GTC-DC

Build, Train, and Deploy Deep Learning Models Faster in the Cloud with Amazon SageMaker

Wenming Ye @wenmingye
Sr. Solutions Architect (AI/ML)
Amazon Web Services

aws SUMMIT

Agenda

Deep learning trend (5 min)

DL architectures: CNN & BERT (35 min)

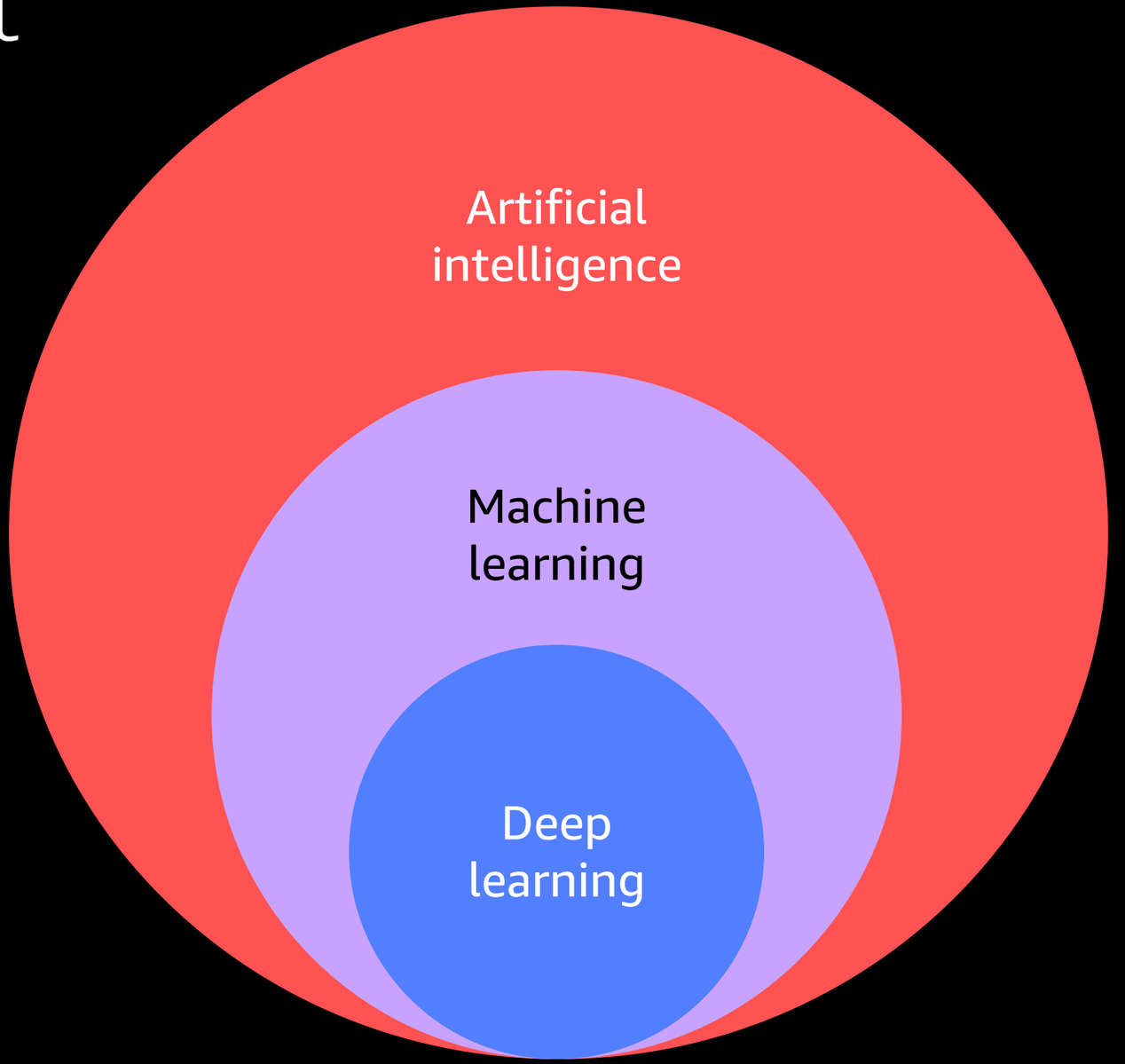
P3 & G4 instances details (5 min)

Lab 1: object detection (SSD) (40 min)

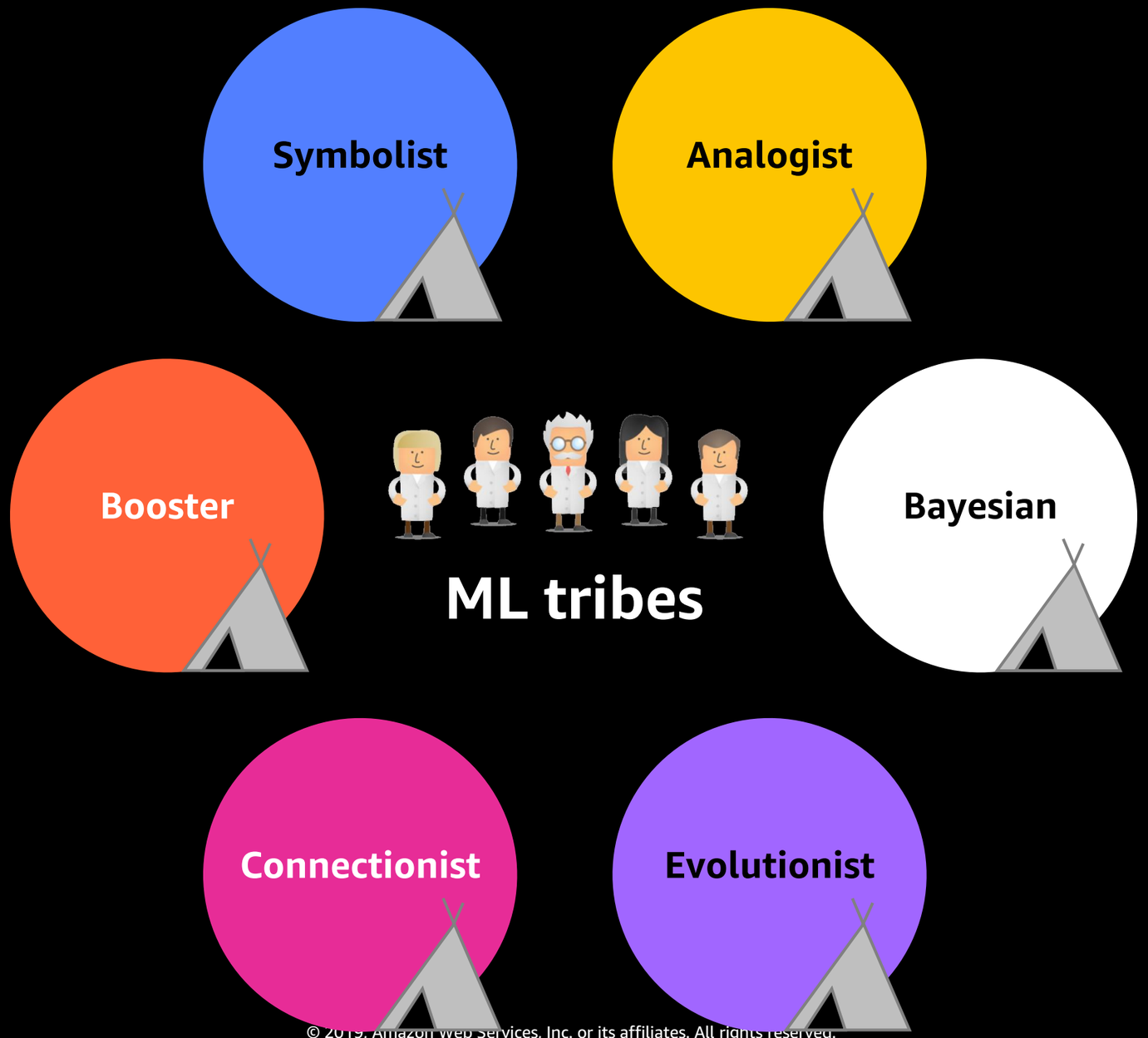
Lab 2: sentiment analysis (BERT) (30 min)

Learning resources (5 min)

