Real-Life Machine Learning on AWS

Julien Simon
Global Evangelist, AI & Machine Learning, Amazon Web Services
@julsimon

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

aws

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

aws

The Amazon ML stack: broadest & deepest set of capabilities

AI SERVICES

Vision


REKOGNITION IMAGE


REKOGNITION VIDEO

Speech


TEXTRACT


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 NFERENCE

aws

Hyper-Personalisation

AI System

Peak unified and ingested Footasylum's data in the AI System to create a "single customer view".

Machine Learning Algorithms

Peak's AI & ML powered algorithms draw insights based on the past transactional and behavioral data of each customer.

Using Amazon SageMaker powered by Intel C5 instances

Segment-based Factors

- Predicted Customer Value
- In-market scores
- Churn risk
- Brand Preferences
- Style Preferences

AI Solutions Include:

- **Segmentation**
- **Churn prediction**
- **Great Customer Prediction**
- **Customer Lifetime Value**
- **Preference scoring**
- **In-market prediction**

Uplift in Revenue ⬆

Highly-targeted, hyper-personalised marketing. Immediate results, with a 28% uplift in revenue per email sent.

Finding missing children

THORN

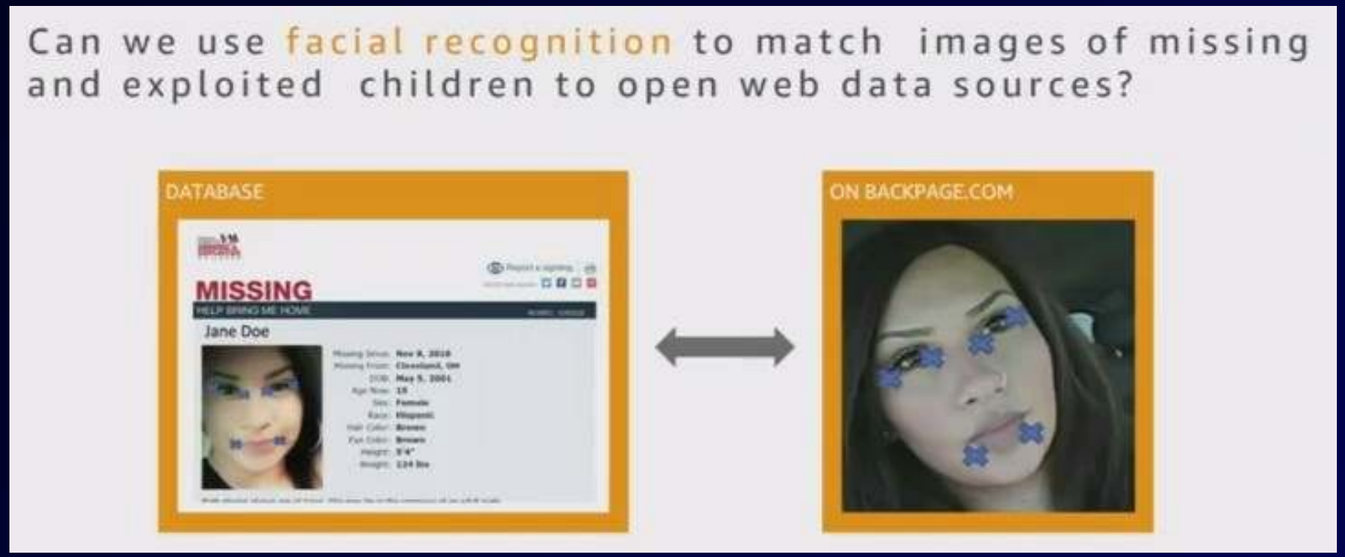
Thorn is a global nonprofit organization fighting child sexual exploitation and child traffickers.

Thorn have built an age progressed facial recognition service using Deep Learning and C5 instances.

Thorn can apply 5,000 data points to a single face and make a positive match in 200 milliseconds, compared to 20 minutes previously.

This tool is used by 5,300 officers in over 18 countries and identifies an average of 5 kids per day.

