AWS
re:Invent

CMP307-R

Optimize ML training and inferencing using Amazon EC2

Wen-ming Ye
Senior Solutions Architect AI/ML
Amazon Web Services

Rachel Hu
Applied Scientist
Amazon Web Services

AWS re:Invent

aws

Agenda

Deep learning trend (5 min)

DL architectures: CNN & BERT (30 min)

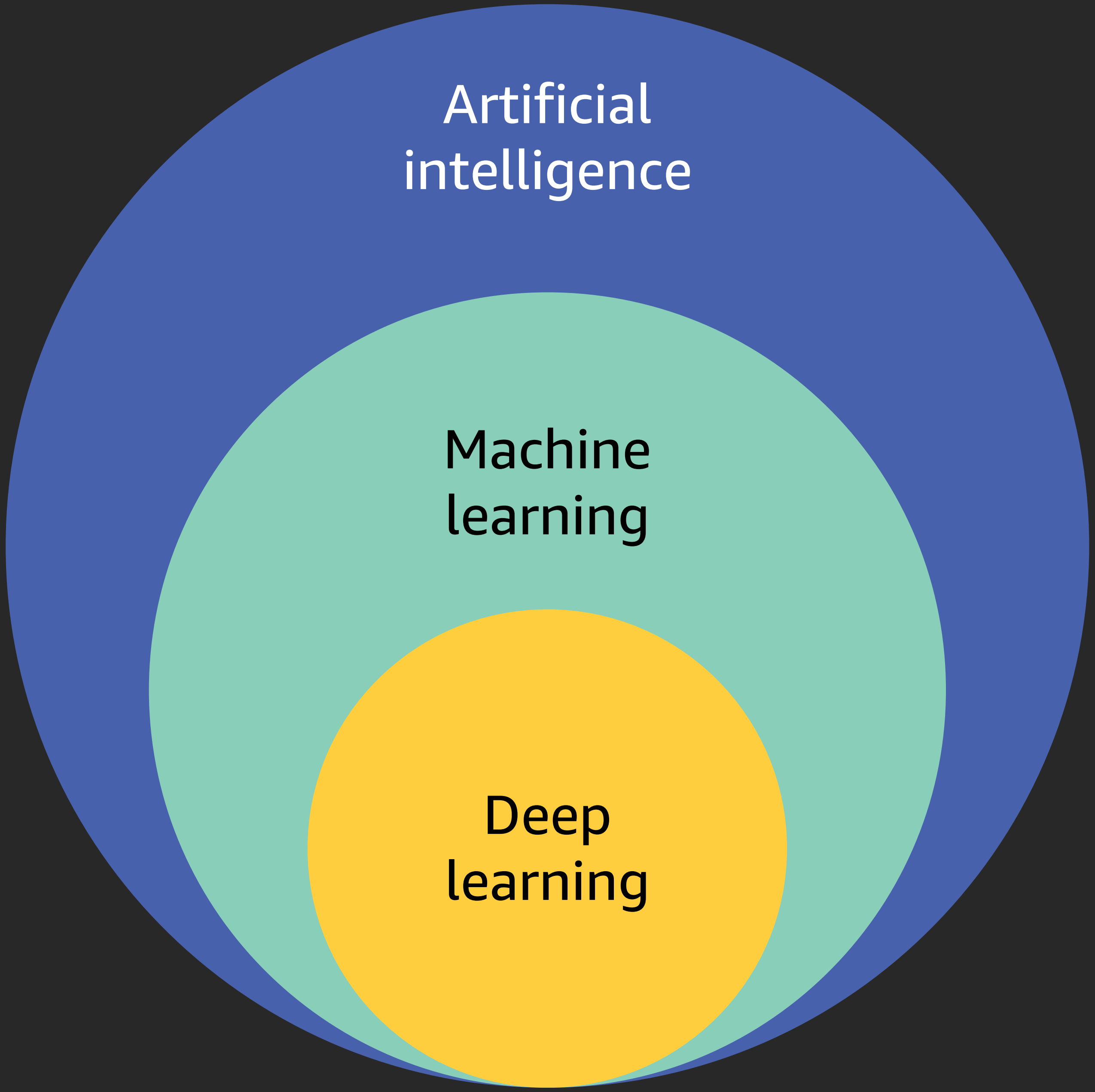
P3 & G4 instances details (5 min)

