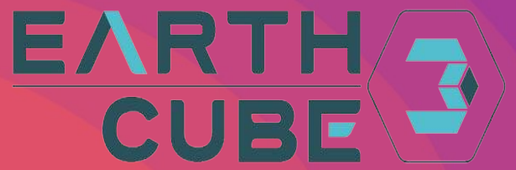
Deep Learning on AWS with TensorFlow and Apache MXNet

Julien Simon
Global Evangelist, AI & Machine Learning
@julsimon

Renaud ALLIOUX
CTO, Earthcube



aws SUMMIT

The Amazon ML Stack: Broadest & Deepest Set of Capabilities

AI SERVICES

Vision


REKOGNITION IMAGE


REKOGNITION VIDEO



TEXTRACT

Speech


POLLY


TRANSCRIBE

Language


TRANSLATE


COMPREHEND

COMPREHEND MEDICAL

Chatbots


LEX

Forecasting



FORECAST

Recommendations



PERSONALIZE

ML SERVICES


AMAZON SAGEMAKER

BUILD

Pre-built algorithms & notebooks

Data labeling (GROUND TRUTH)

Algorithms & models (AWS MARKETPLACE)

TRAIN

One-click model training & tuning

Optimization (NEO)

Models without training data (REINFORCEMENT LEARNING)

DEPLOY

One-click deployment & hosting

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

TensorFlow

mxnet

PYTORCH

Interfaces

GLUON

Keras

Infrastructure


EC2 P3 & P3dn


EC2 C5


FPGAs


GREENGRASS


ELASTIC INFERENCE

aws

The Amazon ML Stack: Broadest & Deepest Set of Capabilities

AI SERVICES

Vision

- REKOGNITION IMAGE
- REKOGNITION VIDEO
- TEXTRACT (NEW)

Speech

- POLLY
- TRANSCRIBE

Language

- TRANSLATE
- COMPREHEND
- COMPREHEND MEDICAL (NEW)

Chatbots

- LEX

Forecasting

- FORECAST (NEW)

Recommendations

- PERSONALIZE (NEW)

ML SERVICES

AMAZON SAGEMAKER

BUILD

Pre-built algorithms & notebooks (NEW)

Data labeling (GROUND TRUTH) (NEW)

Algorithms & models (AWS MARKETPLACE)

TRAIN

One-click model training & tuning

Optimization (NEO) (NEW)

Models without training data (REINFORCEMENT LEARNING)

DEPLOY

One-click deployment & hosting (NEW)

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

- TensorFlow
- mxnet
- PYTORCH

Interfaces

- GLUON
- Keras

Infrastructure

- EC2 P3 & P3dn
- EC2 C5
- FPGAs
- GREENGRASS
- ELASTIC INFERENCE (NEW)

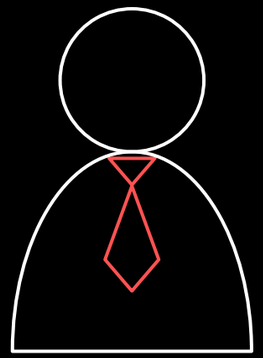
aws

AWS is framework agnostic

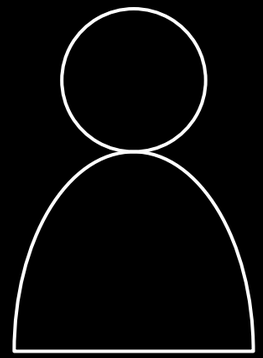
Choose from popular frameworks

TensorFlow mxnet PYTORCH Chainer Caffe2 ONNX Keras GLUON

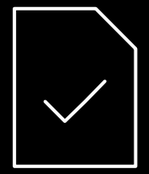
Run them fully managed



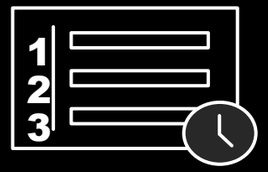
Or run them yourself



Amazon SageMaker:
Build, Train, and Deploy ML Models at Scale

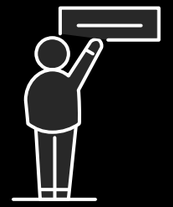


Collect and prepare training data



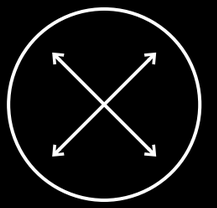
Choose and optimize your ML algorithm

Set up and manage environments for training



Train and Tune ML Models

Deploy models in production



Scale and manage the production environment

intuit

F1

NFL

tinder

KIA

Liberty Mutual.

