Opening keynote

Julien Simon

Global Evangelist, AI & Machine Learning, AWS
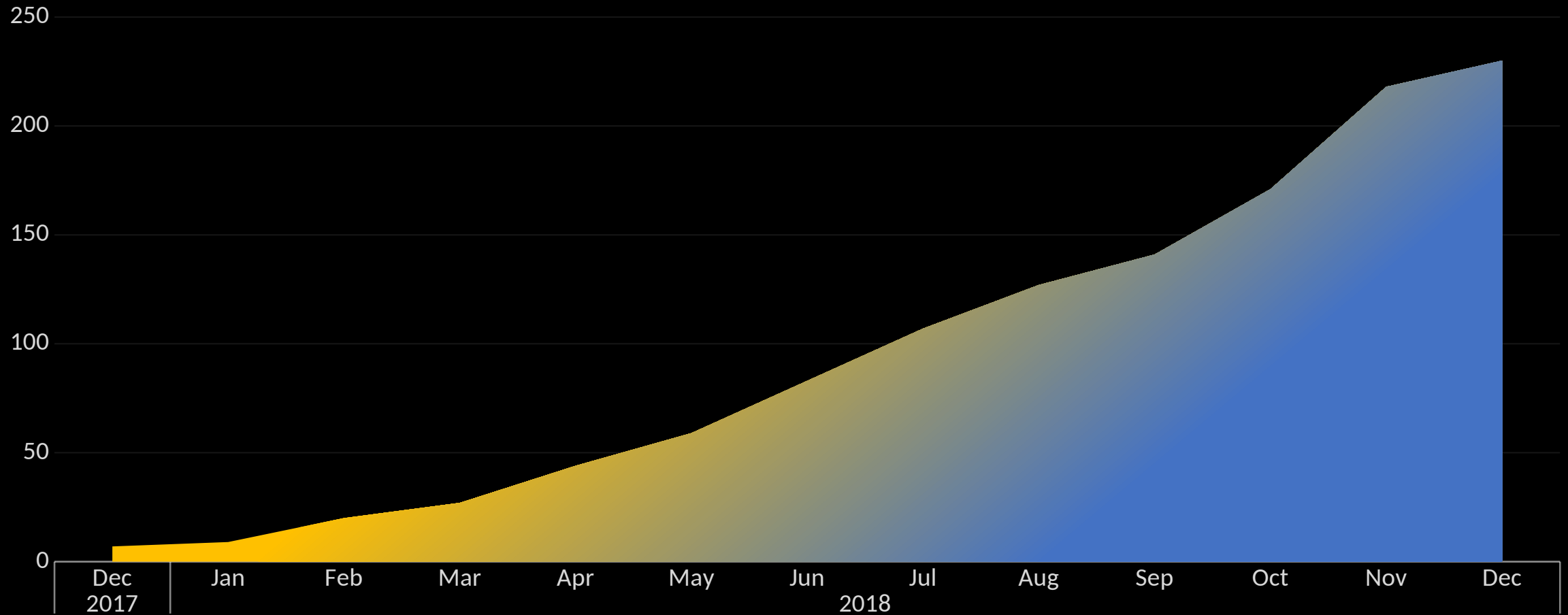
@julsimon

https://medium.com/@julsimon

Our mission at AWS

Put machine learning in the
hands of every developer

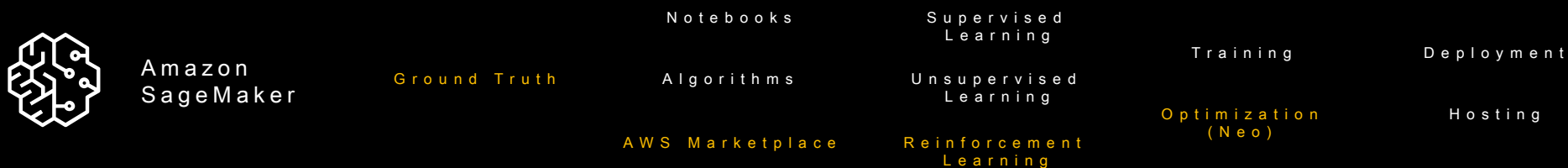# Over 200 launches since re:Invent 2017

The Amazon ML Stack

		Vision			Speech		Language		Chatbots	Forecasting	Recommendations

AI SERVICES
(App developers with little knowledge of ML)

| | Amazon Rekognition Image | Amazon Rekognition Video | Amazon Textract | Amazon Polly | Amazon Transcribe | Amazon Translate | Amazon Comprehend & Comprehend Medical | Amazon Lex | Amazon Forecast | Amazon Personalize |

ML SERVICES
(ML developers and data scientists)

