
aws
Using AWS Cloud for ML
Neil Mackin, AWS
nemackin@amazon.com


Our mission at AWS
Put machine learning in the hands
of every developer
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark
aws

Our unique approach

Customer-focused
90%+ of our ML roadmap is defined by customers

Pace of innovation
200+ new ML launches in the last year

Breadth & depth
More AI and ML services in production than any other provider

Multi-framework
Support for the most popular frameworks

Security & analytics
Deepest set of security and encryption features, with robust analytics capabilities

Embedded R&D
Customer-centric approach to advancing the state of the art

More machine learning happens on AWS than anywhere else

10,000+
customers have used machine learning on AWS

81% of deep learning in the cloud runs on AWS

85% of TensorFlow projects in the cloud run on AWS

AWS holds the top spots on Stanford's deep learning benchmark, DAWN, for fastest training time, lowest cost, lowest inference latency

Technology | Culture

Technology | Bringing AI into your digital transformation requires a new "stack" that makes it easier to put ML to work

The Amazon ML stack: Broadest & deepest set of capabilities

AI SERVICES

Easily add intelligence to applications without machine learning skills

Vision | Documents | Speech | Language | Chatbots | Forecasting | Recommendations

ML SERVICES

Build, train, and deploy machine learning models fast

Data labeling | Pre-built algorithms & notebooks | One-click training and deployment

ML FRAMEWORKS & INFRASTRUCTURE

Flexibility & choice, highest-performing infrastructure

Support for ML frameworks | Compute options purpose-built for ML

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

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks: TensorFlow, mxnet, PYTORCH

Interfaces: GLUON, Keras

Infrastructure: EC2 P3 & P3N, EC2 C5, FPGAs, GREENGRASS, ELASTIC INFERENCE

Amazon Rekognition

Deep learning-based image and video analysis

Object, Scene & Activity Recognition

Facial Recognition

Facial Analysis

Person Tracking

Unsafe Content Detection

Celebrity Recognition

Text in Images

aws

Amazon Rekognition Images

Given this photo of Andy Jassy … Tell me if you find him in this image

Rekognition: Confidence level that source face is in target photo: 98 %

Amazon Rekognition Images

```
bucket = 'my-photos-bucket'
key_s = 'andy_jassy_headshot.png'
key_t = 'interview.png'

rek_client=boto3.client('rekognition', 'ap-southeast-2')

response = rek_client.compare_faces( SimilarityThreshold = 75,
   SourceImage={ 'S3Object': { 'Bucket': bucket, 'Name': key_s, } },  TargetImage={
'S3Object': { 'Bucket': bucket, 'Name': key_t, } },)

confidence = response['FaceMatches'][0]['Similarity']

print('Confidence level that source face is in target photo: ',confidence,'%')
```


Amazon Rekognition Images

```
bucket = 'my-photos-bucket'
key_s = 'andy_jassy_headshot.png'
key_t = 'interview.png'

rek_client=boto3.client('rekognition', 'ap-southeast-2')

response = rek_client.compare_faces( SimilarityThreshold = 75,
   SourceImage={ 'S3Object': { 'Bucket': bucket, 'Name': key_s, } },  TargetImage={
'S3Object': { 'Bucket': bucket, 'Name': key_t, } },)

confidence = response['FaceMatches'][0]['Similarity']

print('Confidence level that source face is in target photo: ',confidence,'%')
```

Build applications that understand text

APPLICATION SERVICES

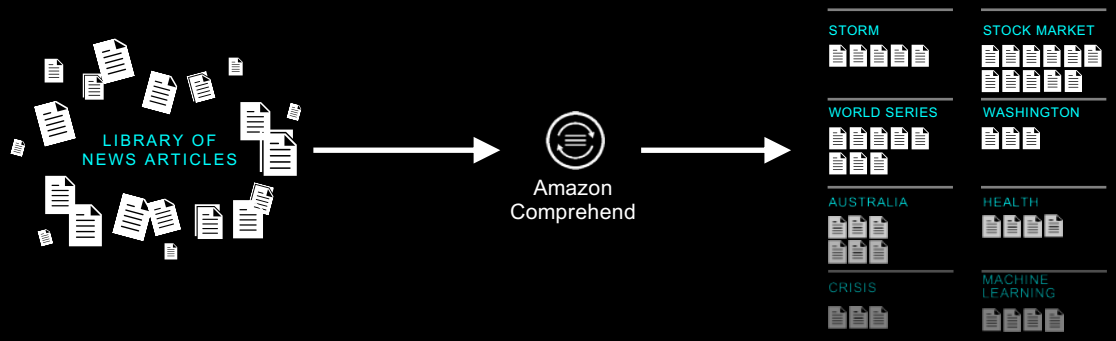
AMAZON REKOGNITION | AMAZON REKOGNITION VIDEO | AMAZON POLLY | AMAZON TRANSCRIBE | AMAZON TRANSLATE | **AMAZON COMPREHEND** | AMAZON LEX

