AI and Machine Learning on AWS

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

Our Approach for Machine Learning

Customer-focused

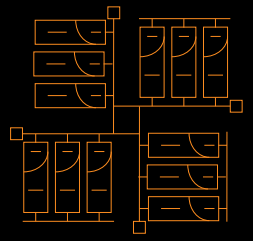
90%+ of our ML roadmap is defined by customers

Pace of innovation

200+ new ML launches and major feature updates in the last year

Breadth and depth

A wide range of AI and ML services



Multi-framework

Support for the most popular frameworks

Security and analytics

Deep set of security and encryption features, with robust analytics capabilities

Embedded R&D

Customer-centric approach to advancing the state of the art

The Amazon ML Stack

AI SERVICES

(App developers with little knowledge of ML)

Vision


Amazon Rekognition Image


Amazon Rekognition Video

Speech


Amazon Textract


Amazon Polly


Amazon Transcribe

Language


Amazon Translate


Amazon Comprehend & Comprehend Medical

Chatbots


Amazon Lex

Forecasting


Amazon Forecast

Recommendations


Amazon Personalize

ML SERVICES

(ML developers and data scientists)


Amazon SageMaker

Notebooks

Algorithms

AWS Marketplace

Ground Truth

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

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

Three challenges facing ML world today

Flexibility & Cost

Optimized TensorFlow

Amazon Elastic Inference

Data

Amazon SageMaker Ground Truth

Amazon SageMaker RL

Ease of use

AWS Marketplace for Machine Learning

Amazon SageMaker Neo

Amazon Textract

Amazon Forecast

Amazon Personalize

Amazon Comprehend Medical

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

Amazon Elastic Inference

2

Data

Amazon SageMaker Ground Truth

Amazon SageMaker RL

3

Ease of use

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


Amazon Rekognition Image


Amazon Rekognition Video

Speech


Amazon Textract


Amazon Polly


Amazon Transcribe

Language


Amazon Translate


Amazon Comprehend & Comprehend Medical

Chatbots


Amazon Lex

Forecasting


Amazon Forecast

Recommendations


Amazon Personalize

ML SERVICES

(ML developers and data scientists)


Amazon SageMaker

Notebooks

Ground Truth

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

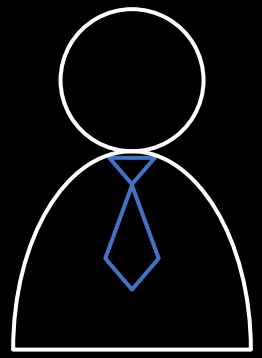

Amazon Inferentia

AWS is framework agnostic

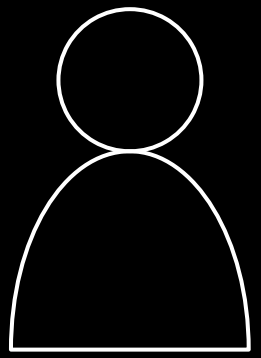
Choose from popular frameworks

TensorFlow mxnet PYTORCH Chainer Caffe2 ONNX Keras GLUON

Run them fully managed



Or run them yourself



AWS: The platform of choice to run TensorFlow

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

Source: Nucleus Research, November 2018

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,
AWS Deep Learning AMIs,
AWS Deep Learning containers

Apache MXNet: deep learning for enterprise developers

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

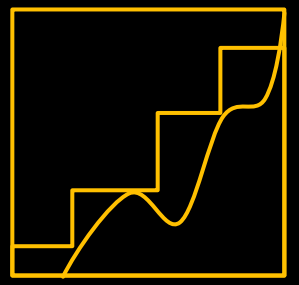
TuSimple

- TuSimple, a leader in self-driving technology, uses Deep Learning to build sophisticated algorithms for computer vision and driving simulation.
- They rely on Apache MXNet to teach computers how to recognize and track objects and to make decisions to avoid collisions and prioritize safety.
- They simulated a billion miles of road driving with a wide range of variables and driving conditions—the largest simulation of its kind in history.

