AIM3

Build, Train, and Deploy Machine Learning Models at Any Scale | Messaging Block

Shyam Srinivasan

aws SUMMIT

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) NEW

DEPLOY

One-click deployment & hosting

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

TensorFlow

mxnet

PYTORCH

intel RL Coach

Interfaces

GLUON

Keras

Infrastructure


EC2 P3 & P3dn


EC2 C5


FPGAs


GREENGRASS

OpenVINO

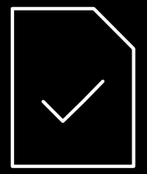
intel inside


ELASTIC INFERENCE NEW

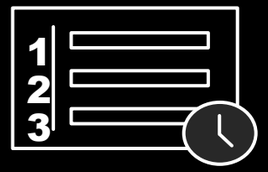
aws

intel XEON PLATINUM inside

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

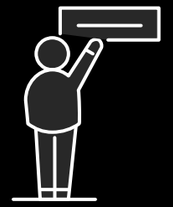


Collect and prepare training data



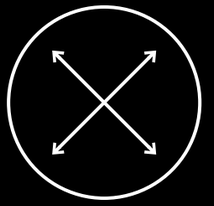
Choose and optimize your ML algorithm

Set up and manage environments for training



Train and Tune ML Models

Deploy models in production

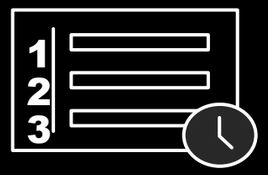


Scale and manage the production environment

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

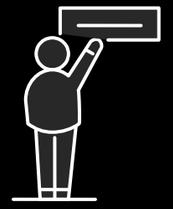
Pre-built notebooks for common problems

Collect and prepare training data



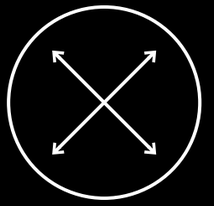
Choose and optimize your ML algorithm

Set up and manage environments for training



Train and Tune ML Models

Deploy models in production



Scale and manage the production environment

aws SUMMIT

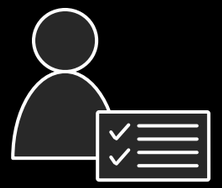
Amazon SageMaker:
Build, Train, and Deploy ML Models at Scale

Pre-built notebooks for common problems

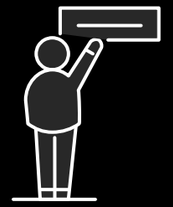
Collect and prepare training data

Built-in, high performance algorithms

Choose and optimize your ML algorithm

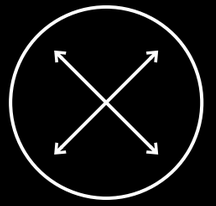


Set up and manage environments for training



Train and Tune ML Models

Deploy models in production



Scale and manage the production environment

aws SUMMIT

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

Pre-built notebooks for common problems

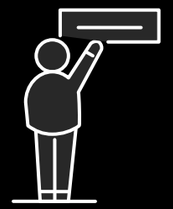
Collect and prepare training data

Built-in, high performance algorithms

Choose and optimize your ML algorithm

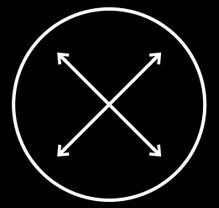
One-click training on the highest performing infrastructure

Set up and manage environments for training



Train and Tune ML Models

Deploy models in production



Scale and manage the production environment

intel RL Coach

intel XEON PLATINUM inside

aws SUMMIT

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

Pre-built notebooks for common problems

Collect and prepare training data

Built-in, high performance algorithms

Choose and optimize your ML algorithm

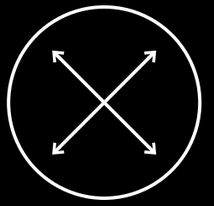
One-click training on the highest performing infrastructure

Set up and manage environments for training

Model Optimization

Train and Tune ML Models

Deploy models in production



Scale and manage the production environment

aws SUMMIT

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

Pre-built notebooks for common problems

Collect and prepare training data

Built-in, high performance algorithms

Choose and optimize your ML algorithm

One-click training on the highest performing infrastructure

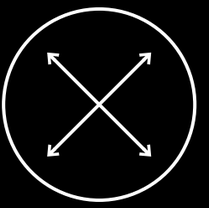
Set up and manage environments for training

Model Optimization

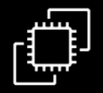
Train and Tune ML Models

One-click Deployment

Deploy models in production



Scale and manage the production environment



EC2 C5

GREENGRASS

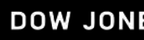
OpenVINO

aws SUMMIT

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

Pre-built notebooks for common problems	Built-in, high performance algorithms	One-click training on the highest performing infrastructure	Model Optimization	One-click Deployment	Fully managed with auto-scaling for 75% less
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



Reducing the cost of Sentiment Analysis

Zignal Labs performs nuanced sentiment analysis on billions of stories per month using AWS. Zignal Labs offers solutions that analyze the entire digital media landscape to deliver instant insights for the company's Fortune 1000 customers. The company built a sentiment-analysis pipeline that uses Amazon SageMaker for machine-learning capabilities and Amazon EC2 C5 instances with Intel Xeon Scalable (Skylake) processors for faster model training and evaluation.

