AWS
BUILDERS' DAY

AWS re:Invent 2018
New Machine Learning Services

Julien Simon
Principal Technical Evangelist, AI & Machine Learning, AWS
@julsimon

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

ML FRAMEWORKS & INFRASTRUCTURE

Frameworks

TensorFlow

mxnet

PYTORCH

Interfaces

GLUON

Keras

Infrastructure


EC2 P3 & P3N


EC2 C5

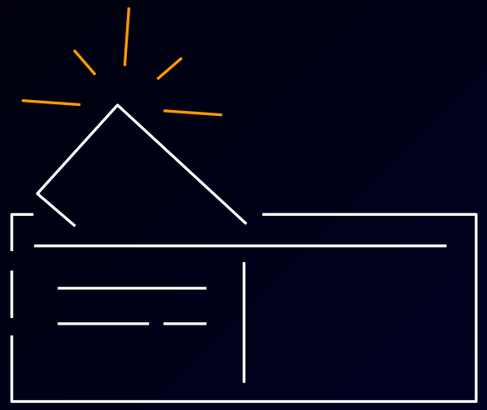

FPGAs


GREENGRASS


ELASTIC NFERENCE

aws

Three areas we are improving for ML developers

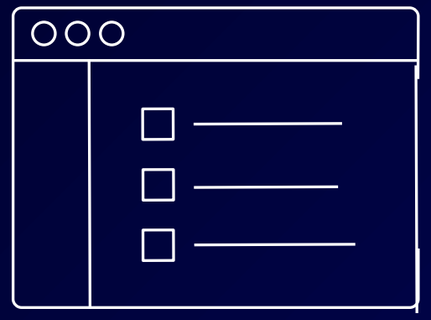


Cost

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

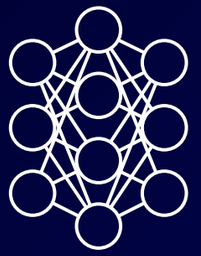
Amazon EC2 P3dn instance

The largest P3 instance, optimized for distributed training

https://aws.amazon.com/blogs/aws/new-ec2-p3dn-gpu-instances-with-100-gbps-networking-local-nvme-storage-for-faster-machine-learning-p3-price-reduction/

Reduce machine learning training time

Better GPU utilization



Support larger, more complex models

KEY FEATURES

- 100Gbps of networking bandwidth
- 8 NVIDIA Tesla V100 GPUs
- 32GB of memory per GPU (2x more P3)
- 96 Intel Skylake vCPUs (50% more than P3) with AVX-512

Most cost efficient platform for TensorFlow

Stock
TensorFlow

65%

scaling efficiency
with 256 GPUs

Most cost efficient platform for TensorFlow

Stock TensorFlow

65%

scaling efficiency with 256 GPUs

NEW

AWS-Optimized TensorFlow

90%

scaling efficiency with 256 GPUs

Available with Amazon SageMaker and the AWS Deep Learning AMIs

aws

Most cost efficient platform for TensorFlow

https://aws.amazon.com/about-aws/whats-new/2018/11/tensorflow-scalability-to-256-gpus/

NEW

Stock TensorFlow

65%

scaling efficiency with 256 GPUs

AWS-Optimized TensorFlow

90%

scaling efficiency with 256 GPUs

Available with Amazon SageMaker and the AWS Deep Learning AMIs

30m

training time

14m

training time

Fastest time for TensorFlow

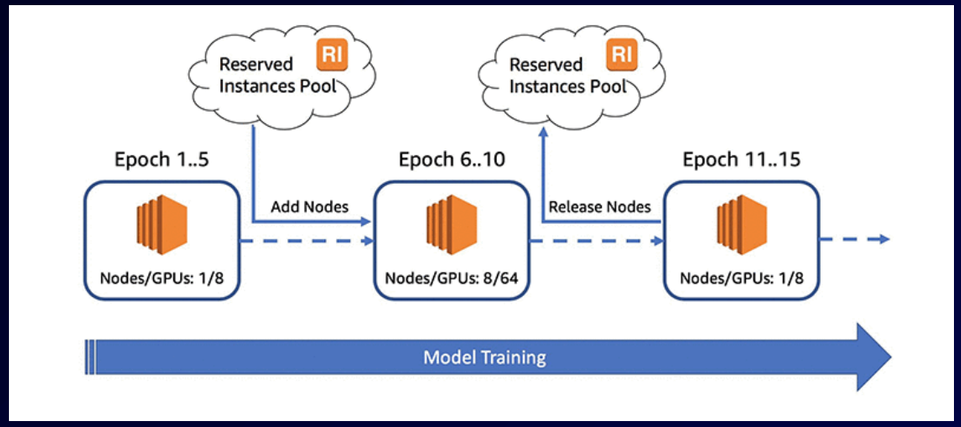
Dynamic training with Apache MXNet and RIs

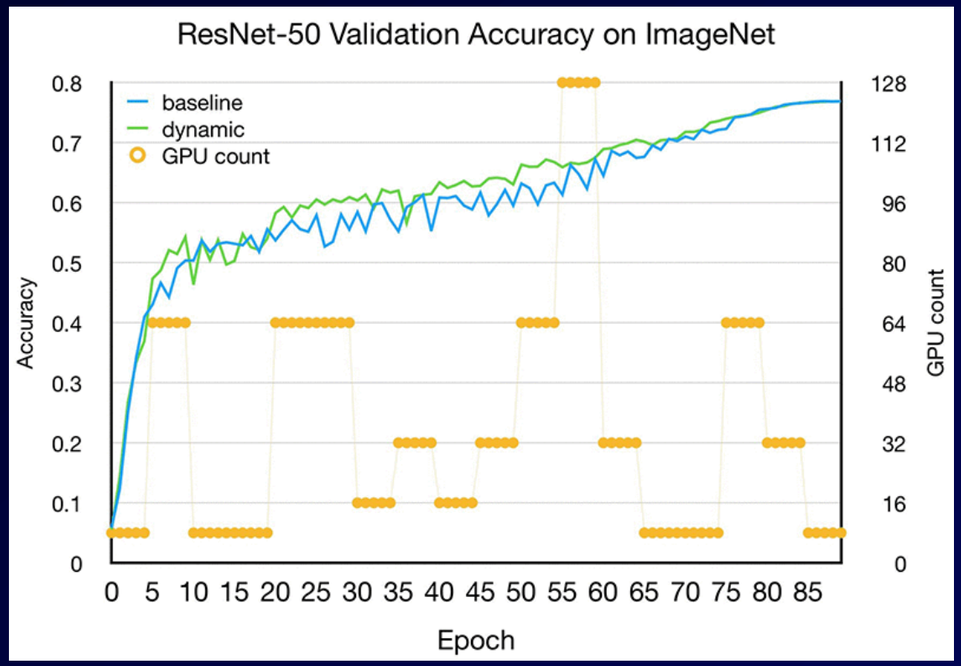
https://aws.amazon.com/blogs/machine-learning/introducing-dynamic-training-for-deep-learning-with-amazon-ec2/