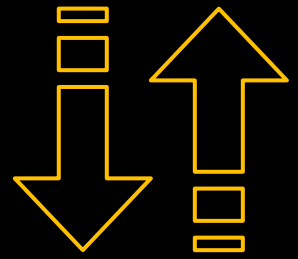
https://www.oreilly.com/ideas/self-driving-trucks-enter-the-fast-lane-using-deep-learning

Amazon Elastic Inference

Lower inference costs up to 75%



Match capacity to demand



Available between 1 to 32 TFLOPS

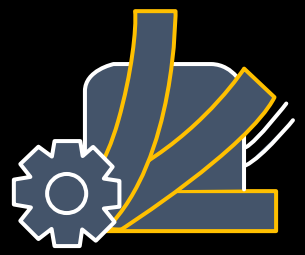
KEY FEATURES

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

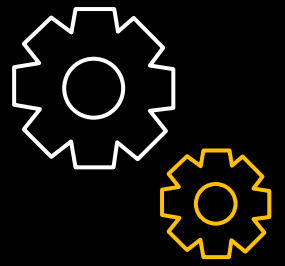
AWS: Best platform to run PyTorch



Flexible

Fast prototyping

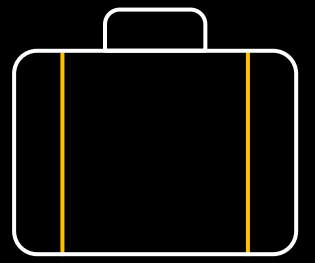
Seamless transition from research to production using Amazon SageMaker



Versatile

Train and run a variety of models, including CNN and LSTM

Train custom models with Facebook's FAIRSeq toolkit on Amazon SageMaker



Portable

Develop models in PyTorch and transfer to other frameworks like MXNet for inference using ONNX