90%.

Reduction of operations cost

"Using Amazon SageMaker and Amazon EC2 C5 instances, it's easier than ever for us to deliver high-quality sentiment analysis."

Jonathan Dodson, Vice President of Engineering, Zignal Labs

Over 150 algorithms and models

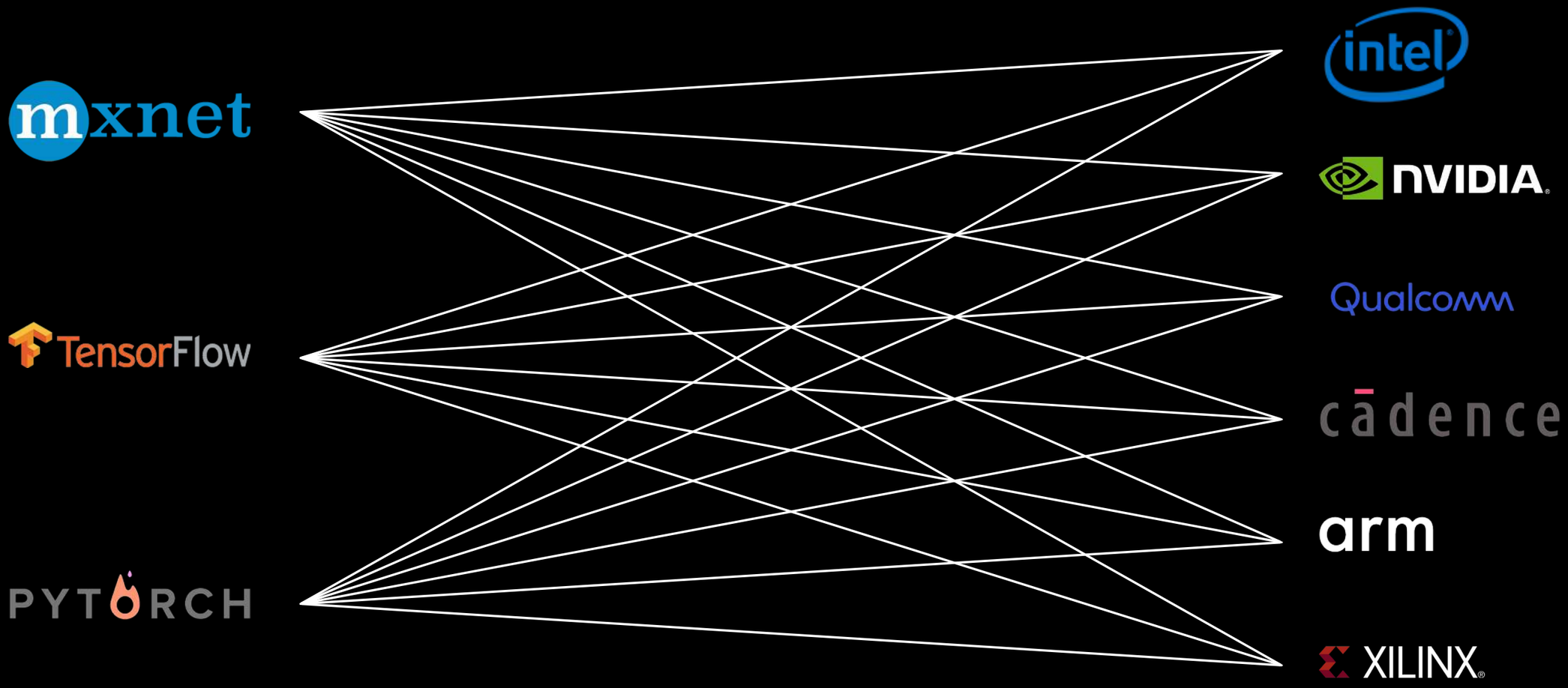
SELECTED VENDORS

AVAILABLE ALGORITHMS & MODELS

Natural Language Processing	Grammar & Parsing	Text OCR	Computer Vision	Named Entity Recognition	Video Classification
Speech Recognition	Text-to-Speech	Speaker Identification	Text Classification	3D Images	Anomaly Detection
Text Generation	Object Detection	Regression	Text Clustering	Handwriting Recognition	Ranking