Use a variable number of instances for distributed training

No loss of accuracy

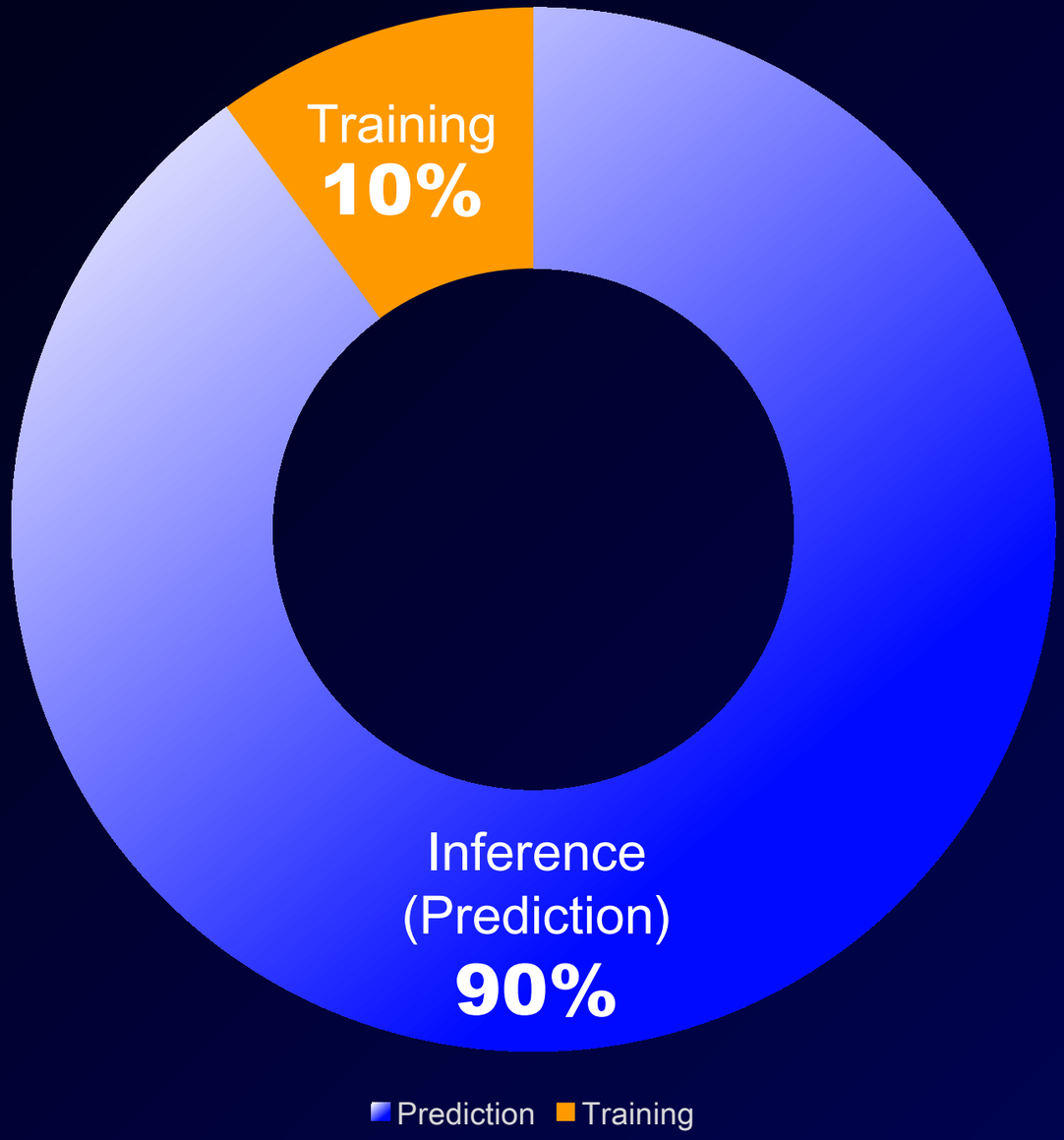
Coming soon
spot instances,
additional frameworks



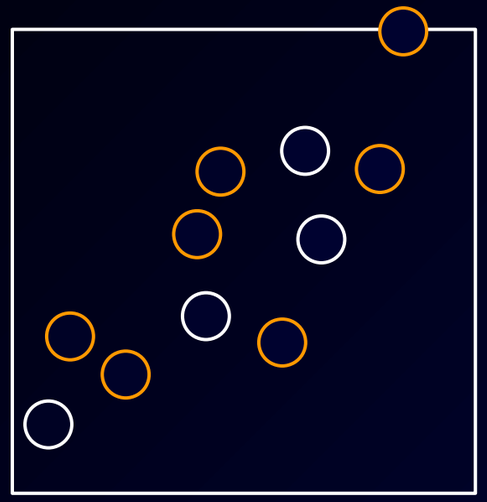


Training gets a lot of attention,
but what about inference?

