Build Intelligent Applications with Machine Learning on AWS

Shyam Srinivasan
Sr. Product Marketing Manager, AWS

Machine Learning at Amazon: A long heritage

Personalized Recommendations

Fulfillment automation & Inventory Management

Drones

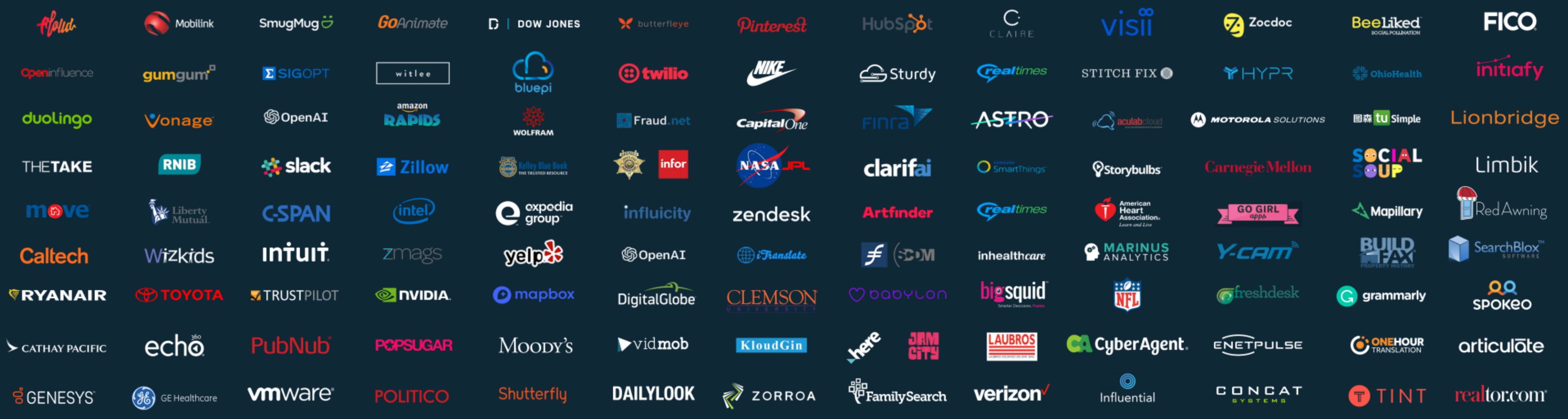
Voice driven Interactions

Inventing entirely new Customer experiences

ML @ AWS: OUR MISSION

Put machine learning in the hands of every developer and data scientist

Tens of thousands of customers running ML on AWS



aws

The Amazon Machine Learning Stack

APPLICATION SERVICES

- Rekognition
- Rekognition Video
- Polly
- Transcribe
- Translate
- Comprehend
- Lex

PLATFORMS

- Amazon SageMaker
- AWS DeepLens

FRAMEWORKS

- TensorFlow
- mxnet
- PYTORCH
- Caffe2
- Chainer
- HOROVOD
- GLUON
- Keras

aws

Complete Control over Frameworks & Infrastructure

For the data scientist, AI researcher, or advanced ML practitioner

FRAMEWORKS & INTERFACES

TensorFlow mxnet PYTORCH Caffe2 Chainer HOROVOD GLUON K Keras

INFRASTRUCTURE (GPU)

P3

NVIDIA
Tesla V100 GPUs
(14x faster than P2)

5,120 Tensor cores	**1 Petaflop of compute**
128GB of memory	**NVLink 2.0**

Machine Learning AMIs

INFRASTRUCTURE (CPU)

C5

Intel Xeon
3.0 GHz Skylake CPU
(25% better perf/price than C4)

intel XEON PLATINUM inside

72 vCPUs	**AVX 512**
144 GB of memory	**Nitro Hypervisor**

Machine Learning AMIs

aws

AWS Deep Learning AMI

Get started quickly with easy-to-launch tutorials

Hassle-free setup and configuration

Pay only for what you use – no additional charge for the AMI

Accelerate your model training and deployment

Support for popular deep learning frameworks

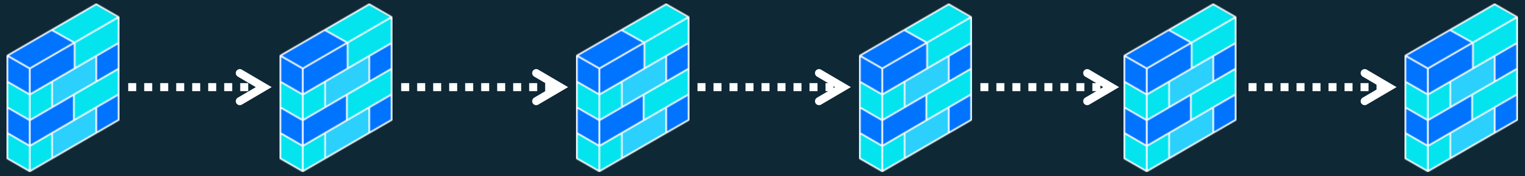
TensorFlow mxnet PYTORCH Caffe2 Chainer

HOROVOD GLUON Keras

zendesk

MATRIX ANALYTICS

The Machine Learning Process



Collect and prepare training data

Choose and optimize your ML algorithm

Set up and manage environments for training

Train and tune model (trial and error)