CONVOY

SIEMENS

DOW JONES

Celgene

SONY

GE Healthcare

aws SUMMIT

AWS Deep Learning AMIs

Preconfigured environments Deep Learning applications

NEW (27/3)
Deep Learning
containers

Conda AMI
For developers who want pre-installed pip packages of DL frameworks in separate virtual environments.

Base AMI
For developers who want a clean slate to set up private DL engine repositories or custom builds of DL engines.

AMI with source code
For developers who want preinstalled DL frameworks and their source code in a shared Python environment.

aws SUMMIT

TensorFlow

aws SUMMIT

TensorFlow

- Open source software library for Machine Learning
- Main API in Python, experimental support for other languages
- Built-in support for many network architectures: FC, CNN, LSTM, etc.
- Support for symbolic execution, as well as imperative execution since v1.7
  (aka “eager execution”)
- Complemented by the Keras high-level API

AWS: The platform of choice to run TensorFlow

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

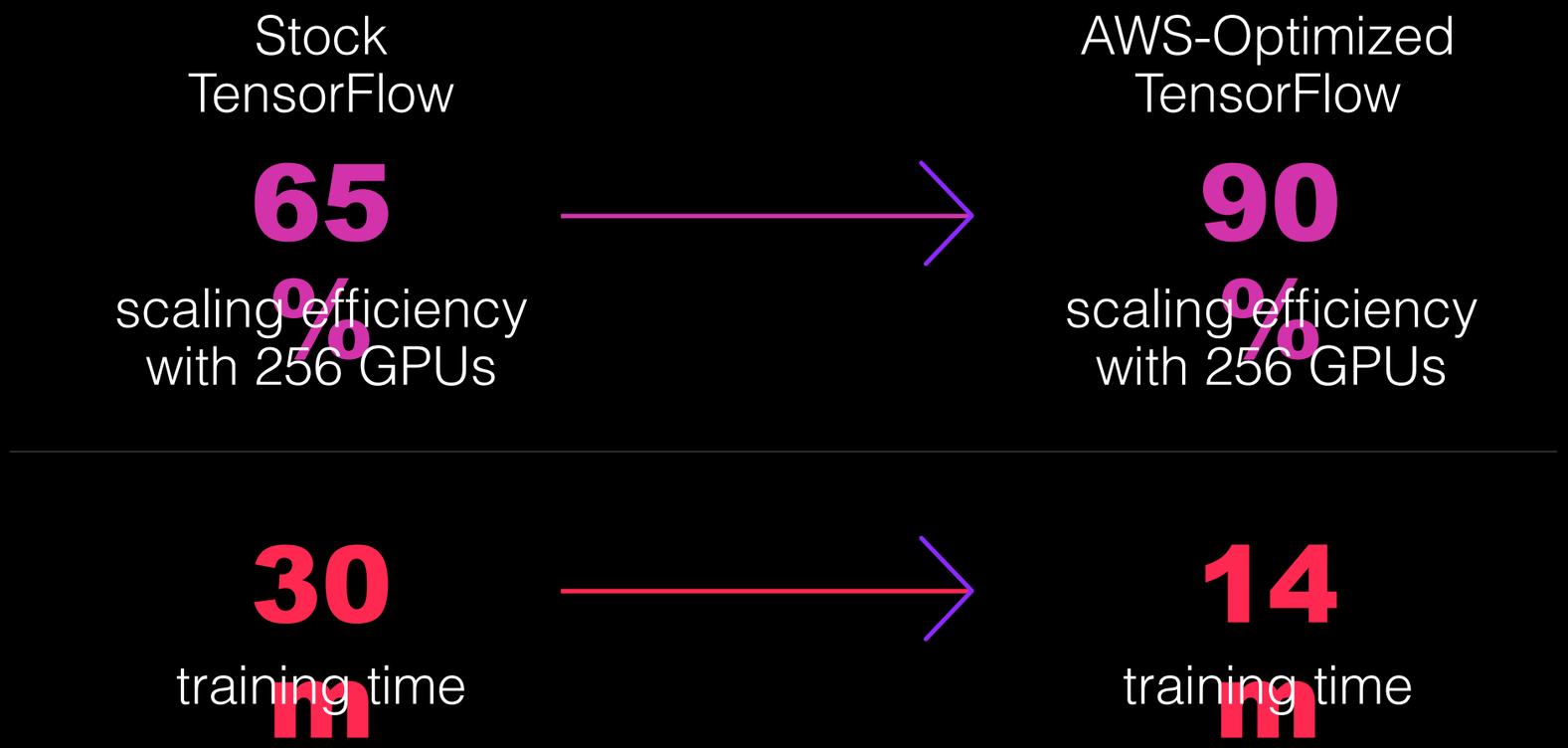
Source: Nucleus Research, November 2018

Optimizing TensorFlow for Amazon EC2 C5 instances

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

Optimizing TensorFlow for Amazon EC2 P3 instances



Apache MXNet

- Open source software library for Deep Learning
- Natively implemented in C++
- Built-in support for many network architectures: FC, CNN, LSTM, etc.
- Symbolic API: Python, Scala, Clojure, R, Julia, Perl, Java (inference only)
- Imperative API: Gluon (Python), with toolkits for computer vision (Gluon CV) and natural language processing (Gluon NLP)

Apache MXNet: deep learning for enterprise developers

Start with off-the-shelf models

- Gluon CV and Gluon NLP
- ONNX compatibility

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
- Model Server

Analyzing satellite images at scale with Tensorflow on AWS

Renaud ALLIOUX
CTO, Earthcube

aws SUMMIT

“Over 95% of collected intelligence data is never looked at”

A senior French MoD official
BFM Business, October 5th, 2018

What we do: AI-enabled GEOINT services

Why Deep Learning with Tensorflow on AWS

Deep Learning: no other computer vision technology allows such performance

Tensorflow: very flexible, especially when using Keras

AWS: scalability of storage and compute

Extensive R&D on Deep Learning models

- Main use: segmentation and object detection
- Custom architectures implemented with Keras
- Ensembling, wide networks (ResNext), capsule and Bayesian Neural networks
- Residual and spatial pyramid pooling layers
- Custom weighted loss functions (eg: weighted cross entropy)

How Earthcube builds on AWS

Data set

- 1.5 million labelled objects
- 3rd party platform: Ingedata.net (hosted on AWS)