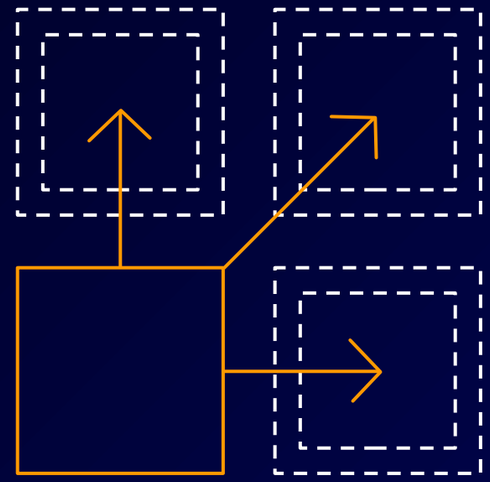
Predictions drive complexity and cost in production



The challenges of prediction in production



One size does not fit all



Elasticity is important

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

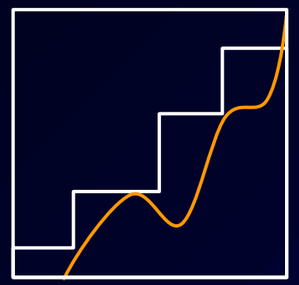
Instance Name	vCPUs	RAM	EBS Bandwidth	Network Bandwidth
c5n.large	2	5.25 GiB	Up to 3.5 Gbps	Up to 25 Gbps
c5n.xlarge	4	10.5 GiB	Up to 3.5 Gbps	Up to 25 Gbps
c5n.2xlarge	8	21 GiB	Up to 3.5 Gbps	Up to 25 Gbps
c5n.4xlarge	16	42 GiB	3.5 Gbps	Up to 25 Gbps
c5n.9xlarge	36	96 GiB	7 Gbps	50 Gbps
c5n.18xlarge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

aws

Amazon Elastic Inference

https://aws.amazon.com/blogs/aws/amazon-elastic-inference-gpu-powered-deep-learning-inference-acceleration/

Lower inference costs up to 75%



Match capacity to demand

Available between 1 to 32 TFLOPS

KEY FEATURES

Integrated with Amazon EC2, Amazon SageMaker, and Amazon DL AMIs

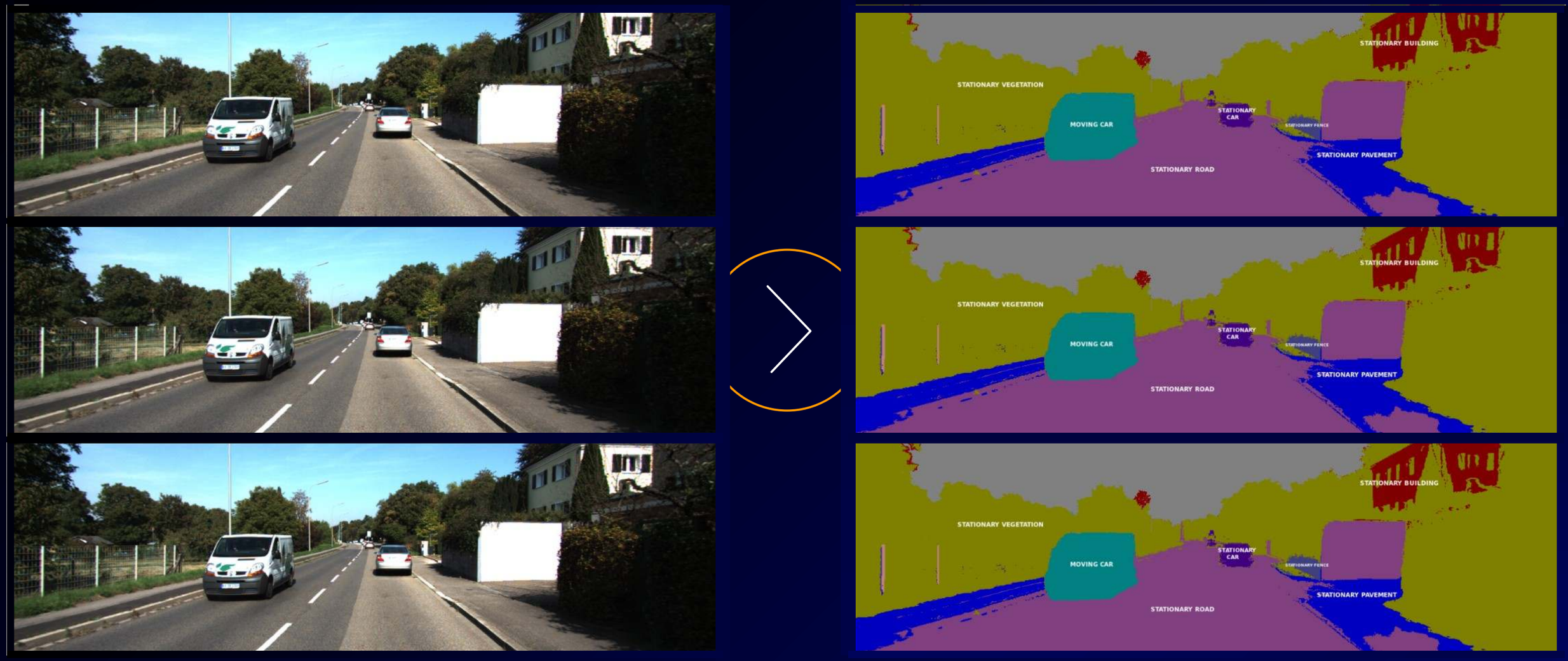
Support for TensorFlow, Apache MXNet, and ONNX with PyTorch coming soon

Single and mixed-precision operations

Making it easier to obtain high quality labeled data

Successful models require high-quality data

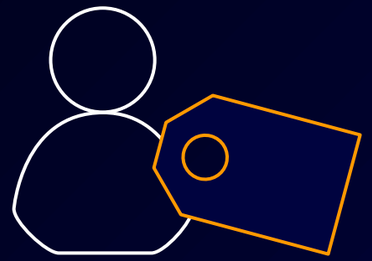
Successful models require high-quality data



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



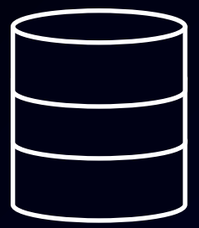
Easily integrate human labelers

Get accurate results

KEY FEATURES

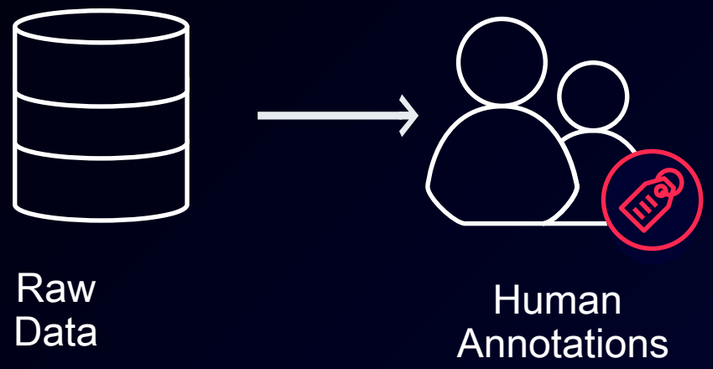
- Automatic labeling via machine learning
- Ready-made and custom workflows for image bounding box, segmentation, and text
- Private and public human workforce
- Label management