Amazon Comprehend

- Natural language processing (NLP) service that finds insights and relationships in text
- Detect sentiment & language, model documents by topic
- Identify entities, key phrases, and syntax
- Financial services, healthcare, insurance, government, document classification & search, customer analytics ++

Amazon Comprehend

Discover insights and relationships in text



Amazon Comprehend – extract insights from text

Amazon.com, Inc. is located in Seattle, WA and was founded July 5th, 1994 by Jeff Bezos. Our customers *love buying everything* from books to blenders at great prices

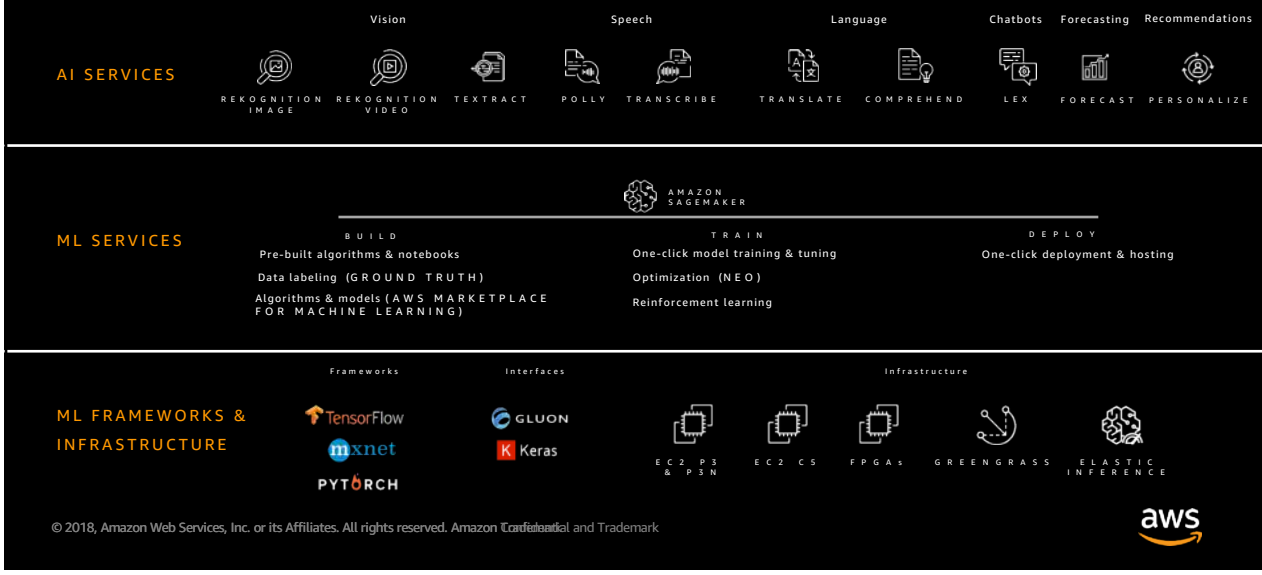
Named entities
- Amazon.com: Organization
- Seattle, WA: Location
- July 5th,1994: Date
- Jeff Bezos: Person