Deploy model in production

Scale and manage the production environment

Amazon SageMaker

Pre-built notebooks for common problems

Choose and optimize your ML algorithm

Set up and manage environments for training

Train and tune model (trial and error)

Deploy model in production

Scale and manage the production environment

BUILD

Amazon SageMaker

Pre-built notebooks for common problems

Built-in, high performance algorithms

ALGORITHMS

- K-Means Clustering
- Principal Component Analysis
- Neural Topic Modelling
- Factorization Machines
- Linear Learner – Regression
- DeepAR
- XGBoost
- Latent Dirichlet Allocation
- Image Classification (ResNet)
- Sequence2Sequence
- Linear Learner – Classification
- BlazingText Word2Vec

FRAMEWORKS

Apache MXNet
TensorFlow

Caffe2, CNTK, PyTorch, Torch, Chainer

BUILD

Set up and manage environments for training

Train and tune model (trial and error)

Deploy model in production

Scale and manage the production environment

Designing Better Algorithms

- Algorithms for “infinite scale”
- Distributed by default
- Train on a data stream
- Single pass training
- Not memory bound
- Checkpoint for re-training

Amazon SageMaker

Pre-built notebooks for common problems

Built-in, high performance algorithms

One-click training

Train and tune model (trial and error)

Deploy model in production

Scale and manage the production environment

BUILD

TRAIN

Amazon SageMaker

Pre-built notebooks for common problems

Built-in, high performance algorithms

BUILD

One-click training

Automatic Model Tuning

TRAIN

Deploy model in production

Scale and manage the production environment

Amazon SageMaker:
Custom ML Models Using Your Own Data

Train with
your own
algorithms

Train with
a framework
of your choice

Train with
SageMaker
algorithms

Optimize
your models

Host models

A/B test

“How can we make our algorithms as fast as those in SageMaker?”

NEW

SageMaker Streaming For Custom Algorithms

ACCELERATE YOUR OWN ALGORITHMS BY STREAMING LARGE VOLUMES OF TRAINING DATA FROM AMAZON S3

Stream data to your own algorithm

Quicker time to start training

Faster training

Lower cost training

TensorFlow

Additional frameworks coming soon

Amazon SageMaker

BUILD

Pre-built notebooks for common problems

Built-in, high performance algorithms

TRAIN

One-click training

Automatic Model Tuning

DEPLOY

One-click deployment

Scale and manage the production environment

Amazon SageMaker

Pre-built notebooks for common problems

Built-in, high performance algorithms

BUILD

One-click training

Automatic Model Tuning

TRAIN

One-click deployment

Fully managed hosting with auto-scaling

DEPLOY

Amazon SageMaker:
Custom ML Models Using Your Own Data

Train with
your own
algorithms

Train with
a framework
of your choice

Train with
SageMaker
algorithms

Optimize
your models

Host models

A/B test

Amazon SageMaker:
Elastic Hosting For Custom Models

Easy to integrate API endpoint	Low latency
Auto-scaling	Fault tolerant, multi-AZ

But…

What if you need to batch process?

What if your files are big?

NEW

Amazon SageMaker Batch Transform

RUN FULLY MANAGED, HIGH-THROUGHPUT BATCH TRANSFORM JOBS WITH A SIMPLE API CALL

Process data dumps in a batch

Process large files more easily

Batch test models before hosting

Test models before deployment at the edge

Use same model for batch and real-time

Reuse pre-processing pipelines between training and prediction

Fully managed batch transforms

Amazon SageMaker Demo

Build, Train and Deploy Machine Learning Models Quickly & Easily

The Amazon Machine Learning Stack

APPLICATION SERVICES

Rekognition | Rekognition Video | Polly | Transcribe | Translate | Comprehend | Lex