Amazon SageMaker

	Ground Truth	Notebooks	Supervised Learning	Training	Deployment
		Algorithms	Unsupervised Learning	Optimization (Neo)	Hosting
	AWS Marketplace		Reinforcement Learning	

ML FRAMEWORKS & INFRASTRUCTURE
(ML researchers and academics)

Frameworks

- TensorFlow
- mxnet
- PYTORCH
- Chainer
- HOROVOD

Interfaces

- GLUON
- K Keras

Infrastructure

| Amazon EC2 P3 & P3DN | Amazon EC2 C5 | FPGAs | AWS Greengrass | Amazon Elastic Inference | Amazon Inferentia |

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

Machine Learning University

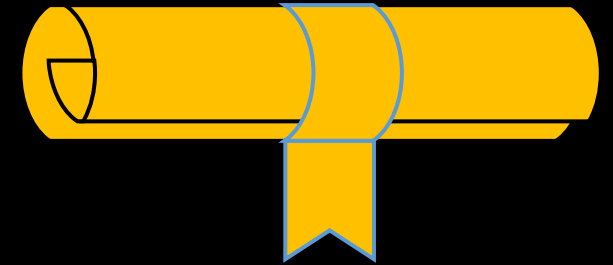https://aws.training/machinelearning

Uses the same
materials used to train
Amazon developers

Foundational knowledge
with
real-world application

Structured
courses and
specialist certification

We're focused on solving the toughest challenges that hold back success with machine learning

Three challenges facing ML world today

1

Flexibility & Cost

Optimized TensorFlow

MXNet Dynamic Training

Amazon Elastic Inference

2

Data

Amazon SageMaker Ground Truth

Amazon SageMaker RL

3

Ease of use

AWS Marketplace for Machine Learning

Amazon SageMaker Neo

Amazon Textract

Amazon Forecast

Amazon Personalize

Amazon Comprehend Medical

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

Let's look under the hood into some of the new features in each layer of the stack.

Three challenges we're focused on today

1
Flexibility & Cost
Optimized TensorFlow

MXNet Dynamic Training

Amazon Elastic Inference

2
Data
Amazon SageMaker Ground Truth

Amazon SageMaker RL

3
Ease of use
AWS Marketplace for Machine Learning

Amazon SageMaker Neo

Amazon Textract

Amazon Forecast

Amazon Personalize

Amazon Comprehend Medical

The Amazon ML Stack

AI SERVICES
(App developers with little knowledge of ML)

	Vision		Speech		Language		Chatbots	Forecasting	Recommendations
	Amazon Rekognition Image	Amazon Rekognition Video	Amazon Polly	Amazon Transcribe	Amazon Translate	Amazon Comprehend & Comprehend Medical	Amazon Lex	Amazon Forecast	Amazon Personalize
			Amazon Textract						

ML SERVICES
(ML developers and data scientists)

Amazon SageMaker

Ground Truth

Notebooks
Algorithms
AWS Marketplace

Supervised Learning
Unsupervised Learning
Reinforcement Learning

Training
Optimization (Neo)

Deployment
Hosting

ML FRAMEWORKS & INFRASTRUCTURE
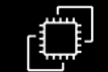(ML researchers and academics)

Frameworks
TensorFlow
mxnet
PYTORCH
Chainer
HOROVOD

Interfaces
GLUON
K Keras

Infrastructure
Amazon EC2 P3 & P3DN
Amazon EC2 C5
FPGAs
AWS Greengrass
Amazon Elastic Inference
Amazon Inferentia

aws INNOVATE
ONLINE CONFERENCE | AI EDITION
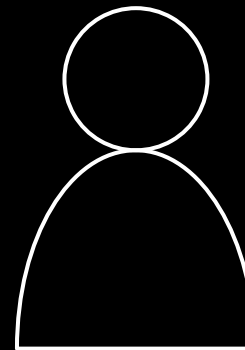
AWS is framework agnostic

Choose from popular frameworks

TensorFlow   mxnet   PYTORCH   Chainer   Caffe2   ONNX   Keras   GLUON

Run them fully managed   Or run them yourself

aws INNOVATE
ONLINE CONFERENCE | AI EDITION
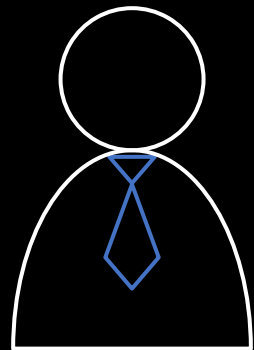
AWS: The platform of choice to run TensorFlow