DL in context

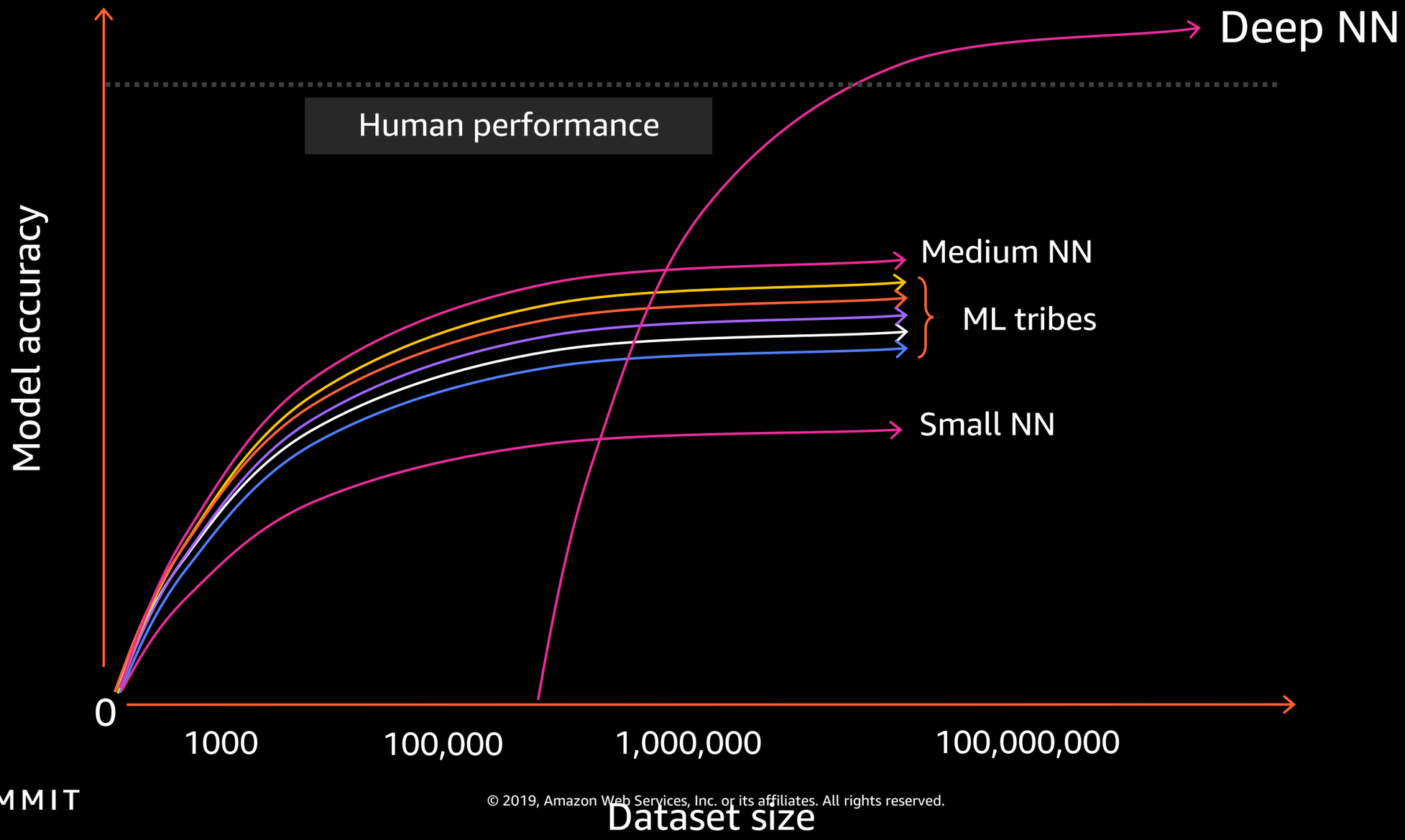


aws SUMMIT



aws SUMMIT

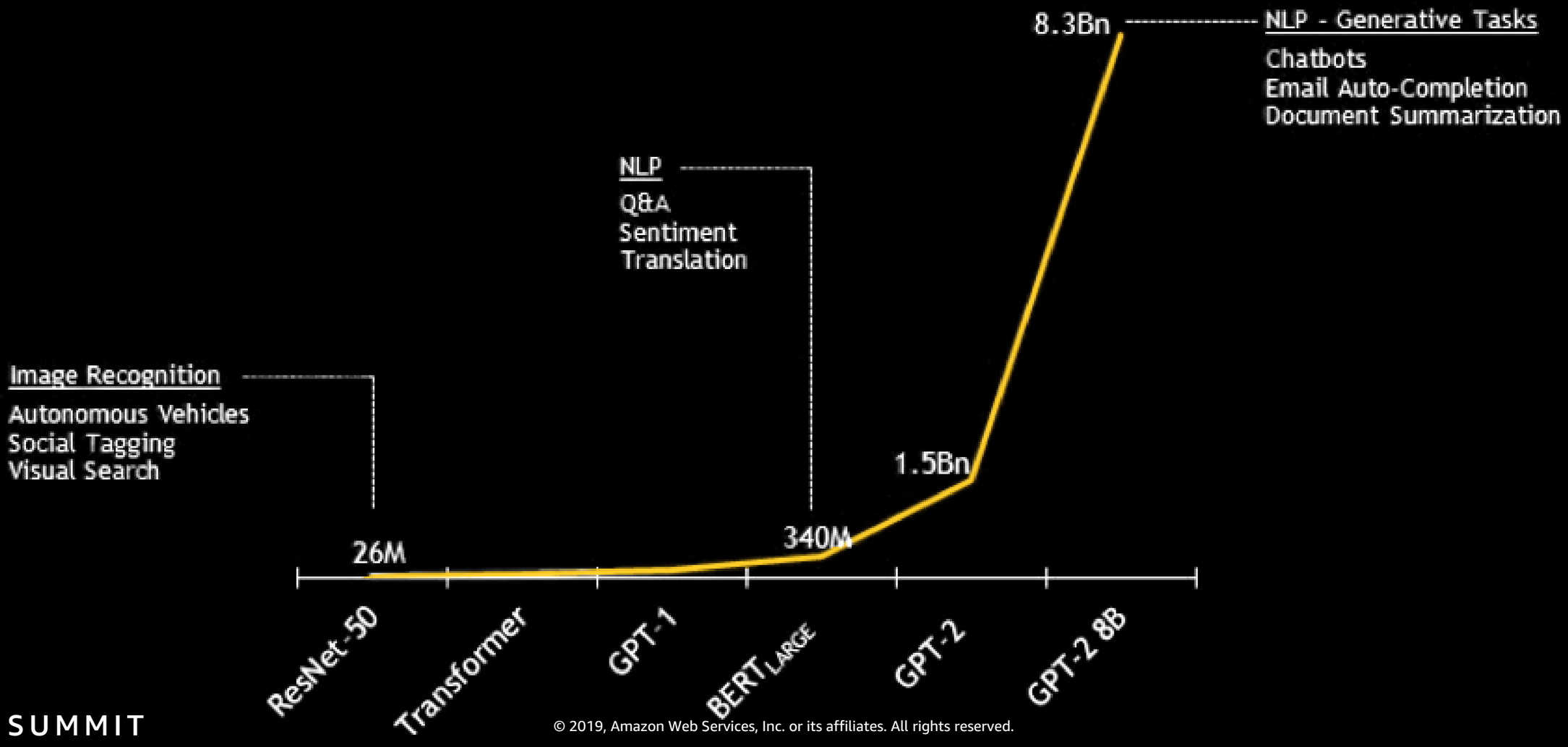
Learning at scale



aws SUMMIT

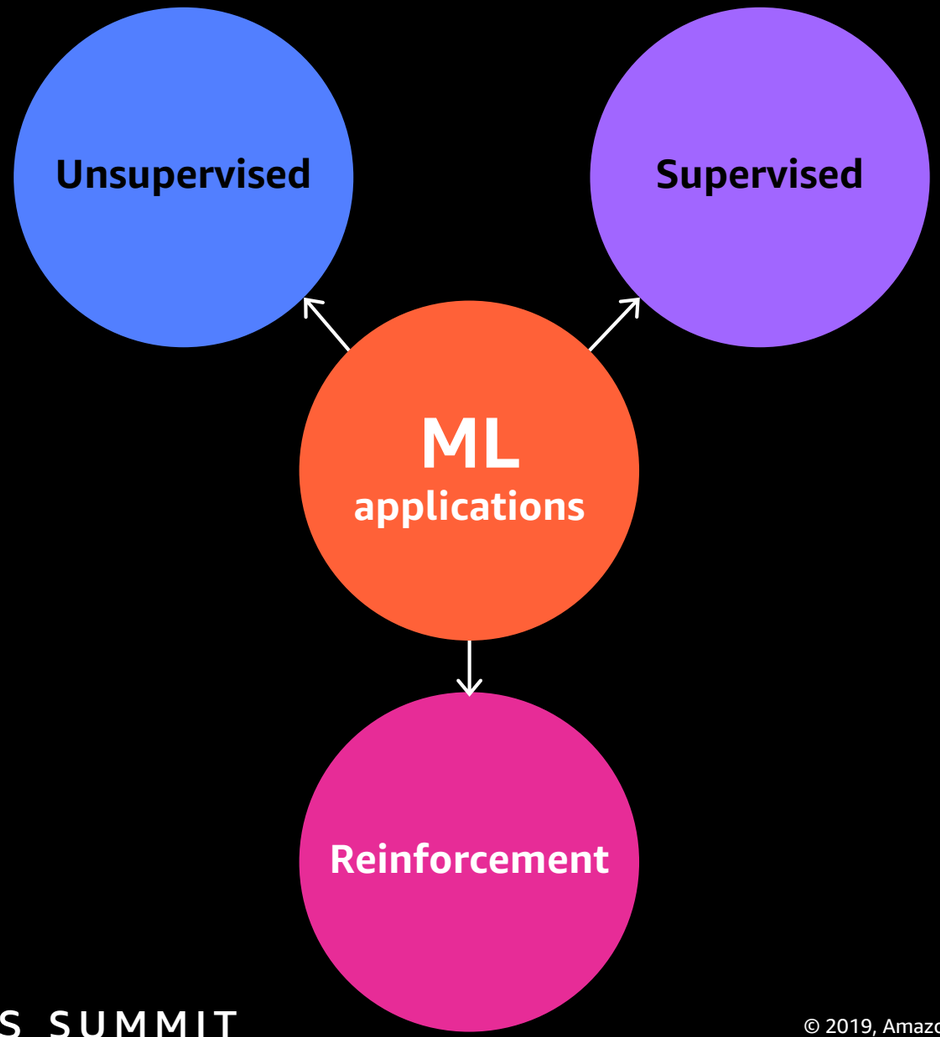
Exploding model complexity

Number of parameters by network



aws SUMMIT

Three types of learning



aws SUMMIT

THE AWS ML STACK

Broadest and deepest set of capabilities

AI Services

VISION			SPEECH		LANGUAGE		CHATBOTS	FORECASTING	RECOMMENDATIONS
 REKOGNITION IMAGE	 REKOGNITION VIDEO	 TEXTRACT	 POLLY	 TRANSCRIBE	 TRANSLATE	 COMPREHEND	 LEX	 FORECAST	 PERSONALIZE

ML Services

 **Amazon SageMaker**

Ground Truth	Notebooks	Algorithms + Marketplace	Reinforcement Learning	Training	Optimization	Deployment	Hosting

ML Frameworks + Infrastructure

FRAMEWORKS	INTERFACES
TensorFlow, mxnet, PYTORCH	GLUON, Keras

INFRASTRUCTURE						
 EC2 P3 & P3DN	 EC2 G4	 EC2 C5	 FPGAS	 GREENGRASS	 ELASTIC INFERENCE	 INFERENTIA

aws SUMMIT

Our deep experience with AI/ML differentiates our approach

Amazon has invested in AI/ML since our inception, and we share our knowledge and capabilities with our customers

Checkout-free shopping using deep learning

Product recommendation engine

Natural language processing-supported contact centers

Robot-enabled fulfillment centers

ML-driven supply chain and capacity planning

New product categories

More machine learning happens on AWS than anywhere else

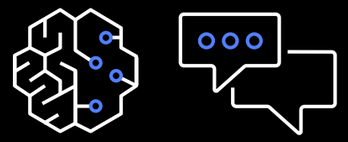
aws SUMMIT

Machine learning use cases

Applications that benefit from accelerated compute

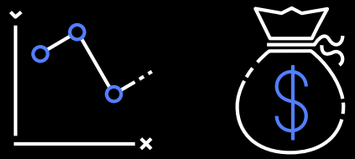
Machine learning/AI

Natural language processing



Image/ video analysis

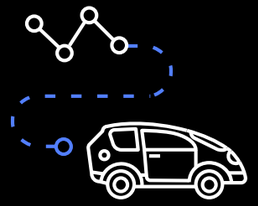
Financial services



Healthcare & life sciences

Manufacturing

Autonomous vehicle systems



Recommendation systems

Retail

Travel and hospitality

Energy

Scenarios and DL architecture

Architecture

Vision: convolutional neural network (CNN)

Language: bidirectional transformers for NLP (BERT)

CNN scenarios

- Image classification
- Object detection
- Image segmentation
- Visual search
- GANs for item generation