Key phrases
- Our customers
- books
- blenders
- great prices

Sentiment
Positive

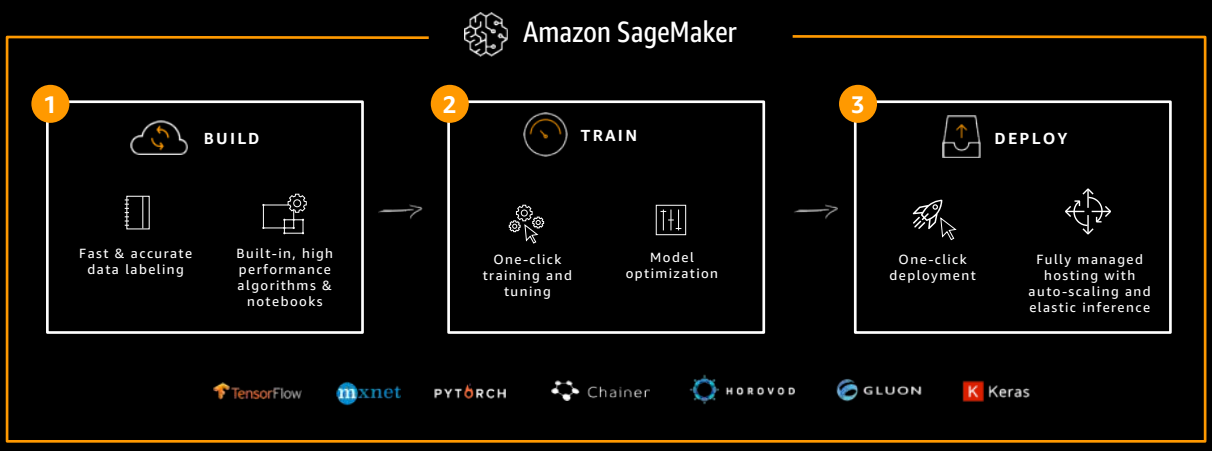
Language
English

The Amazon ML stack: Broadest & deepest set of capabilities



Free Data Scientists to do Data Science

Custom machine learning for your business



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

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

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

Pre-built notebooks for common problems

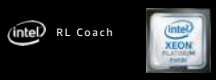
Collect and prepare training data

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

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

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	
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	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




DOW JONES

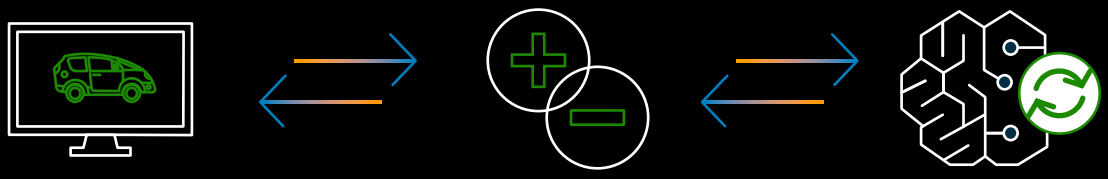
Amazon SageMaker Built-in Algorithms

- Image Classification
- Object Detection
- K-Nearest Neighbors (k-NN)
- Linear Learner
- Factorization Machines
- XGBoost
- Sequence2Sequence
- Principal Component Analysis (PCA)
- Latent Dirichlet Allocation (LDA)
- Neural Topic Model (NTM)
- DeepAR Forecasting
- Semantic Segmentation
- BlazingText
- Random Cut Forest

Three new ones:

- IP Insights: built-in algorithms for detecting suspicious IP addresses
- Object2Vec : Low dimensional embeddings for high dimensional objects
- K-means: clustering unsupervised grouping

How does RL work?



Simulation environment

Scoring function

RL algorithm

USE CASES

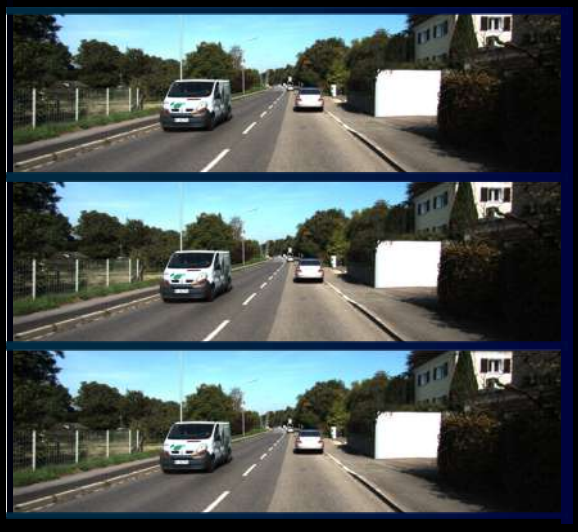
Supply chain simulation, manufacturing process, robot manipulation, autonomous car, drone navigation…