Adobe
AUTODESK
edmunds
Expedia

GoodData
grammarly
intuit
Pinterest

Rakuten Intelligence
SIEMENS
THOMSON REUTERS
Zocdoc

hudl
Cox Automotive
News Corp Australia
zendesk

85% of all TensorFlow workloads in the cloud runs on AWS

Source: Nucleus Research, November 2018

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

The best place to run TensorFlow

Stock
TensorFlow

65%

scaling efficiency
with 256 GPUs

The best place to run TensorFlow

Stock
TensorFlow

65%

scaling efficiency
with 256 GPUs

→

AWS-Optimized
TensorFlow

90%

scaling efficiency
with 256 GPUs

HOROVOD

Available with
Amazon SageMaker
and the AWS Deep
Learning AMIs

Apache MXNet: deep learning for enterprise developers

SAMSUNG
SAMSUNG SDS

ByteDance

NTT docomo

Celgene

cimpress

SK telecom

NVIDIA.

intel

EAGLEVIEW®

WOLFRAM

图森 tu Simple

gumgum·

CURALATE

pixm

Finn AI

Georgian Partners

Σ SIGOPT

LOHIKA

comet

TEAMWORK

Start with off-the-shelf toolkits
- Gluon CV and Gluon NLP

Fast and scalable training
- Keras-MXNet up to 2x faster than Keras-TensorFlow
- Near-linear scalability up to 256 GPUs
- Dynamic training

Easy deployment
- Java/Scala APIs
- MXNet Model Server

Apache MXNet customer example: Chick-fil-A
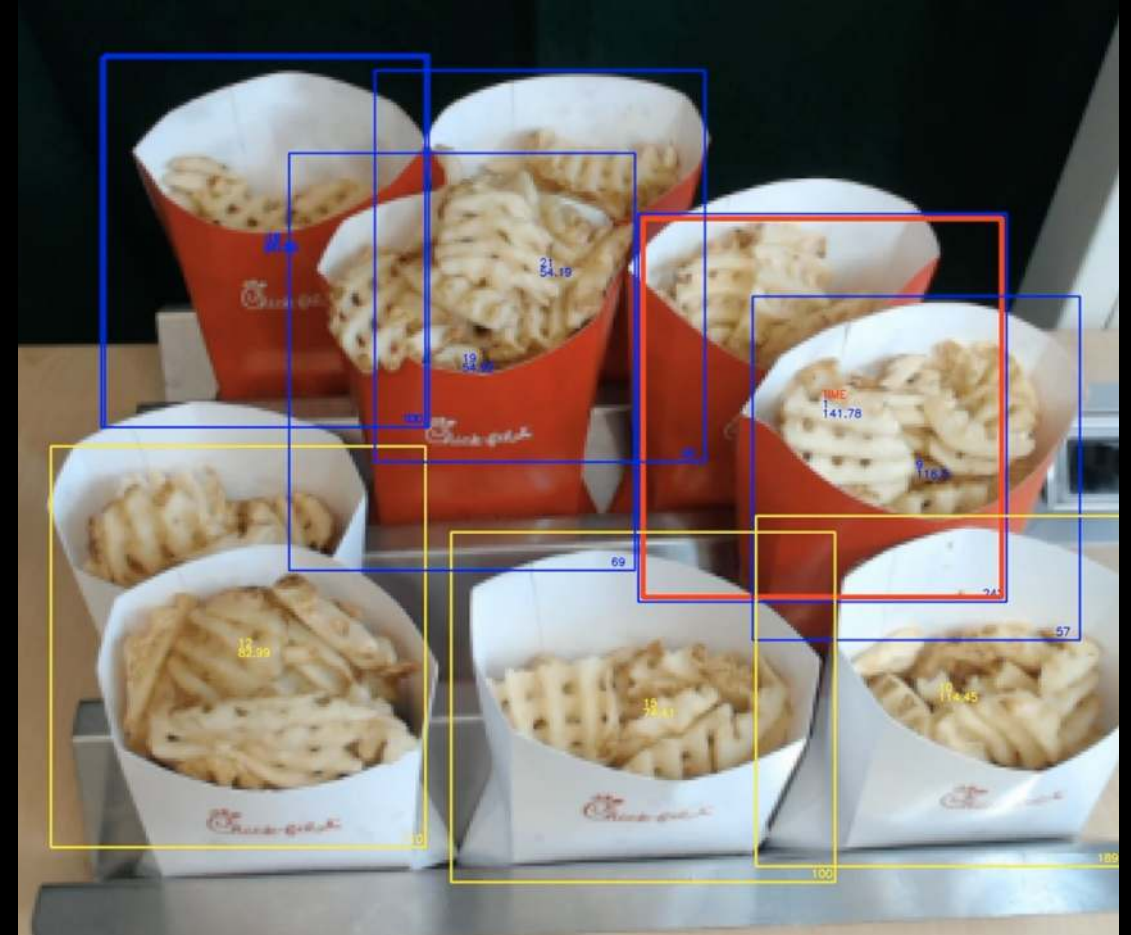
TRACK WAFFLE FRY FRESHNESS
Identify fries that have exceeded the hold time