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

Three challenges we're focused on today

1

Flexibility & Cost

Optimized TensorFlow

Amazon Elastic Inference

2

Data

Amazon SageMaker Ground Truth

Amazon SageMaker RL

3

Ease of use

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


Amazon Rekognition Image


Amazon Rekognition Video

Speech


Amazon Textract


Amazon Polly


Amazon Transcribe

Language


Amazon Translate


Amazon Comprehend & Comprehend Medical

Chatbots


Amazon Lex

Forecasting


Amazon Forecast

Recommendations


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

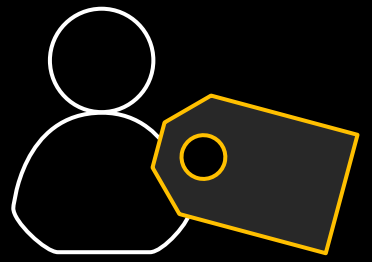

Amazon Inferentia

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

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

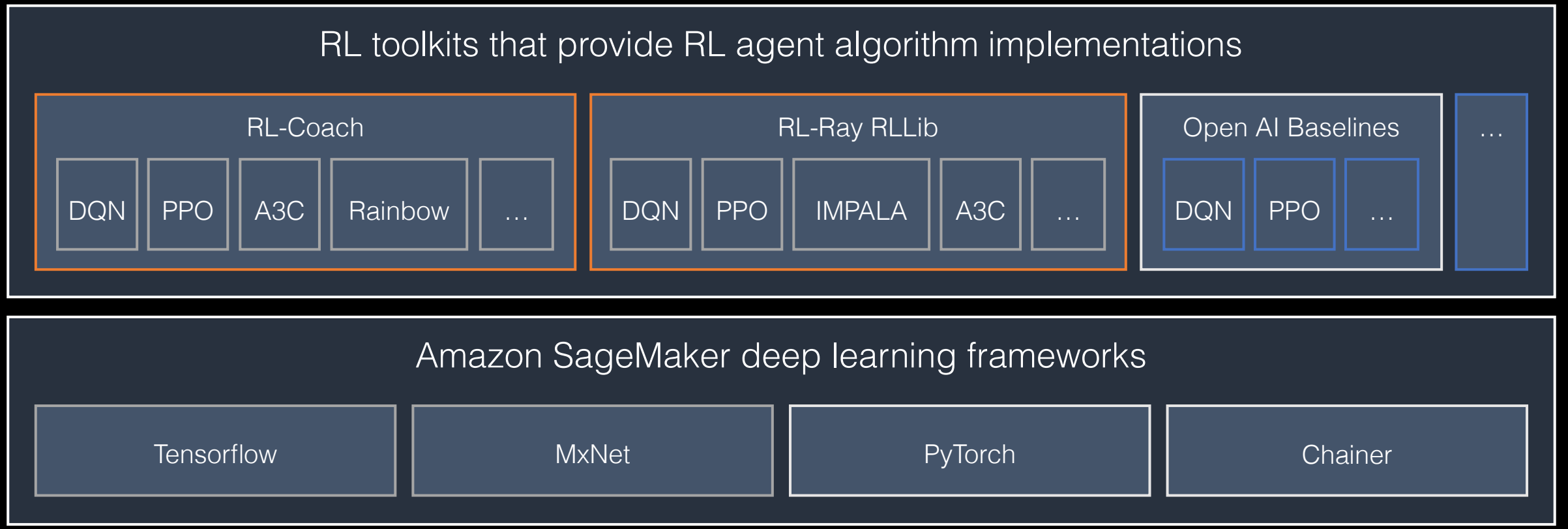
Trial and error

Observe results

Optimize learning strategy to maximize long-term reward

Model learns how to make complex decisions

Amazon SageMaker RL



Customers are using Amazon SageMaker RL

GE Healthcare

HONDA

SIXT

mixi

amazon

SyntheticGestalt

Scientific Research by Artificially Intelligent Agents

Tradelegs

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

Flexibility & Cost

Optimized TensorFlow

Amazon Elastic Inference

Data

Amazon SageMaker Ground Truth

Amazon SageMaker RL