NLU scenarios

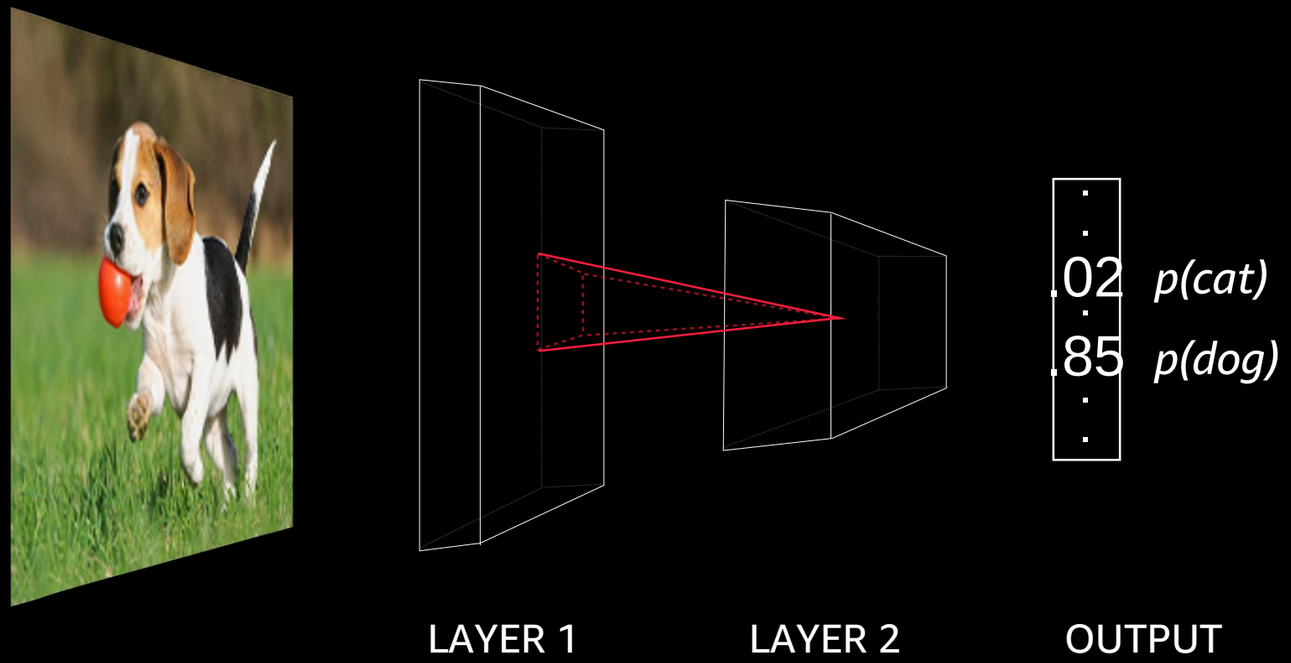
- Classification, topic modeling
- Sentiment analysis
- Text generation
- Entity recognition
- Translation, Q&A