USES DEEP LEARNING
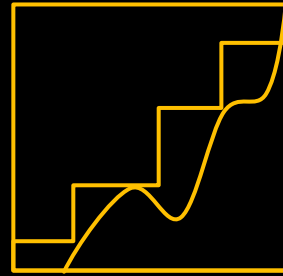Computer vision model for object detection and tracking

BUILT THE MODEL USING MXNet
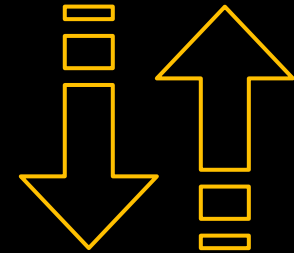A team of enterprise developers with no ML expertise

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

Amazon Elastic Inference

Lower inference costs
up to 75%

Match capacity
to demand

Available between 1 to 32
TFLOPS

K E Y
F E A T U R E S

Integrated with
Amazon EC2,
Amazon SageMaker,
and Amazon DL AMIs

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

Single and
mixed-precision
operations

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

Helping developers learn computer vision

AWS DeepLens: the world's first deep learning-enabled video camera for developers

- Purpose-built for ML-skills development

- Fully programmable & customizable

- Build custom Amazon SageMaker models

- 10-minutes to your first deep learning project

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

Three challenges we're focused on today

1

Flexibility & Cost

Optimized TensorFlow

MXNet Dynamic Training

Amazon Elastic Inference

2

Data

Amazon SageMaker Ground Truth

Amazon SageMaker RL

3

Ease of use

AWS Marketplace for Machine Learning

Amazon SageMaker Neo

Amazon Textract

Amazon Forecast

Amazon Personalize

Amazon Comprehend Medical

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

The Amazon ML Stack

AI SERVICES
(App developers with little knowledge of ML)

Vision			Speech		Language		Chatbots	Forecasting	Recommendations
Amazon Rekognition Image	Amazon Rekognition Video	Amazon Textract	Amazon Polly	Amazon Transcribe	Amazon Translate	Amazon Comprehend & Comprehend Medical	Amazon Lex	Amazon Forecast	Amazon Personalize

ML SERVICES
(ML developers and data scientists)

Amazon SageMaker

Ground Truth

Notebooks

Algorithms

AWS Marketplace

Supervised Learning

Unsupervised Learning

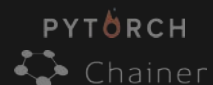Reinforcement Learning

Training

Optimization (Neo)

Deployment

Hosting

ML FRAMEWORKS & INFRASTRUCTURE
(ML researchers and academics)

Frameworks
TensorFlow
mxnet
PYTORCH
Chainer
HOROVOD

Interfaces
GLUON
Keras

Infrastructure
Amazon EC2 P3 & P3DN
Amazon EC2 C5
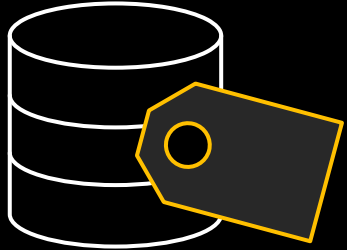FPGAs
AWS Greengrass
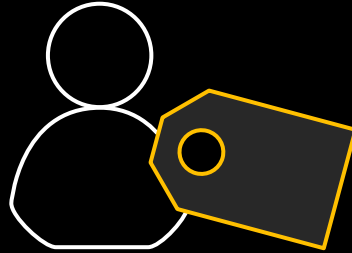Amazon Elastic Inference
Amazon Inferentia

Labeling data sets is too time-consuming

Amazon SageMaker Ground Truth
Label machine learning training data easily and accurately

Quickly label
training data

Easily integrate
human labelers

Get accurate
results

**KEY
FEATURES**

Automatic labeling via
machine learning

Ready-made and
custom workflows

Private and public
human workforce

Label
management

Tyson

CHANGE
HEALTHCARE

Automagi

gumgum

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

How do you teach machine learning models to make decisions when there is no training data?
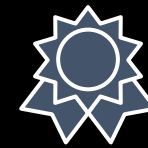
Reinforcement Learning

Learn by interacting with the real world

Model the real-world problem as a simulation environment

Trial and error
Observe results

Optimize learning strategy to maximize long-term reward

Model learns how to make complex decisions

Amazon SageMaker RL

RL toolkits that provide RL agent algorithm implementations

RL-Coach

| DQN | PPO | A3C | Rainbow | ... |

RL-Ray RLLib

| DQN | PPO | IMPALA | A3C | ... |

Open AI Baselines

| DQN | PPO | ... |

...