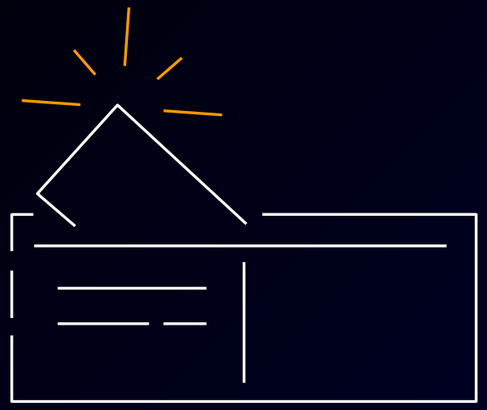
Thorn are also now using Amazon Rekognition, our fully-managed image analysis service.



https://www.youtube.com/watch?v=ITlv273XUNM

aws

Three areas we are improving for ML developers

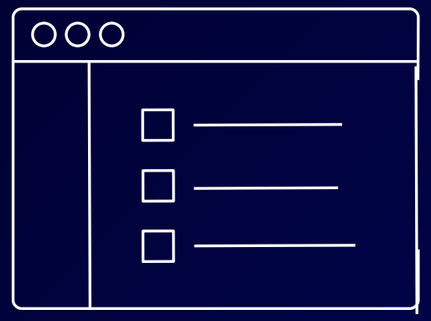


Cost

We're improving both training and inference speed & cost

Data

Preparing data for ML is major expensive, complex, and time consuming

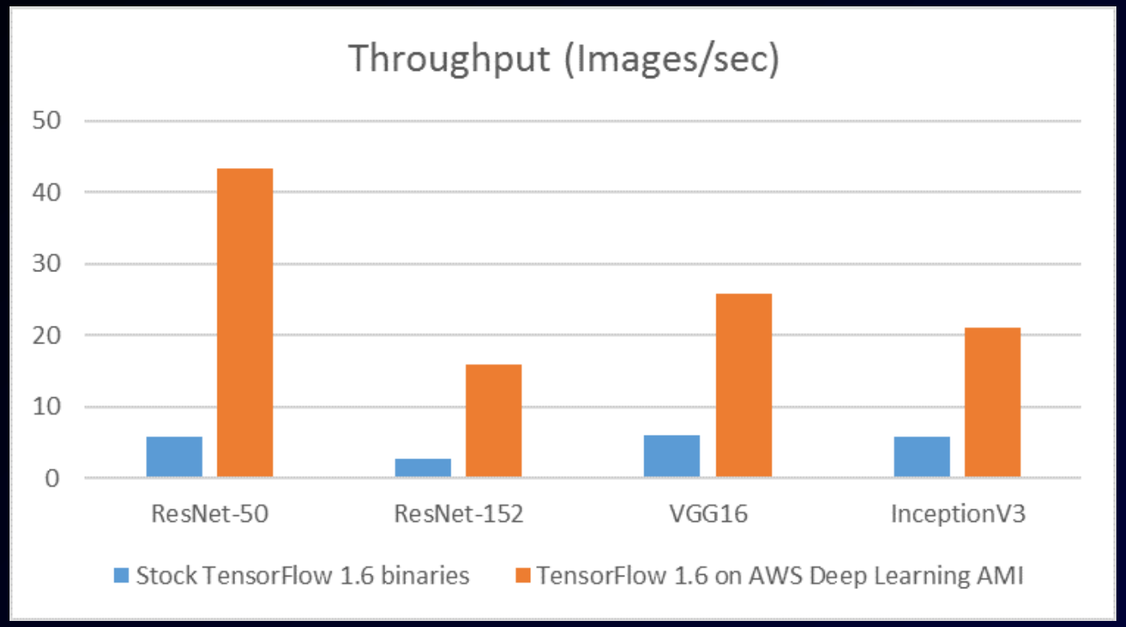


Ease of use

We continue to want to reduce the barrier of entry to ML for all developers

Improving Training & Inference Cost

Optimizing TensorFlow training on C5



https://aws.amazon.com/blogs/machine-learning/faster-training-with-optimized-tensorflow-1-6-on-amazon-ec2-c5-and-p3-instances/ (March 2018)

Training a ResNet-50 benchmark with the synthetic ImageNet dataset using our optimized build of TensorFlow 1.11 on a c5.18xlarge instance type is 11X faster than training on the stock binaries.

https://aws.amazon.com/about-aws/whats-new/2018/10/chainer4-4_theano_1-0-2_launch_deep_learning_ami/ (October 2018)