Amazon SageMaker: Build, train, and deploy ML

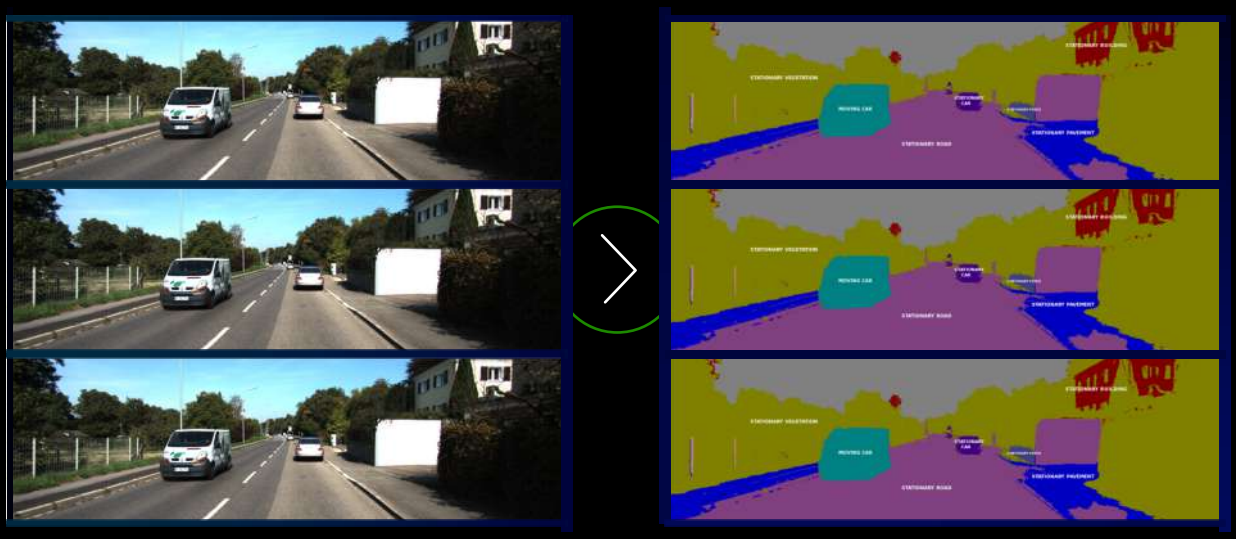
Collect and prepare training data

Pre-built notebooks for common problems	Built-in, high performance algorithms	One-click training	Optimization	One-click deployment	Fully managed with auto-scaling
Choose your ML algorithm	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

Successful models require high-quality data



Successful models require high-quality data



Amazon SageMaker ground truth
Label machine learning training data easily and accurately

Quickly label training data

Easily integrate human labelers

Get accurate results

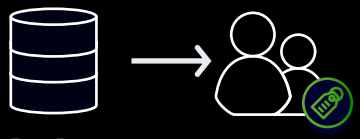
KEY FEATURES

- Automatic labeling via machine learning
- Ready-made and custom workflows for image bounding box, segmentation, and text
- Private and public human workforce
- Label management