Amazon SageMaker deep learning frameworks

| Tensorflow | MxNet | PyTorch | Chainer |

SageMaker supported　　Customer BYO

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

Customers are using Amazon SageMaker RL

Helping developers learn RL

AWS DeepRacer: a fully autonomous 1/18th-scale race car

- Build machine learning models in Amazon SageMaker
- Train, test, and iterate on the track using the AWS DeepRacer 3D racing simulator
- Compete in the world's first global autonomous racing league, either at AWS Summits or at virtual events
- Race for prizes and a chance to advance to win the coveted AWS DeepRacer Cup

Three challenges we're focused on today

1

Flexibility & Cost

Optimized TensorFlow

MXNet Dynamic Training

Amazon Elastic Inference

2

Data

Amazon SageMaker Ground Truth

Amazon SageMaker RL

3

Ease of use

AWS Marketplace for Machine Learning

Amazon SageMaker Neo

Amazon Textract

Amazon Forecast

Amazon Personalize

Amazon Comprehend Medical

Deploying models across multiple platforms is too time-consuming

We all want machine learning everywhere
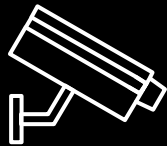
Autonomous
vehicles

Smart
agriculture

Predictive
maintenance

Robotics

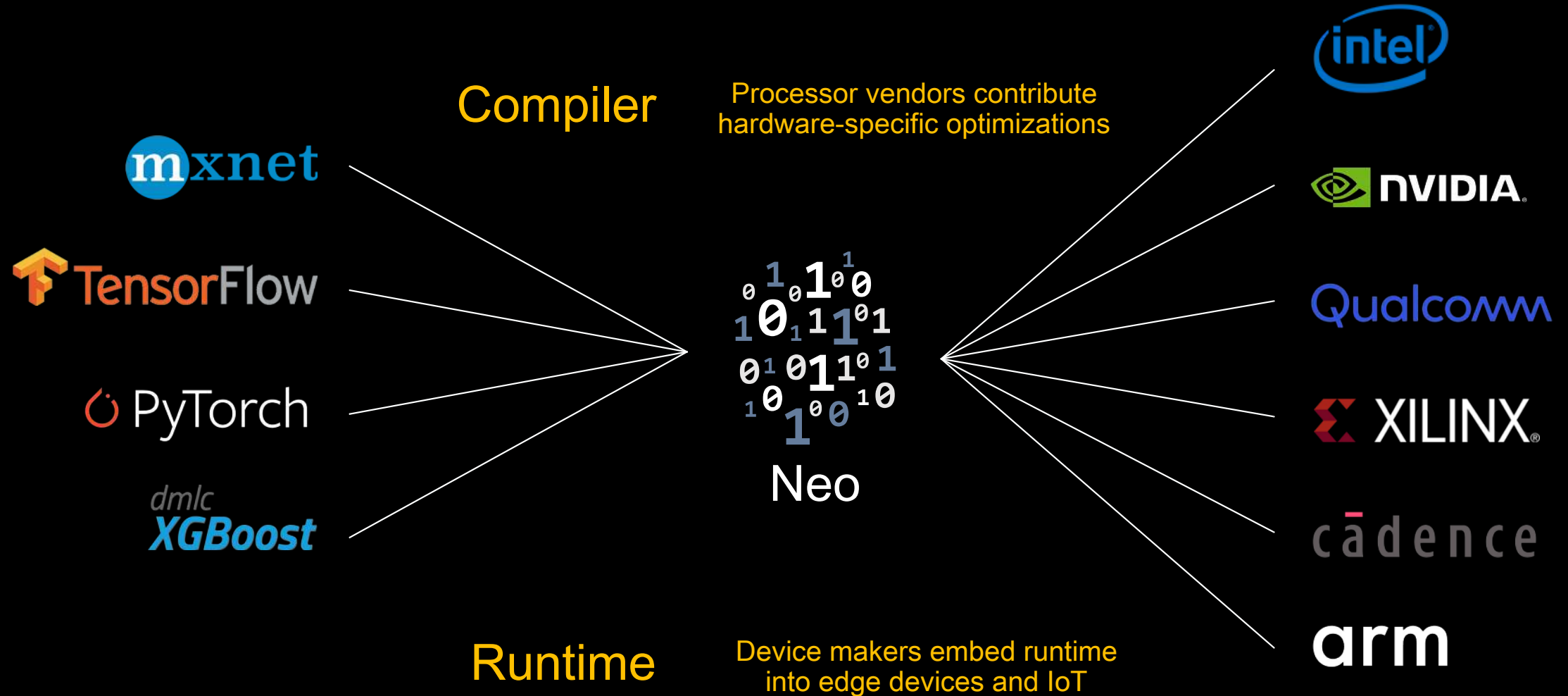Speech and sound
recognition

Video
security

Anomaly
detection

More

Neo: train once, run anywhere

The Amazon ML Stack

	Vision			Speech		Language		Chatbots	Forecasting	Recommendations

AI SERVICES

(App developers with little knowledge of ML)

| Amazon Rekognition Image | Amazon Rekognition Video | Amazon Textract | Amazon Polly | Amazon Transcribe | Amazon Translate | Amazon Comprehend & Comprehend Medical | Amazon Lex | Amazon Forecast | Amazon Personalize |