Ease of use

Amazon SageMaker Neo

Amazon Textract

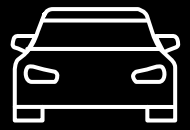
Amazon Forecast

Amazon Personalize

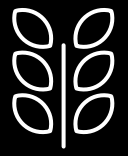
Amazon Comprehend Medical

Deploying models across multiple platforms is too time-consuming

We all want machine learning everywhere



Autonomous vehicles

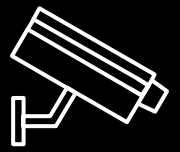


Smart agricultur e

Predictive maintenance

Robotics

Speech and sound recognition

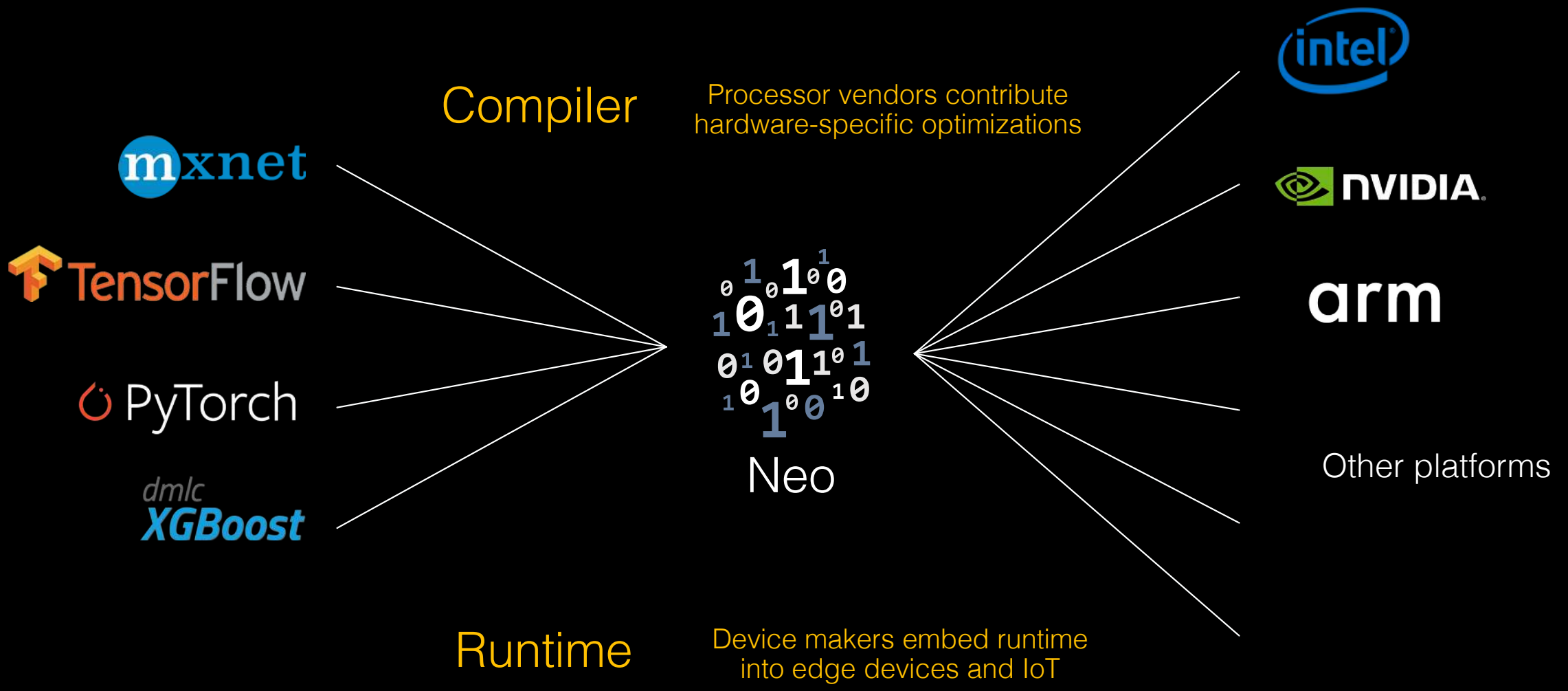


Video security

Anomaly detection

More

Neo: train once, run anywhere



The Amazon ML Stack

AI SERVICES

(App developers with little knowledge of ML)

Vision


Amazon Rekognition Image


Amazon Rekognition Video

Speech


Amazon Textract


Amazon Polly


Amazon Transcribe

Language


Amazon Translate


Amazon Comprehend & Comprehend Medical

Chatbots


Amazon Lex

Forecasting


Amazon Forecast

Recommendations


Amazon Personalize

ML SERVICES

(ML developers and data scientists)


Amazon SageMaker

Notebooks

Ground Truth

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

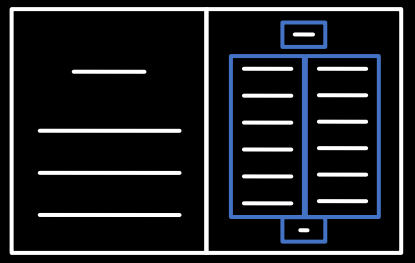

AWS Greengrass


Amazon Elastic Inference

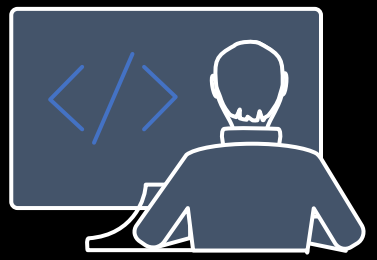

Amazon Inferentia

Amazon Textract

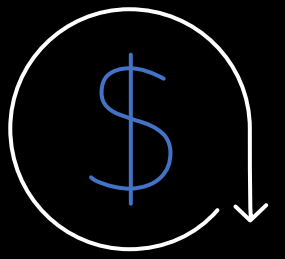
Extract text and data from virtually any document