How it works

Raw Data

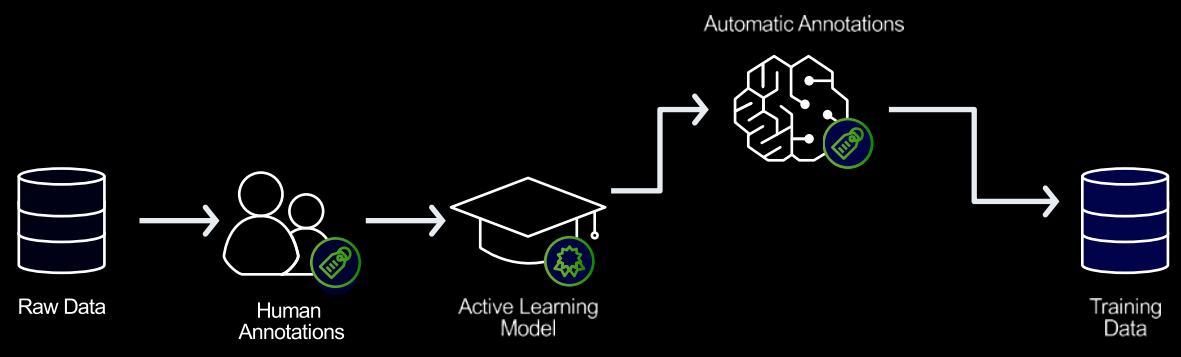
How it works



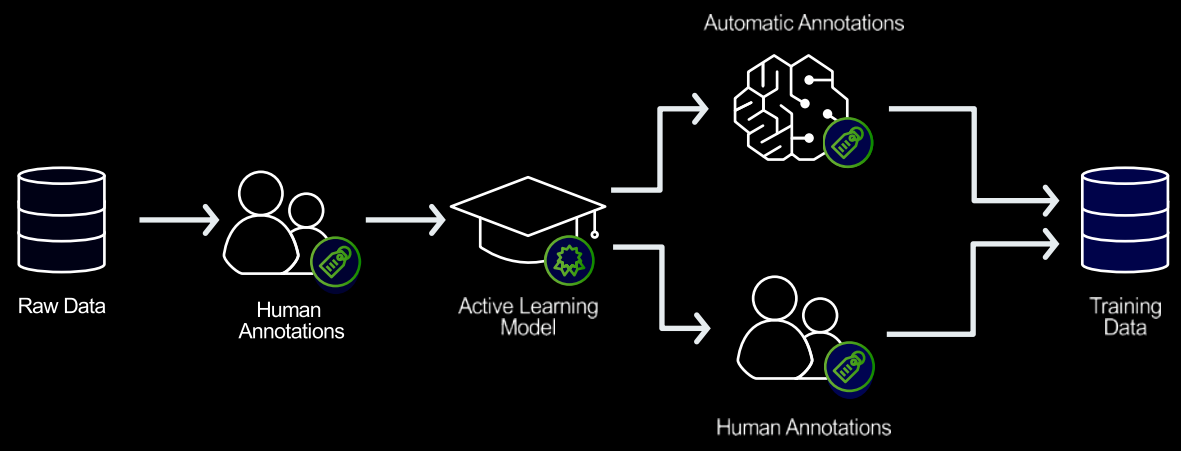
Raw Data

Human Annotations

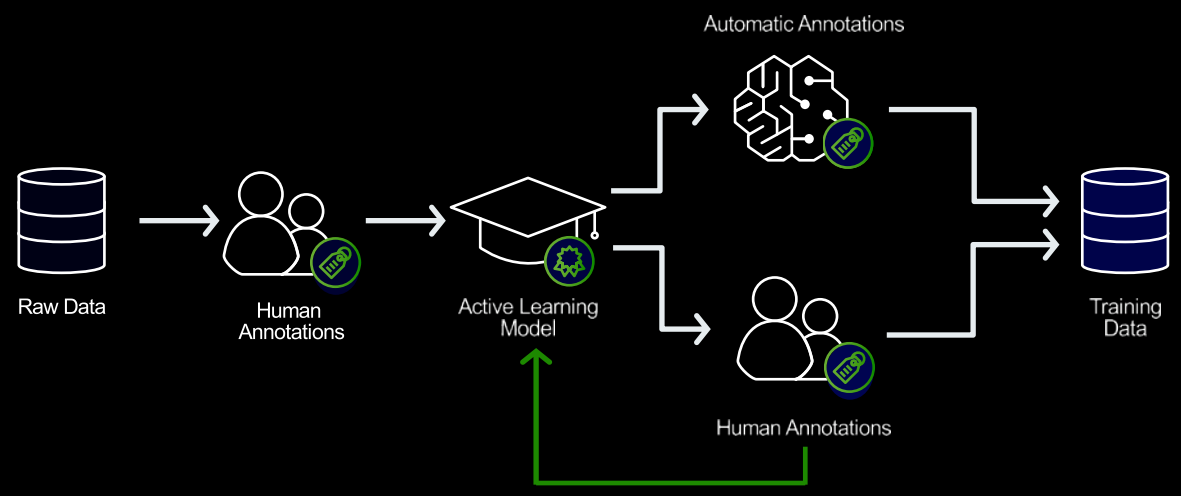
How it works



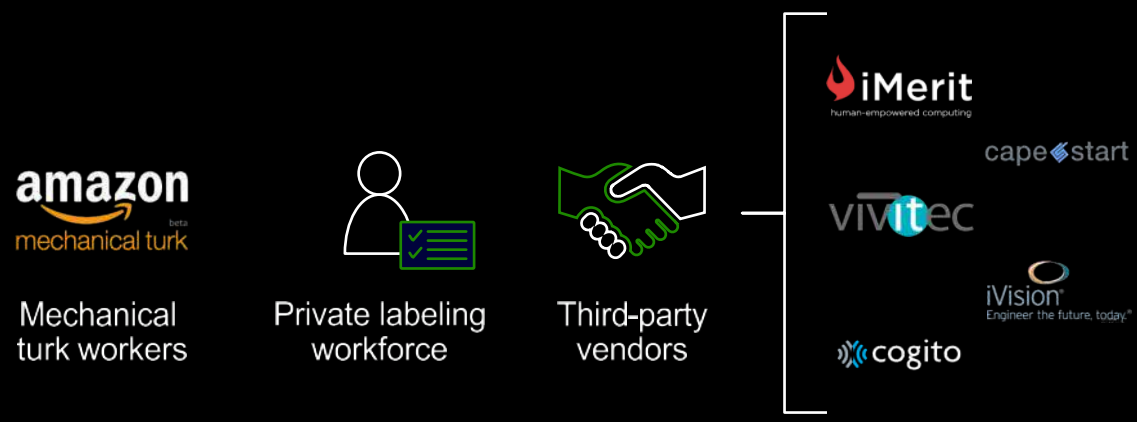
How it works



How it works



Creating training data



Real-time Fraud Detection with Amazon SageMaker

AI and ML at **Intuit** have three areas of focus.

✓ Smart Products

✓ Fraud detection and prevention

✓ Customer Care and Expert advice

In order to keep fraudsters out of their systems and data, Intuit always stays several moves ahead by leveraging AI/ML-generated insights from data that can determine **real-time fraud detection** in TurboTax: Specifically

- Account take-over detection at login
- Identity theft detection at filing

aws

Saving lives with Amazon SageMaker

Harnessing data and analytics across hardware, software and biotech, GE Healthcare is transforming healthcare by delivering better outcomes for providers and patients.

- Amazon SageMaker allows GE Healthcare to access powerful Artificial Intelligence tools and services to advance improved patient care.

"The scalability of Amazon SageMaker, and its ability to integrate with native AWS services, adds enormous value for us. We are excited about how our continued collaboration between the GE Health Cloud and Amazon SageMaker will drive better outcomes for our healthcare provider partners and deliver improved patient care."

- Sharath Pasupunuti, AI Engineering Leader

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	TRAIN	DEPLOY
