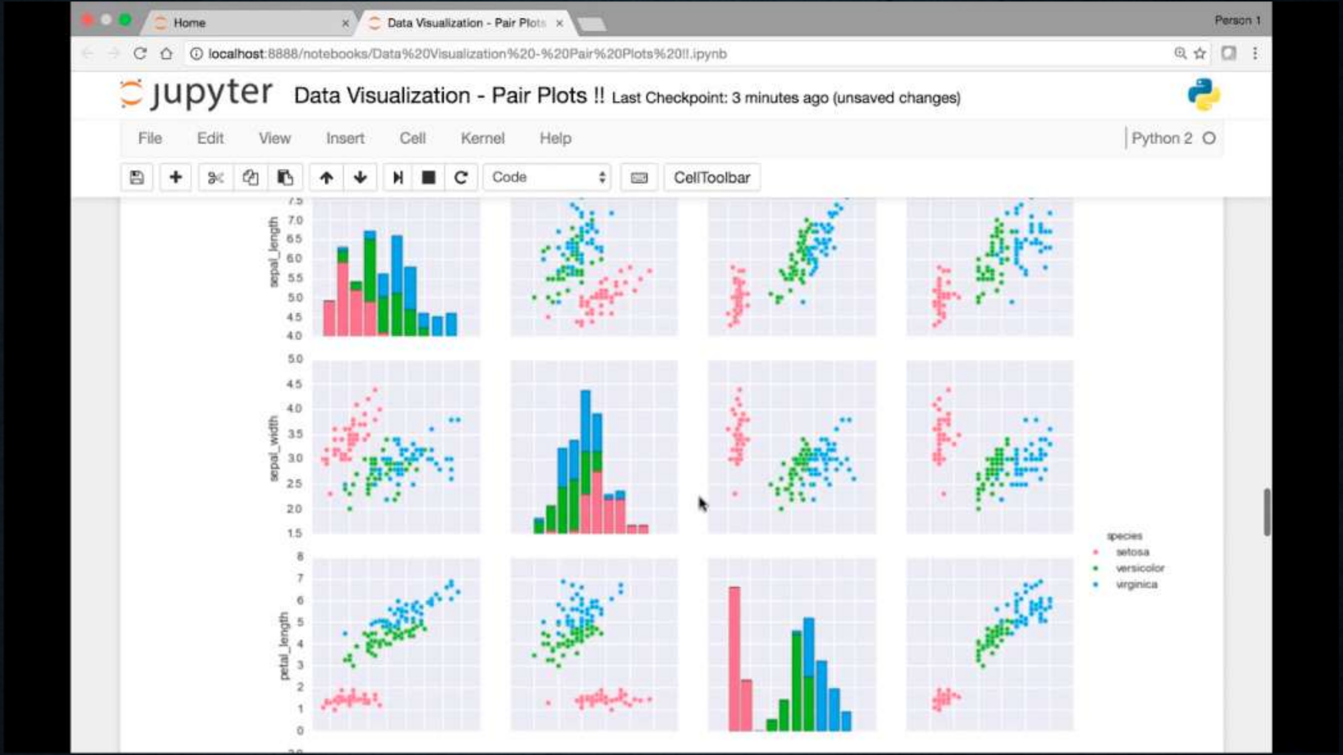
Above threshold

Active Learning Model

Automatic annotation job on Amazon SageMaker

Training data for the automatic annotation model

Managed Notebooks: Exploration, Experimentation



Your Choice. The Right Tools.