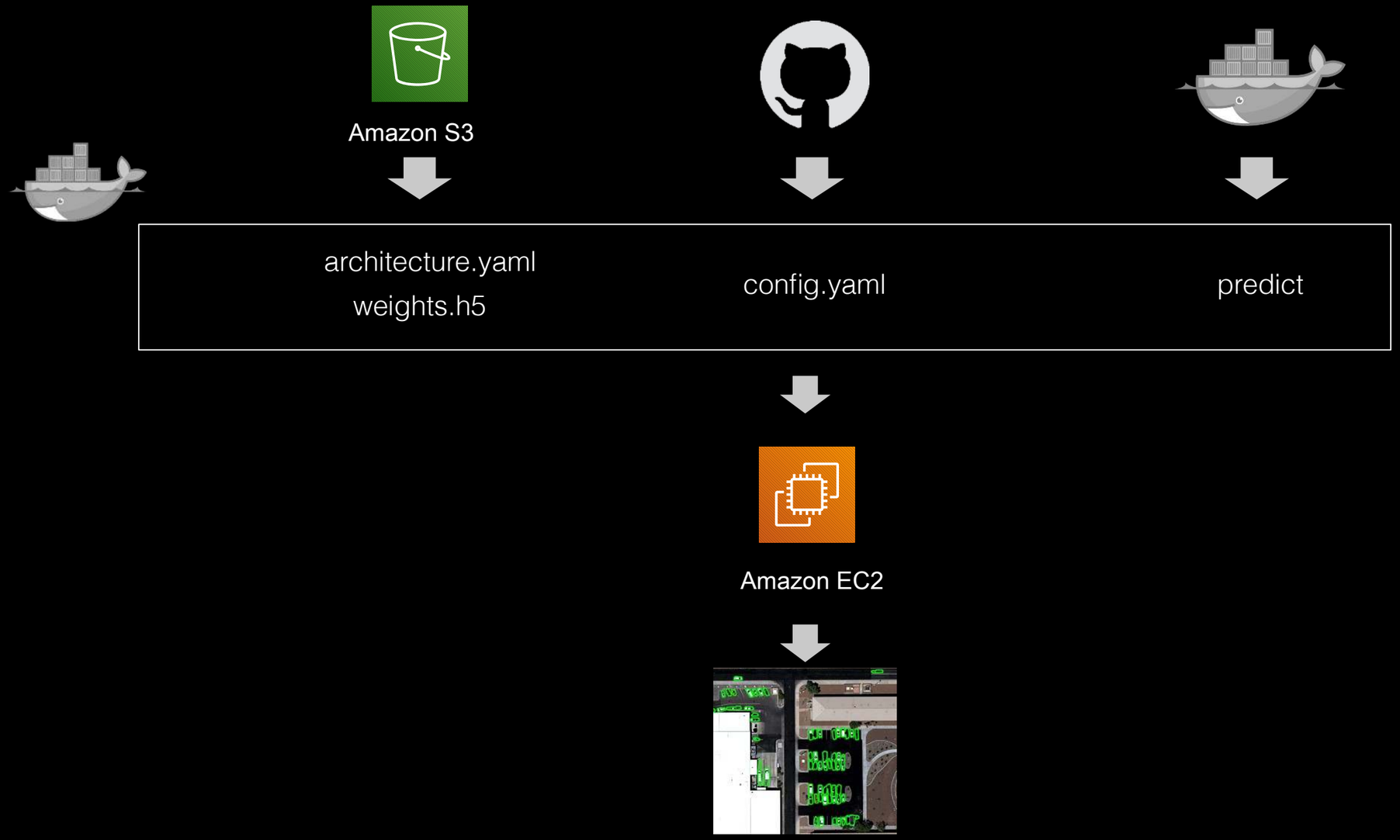
Training and inference

- Training on up to 1,000 images : 4 billion pixels each
- AMI based on the AWS Deep Learning AMI
- Amazon EC2 GPU instances

Deployment

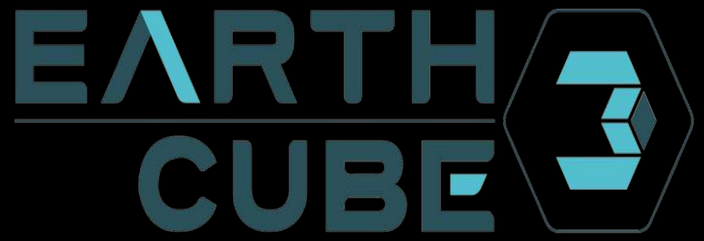
- Docker containers on Amazon EC2 (or on-premise)
- Celery

Cloud native deployment using Celery



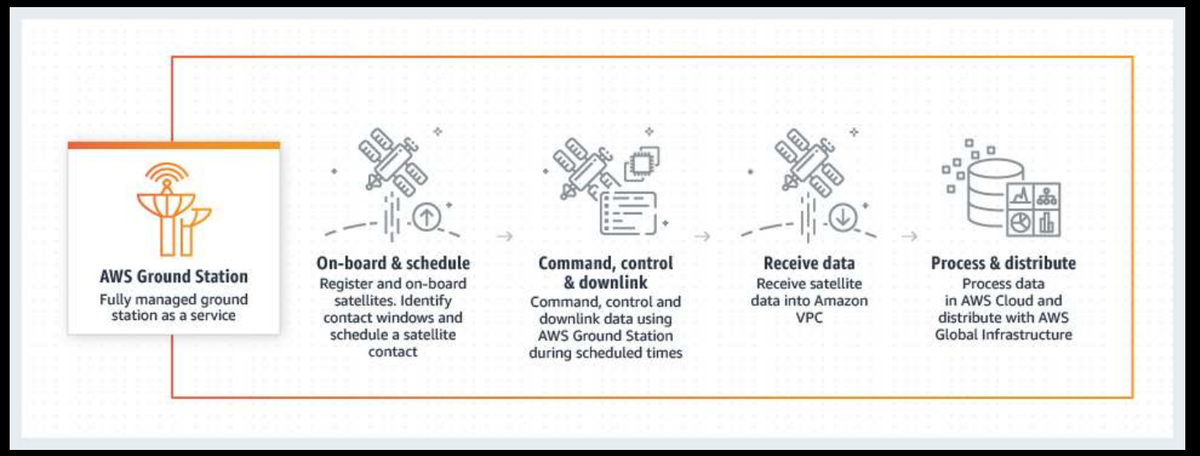
Lessons learned

- More data beats the most clever algorithm
- More data requires more infrastructure
- Without scalable infrastructure, there is no AI
  - Labeling, storing and versioning datasets and models
  - Training and prediction at scale
  - Automating deployments
- AWS is the key that unlocked AI for Earthcube