Pre-built algorithms & notebooks	One-click model training & tuning	One-click deployment & hosting
Data labeling (GROUND TRUTH)	Optimization (NEO)	
Algorithms & models (AWS MARKETPLACE FOR MACHINE LEARNING)	Reinforcement learning	

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks: TensorFlow, mxnet, PYTORCH
Interfaces: GLUON, Keras
Infrastructure: EC2 P3 & P3N, EC2 C5, FPGAs, GREENGRASS, ELASTIC INFERENCE

aws

Highest-performing infrastructure for your business

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks: TensorFlow, mxnet, PYTORCH
Interfaces: GLUON, Keras
Infrastructure: EC2 P3 & P3N, EC2 C5, FPGAs, GREENGRASS, ELASTIC INFERENCE

Build custom algorithms using the ML frameworks

Fastest and lowest-cost compute options for ML workloads

Elastic compute to provision just-right compute for your ML workloads

aws

The best place to run TensorFlow

Amazon SageMaker is the best place to run TensorFlow in the cloud

- Fully-managed training and hosting
- Near-linear scaling across 100s of GPU
- 75% lower inference costs with Amazon Elastic Inference
- 3x faster network throughput with EC2 P3

STOCK TENSORFLOW

65%

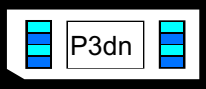
AWS-OPTIMIZED TENSORFLOW

90%

Scaling efficiency with 256 GPUs

aws

Bottom Layer: Frameworks & interfaces



Model	p3dn.24xlarge
NVIDIA V100 Tensor Core GPUs	8
GPU Memory	256 GB
NVIDIA NVLink	300 GB/s
vCPUs	96
Main Memory	768 GiB
Local Storage	2 x 900 GB NVMe SSD
Network Bandwidth	100 Gbps
EBS-Optimized Bandwidth	14 Gbps

AWS Deep Learning AMI

Caffe2

mxnet

TensorFlow

Keras

Chainer

Microsoft CNTK

PYTORCH

torch

GLUON

aws

Thank you