How it works

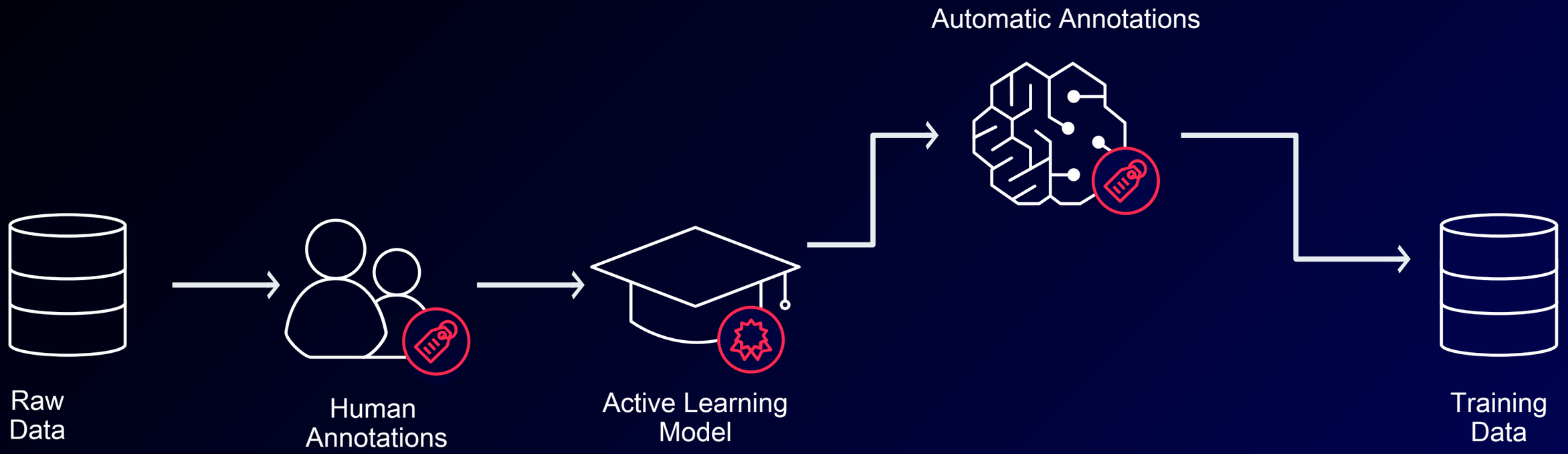


Raw
Data

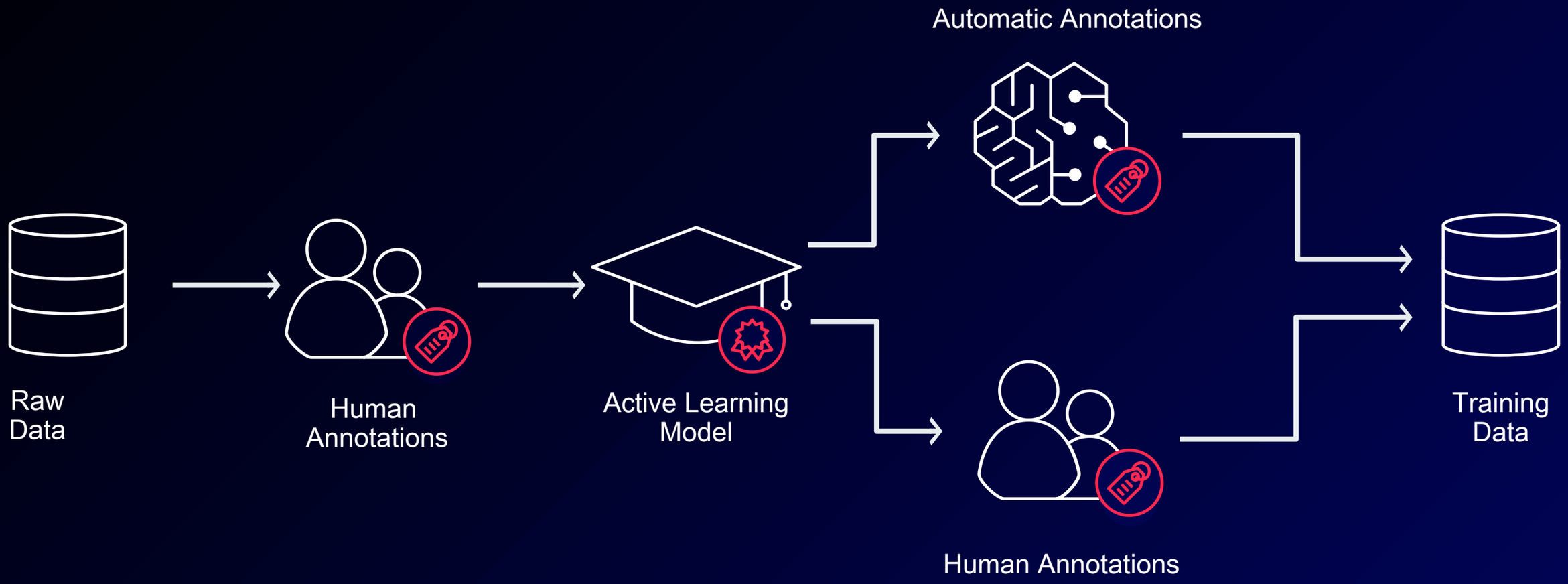
How it works



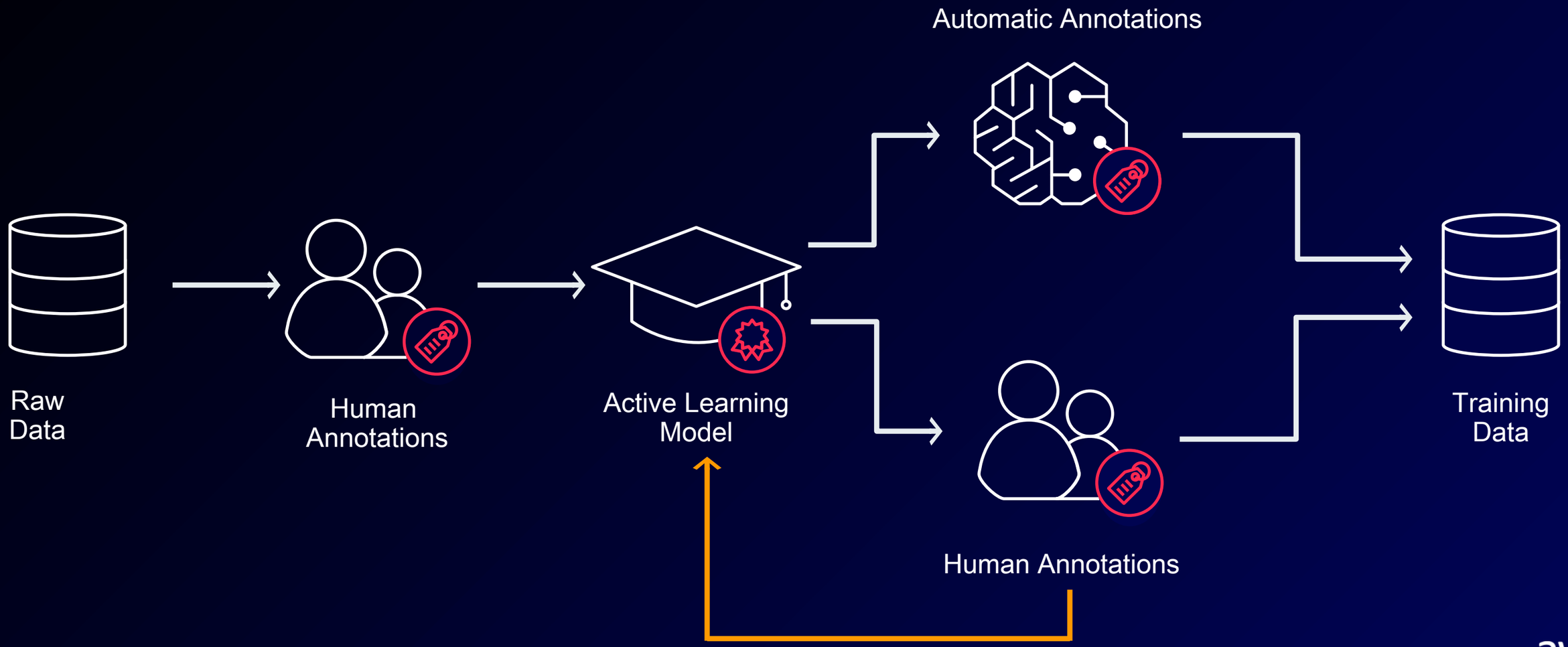
How it works



How it works



How it works



Creating training data

amazon mechanical turk beta

500,000+ workers

Private labeling workforce

Third-party vendors