Convolution neural network

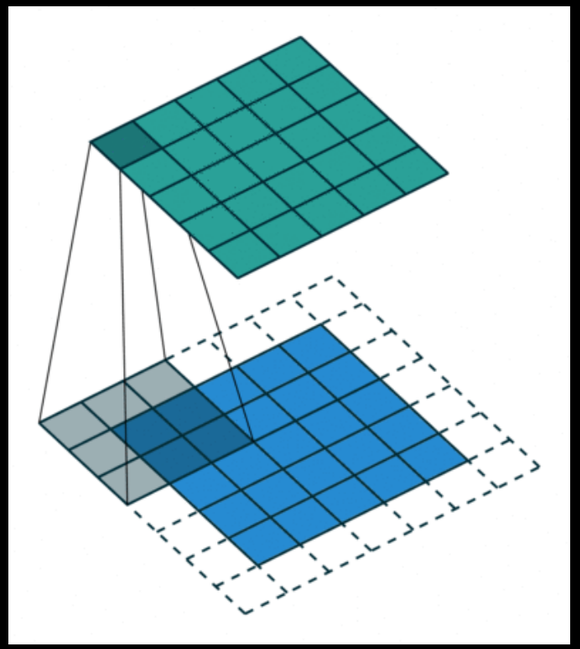
aws SUMMIT

Deep learning in computer vision

Explore spatial information with convolution layers



Convolutional neural network



https://leonardoaraujosantos.gitbooks.io/artificial-inteligence

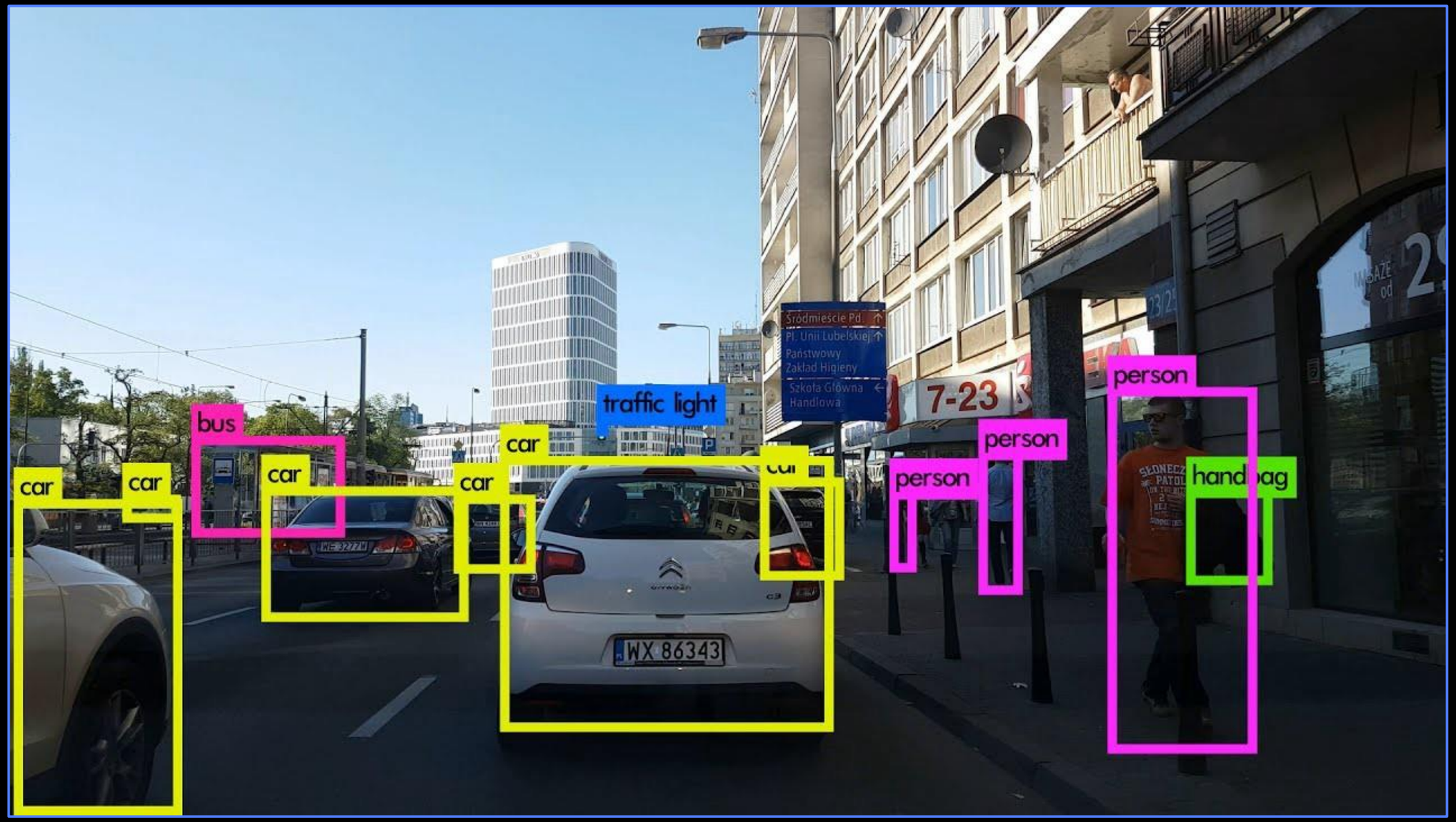
Demo: convolution neural network

http://scs.ryerson.ca/~aharley/vis/conv

Demo: object detection

aws SUMMIT

Object detection



Amazon Rekognition Video

Object and Activity Detection

Person Tracking

Face Recognition

Real-time Live Stream

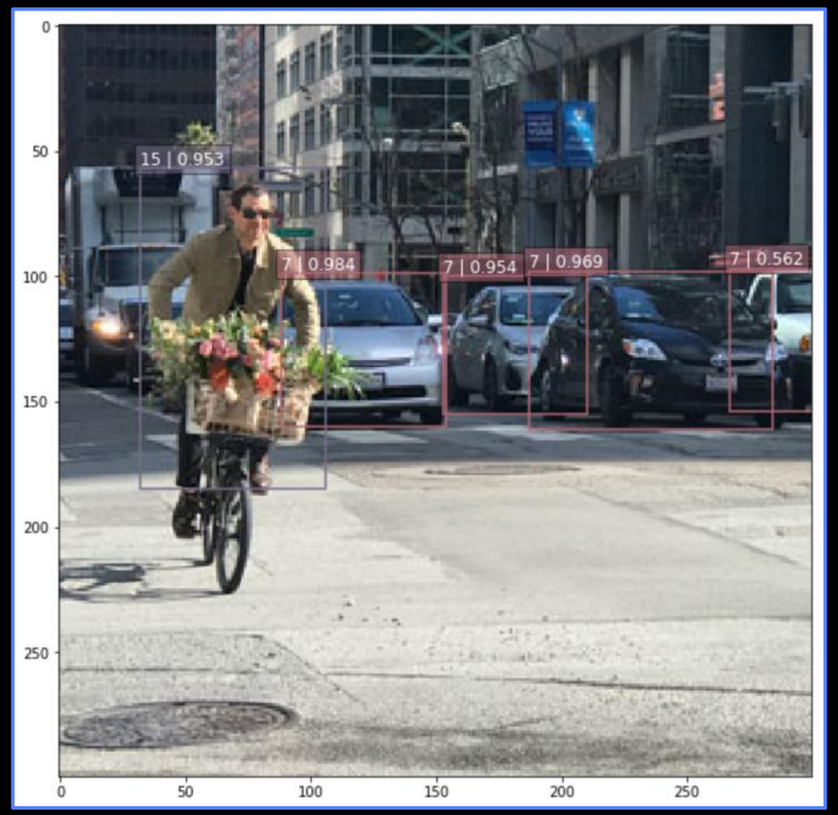
Content Moderation

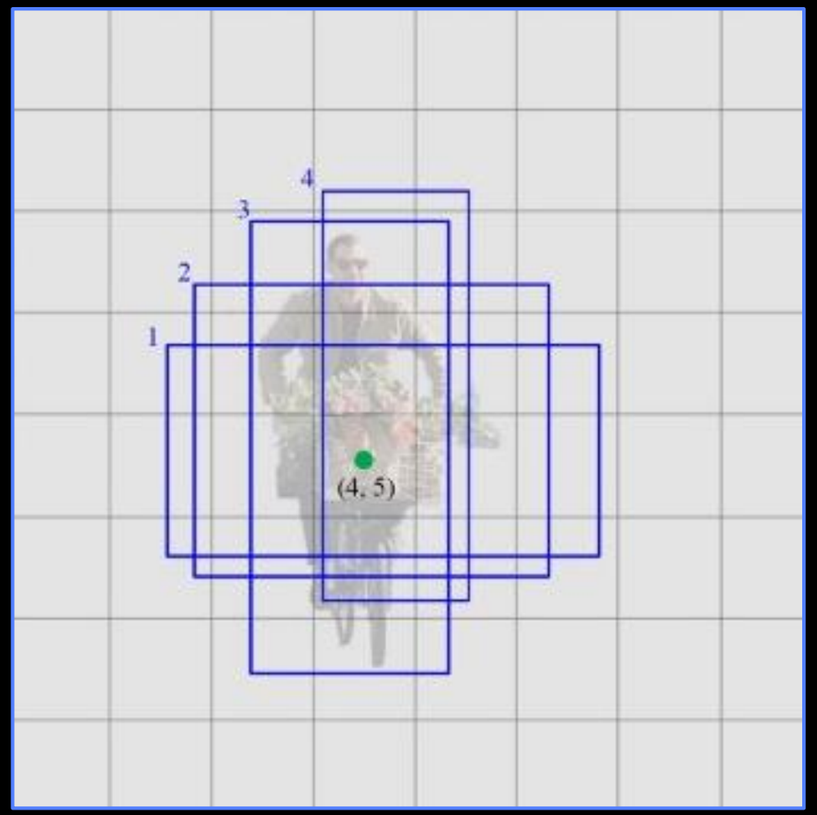
Celebrity Recognition


Traffic Cam 352 - Riverview West
Northbound
03:15:05:41
Mon, Nov 27, 2017


LIVE STREAMING
Front Door Camera 1

Single shot detector



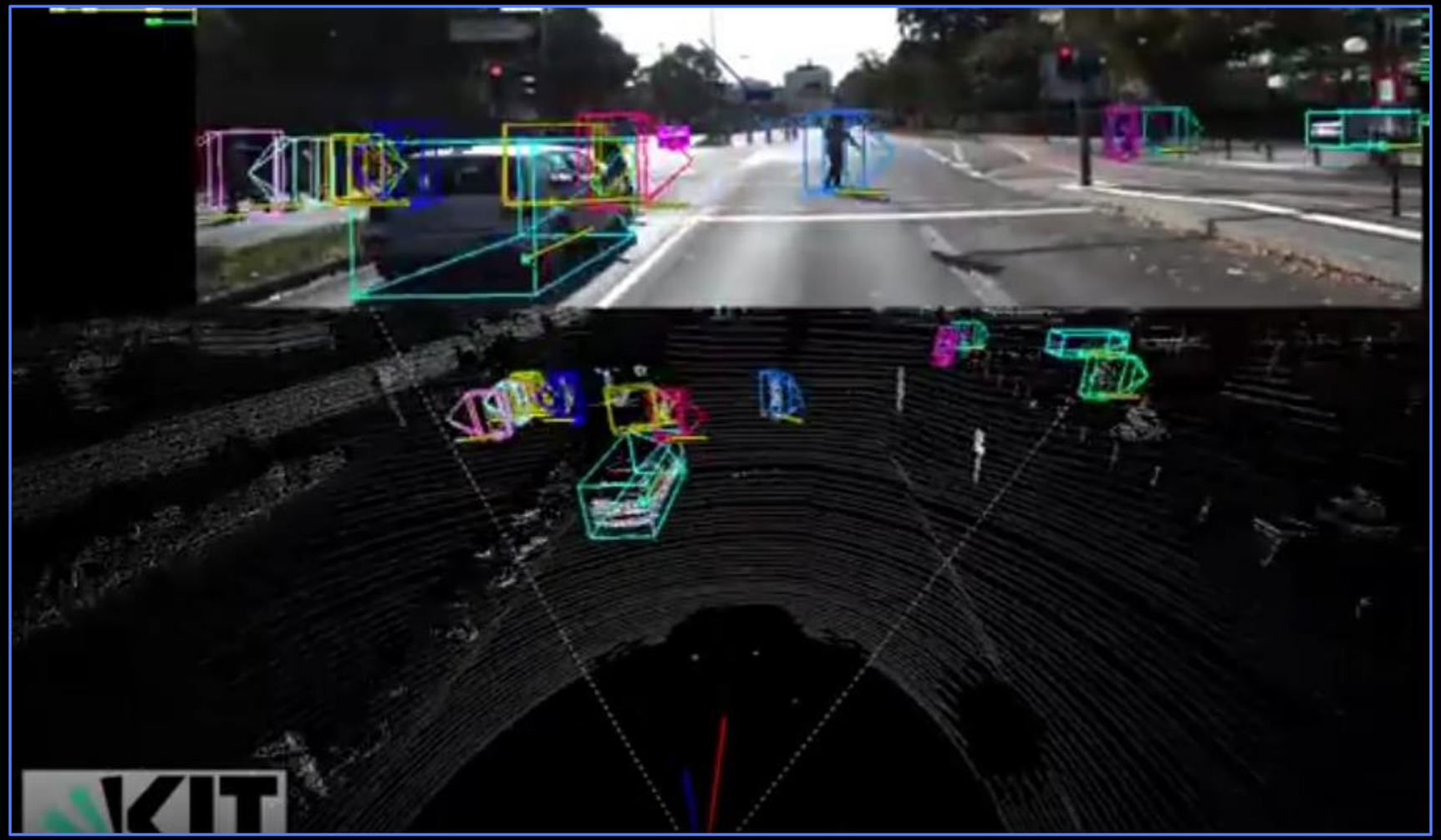


Demo: image segmentation

aws SUMMIT

Image segmentation

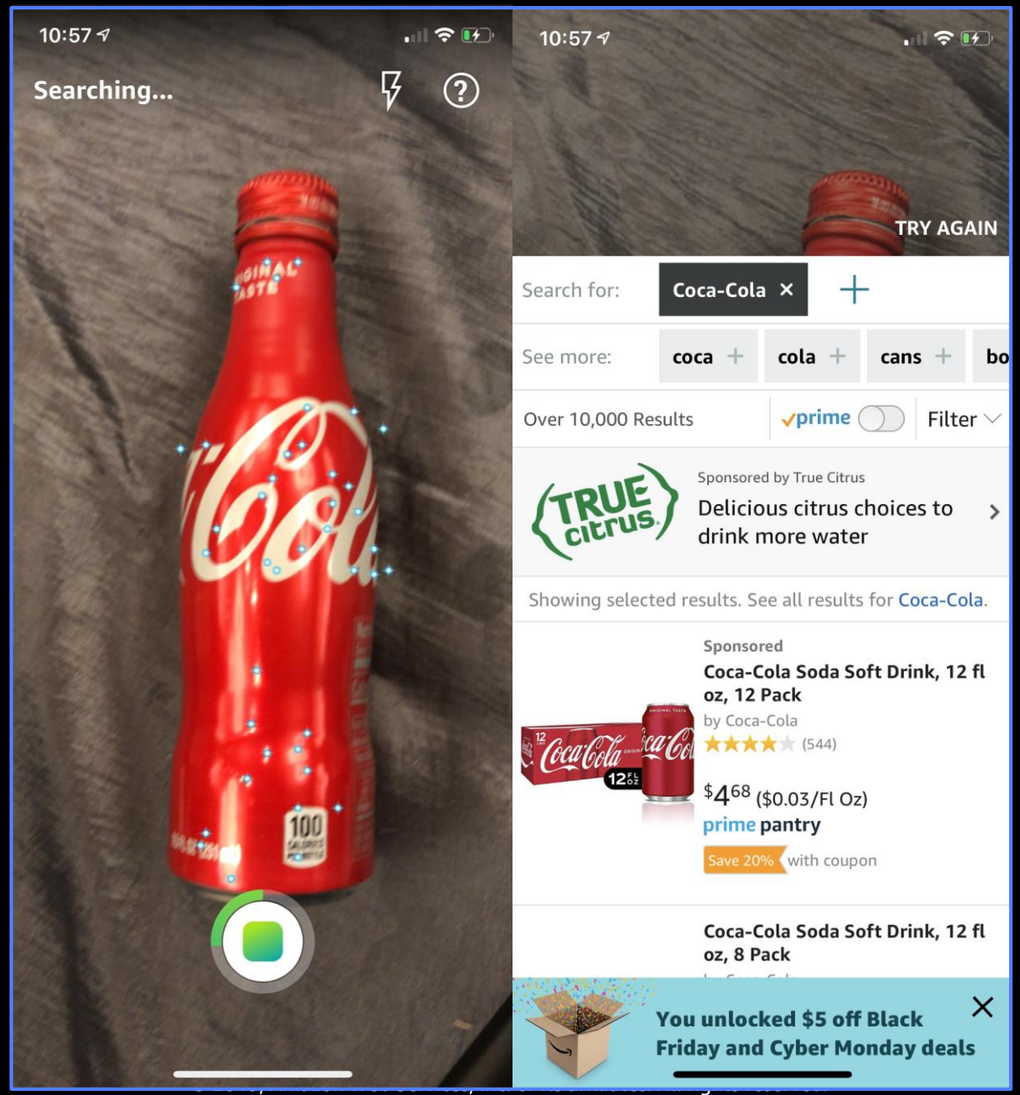
KITTI



Visual search

aws SUMMIT

Visual search



Pipeline stages

Image query processing

Data normalization/augmentation

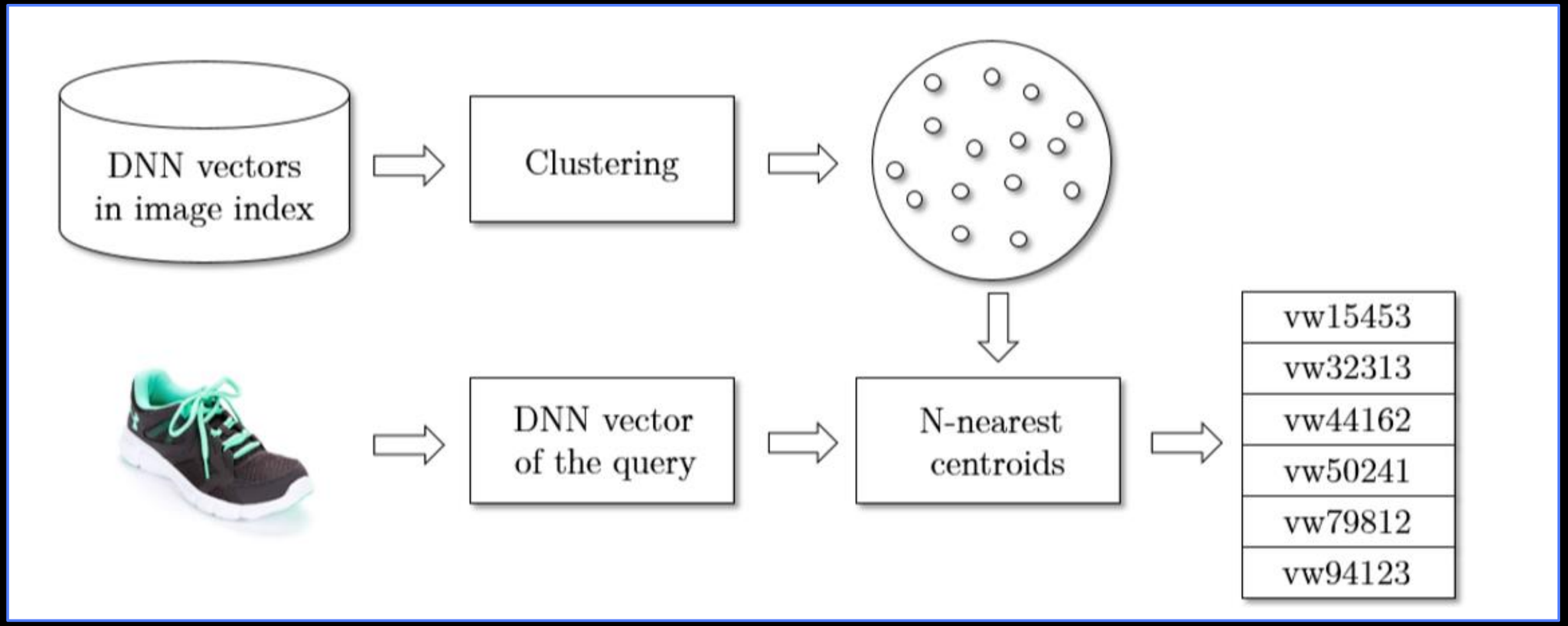
Embedding

DNN model(s)

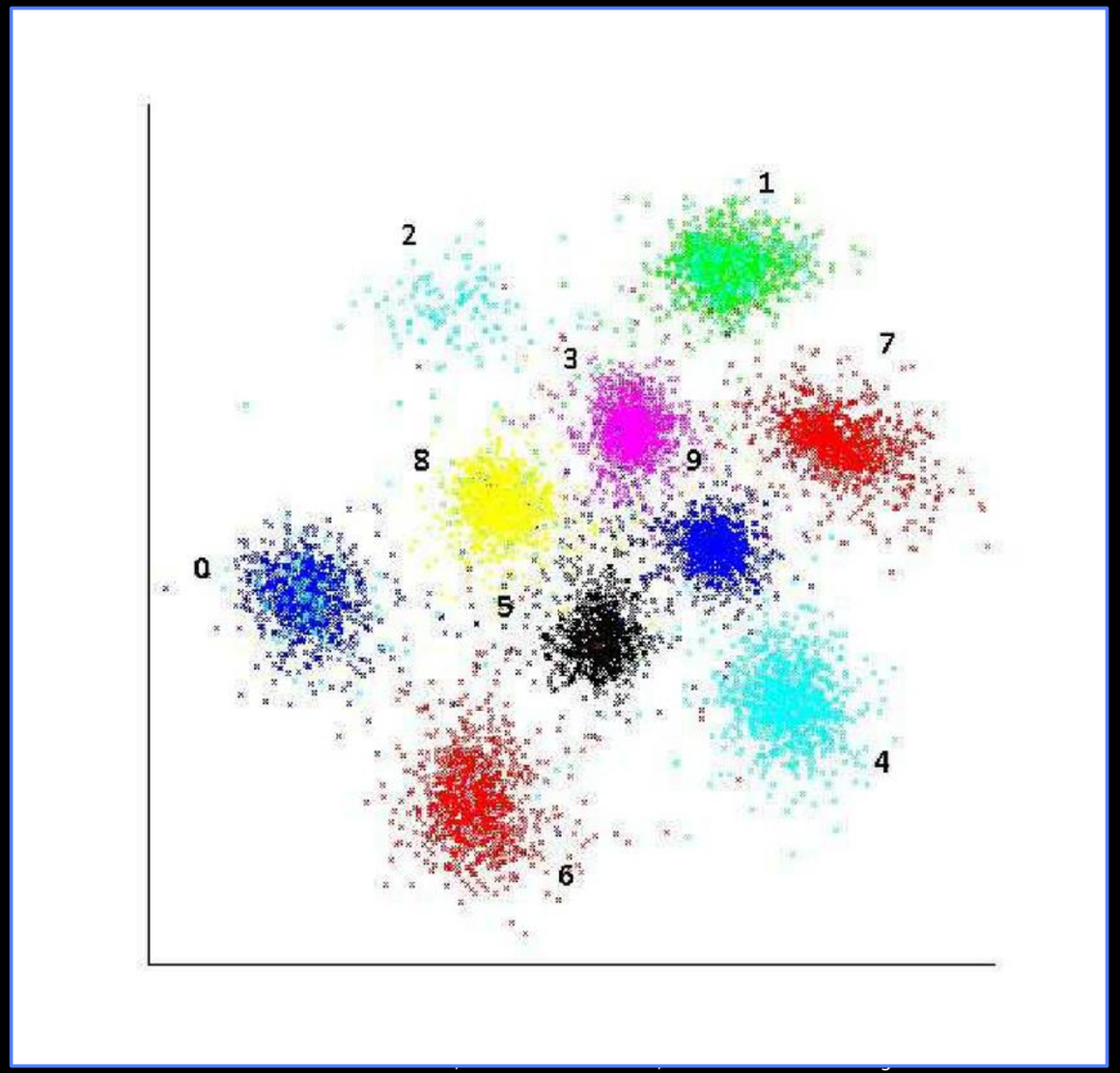
kNN + ranking

Post-processing, de-dup

Architecture



Embedding = learned representation space



Demo: Image Embedding

aws SUMMIT

Domains

Domain	Purpose
Generic	Optimize for a broad range of image classification tasks. If none of the other domains are appropriate, or you are unsure of which domain to choose, select the generic domain.