NATIVE SUPPORT FOR MOST POPULAR FRAMEWORKS

TensorFlow
mxnet
PYTORCH
GLUON
scikits learn
K Keras
Spark ML
Chainer

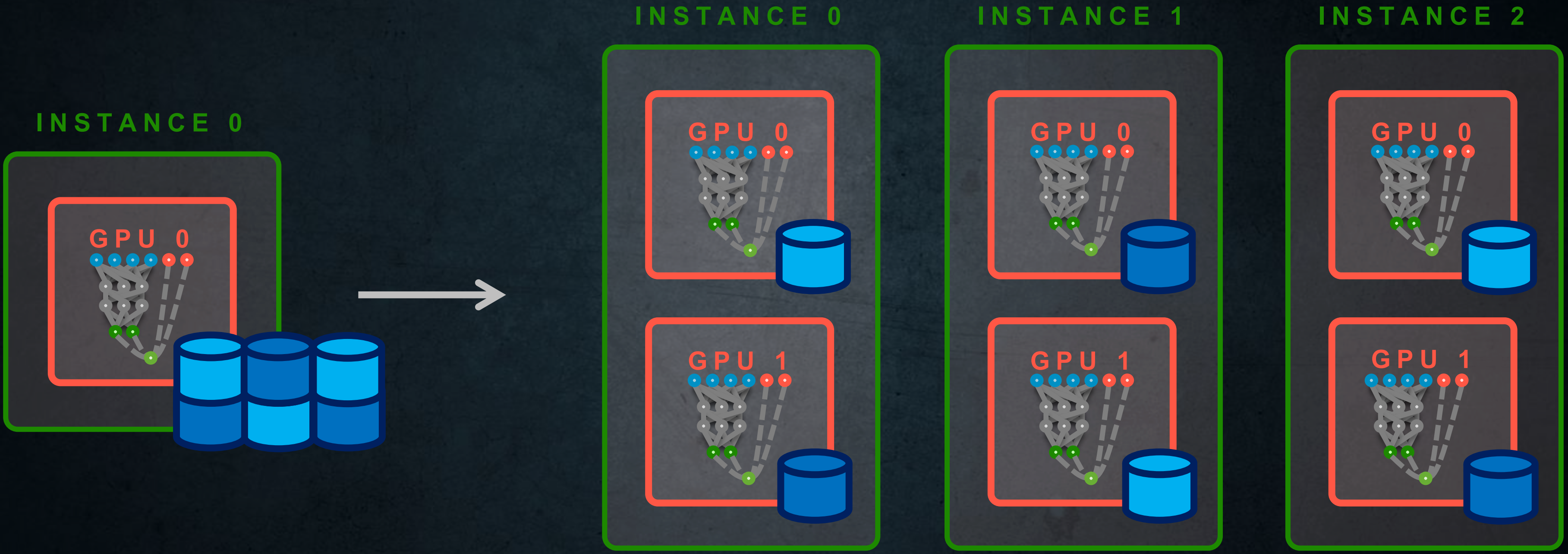
SAGEMAKER OPTIMIZED ALGORITHMS

- 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

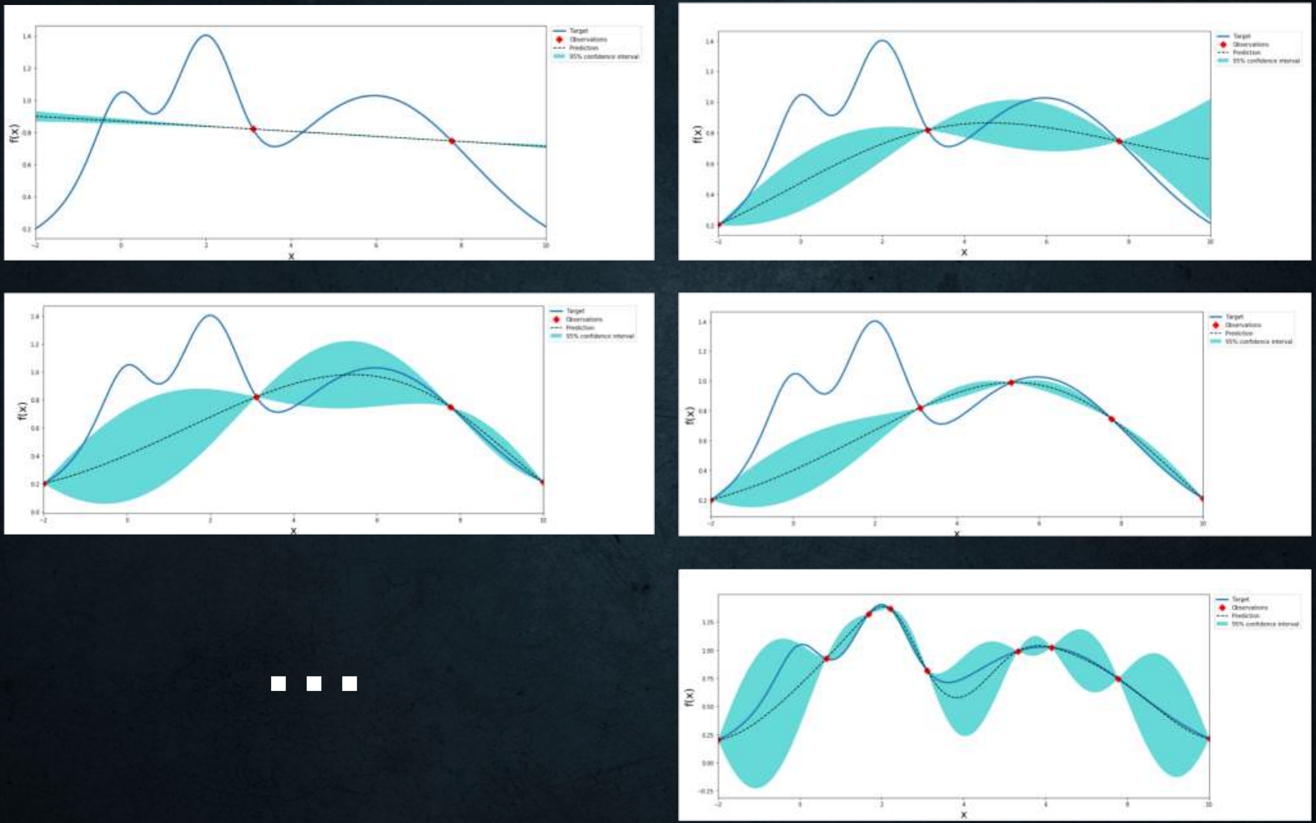
aws marketplace INTEGRATION

DEEPVISION
H2O.ai
figure eight
TIBCO
D.M
DIMENSIONALMECHANICS
Open for Innovation KNIME
wipro
Knowledgent

Zero Setup Training

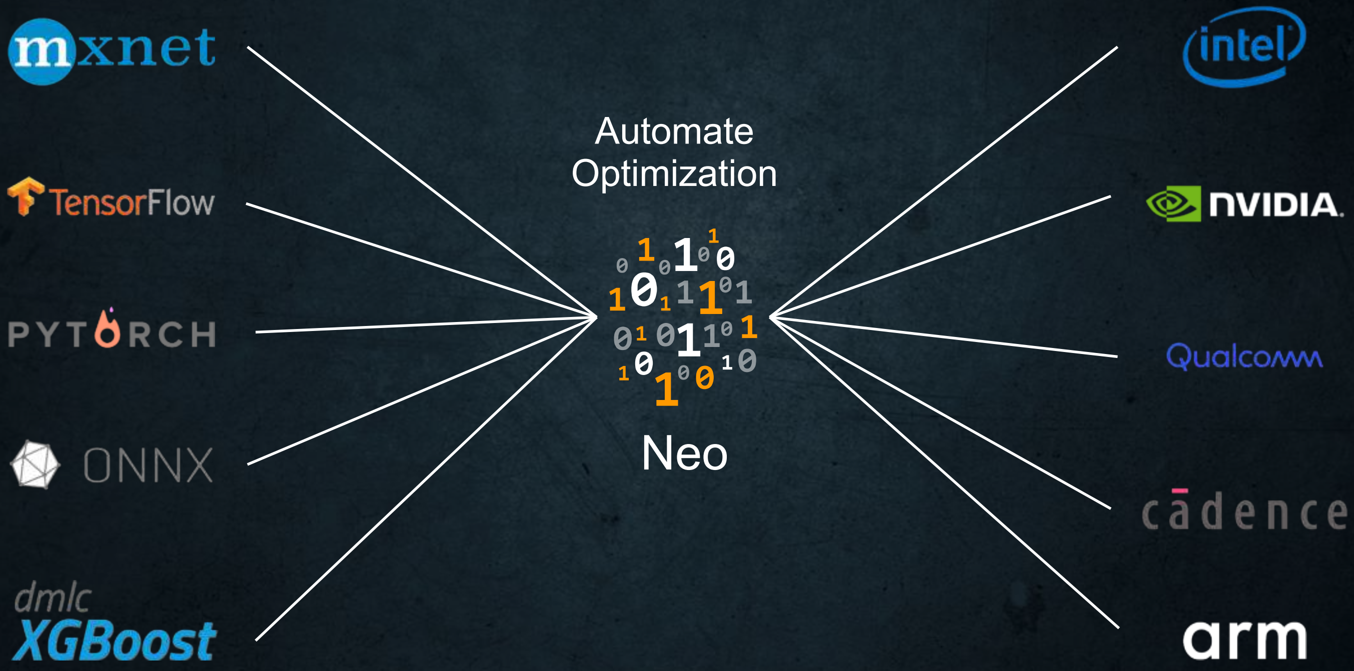


Automatic Model Tuning



Amazon SageMaker Neo

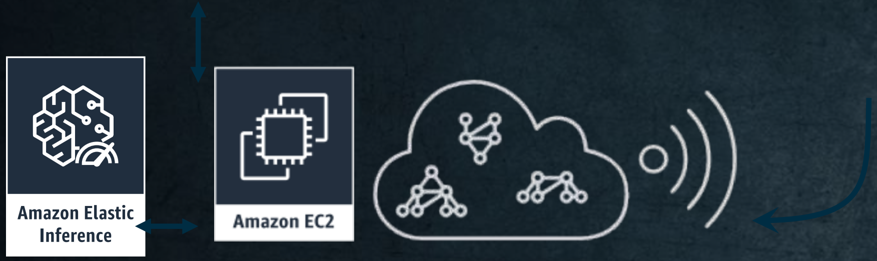
Train once, run everywhere with 2x performance and no accuracy lost



Automate Optimization

Neo

Simplified Model Deployment and Hosting



SageMaker Hosted Endpoints

- Auto-scaling
- Performance monitoring
- A/B Testing
- Elastic Inference support
- Suited for real-time and batch workloads

Custom machine learning for your business

AMAZON SAGEMAKER

REDUCE COSTS

70%

cost reduction for data labeling using Ground Truth

75%

cost reduction for inference with Elastic Inference

INCREASE PERFORMANCE

10x

better algorithm performance

2x

performance increases from model optimization with Neo

EASE-OF-USE

One-click

model training & deployment

Train once

run anywhere

The Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

Vision

REKOGNITION IMAGE

REKOGNITION VIDEO

TEXTRACT

Speech

POLLY

TRANSCRIBE

Language

TRANSLATE

COMPREHEND

Chatbots

LEX

Forecasting

FORECAST

Recommendations

PERSONALIZE

ML SERVICES

AMAZON SAGEMAKER

BUILD

Pre-built algorithms & notebooks

Data labeling (GROUND TRUTH)