Extract data quickly and accurately



Eliminate manual effort

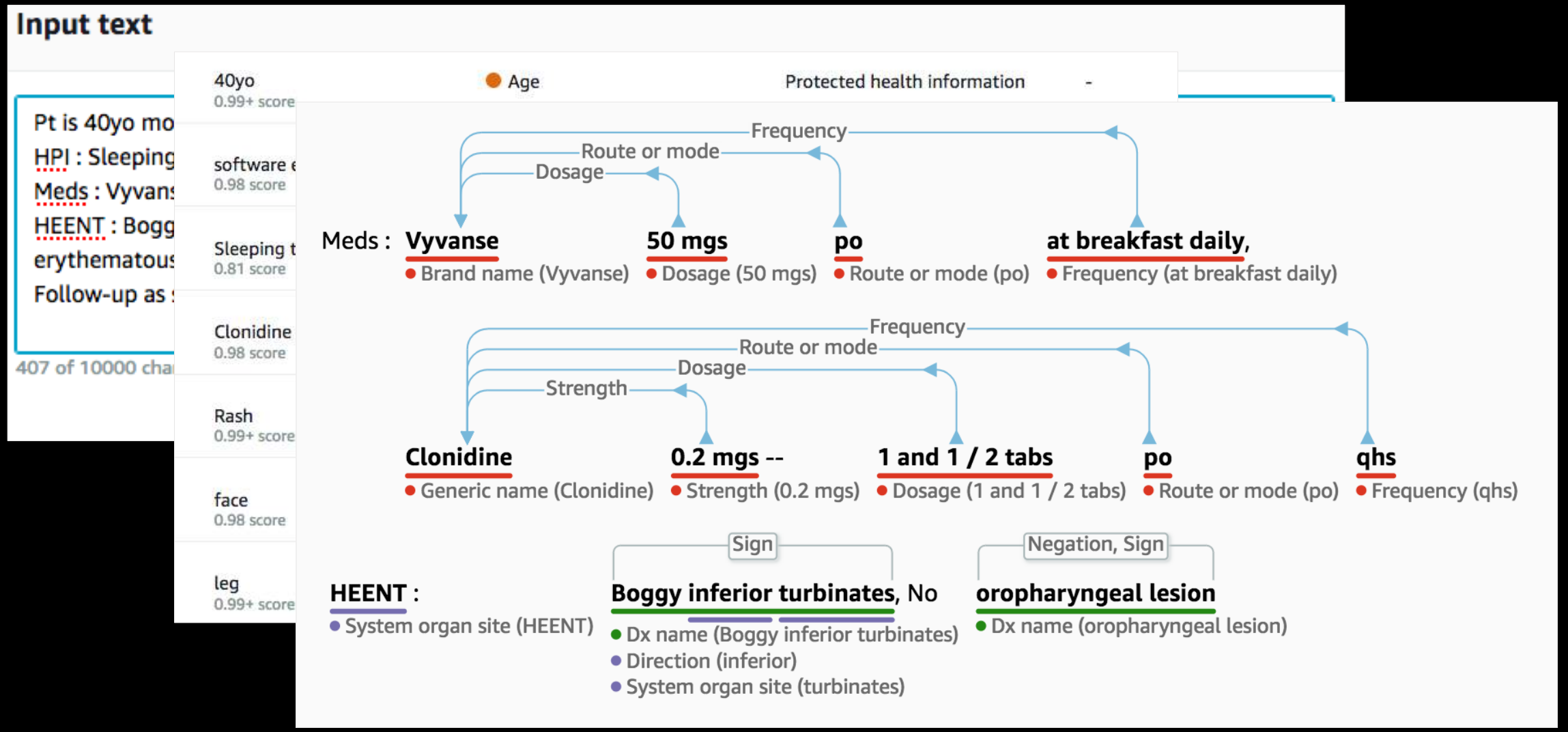


Lower document processing costs

KEY FEATURES

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

Amazon Comprehend Medical

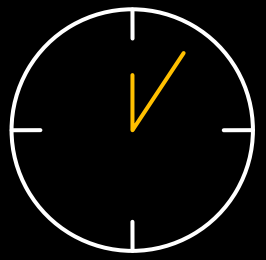


© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

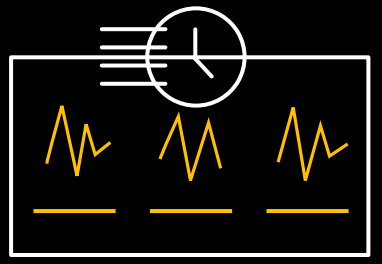
Amazon Personalize

Improve customer experiences with personalization and recommendations

Deliver high quality recommendations



Real-time



Deliver personalization in days, not months

Works with any product or content

KEY FEATURES

Context-aware Recommendations

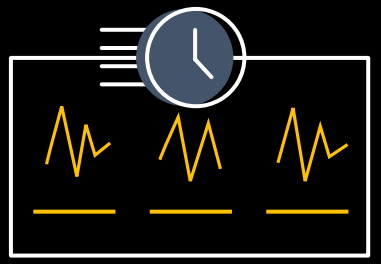
Automated machine learning

Continuous learning to improve performance

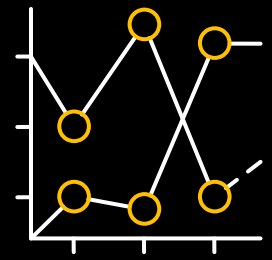
Amazon Forecast (preview)

Improve forecasting accuracy by up to 50% at 1/10th the cost

Accurate forecasts



Get to results quickly



Works with any historical time-series

KEY FEATURES

- Consider multiple time-series at once
- Automatic machine learning
- Evaluate model accuracy
- Visualize forecasts & import results into business apps
- Schedule forecasts and model retraining

Three challenges facing ML world today

Flexibility & Cost

Optimized TensorFlow

Amazon Elastic Inference

Data

Amazon SageMaker Ground Truth

Amazon SageMaker RL

Ease of use

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


Amazon Rekognition Image


Amazon Rekognition Video

Speech


Amazon Textract


Amazon Polly


Amazon Transcribe

Language


Amazon Translate


Amazon Comprehend & Comprehend Medical

Chatbots


Amazon Lex

Forecasting


Amazon Forecast

Recommendations


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

Thank you!

Julien Simon

Global Evangelist, AI & Machine Learning, AWS

@julsimon

https://medium.com/@julsimon