Food	Optimized for photographs of dishes as you would see them on a restaurant menu. If you want to classify photographs of individual fruits or vegetables, use the food domain.
Landmarks	Optimized for recognizable landmarks, both natural and artificial. This domain works best when the landmark is clearly visible in the photograph. This domain works even if the landmark is slightly obstructed by people in front of it.
Retail	Optimized for images that are found in a shopping catalog or shopping website. If you want high precision classifying between dresses, pants, and shirts, use this domain.
Adult	Optimized to better define adult content and non-adult content. For example, if you want to block images of people in bathing suits, this domain allows you to build a custom classifier to do that.
Compact domains	Optimized for the constraints of real-time classification on mobile devices. The models generated by compact domains can be exported to run locally.

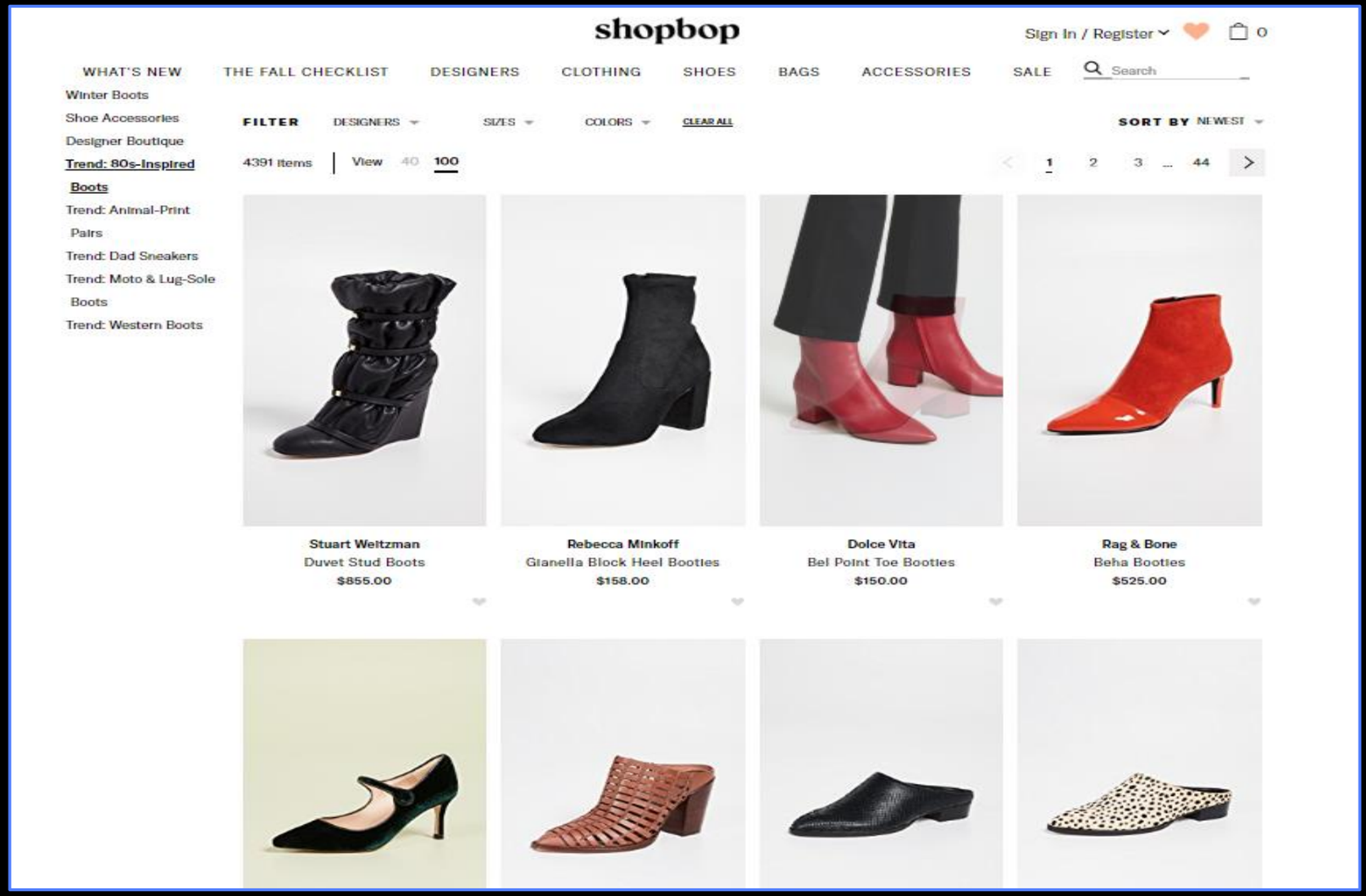
Generative adversarial networks (GANs)

aws SUMMIT

GAN overview

GANs at celebrity faces

Helping ShopBop to design new shoes using progressive GANs



Generative adversarial networks (GANs)

aws SUMMIT

BERT: SOTA for language modeling

aws SUMMIT

Natural language processing example

Question answering

Question: Who shall use GluonNLP?