Algorithms & models (AWS MARKETPLACE FOR MACHINE LEARNING)

TRAIN

One-click model training & tuning

Optimization (NEO)

Reinforcement learning

DEPLOY

One-click deployment & hosting

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

TensorFlow

mxnet

PYTORCH

Interfaces

GLUON

Keras

Infrastructure

EC2 P3 & P3N

EC2 C5

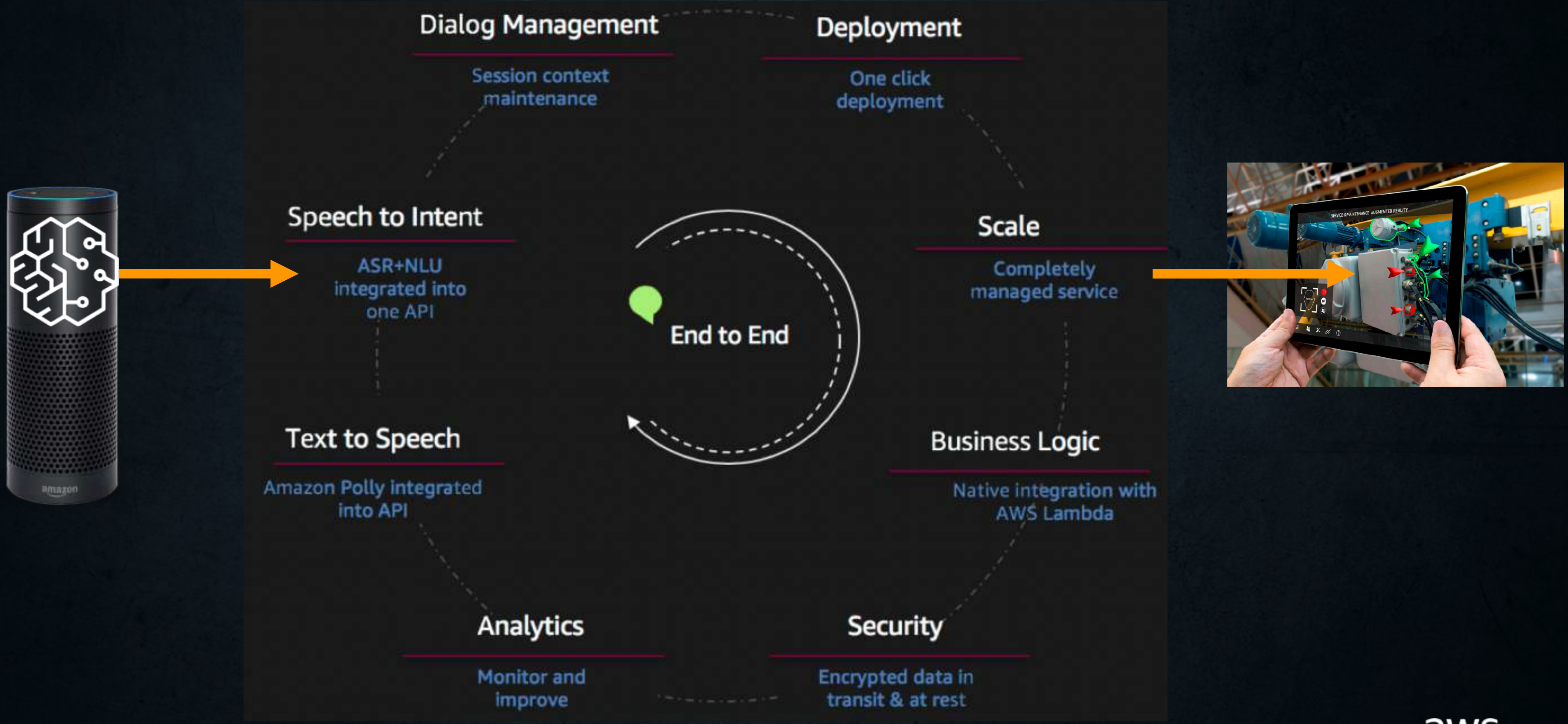
FPGAs

GREENGRASS

ELASTIC INFERENCE

aws

Chatbots with Amazon Lex



End to End

Dialog Management
Session context maintenance

Deployment
One click deployment

Scale
Completely managed service

Business Logic
Native integration with AWS Lambda

Security
Encrypted data in transit & at rest

Analytics
Monitor and improve

Text to Speech
Amazon Polly integrated into API

Speech to Intent
ASR+NLU integrated into one API

Vision: Amazon Rekognition

IMAGES

Object and Scene Detection

Facial Analysis

Face Recognition

Unsafe Image Detection

Celebrity Recognition

Text in Image

VIDEO

Person Tracking

Real-time Live Stream

Vision: Amazon Rekognition

IMAGES

Object and Scene Detection

Facial Analysis

Face Recognition

Unsafe Image Detection

Celebrity Recognition

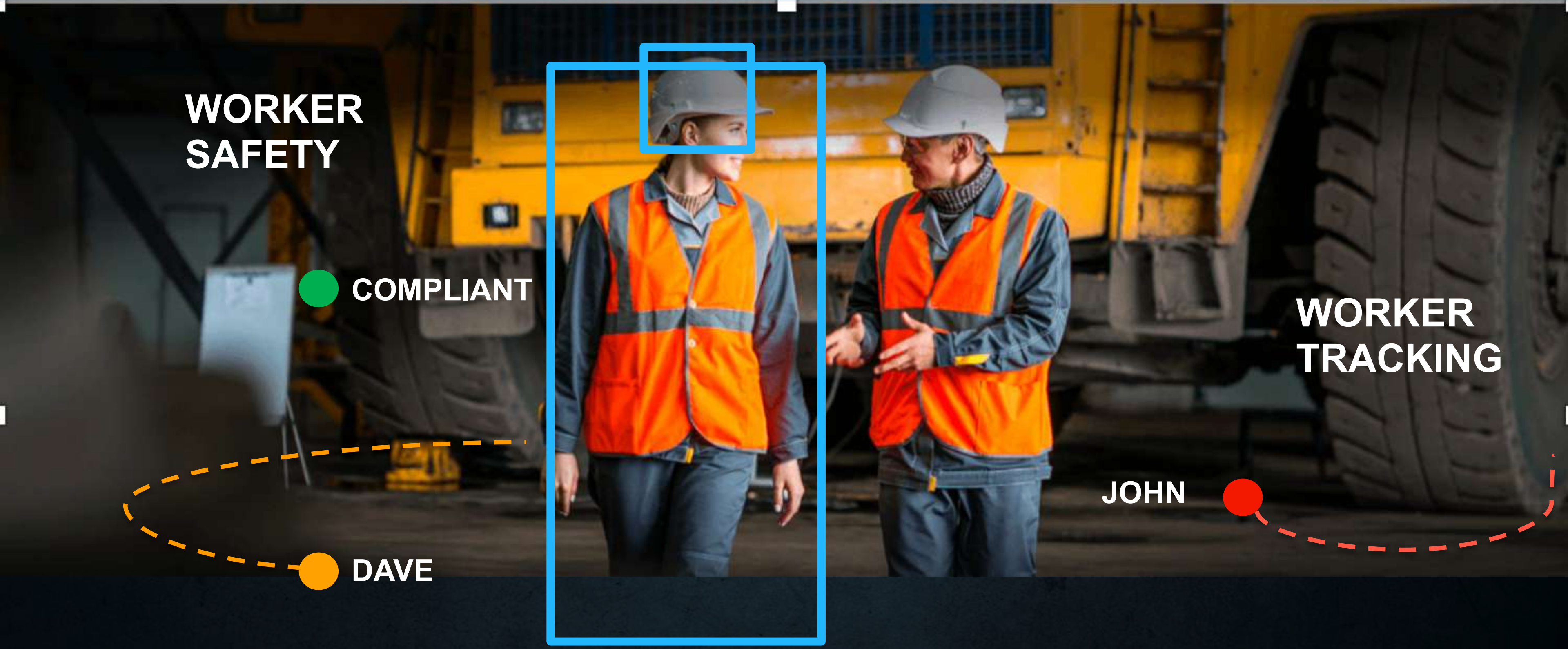
Text in Image

VIDEO

Person Tracking

Real-time Live Stream

The Connected Worker



WORKER SAFETY

● COMPLIANT

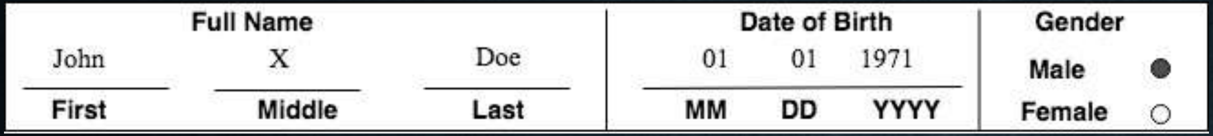
● DAVE

WORKER TRACKING

JOHN ●

Vision: Amazon Textract

Form Extraction simplified



Full Name			Date of Birth			Gender	
John	X	Doe	01	01	1971	Male	●
First	Middle	Last	MM	DD	YYYY	Female	○

- ✓ Logical groupings captured
- ✓ Relationships captured
- ✓ Glyphs captured

Output

Full Name:
First: John
Middle: X
Last: Doe

Date of Birth:
MM: 01
DD: 01
YYYY: 1971

Gender:
Male: True
Female: False

NLP: Amazon Comprehend

Amazon.com, Inc. is located in Seattle, WA and was founded July 5th, 1994 by Jeff Bezos. Our customers *love buying everything* from books to blenders at great prices

Named Entities

- Amazon.com: Organization
- Seattle, WA : Location
- July 5th,1994: Date
- Jeff Bezos : Person

Keyphrases

- Our customers
- books
- blenders
- great prices

Sentiment

- *Positive*

Language

- English

More Natural Language Processing...