Amazon EC2 C5n instance

https://aws.amazon.com/blogs/aws/new-c5n-instances-with-100-gbps-networking/

Intel Xeon Platinum 8000

Up to 3.5GHz single core speed

Up to 100Gbit networking

Based on Nitro hypervisor for bare metal-like performance

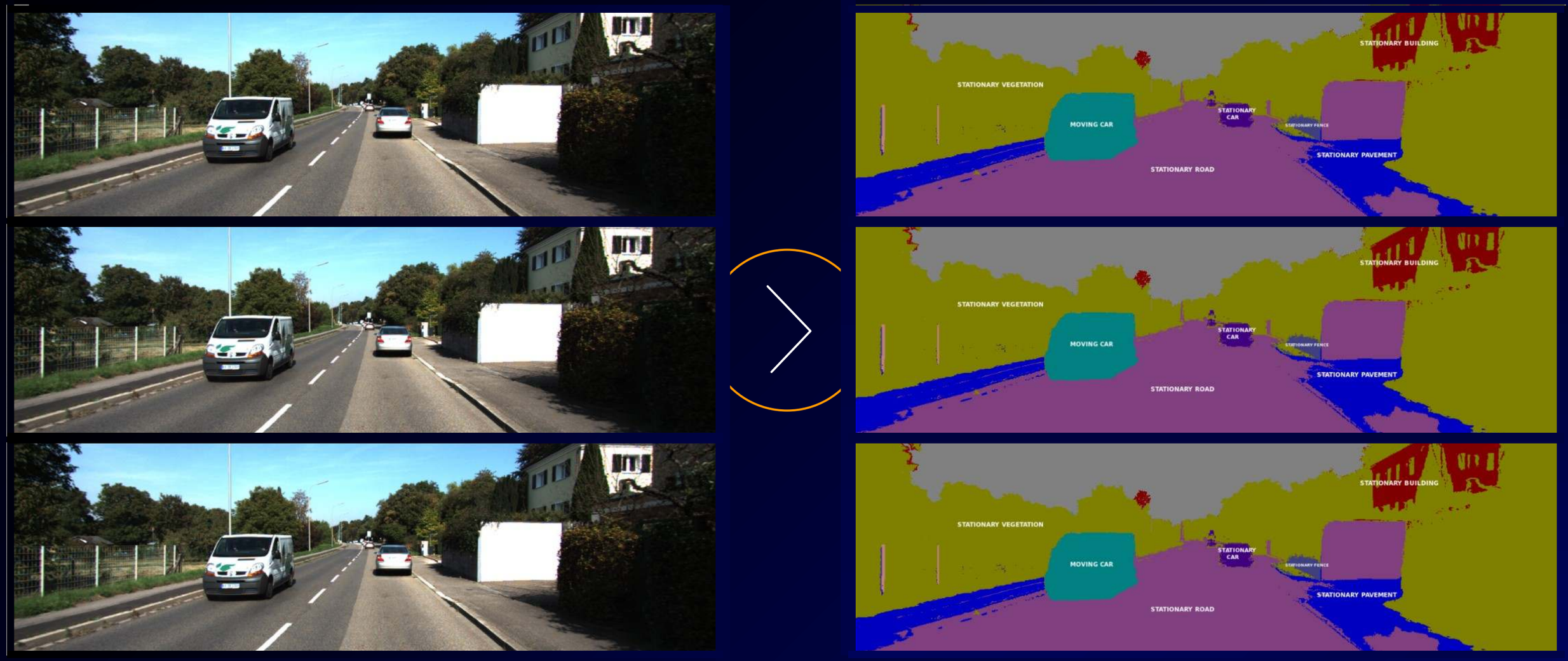
Instance Name	vCPUs	RAM	EBS Bandwidth	Network Bandwidth
c5n.large	2	5.25 GiB	Up to 3.5 Gbps	Up to 25 Gbps
c5n.xlarge	4	10.5 GiB	Up to 3.5 Gbps	Up to 25 Gbps
c5n.2xlarge	8	21 GiB	Up to 3.5 Gbps	Up to 25 Gbps
c5n.4xlarge	16	42 GiB	3.5 Gbps	Up to 25 Gbps
c5n.9xlarge	36	96 GiB	7 Gbps	50 Gbps
c5n.18xlarge	72	192 GiB	14 Gbps	100 Gbps

aws

Making it easier to obtain high quality labeled data

Successful models require high-quality data

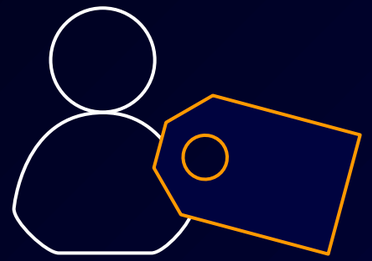
Successful models require high-quality data



Amazon SageMaker Ground Truth

https://aws.amazon.com/blogs/aws/amazon-sagemaker-ground-truth-build-highly-accurate-datasets-and-reduce-labeling-costs-by-up-to-70