Passage context: GluonNLP provides implementations of the state-of-the-art (SOTA) deep learning models in NLP, and build blocks for text data pipelines and models. It is designed for, engineers, researchers, and students to fast prototype research ideas and products based on these models.

Representation learning in NLP

Word embeddings

Vector representations of words

Word2Vec (shallow word embeddings)

Training

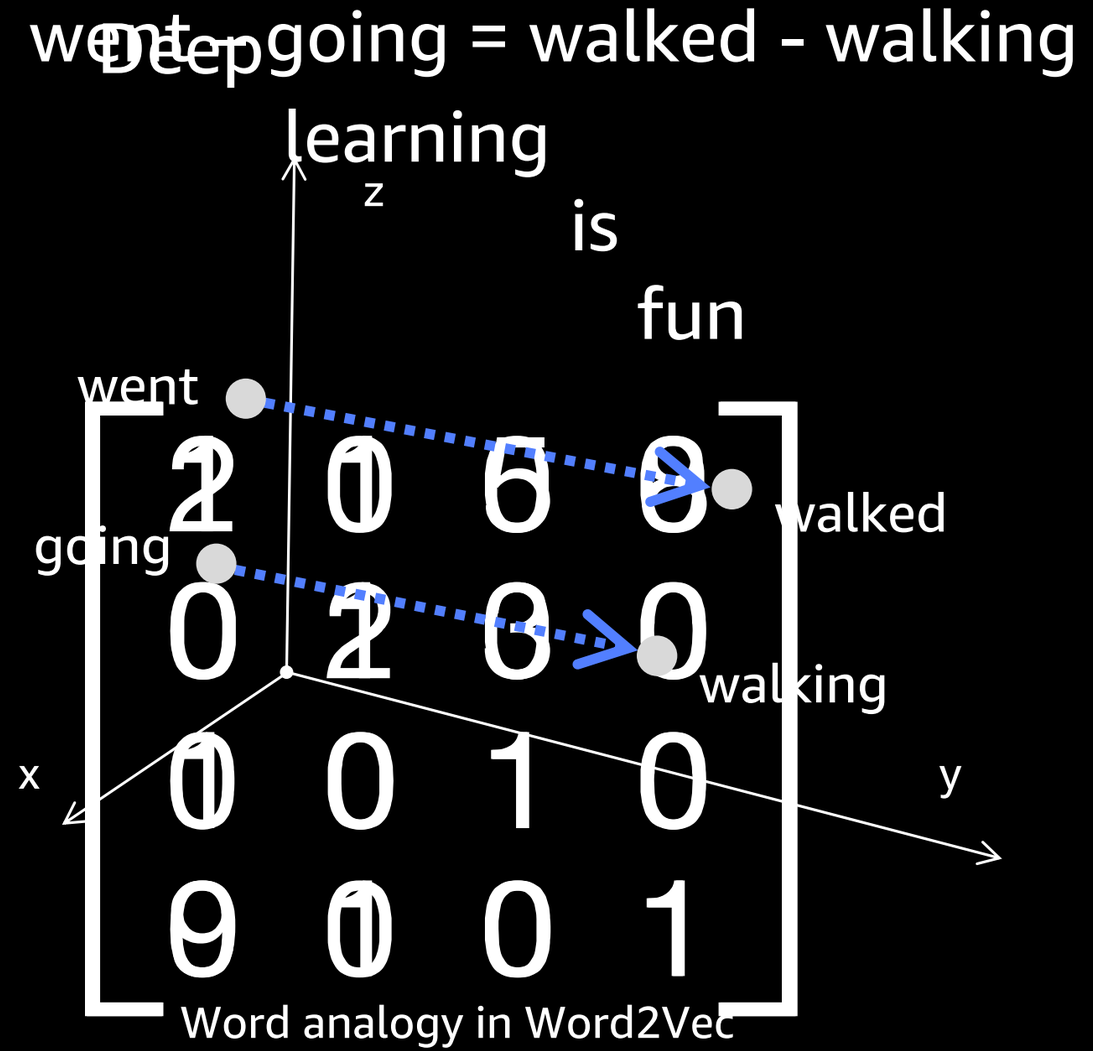
- Models central words given context words

Deep **learning** is fun!

P(learning | deep, is, fun)

Prediction

- Inferences via vector lookups



aws SUMMIT

Representation learning with BERT

Word embeddings

Vector representations of words

Word2Vec (shallow)

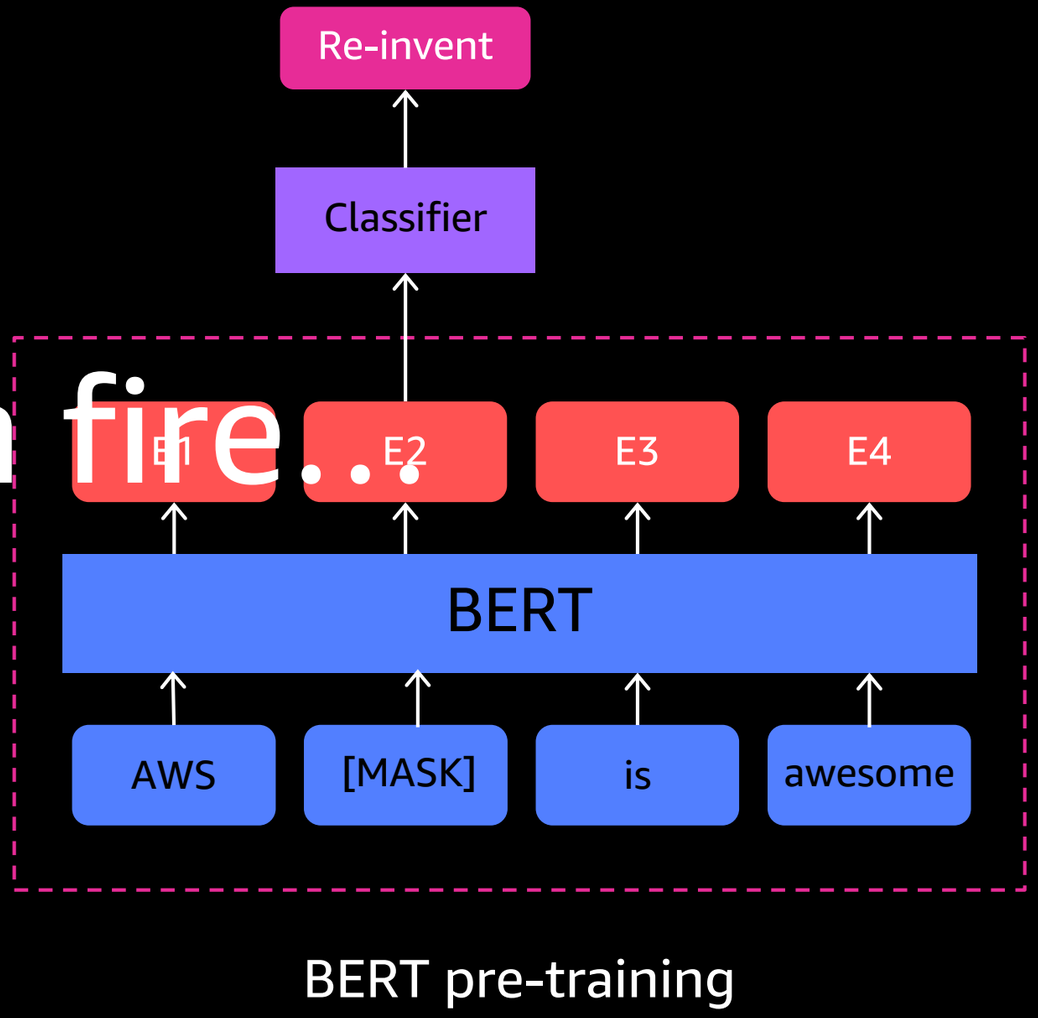
BERT (deep)

Bidirectional, “contextual”, deep

Masked language modeling

AWS [MASK] is awesome.

Outputs: P(re-invent | AWS, [MASK], is, awesome)



BERT pre-training

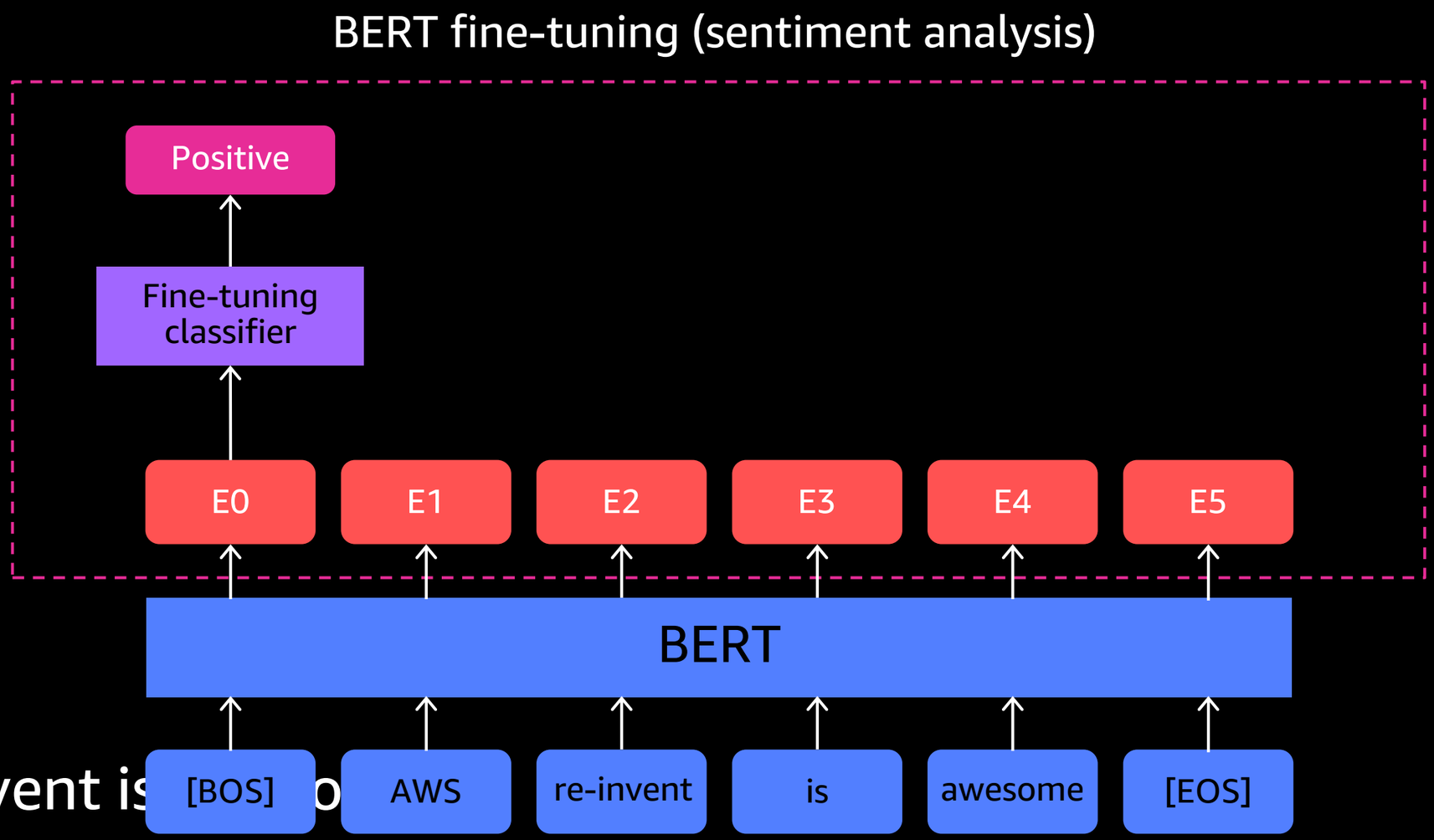
BERT fine-tuning

Sentiment analysis

Output: positive

Embedding:

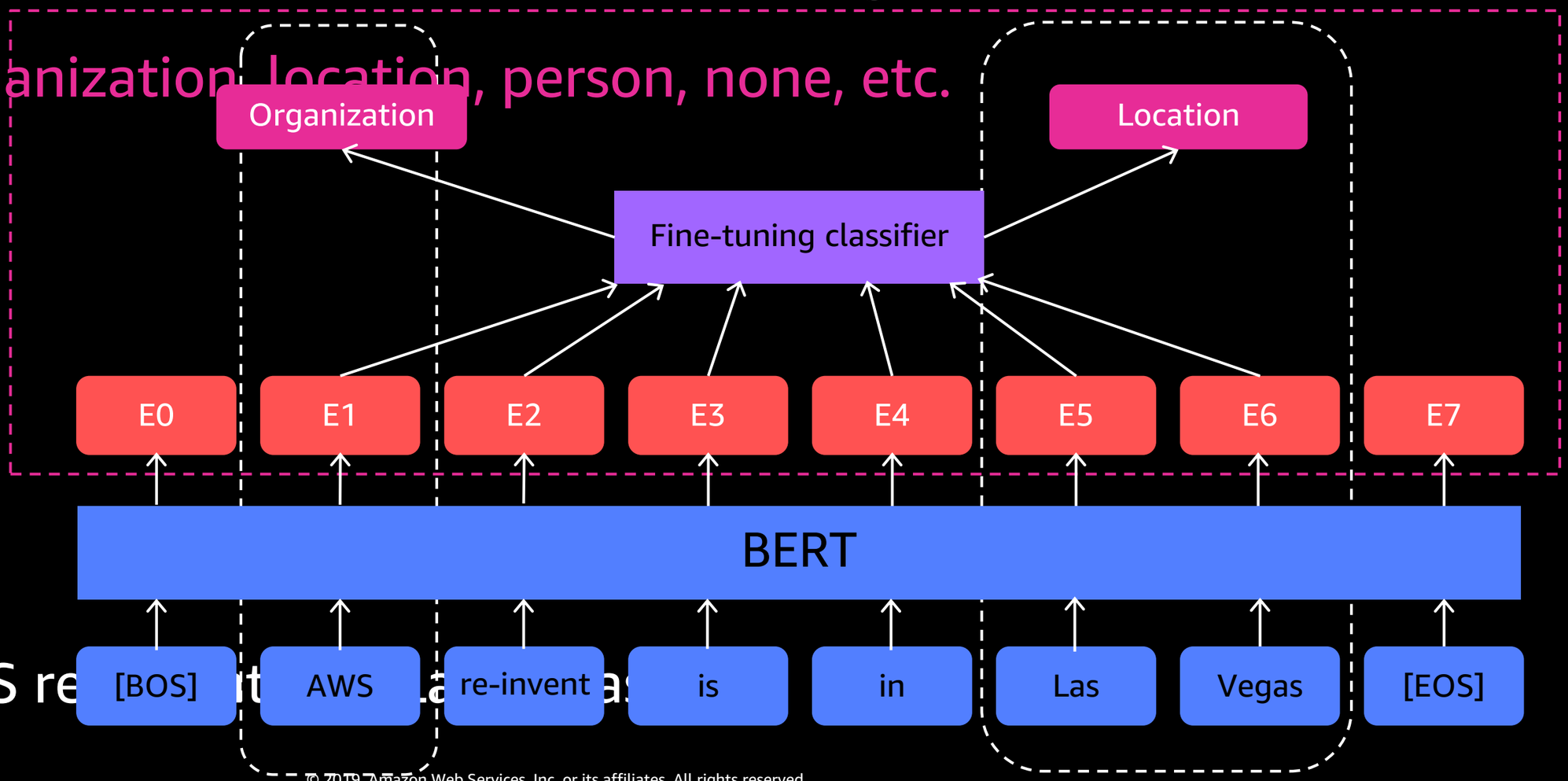
Input: AWS re-invent is



aws SUMMIT

BERT fine-tuning

Name entity recognition (NER)

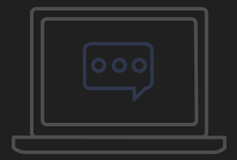


GluonNLP: a natural language toolkit

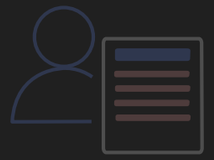
- State-of-the-art models
- Fast development
- Easy deployment

Multiple built-in NLP tasks

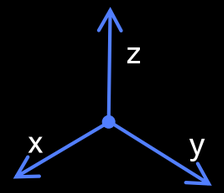
Sentiment analysis



Text generation



Named entity recognition



Representation learning

Machine translation

Question answering

Language modeling

aws SUMMIT

GluonNLP: a natural language toolkit

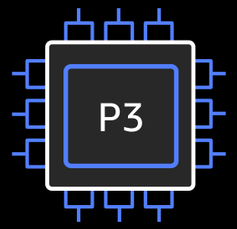
State-of-the-art models (pre-trained and end-to-end)

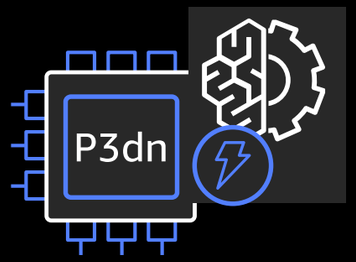
BERT, XLNet, GPT-2, Transformer-XL, FastText, etc.

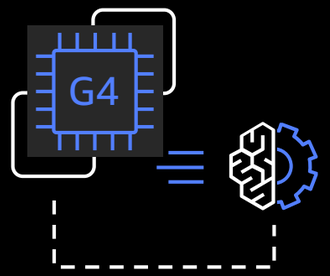
```
model, vocab = gluonnlp.model.get_model(model_name, dataset_name)
```

	Gluonnlp
Stanford sentiment treebank	**95.3 (+1.8%)**
Stanford question answering dataset	**91.0 (+2.5%)**
Recognizing textual entailment	**73.6 (+7.2%)**

Accelerated compute portfolio for machine learning







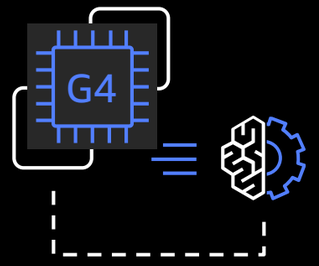
ML training

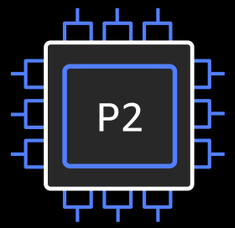
P3/P3dn GPU compute instance

- Up to 1 PetaFLOP of compute with 8x NVIDIA V100 GPUs
- Up to 256 GB of GPU memory
- Up to 100 Gbps of networking
- Designed to handle large distributed training jobs for fastest time to train

G4: GPU compute instance

- Up to 520 TeraFLOPs of compute with 8x NVIDIA T4 GPUs
- Cost-effective small-scale training jobs





ML inference

G4: GPU compute instance

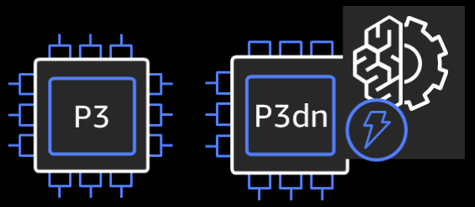
- Up to 1030 TOPs of compute with 8x NVIDIA T4 GPUs
- Increased performance, lower latency and reduced cost per inference compared to previous GPU based instances

P2: GPU compute instance

- Up to 160 TeraFLOPs of compute with 16x NVIDIA K80 GPUs
- General purpose GPU compute

P3 instances

The fastest, most powerful GPU instances in the cloud



Ideal for workloads needing massive parallel processing power

Training machine learning model

Running HPC simulations

Rendering 3D models

Video encoding

Up to eight NVIDIA Tesla V100 GPUs

1 PetaFLOPs of computational performance *—up to 14x better than P2*

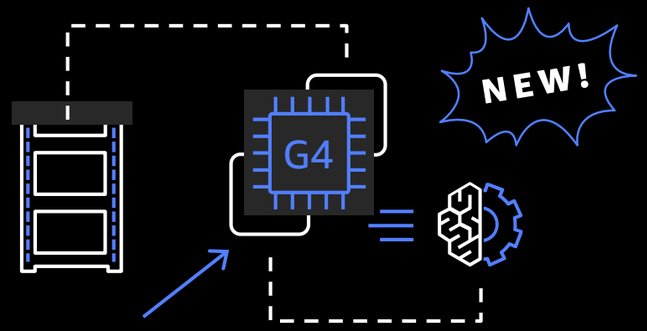
300 GB/s GPU-to-GPU communication (NVLink) *—9X better than P2*

Support all ML frameworks and model types

Available as on-demand, reserved and spot instances with up to 70% discount

Instance size	GPUs	GPU memory	GPU peer to peer	vCPUs	Memory (GB)	Network bandwidth	Amazon EBS bandwidth	On-demand price/hr.*	1-yr RI effective hourly*	3-yr RI effective hourly*
P3.2xlarge	1	16 GB	No	8	61	Up to 10 Gbps	1.7 Gbps	$3.06	$1.99 (35% disc.)	$1.05 (60% disc.)
P3.8xlarge	4	64 GB	NVLink	32	244	10 Gbps	7 Gbps	$12.24	$7.96 (35% disc.)	$4.19 (60% disc.)
P3.16xlarge	8	128 GB	NVLink	64	488	25 Gbps	14 Gbps	$24.48	$15.91 (35% disc.)	$8.39 (60% disc.)
P3dn.24xlarge	8	256 GB	NVLink	96	768	100 Gbps	14 Gbps	$31.21	$18.30 (41% disc.)	$9.64 (69% disc.)