TEXT-TO-SPEECH

Amazon Polly

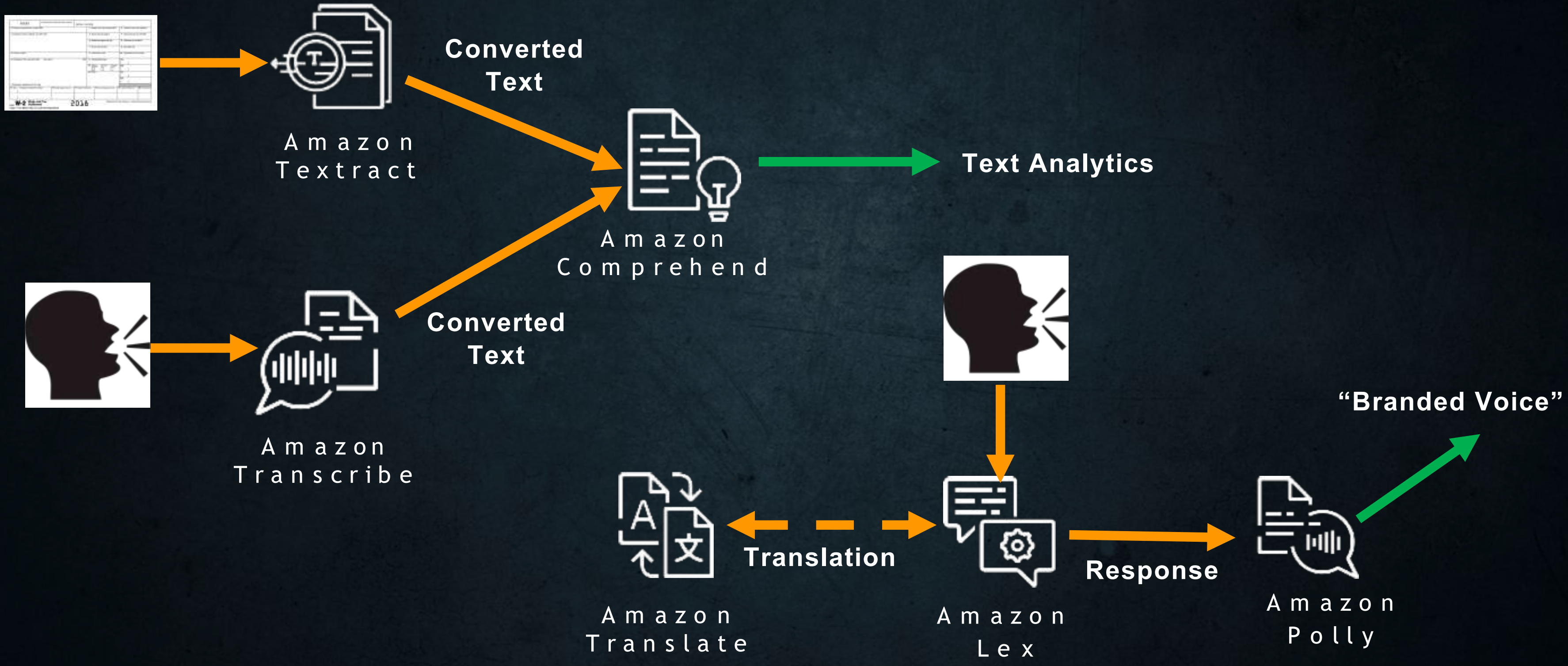
SPEECH-TO-TEXT

Amazon Transcribe

LANGUAGE TRANSLATION

Amazon Translate

Solution Acceleration: Intelligent Building Blocks



AutoML: Tailored Models Automated

FORECASTING

Amazon Forecast

RECOMMENDATIONS

Amazon Personalize

Computer Vision on AWS Jumpstart

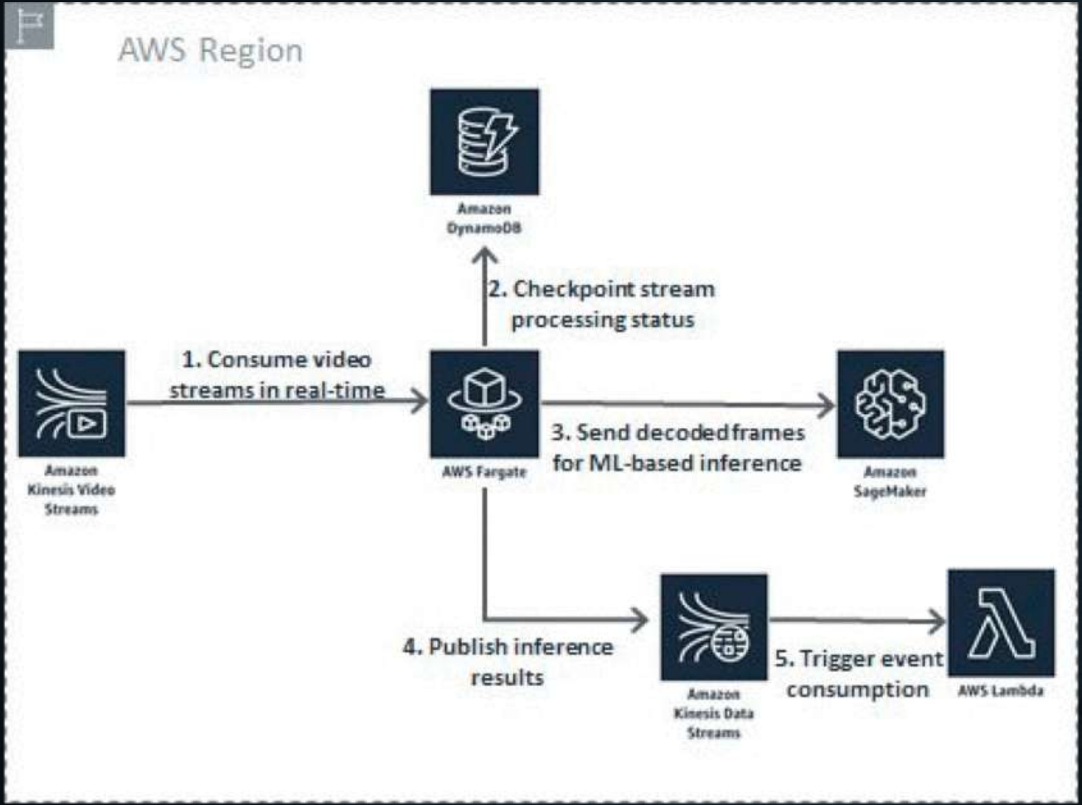
Jumpstart Workshop Menu

- Object Detection
- TBD…

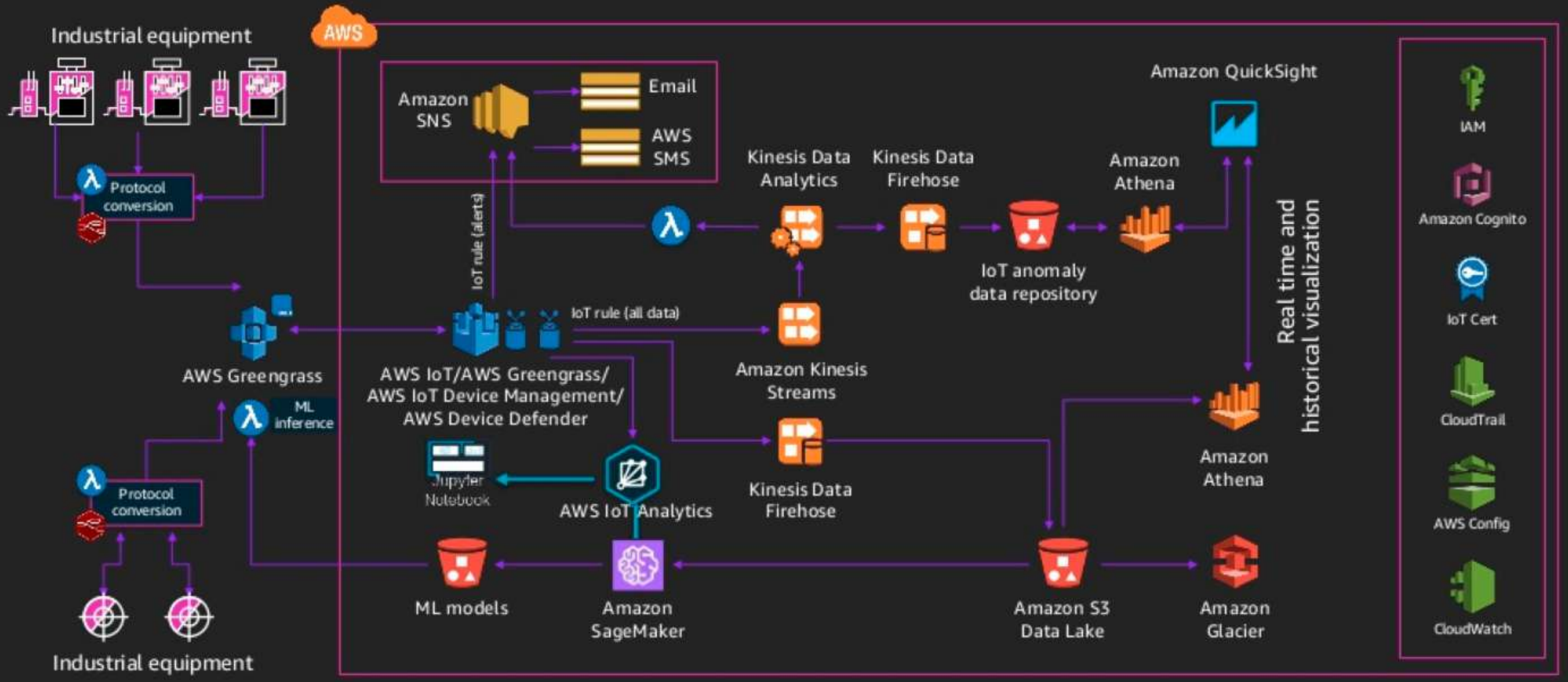
Object Detection Workshop

Reference Architectures

Cloud Inference on Streaming Video


AWS Region
Amazon DynamoDB
2. Checkpoint stream processing status
Amazon Kinesis Video Streams
1. Consume video streams in real-time
AWS Fargate
3. Send decoded frames for ML-based inference
Amazon SageMaker
4. Publish inference results
Amazon Kinesis Data Streams
5. Trigger event consumption
AWS Lambda