iMerit human-empowered computing

capestart

vivitec

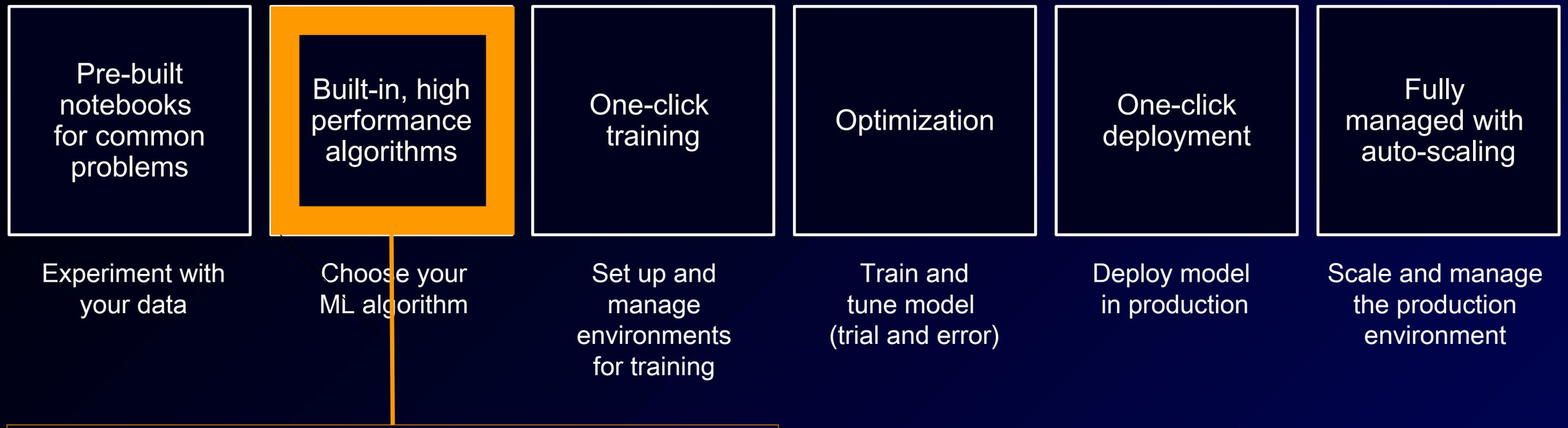
iVision Engineer the future, today.®

cogito

aws

Driving Ease of Use

Amazon SageMaker: build, train, and deploy ML



Recommendation with Factorization Machines

Time-series with Deep AR

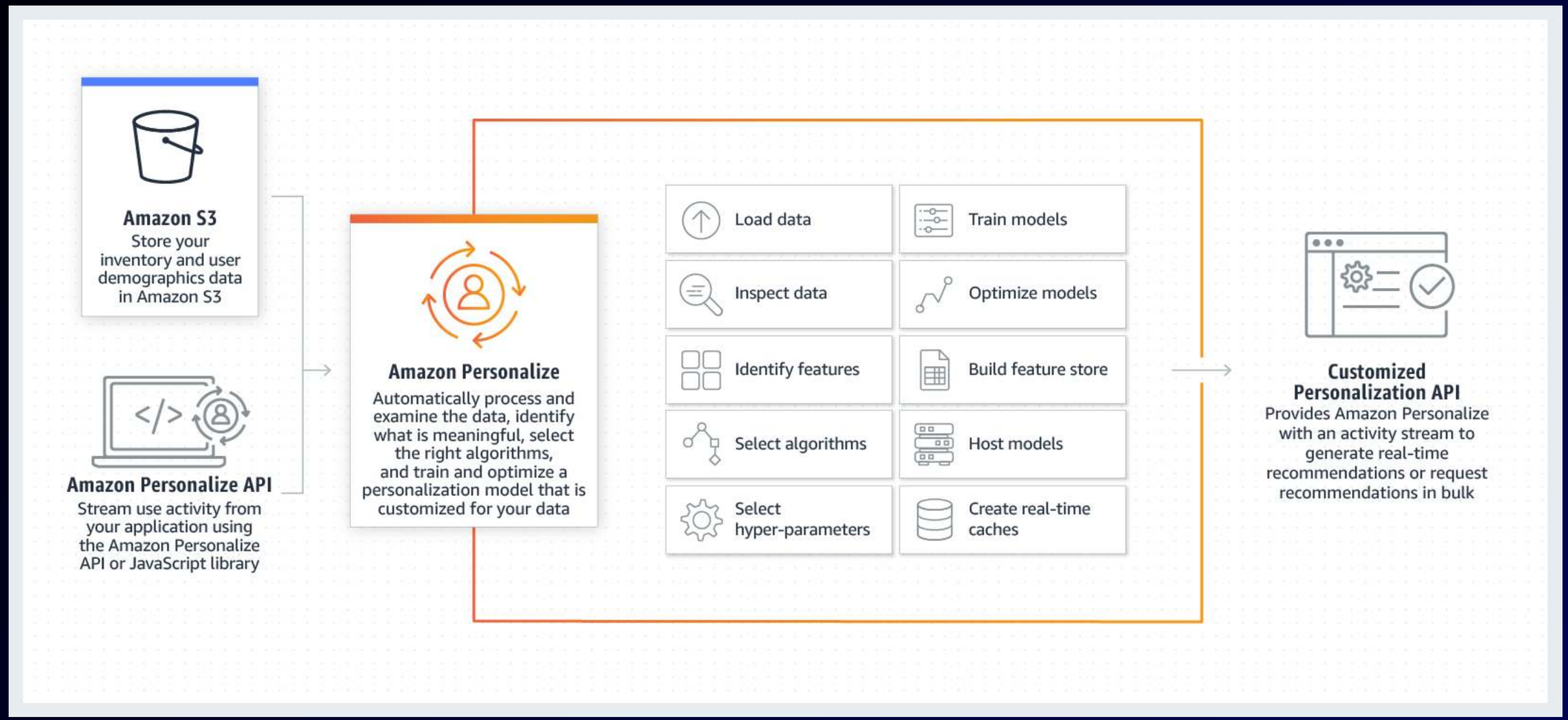