aws SUMMIT

Optimization is extremely complex



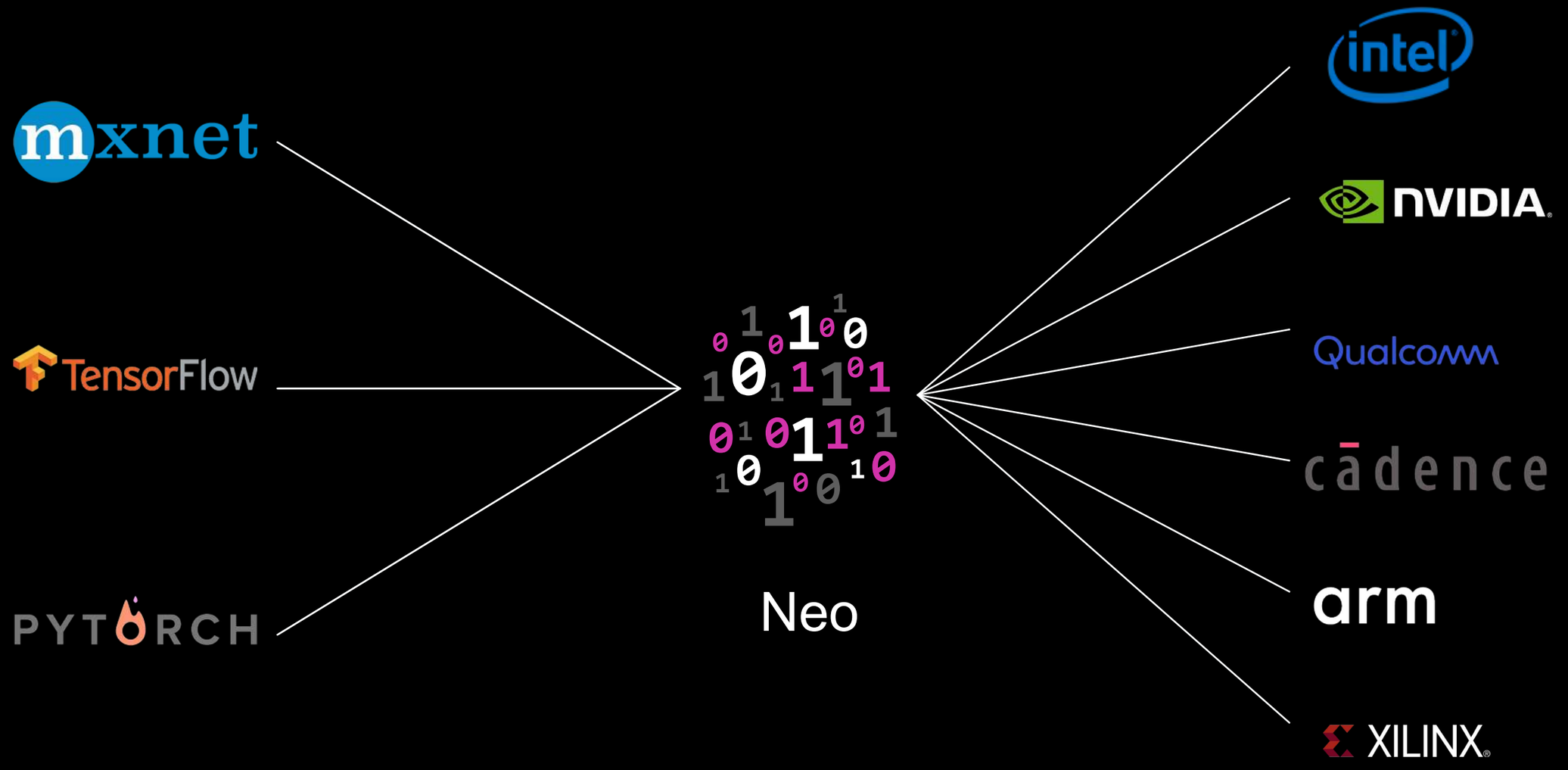
NEW

Amazon SageMaker Neo

Train once, run anywhere with 2x the performance

aws SUMMIT

Amazon SageMaker Neo: Train once, run anywhere



Amazon SageMaker Neo

Train once, run anywhere with 2x the performance

Get accuracy
and performance

Automatic
optimization

mxnet
TensorFlow
PYTORCH

Broad framework
support

intel
NVIDIA
Qualcomm
arm
cadence
XILINX

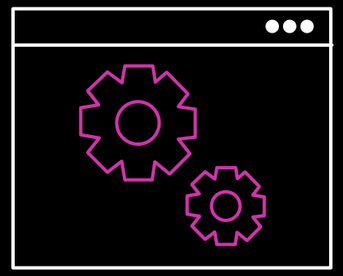
Broad hardware
support

KEY
FEATURES

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

Amazon SageMaker RL

Reinforcement learning for every developer and data scientist



Fully managed

mxnet
TensorFlow
intel riselab UC Berkeley

Broad support for frameworks

Broad support for simulation environments

KEY FEATURES

2D & 3D physics environments and OpenGym support

Support Amazon Sumerian, AWS RoboMaker and the open source Robotics Operating System (ROS) project

Example notebooks and tutorials

GE Healthcare

amazonrobotics

Tradelegs

SyntheticGestalt

mixi

AWS is the best platform for Apache MXNet

Start with high quality, pre-trained models

- Gluon CV and Gluon NLP

Refine with fast, scalable training

- Keras-MXNet up to 2x faster than Keras-TensorFlow
- Near-linear scalability up to 256 GPUs
- Optimized for training on C5 Instances Intel Xeon Scalable (Skylake) Processors
- Dynamic training

Deploy using familiar tools

- Java/Scala APIs
- MXNet Model Server

aws SUMMIT

intel XEON PLATINUM inside