The GEOINT community is already using AWS

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



https://aws.amazon.com/ground-station

Next steps

Experiment with Amazon SageMaker to abstract training infrastructure

Try the Apache MXNet backend with Keras (instead of Tensorflow), as it's often twice as fast:
we'd love to train for 7 days instead of 15

TensorFlow and Apache MXNet on Amazon SageMaker

aws SUMMIT

TensorFlow on Amazon SageMaker: a first-class citizen

- Built-in containers for training and prediction.
  - Code available on Github: https://github.com/aws/sagemaker-tensorflow-containers
  - Build it, run it on your own machine, customize it, push it to Amazon ECR, etc.
  - Supported versions: 1.4.1, 1.5.0, 1.6.0, 1.7.0, 1.8.0, 1.9.0, 1.10.0, 1.11.0, 1.12.0
- Advanced features
  - Local mode: train on the notebook instance for faster experimentation
  - Script mode: use the same TensorFlow code as on your local machine (1.11.0 and up)
  - Distributed training: zero setup!
  - Pipe mode: stream large datasets directly from Amazon S3
  - TensorBoard: visualize the progress of your training jobs
  - Keras support (tf.keras.* and keras.*)

Apache MXNet on Amazon SageMaker: a first-class citizen

- Built-in containers for training and prediction.
  - Code available on Github: https://github.com/aws/sagemaker-mxnet-container
  - Build it, run it on your own machine, customize it, push it to Amazon ECR, etc.
  - Supported versions: 0.12.1, 1.0.0, 1.1.0, 1.2.1, 1.3.0

- Advanced features
  - Local mode: train on the notebook instance for faster experimentation
  - Script mode: use the same TensorFlow as on your local machine
  - Distributed training: zero setup!
  - Pipe mode: stream large datasets directly from Amazon S3
  - Keras support (tf.keras.* and keras.*)