AWS industrial IoT reference architecture

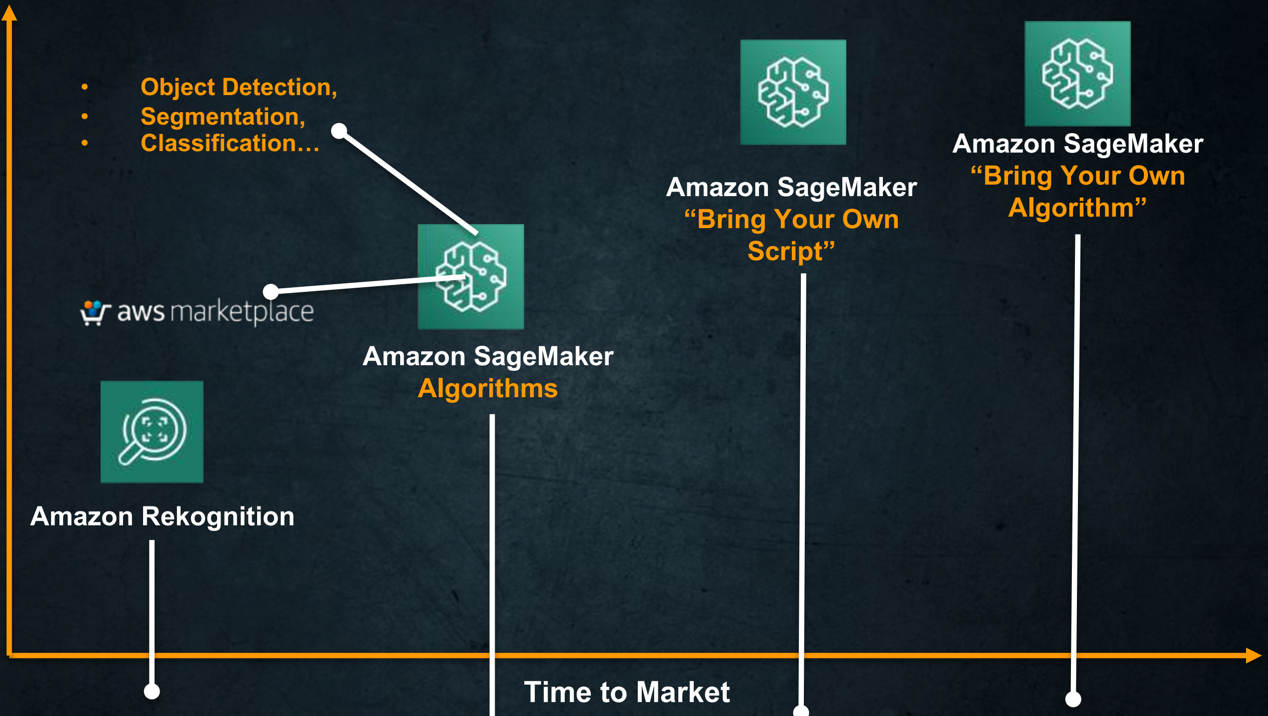


AWS re:Invent

aws

Workshop Objectives and Agenda

Select the Right Strategy



Scope of Use Cases

- Object Detection,
- Segmentation,
- Classification…

aws marketplace

Amazon Rekognition

Amazon SageMaker
Algorithms

Amazon SageMaker
“Bring Your Own
Script”

Amazon SageMaker
“Bring Your Own
Algorithm”

Time to Market

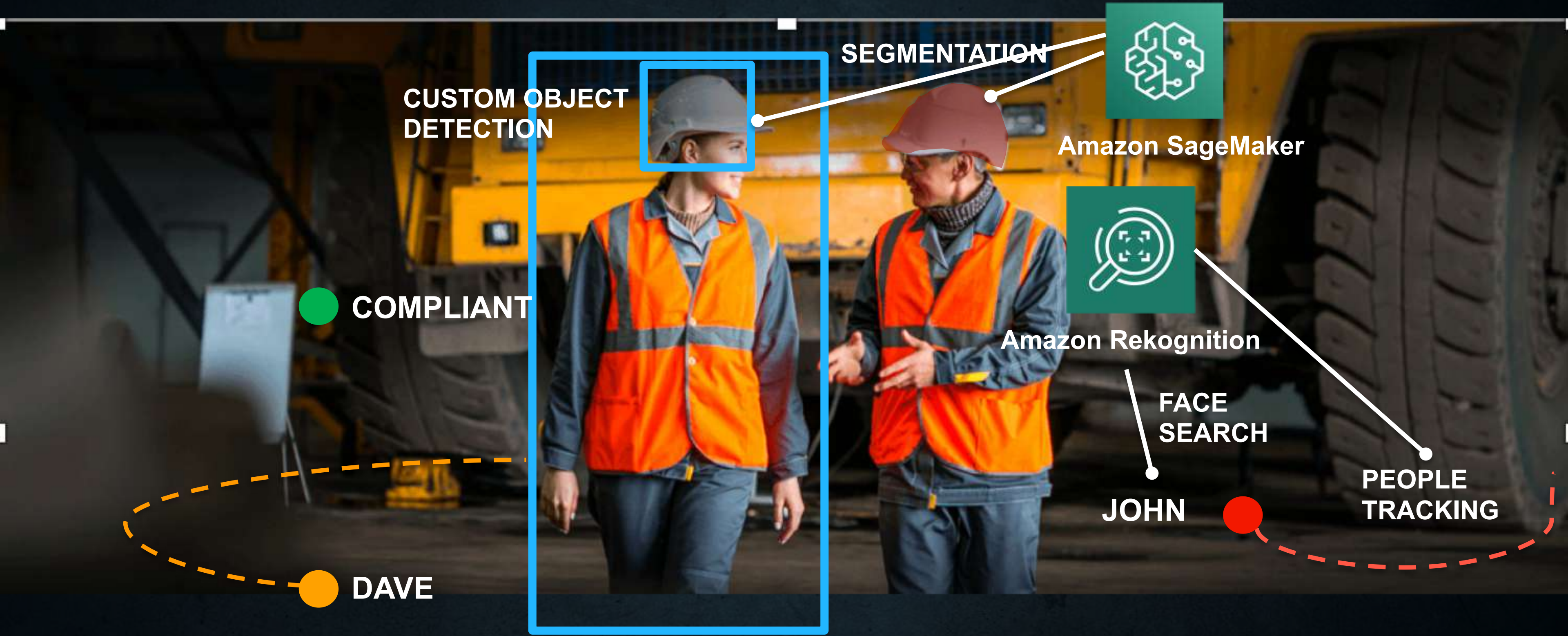
Fully managed APIs. **No ML development required**

Full control over data sets, and hyperparameter tuning. **No coding.**

Full control over running TensorFlow, PyTorch, MXNet, Sklearn… scripts.
No need for managing SageMaker compliant containers.

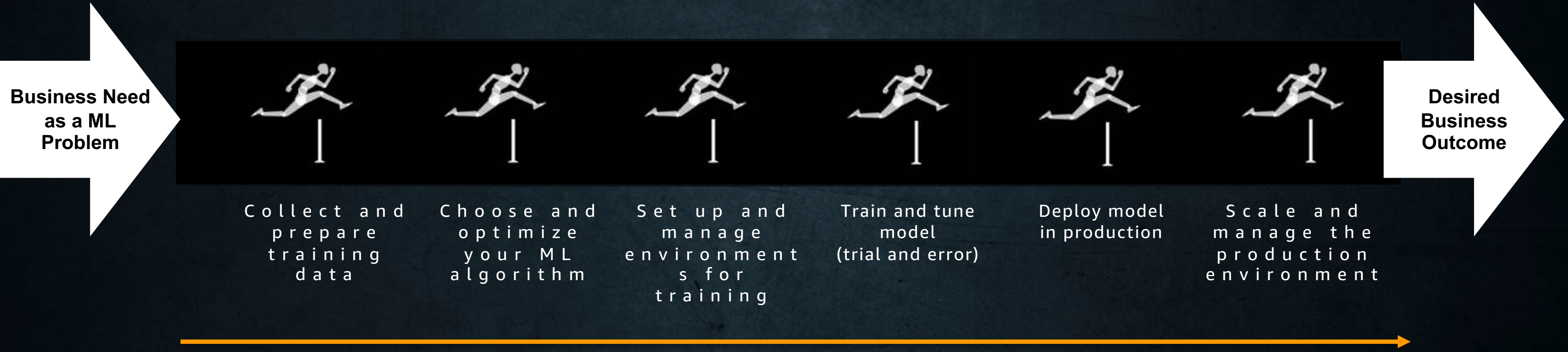
Integrate any algorithm **and benefit from automation.**

Right Tool for the Use Case



Amazon SageMaker

BRINGING MACHINE LEARNING TO ALL DEVELOPERS



Business Need as a ML Problem

Collect and prepare training data

Choose and optimize your ML algorithm

Set up and manage environments for training

Train and tune model (trial and error)