ML SERVICES

(ML developers and data scientists)

Amazon SageMaker

Ground Truth

Notebooks
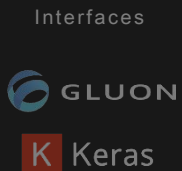
Algorithms

AWS Marketplace

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Training

Optimization (Neo)

Deployment

Hosting

ML FRAMEWORKS & INFRASTRUCTURE

(ML researchers and academics)

Frameworks

TensorFlow

mxnet

PYTORCH

Chainer

HOROVOD

Interfaces

GLUON

Keras

Infrastructure

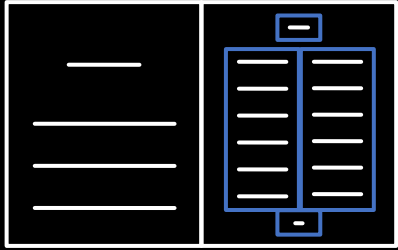Amazon EC2 P3 & P3DN

Amazon EC2 C5

FPGAs

AWS Greengrass

Amazon Elastic Inference

Amazon Inferentia

Amazon Textract (preview)
Extract text and data from virtually any document

Extract data quickly and accurately
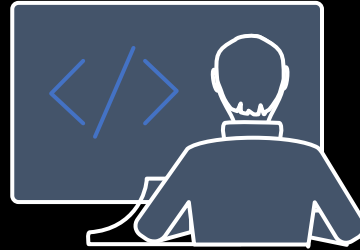
Eliminate manual effort

Lower document processing costs

KEY FEATURES

| Optical Character Recognition (OCR) | Key-value pair detection | Table detection | Adjustable confidence thresholds | Bounding box coordinates | No ML experience required |

Alfresco®

Cox AUTOMOTIVE

Roche

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

Amazon Personalize (preview)

Amazon Forecast (preview)