Demo: Keras with TF and MXNet

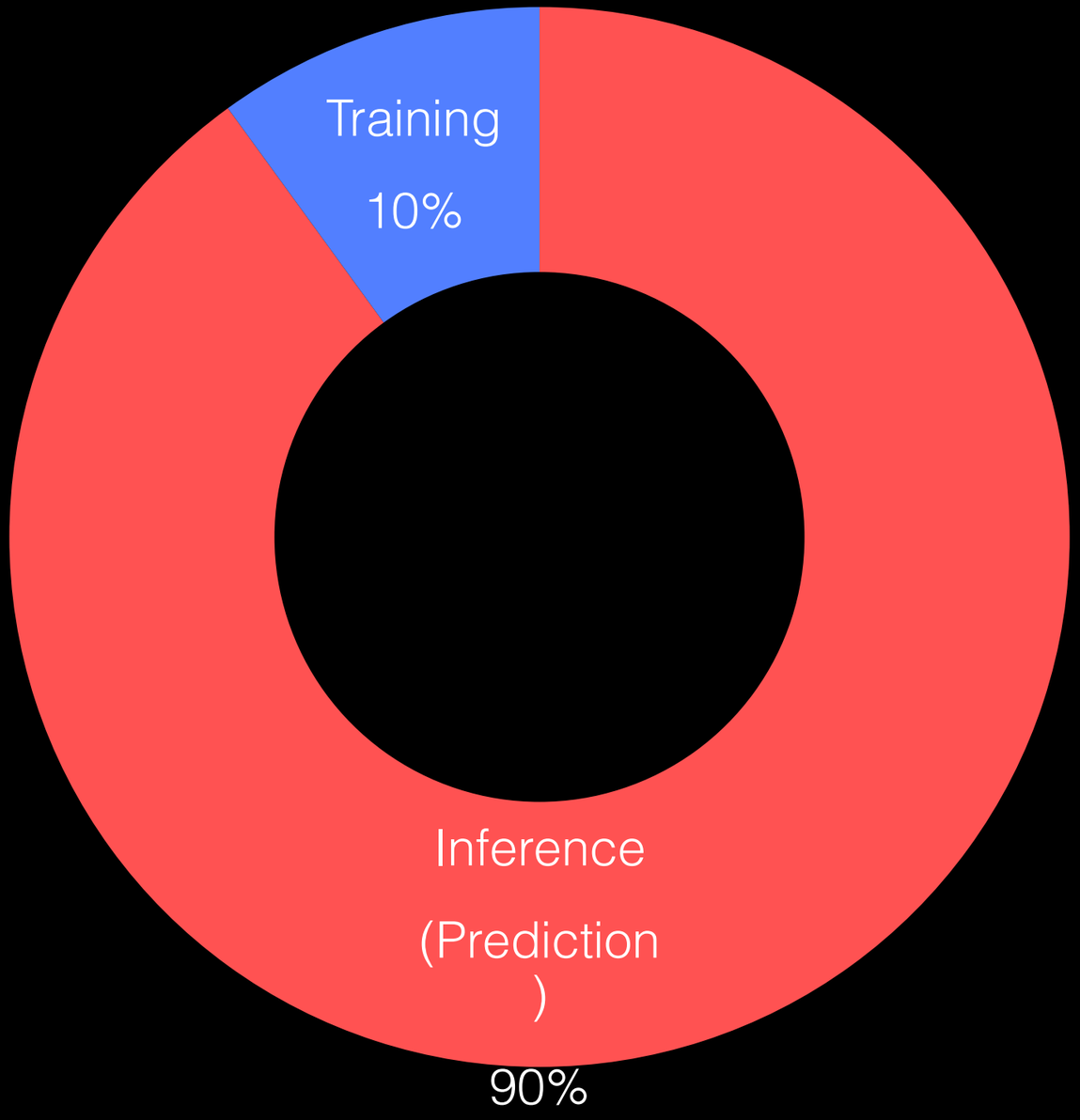
Demo: GluonCV

aws SUMMIT