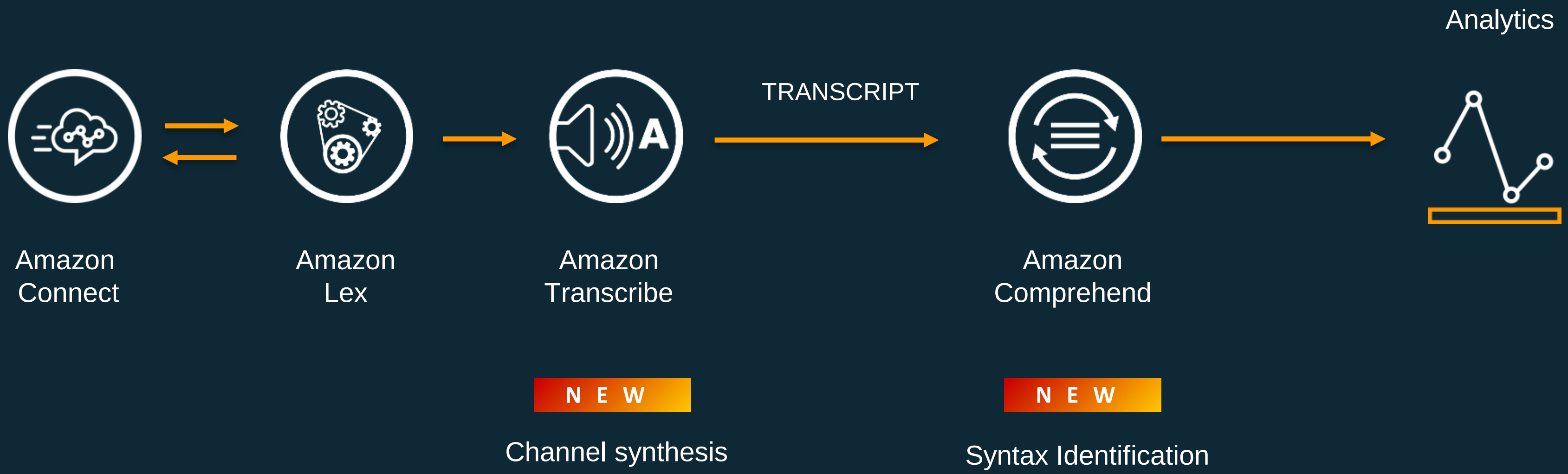
PLATFORMS

Amazon SageMaker | AWS DeepLens

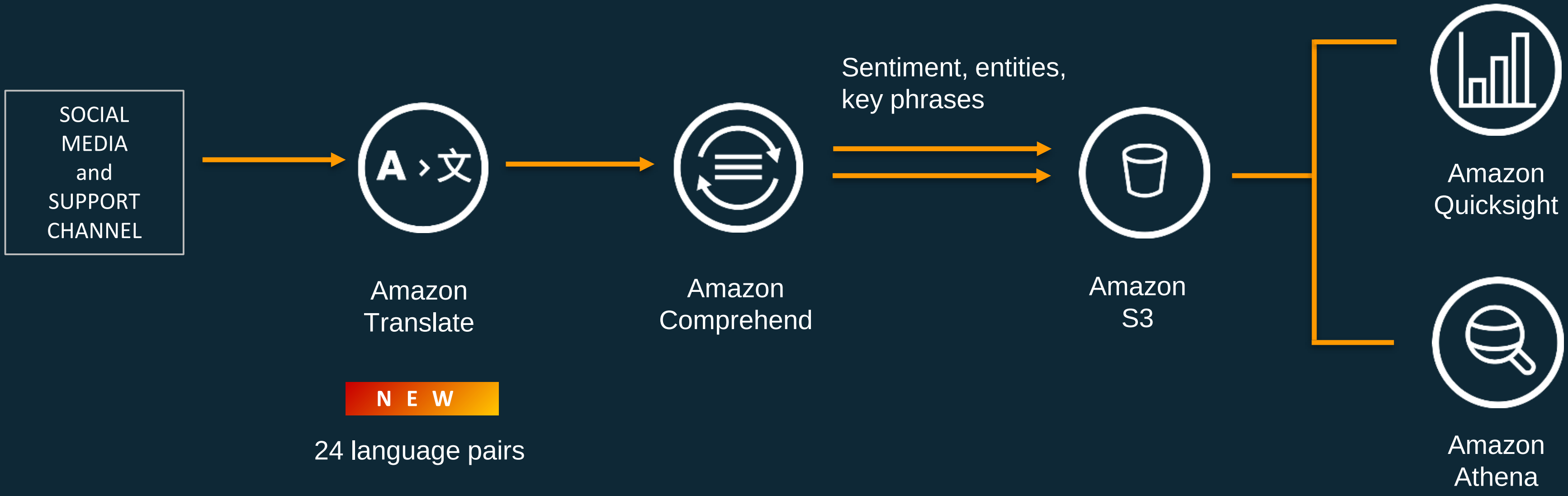
FRAMEWORKS

TensorFlow | mxnet | PYTORCH | Caffe2 | Chainer | HOROVOD | GLUON | Keras

Improving Contact Centers With Artificial Intelligence



Improving Voice Of Customer Analytics With Artificial Intelligence



SOCIAL MEDIA and SUPPORT CHANNEL

Amazon Translate

NEW

24 language pairs

Amazon Comprehend

Sentiment, entities, key phrases

Amazon S3

Amazon Quicksight

Amazon Athena

The Amazon Machine Learning Stack

APPLICATION SERVICES

Rekognition

Rekognition Video

Polly

Transcribe

Translate

Comprehend

Lex

PLATFORMS

Amazon SageMaker

AWS DeepLens

FRAMEWORKS

TensorFlow

mxnet

PYTORCH

Caffe2

Chainer

HOROVOD

GLUON

Keras

Thank You!

ml.aws