AWS G4 GPU instances



Designed for machine learning inferencing, video transcoding, remote graphics workstation and other demanding graphics applications

Up to 8 NVIDIA T4 Tensor Core GPUs

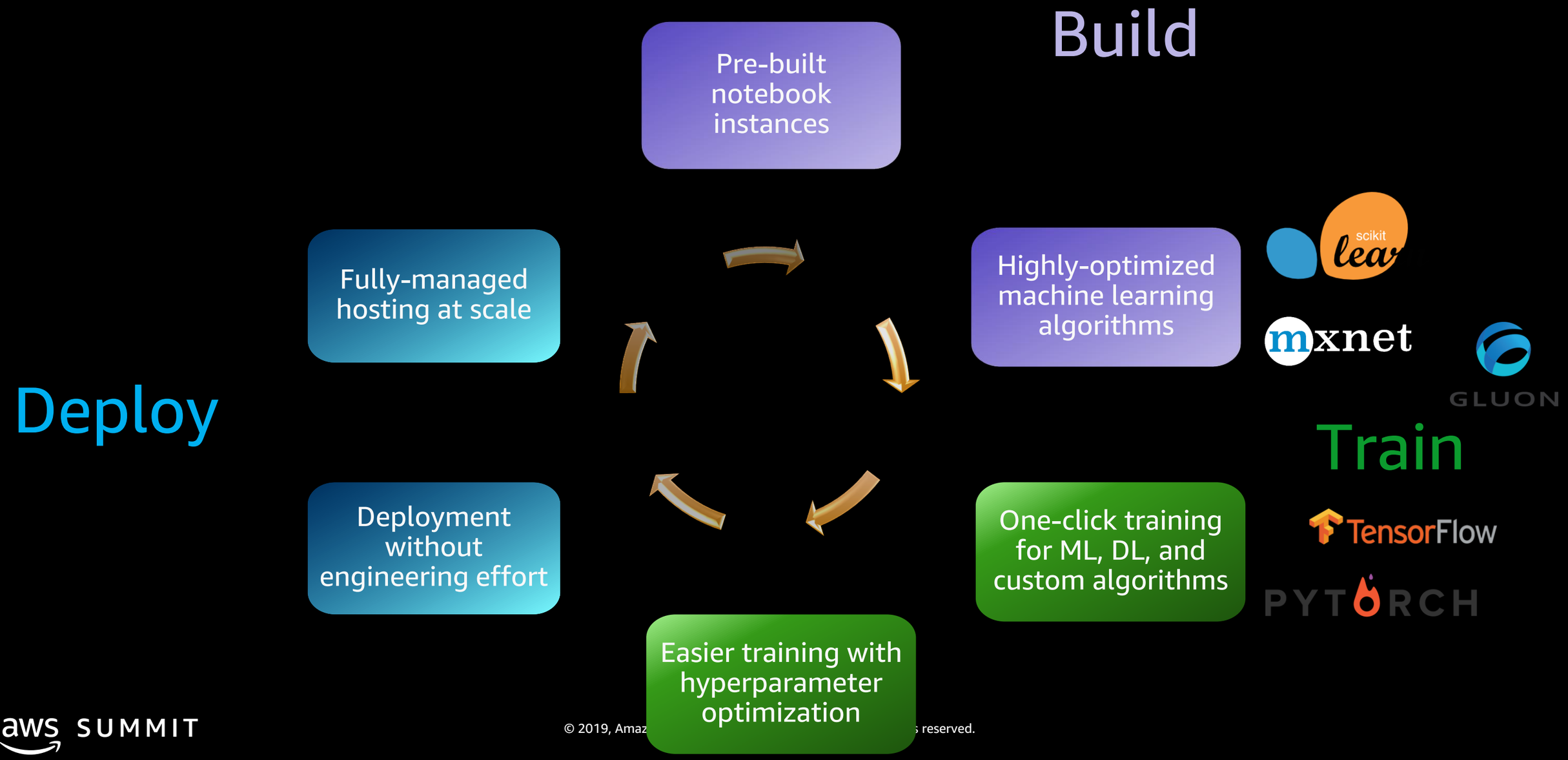
2560 CUDA Cores, 320 Turing Codes including support for Ray-Tracing technology

	Instance size	vCPUs	Memory (GB)	GPU	GPU memory	Storage (GB)	Network bandwidth (Gbps)	EBS bandwidth (GBps)	On-demand price/hr*	1-yr reserved instance effective hourly* (Linux)	3-yr reserved instance effective hourly* (Linux)
Single GPU VMs	g4dn.xlarge	4	16	1	16 GB	125	Up to 25	Up to 3.5	$0.526	$0.316	$0.210
	g4dn.2xlarge	8	32	1	16 GB	225	Up to 25	Up to 3.5	$0.752	$0.452	$0.300
	g4dn.4xlarge	16	64	1	16 GB	225	Up to 25	Up to 3.5	$1.204	$0.722	$0.482
	g4dn.8xlarge	32	128	1	16 GB	1x900	50	7	$2.176	$1.306	$0.870
	g4dn.16xlarge	64	256	1	16 GB	1x900	50	7	$4.352	$2.612	$1.740
Multi GPU VMs	g4dn.12xlarge	48	192	4	64 GB	1x900	50	7	$3.912	$2.348	$1.564
	g4dn.metal**	96	384	8	128 GB	2x900	100	14	Coming soon	Coming soon	Coming soon

Amazon SageMaker

A **fully managed service** that enables **data scientists** and **developers** to quickly and easily **build** machine-learning based models **into production** smart applications.

Amazon SageMaker



Amazon SageMaker: Launch Customers

intuit

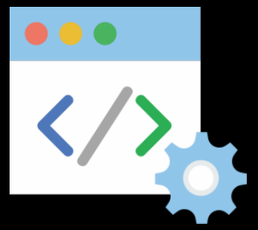
ZipRecruiter®

Hotels.com

THOMSON REUTERS®

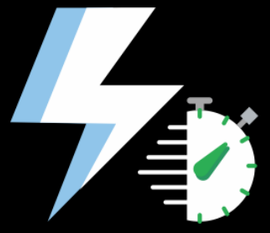
Amazon SageMaker

1



Notebook Instances

2



Algorithms

3

ML Training Service

4

ML Hosting Service

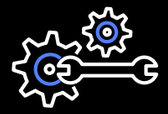
aws SUMMIT

Amazon SageMaker

Bringing machine learning to all developers

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
Pre-built notebooks for common problems	Built-in, high performance algorithms	One-click training	Optimization	One-click deployment	Fully managed with auto-scaling, health checks, automatic handling of node failures, and security checks



End-to-end machine learning platform

Flexible model training







Pay by the second

Amazon SageMaker

Deploy

Fully-managed hosting at scale

Deployment without engineering effort

Build

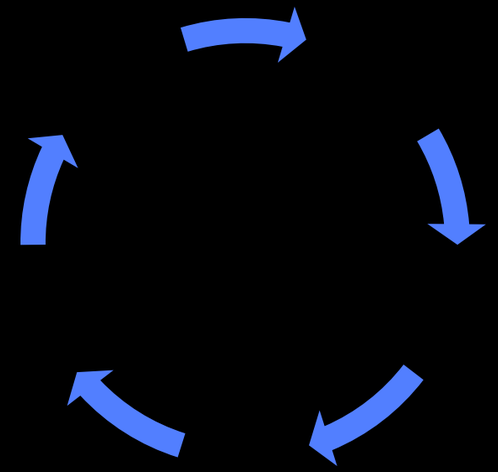
Pre-built notebook instances

Highly-optimized machine learning algorithms

One-click training for ML, DL, and custom algorithms

Train

Easier training with hyperparameter optimization



scikit learn

GLUON

TensorFlow

PyTorch

mxnet

New!

AWS deep learning containers

Quickly set up deep learning environments with optimized, pre-packaged Docker container images

Get best performance automatically; no tuning required

Support TensorFlow and Apache MXNet

Deploy on Amazon ECS, Amazon EKS, or Amazon EC2

Customizable container images

Amazon ECR

AWS Marketplace

Available at no cost from Amazon Elastic Container Registry (Amazon ECR) and AWS Marketplace

aws SUMMIT

Hands-on labs

1. Object detection (SSD)
2. Sentiment analysis (BERT)

Ask lab assistant for access codes.

URL: http://github.com/awshlabs/

Resources

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

http://gluon-nlp.mxnet.io/

http://gluon-cv.mxnet.io/

https://gluon-ts.mxnet.io/

http://d2l.ai/

https://discuss.mxnet.io/

NVIDIA GPUs in the Amazon Cloud: A Best Practice for Artificial Intelligence designed for National Security

(Presented by AWS) - DC91466
Tuesday, November 5 at 12:30pm-1:20pm

Performing near real-time inference on live video streams is notoriously tricky, yet knowing accurate information as quickly as possible is critical for National Security agencies. By combining the scale and agility of AWS with the NVIDIA's technology, agencies can build AI applications that solve tough mission-critical problems that improve national security and strengthen the nation.

AWS, working together with NVIDIA, has built an innovative workflow that leverages the NVIDIA DeepStream SDK, AWS's recently released G4 instance type, and additional AWS services to perform inference and analytics on video. In this session, we will discuss how organizations can build and execute artificial intelligence strategies with NVIDIA in the Amazon cloud which address current and future National Security mission needs.

Learn More: aws.amazon.com/ec2/instance-types/g4/

Brad Kenstler,
Senior Data Scientist, Amazon Web Services

Kate Werling,
Senior Solutions Architect, Amazon Web Services

Scott Junkins,
NVIDIA Deep Learning

aws

Build Today, Fly Tomorrow: How Lockheed Martin is Transforming and Modernizing Its Aircraft Manufacturing with Amazon Web Service

(Presented by AWS) - DC91468
Wednesday, November 6 at 10:00am-10:50am

Digital Transformation has enabled new efficiencies, innovative products, and close customer relationships globally through the effective use of IIoT (Industrial Internet of Things), and cloud technologies to generate models for better decisions.

This session will present how Lockheed Martin's Intelligent Factory is changing how we do business and changing the world through Build Today, Fly Tomorrow. Lockheed Martin has created the defense industry's first intelligent factory to increase visibility of physical and logical production through the product lifecycle. Using a combination of AWS IoT and AI cloud services and emerging technologies, the intelligent factory improves overall equipment effectiveness and capacity planning for multiple lines of business.

Learn more: aws.amazon.com/iot/

Dr. Jeff Daniels,
Director of Internet of Things at Lockheed Martin

Matt Gebhardt,
Engineering Program Manager for Advanced Manufacturing (IIOT), Lockheed Martin

Thomas Cummins,
Senior Partner Solutions Architect - IoT, AWS

aws

Thank you!

Wenming Ye
Twitter: @wenmingye
Amazon Email: wye

aws SUMMIT