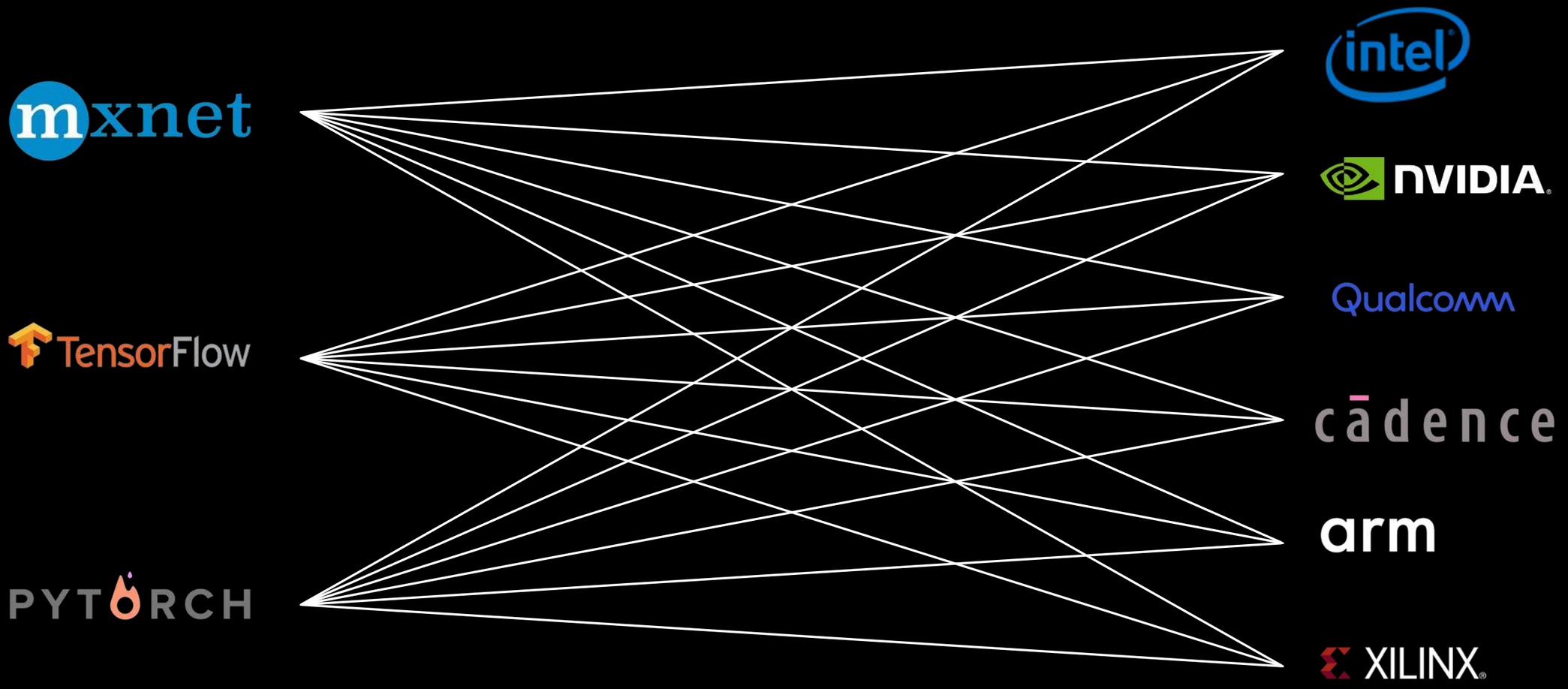
Optimizing prediction with Amazon SageMaker Neo and Amazon Elastic Inference