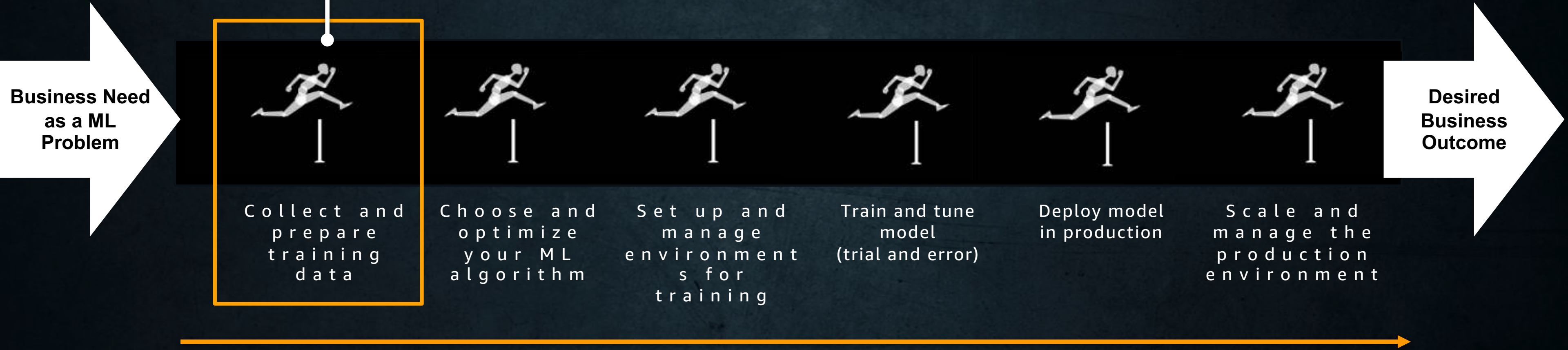
Deploy model in production

Scale and manage the production environment

Desired Business Outcome

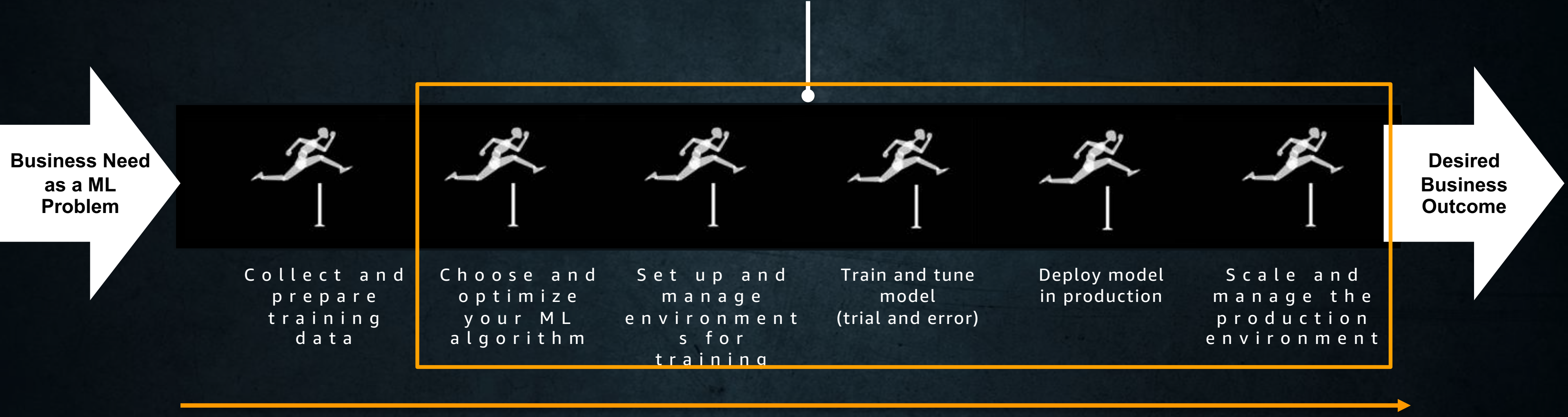
SIMPLIFY THE END-TO-END MACHINE LEARNING PROCESS

Lab 1: Managing a high **quality** training set at **scale** using **SageMaker GroundTruth**


Business Need as a ML Problem
Collect and prepare training data
Choose and optimize your ML algorithm
Set up and manage environments for training
Train and tune model (trial and error)
Deploy model in production
Scale and manage the production environment
Desired Business Outcome

SIMPLIFY THE END-TO-END MACHINE LEARNING PROCESS

Lab 2: Train, tune and deploy a custom object detector (**SSD**) with **zero coding**.



Business Need as a ML Problem

Collect and prepare training data

Choose and optimize your ML algorithm

Set up and manage environments for training

Train and tune model (trial and error)