Historical Data
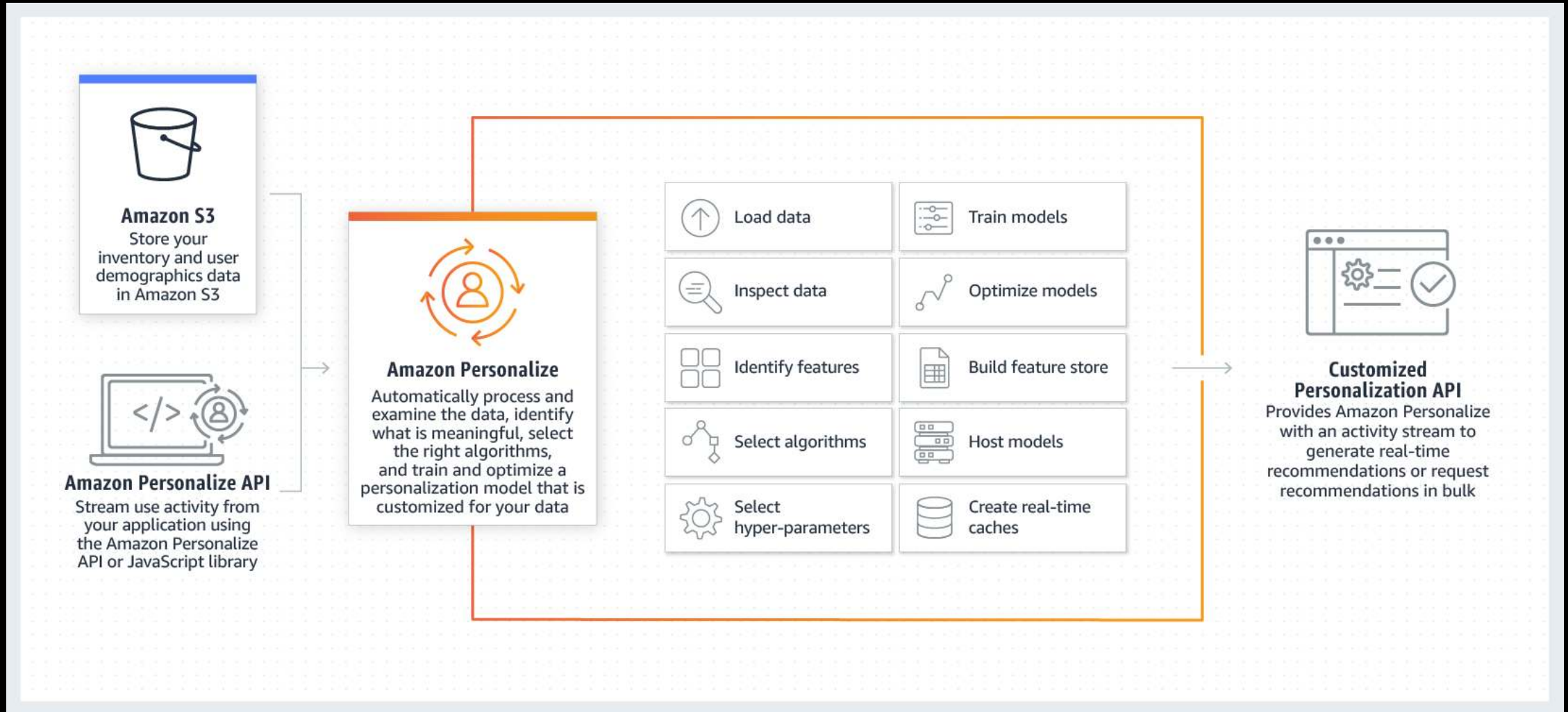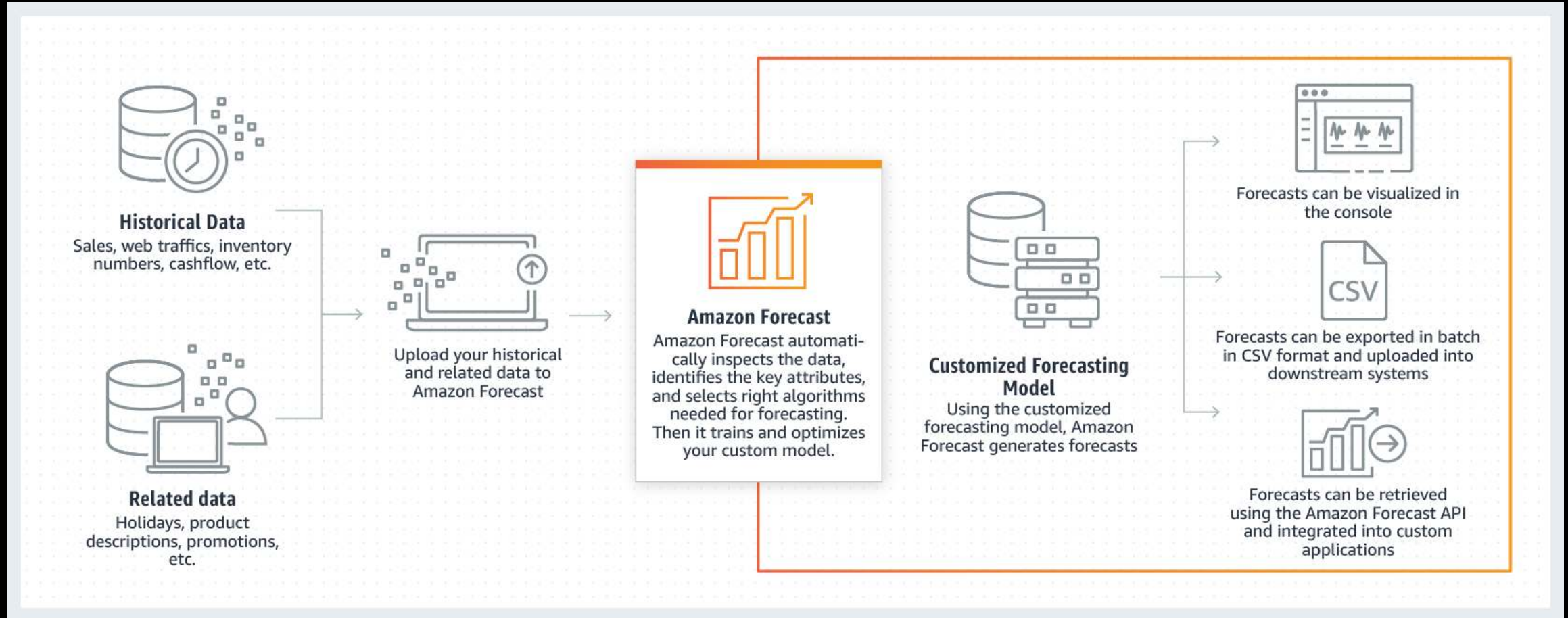Sales, web traffics, inventory numbers, cashflow, etc.

Related data
Holidays, product descriptions, promotions, etc.

Upload your historical and related data to Amazon Forecast

Amazon Forecast
Amazon Forecast automatically inspects the data, identifies the key attributes, and selects right algorithms needed for forecasting. Then it trains and optimizes your custom model.