A lot of expertise is still required

Can we make it simpler?

aws

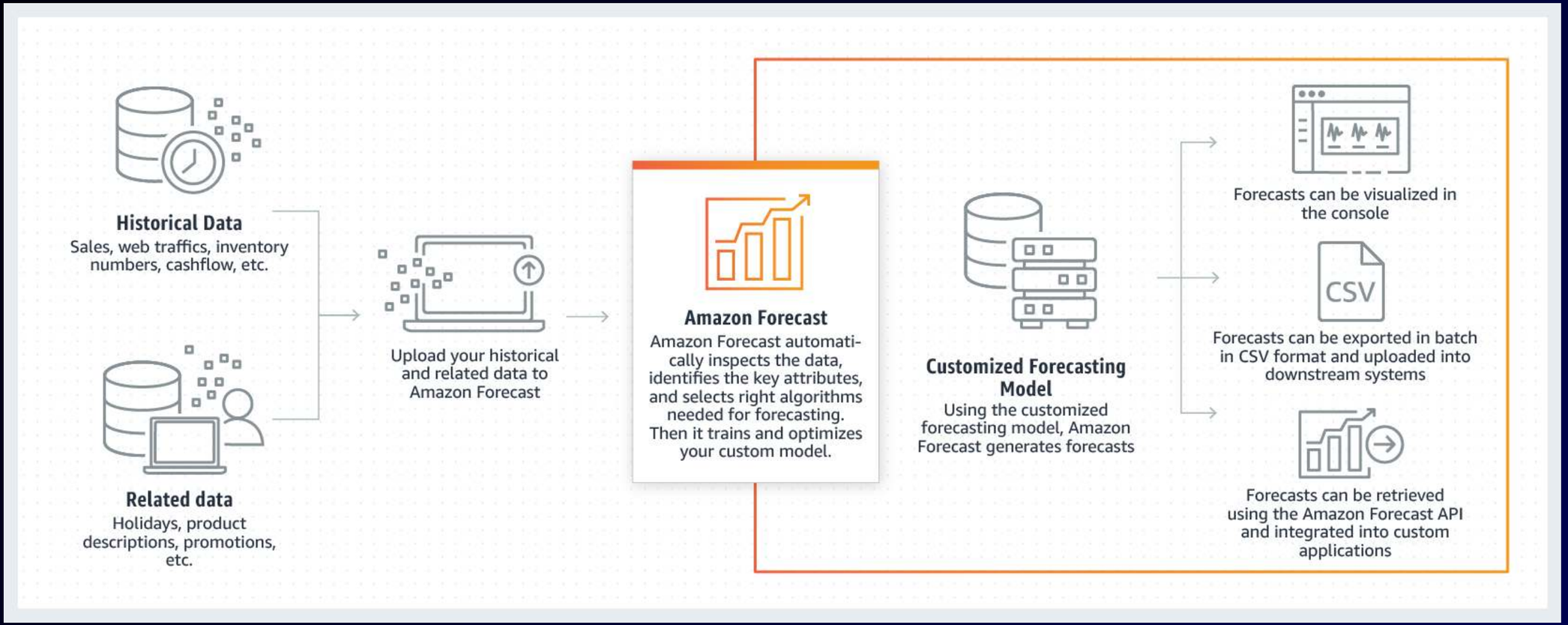
Amazon Personalize

https://aws.amazon.com/blogs/aws/amazon-personalize-real-time-personalization-and-recommendation-for-everyone



Amazon Forecast

https://aws.amazon.com/blogs/aws/amazon-forecast-time-series-forecasting-made-easy/