Quickly label training data



Easily integrate human labelers

Get accurate results

KEY FEATURES

Automatic labeling via machine learning

Ready-made and custom workflows for image bounding box, segmentation, and text

Private and public human workforce

Label management

Creating training data

amazon mechanical turk beta

500,000+ workers

Private labeling workforce

Third-party vendors

iMerit human-empowered computing

capestart

vivitec

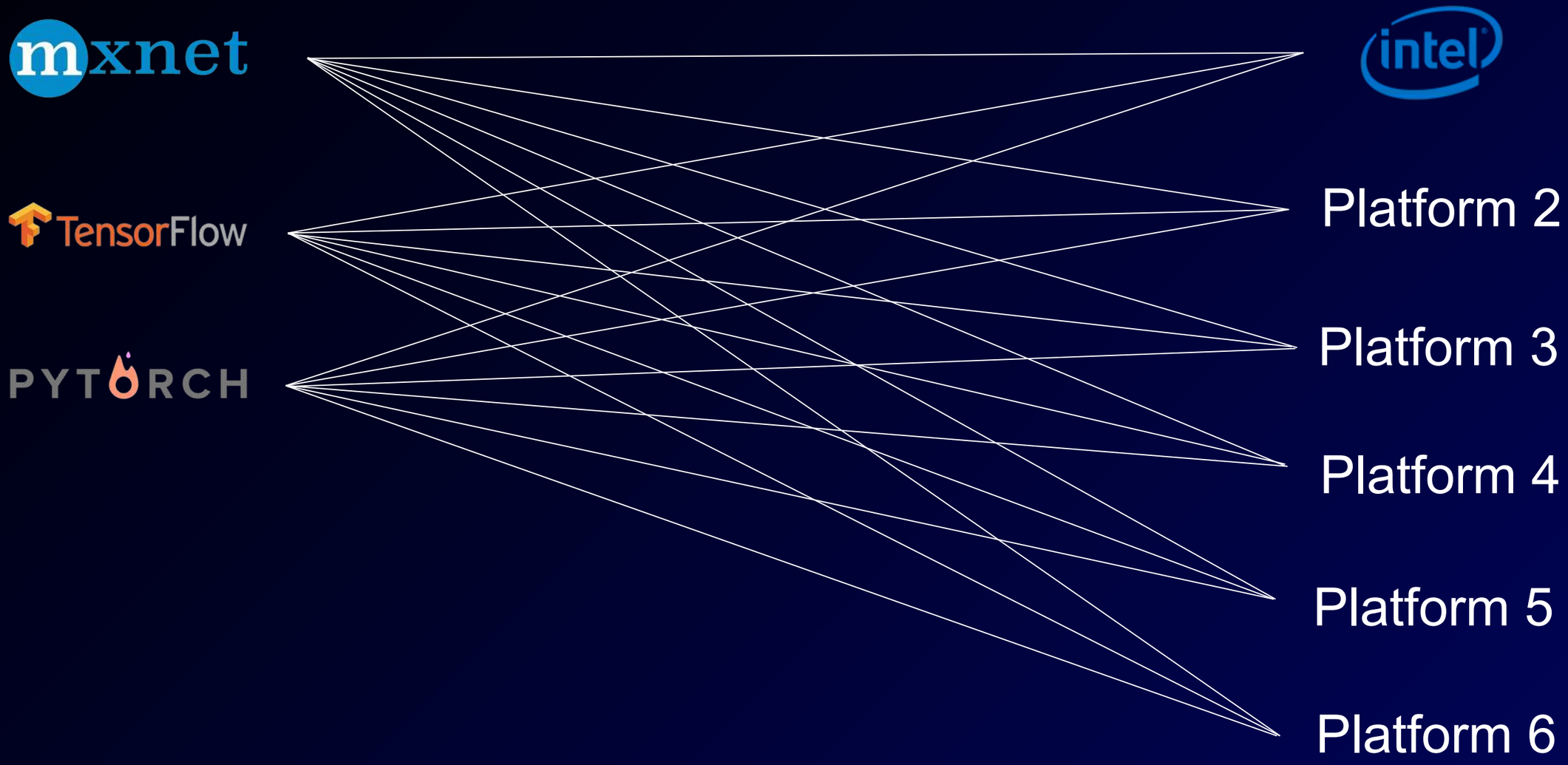
iVision Engineer the future, today.®

cogito

aws

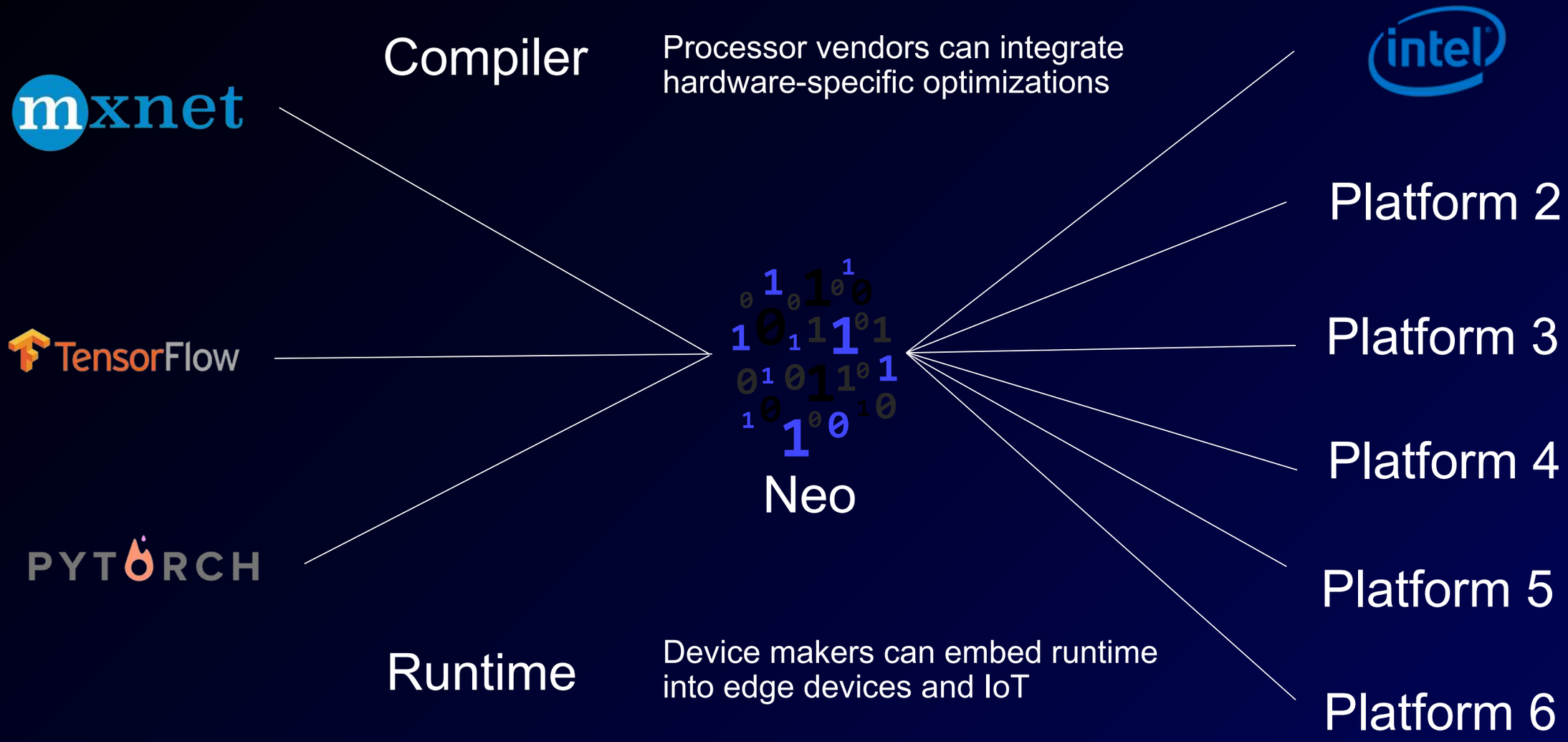
Driving Ease of Use

Model optimization is extremely complex



aws

Train once, run anywhere



Amazon SageMaker Neo

https://aws.amazon.com/blogs/aws/amazon-sagemaker-neo-train-your-machine-learning-models-once-run-them-anywhere/

Get accuracy and up to 2x performance

Automatic optimization

mxnet

TensorFlow

PYTORCH

Broad framework support

intel

and more

Broad hardware support

github.com/neo-ai

Apache Software License

aws

The Amazon ML stack: broadest & deepest set of capabilities

https://ml.aws

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

Machine Learning University

Uses the same materials used to train Amazon developers

Foundational knowledge with real-world application

Structured courses and specialist certification

https://aws.training/machinelearning

Thank you!

Julien Simon
Global Evangelist, AI & Machine Learning, Amazon Web Services

Email: julsimon@amazon.com
Twitter: @julsimon
https://medium.com/@julsimon