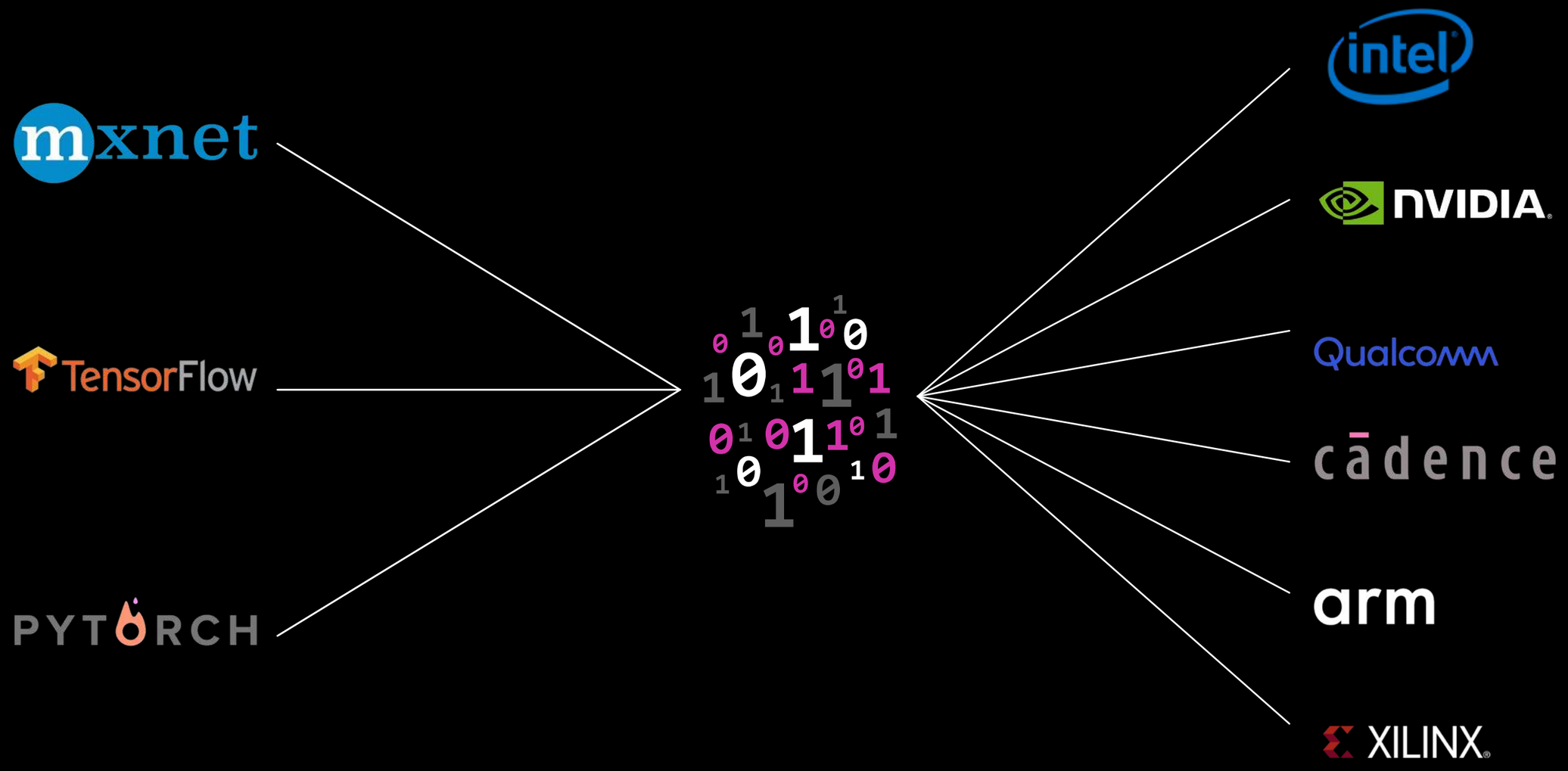
aws SUMMIT

Predictions drive complexity and cost in production



Hardware optimization is extremely complex





Amazon SageMaker Neo

Train once, run anywhere with 2x the performance

Get accuracy
and performance

Automatic
optimization

mxnet
TensorFlow
PYTORCH

Broad framework
support

Qualcomm
intel cadence
NVIDIA
XILINX arm

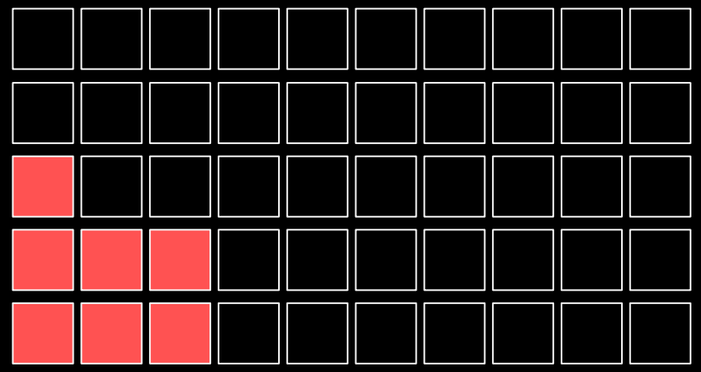
Broad hardware
support

KEY FEATURES

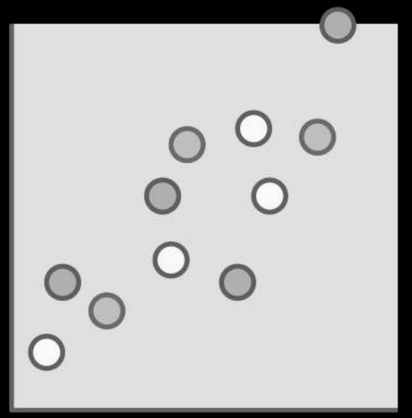
Open-source Neo-AI runtime and compiler
under the Apache software license;
1/10th the size of original frameworks

github.com/neo-ai

Are you making the most of your infrastructure?



Low utilization and high costs

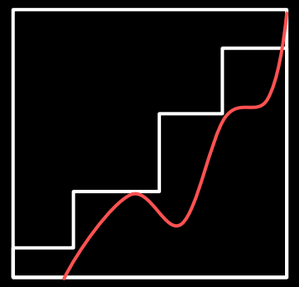


One size does not fit all

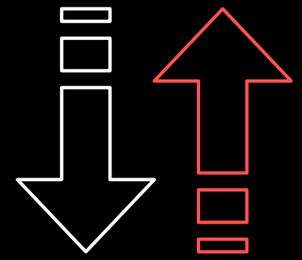
Amazon Elastic Inference

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
and Apache MXNet

Single and
mixed-precision
operations

Getting started

http://aws.amazon.com/free

https://ml.aws
https://aws.amazon.com/sagemaker

https://github.com/aws/sagemaker-python-sdk
https://github.com/awslabs/amazon-sagemaker-examples

https://medium.com/@julsimon
https://gitlab.com/juliensimon/dlnotebooks

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Julien Simon
Global Evangelist, AI and Machine Learning

@julsimon

aws SUMMIT

Please complete the session survey.

aws SUMMIT