AWS Marketplace for Machine Learning

ML algorithms and models available instantly

Browse or search AWS Marketplace

Subscribe in a single click

Available in Amazon SageMaker

KEY FEATURES

SELLERS

Automatic labeling via machine learning

IP protection

Automated billing and metering

BUYERS

Broad selection of paid, free, and open-source algorithms and models

Data protection

Over 150 models and algorithms available

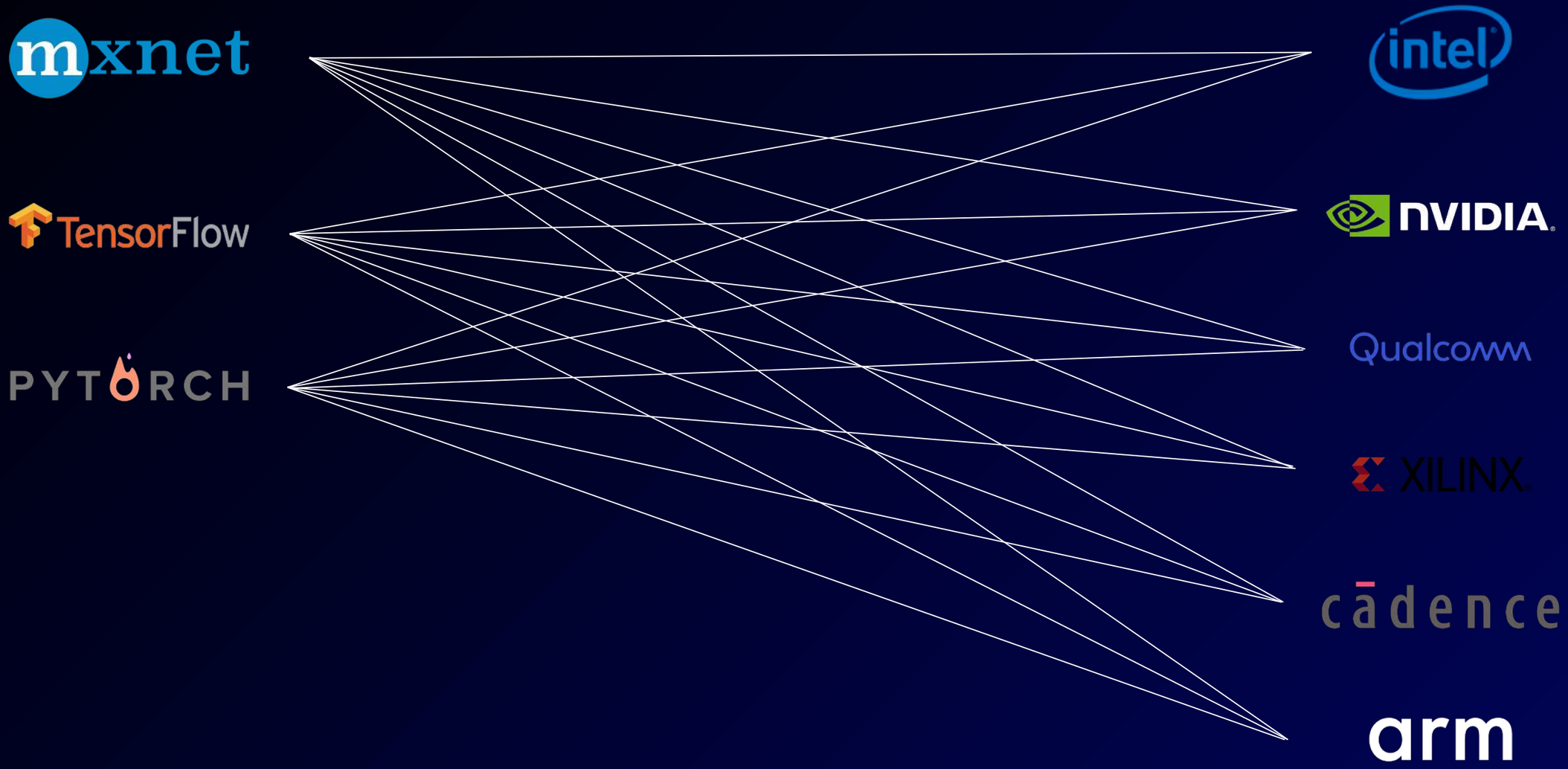
SELECTED VENDORS

SOME OF THE AVAILABLE ALGORITHMS AND 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