Customized Forecasting Model
Using the customized forecasting model, Amazon Forecast generates forecasts

Forecasts can be visualized in the console

CSV
Forecasts can be exported in batch in CSV format and uploaded into downstream systems

Forecasts can be retrieved using the Amazon Forecast API and integrated into custom applications

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

Amazon Comprehend Medical

Three challenges facing ML world today

1

Flexibility & Cost

Optimized TensorFlow

MXNet Dynamic Training

Amazon Elastic Inference

2

Data

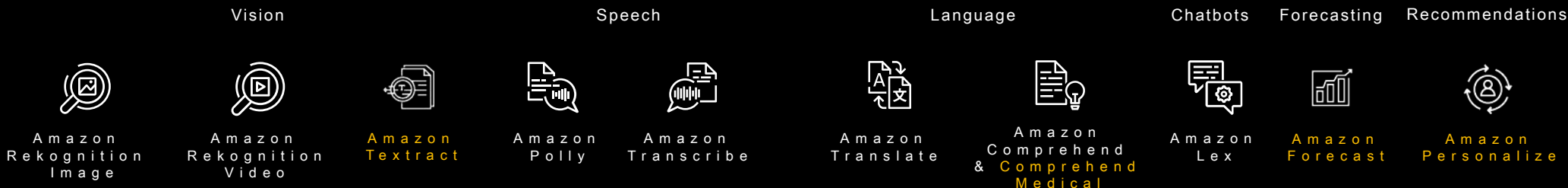Amazon SageMaker Ground Truth

Amazon SageMaker RL

3

Ease of use

AWS Marketplace for Machine Learning

Amazon SageMaker Neo

Amazon Textract

Amazon Forecast
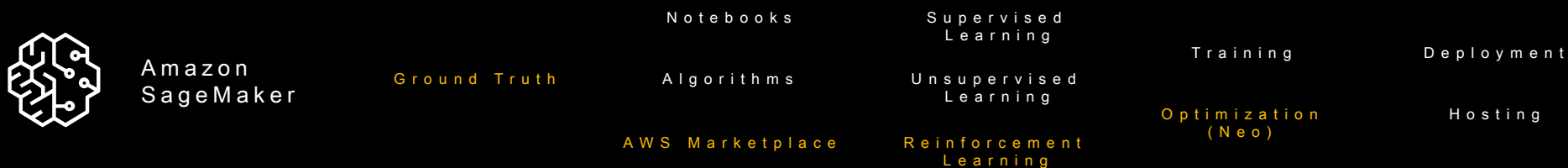
Amazon Personalize

Amazon Comprehend Medical

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

The Amazon ML Stack

	Vision	Speech	Language	Chatbots	Forecasting	Recommendations

AI SERVICES

(App developers with little knowledge of ML)

Amazon Rekognition Image   Amazon Rekognition Video   Amazon Textract   Amazon Polly   Amazon Transcribe   Amazon Translate   Amazon Comprehend & Comprehend Medical   Amazon Lex   Amazon Forecast   Amazon Personalize

---

ML SERVICES

(ML developers and data scientists)

Amazon SageMaker

Ground Truth

Notebooks   Supervised Learning

Algorithms   Unsupervised Learning

AWS Marketplace   Reinforcement Learning

Training   Deployment

Optimization (Neo)   Hosting

ML FRAMEWORKS & INFRASTRUCTURE

(ML researchers and academics)

Frameworks

TensorFlow
mxnet
PYTORCH
Chainer
HOROVOD

Interfaces

GLUON
K Keras

Infrastructure

Amazon EC2 P3 & P3DN   Amazon EC2 C5   FPGAs   AWS Greengrass   Amazon Elastic Inference   Amazon Inferentia

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

Track 1	Track 2	Track 3
An Introduction to Machine Learning (100 Level)	Use Natural Language and Text Analysis to Power More Intelligent Applications (200 Level)	Get Started with Amazon SageMaker: Build, Train, and Deploy Machine Learning Models in the Cloud (200 Level)
Break	Break	Break
Get Started with Amazon SageMaker: Build, Train, and Deploy Machine Learning Models in the Cloud (200 Level)	Power New Application Experiences with Image, Video, and Facial Analysis (200 Level)	Remove Your Biggest Challenge to Getting Started with ML: Using Amazon SageMaker Ground Truth to Label Data at 70% Less Cost (200 Level)
Break	Break	Break
Use AWS DeepRacer to Get Rolling with Machine Learning (200 Level)	Integrate Amazon.com's ML Techniques for Customer Recommendations and Forecasting Workflows into Your Applications (200 Level)	Learn to Build What's Next with Reinforcement Learning on Amazon SageMaker (200 Level)

aws INNOVATE
ONLINE CONFERENCE | AI EDITION

aws INNOVATE

ONLINE CONFERENCE | AI EDITION

Thank you!

Julien Simon

Global Evangelist, AI & Machine Learning, AWS

@julsimon

https://medium.com/@julsimon