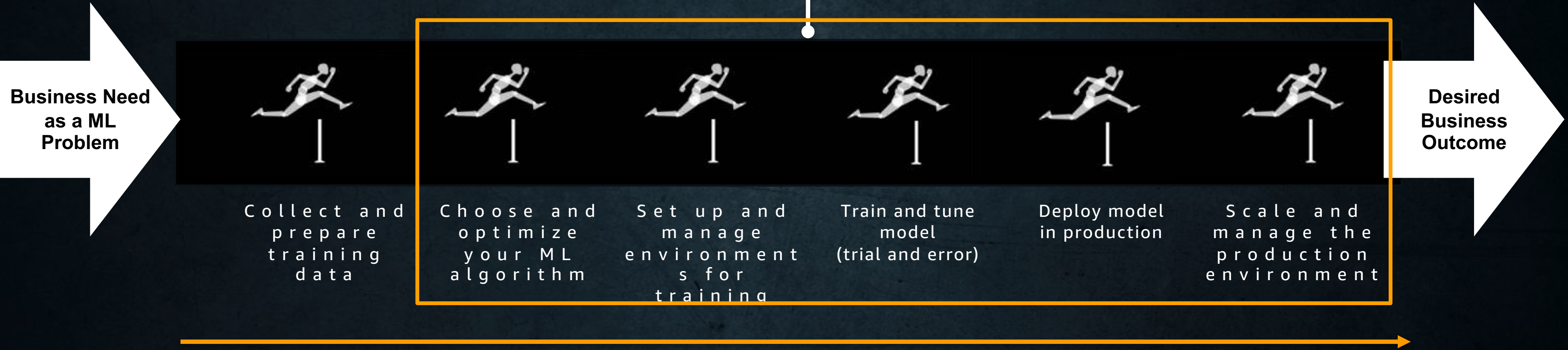
Deploy model in production

Scale and manage the production environment

Desired Business Outcome

SIMPLIFY THE END-TO-END MACHINE LEARNING PROCESS

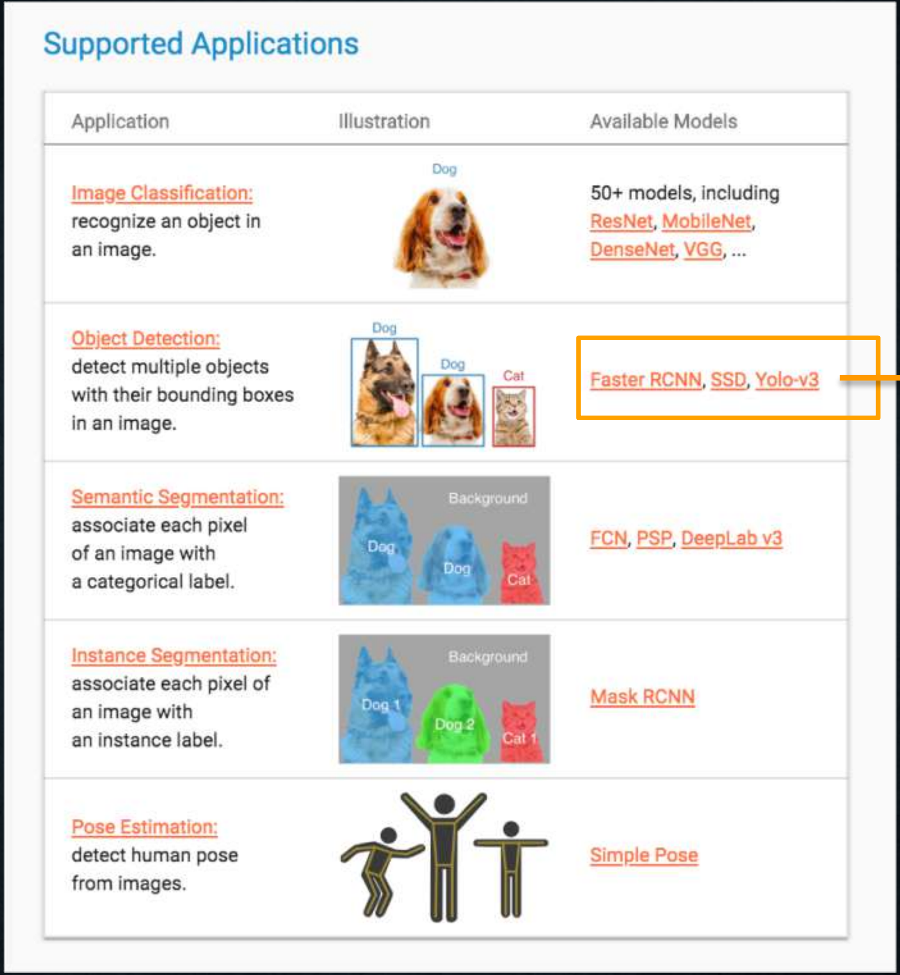
Lab 3: "Bring Your Own Script:" train, tune and deploy a custom object detector (**YOLOv3**) on GluonCV (MXNet).



Business Need as a ML Problem
Collect and prepare training data
Choose and optimize your ML algorithm
Set up and manage environments for training
Train and tune model (trial and error)
Deploy model in production
Scale and manage the production environment
Desired Business Outcome

SIMPLIFY THE END-TO-END MACHINE LEARNING PROCESS

GLUON



Supported Applications

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.	Background Dog Dog Cat	FCN, PSP, DeepLab v3
Instance Segmentation: associate each pixel of an image with an instance label.	Background Dog 1 Dog 2 Cat 1	Mask RCNN
Pose Estimation: detect human pose from images.		Simple Pose

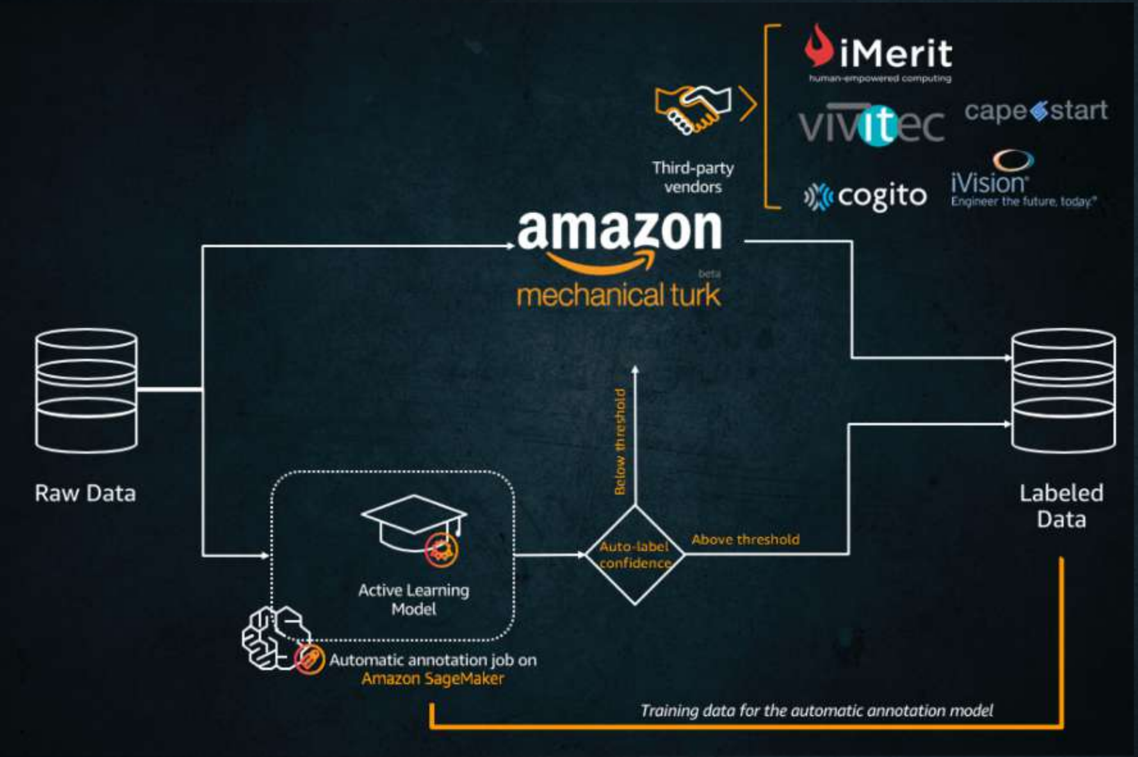
Algorithm variants : For instance, in **Object Detection**, different algorithms offer trade-offs between **accuracy (mAP)** and **latency (fps)**

Future labs

- By feedback and demand! dylatong@amazon.com
- Recognition:
  - Face Search
  - People Tracking
- Textract solutions
- More Amazon SageMaker use cases:
  - Segmentation, Pose Estimation, Similarity Search

Lab 1

Manage a high-quality data set at scale



1. Launch a Notebook Instance.
2. Manage a private workforce.
3. Create an annotation job for Object Detection.
4. Generate a dataset and metadata compatible with Amazon SageMaker algorithms without further data wrangling!

Lab 2

Create a custom object detector with zero coding

SAGEMAKER OPTIMIZED 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

1. Configure a hyperparameter tuning job for an Object Detection Algorithm.
2. Train on GPU
3. Deploy a managed endpoint.
4. Test and visualize!

Lab 3

Bring Your Own Script and automate the ML process

Examples

Training: Only modifications required is to set script certain parameters values from SageMaker container environment variables.

GluonCV YOLOv3 training script
PyTorch Siamese Network training script

Inference: requires overriding programmatic interface implementation.

GluonCV YOLOv3 model serving script
PyTorch Siamese Network model serving script

- **input_fn**: request format pre-processing
- **model_fn:** how to load the model
- **predict_fn:** inference logic
- **output_fn:** response format processing

1. Bring your own YoloV3 script on GluonCV
2. Prepare your data set and environment
3. Explore and prototype locally
4. Automate model tuning, and train
5. Deploy, test and visualize!

https://github.com/dylan-tong-aws/aws-cv-jumpstarter aws.ai

Appendix

How we can help…

ML Solutions Lab

Brainstorming
Custom modeling
Training
Work side-by-side with Amazon experts

Machine Learning Training & Certification

Practical education on ML for new & experienced practitioners
Based on the same material used to train Amazon developers