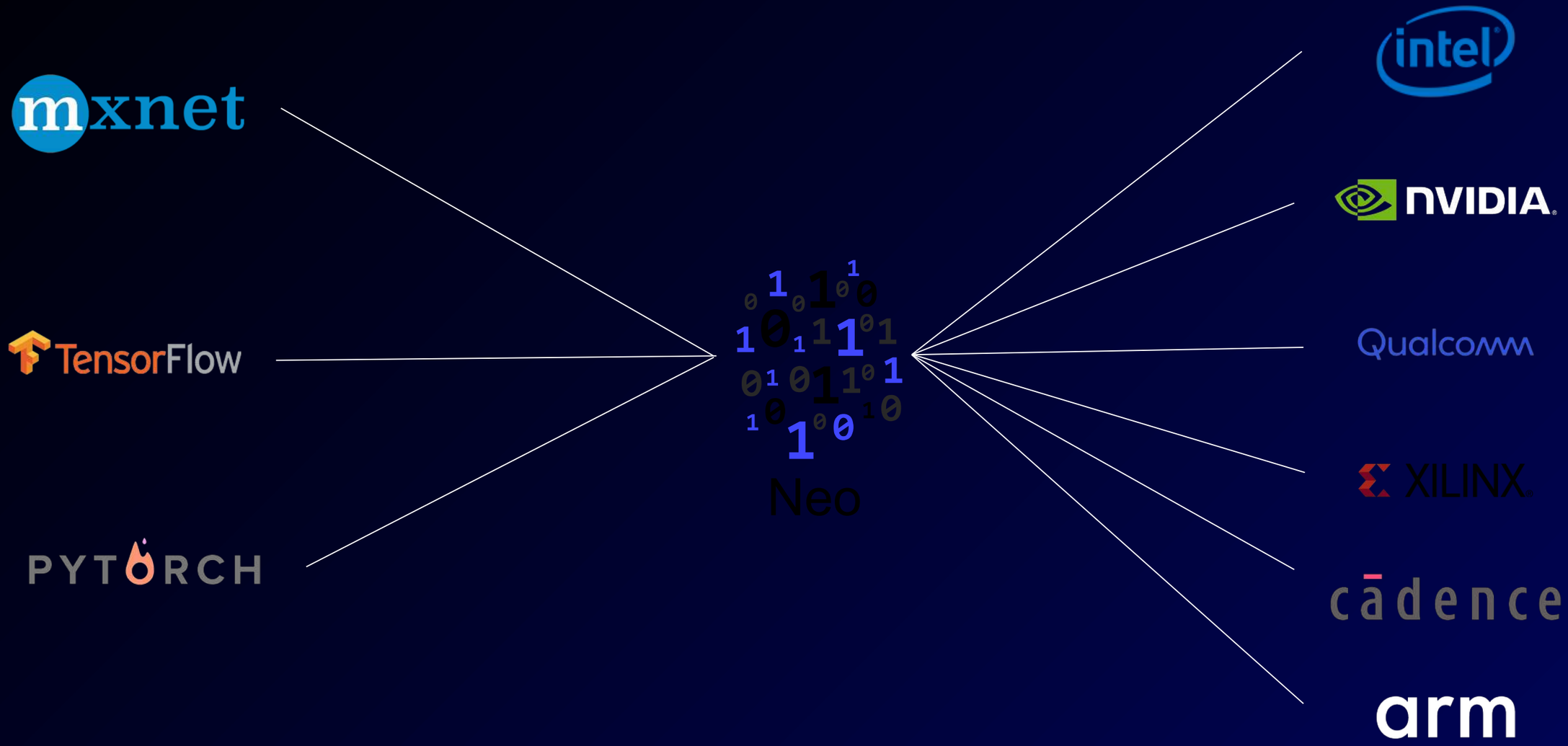
aws

Model optimization is extremely complex



aws

Train once, run anywhere



aws

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

NVIDIA

arm

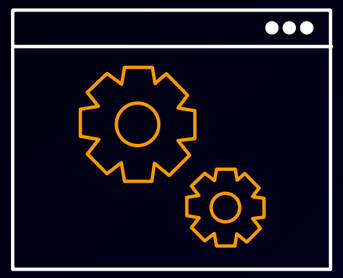
Broad hardware support

aws

What's next for Machine Learning?

Amazon SageMaker RL

Reinforcement learning for every developer and data scientist



Fully managed

Broad support for frameworks

Broad support for simulation environments including SimuLink and MatLab

KEY FEATURES

TensorFlow, Apache MXNet, Intel Coach, and Ray RL support

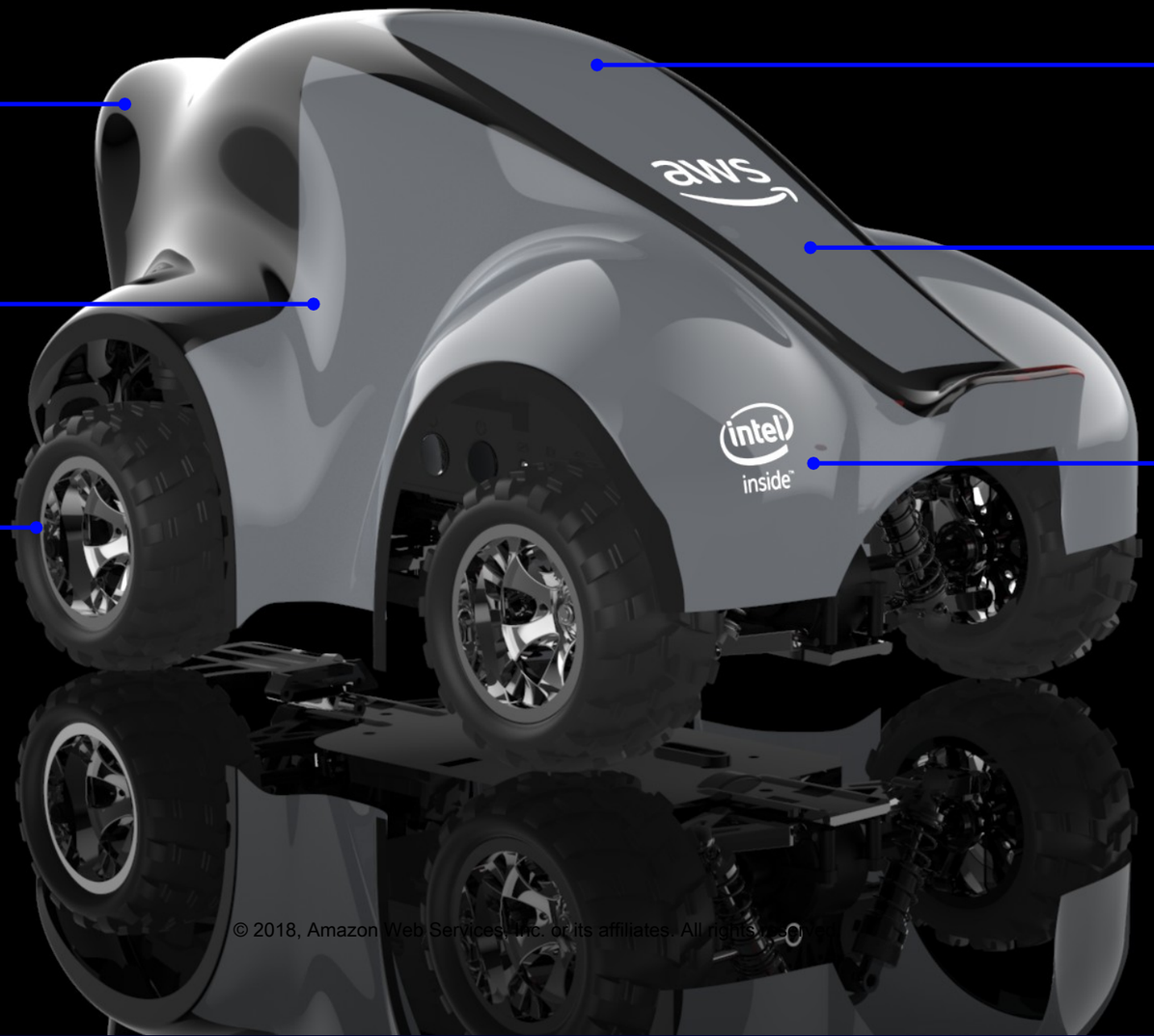
2D & 3D physics environments and OpenAI Gym support

Supports Amazon Sumerian and Amazon RoboMaker

Example notebooks and tutorials

Introducing AWS DeepRacer

Fully autonomous 1/18th scale race car, driven by reinforcement learning



HD video camera

Gyroscope

Four-wheel drive

Dual power for compute and drive

Accelerometer

Dual-core Intel processor

aws

Getting started

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, AWS
@julsimon
https://medium.com/@julsimon/