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

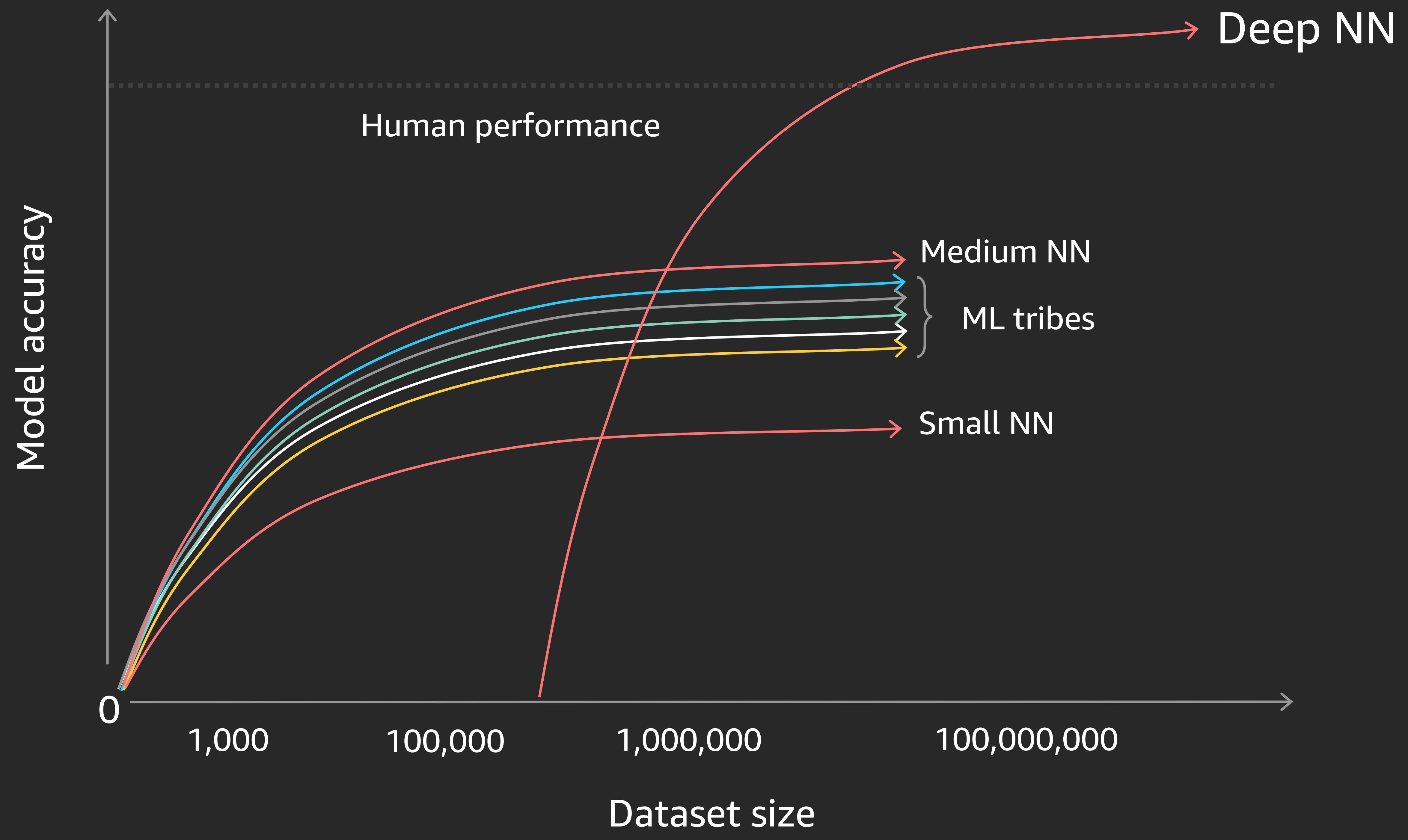
Lab 2: Sentiment analysis (BERT) (30 min)

Learning resources and giveaways (10 min)

DL in context

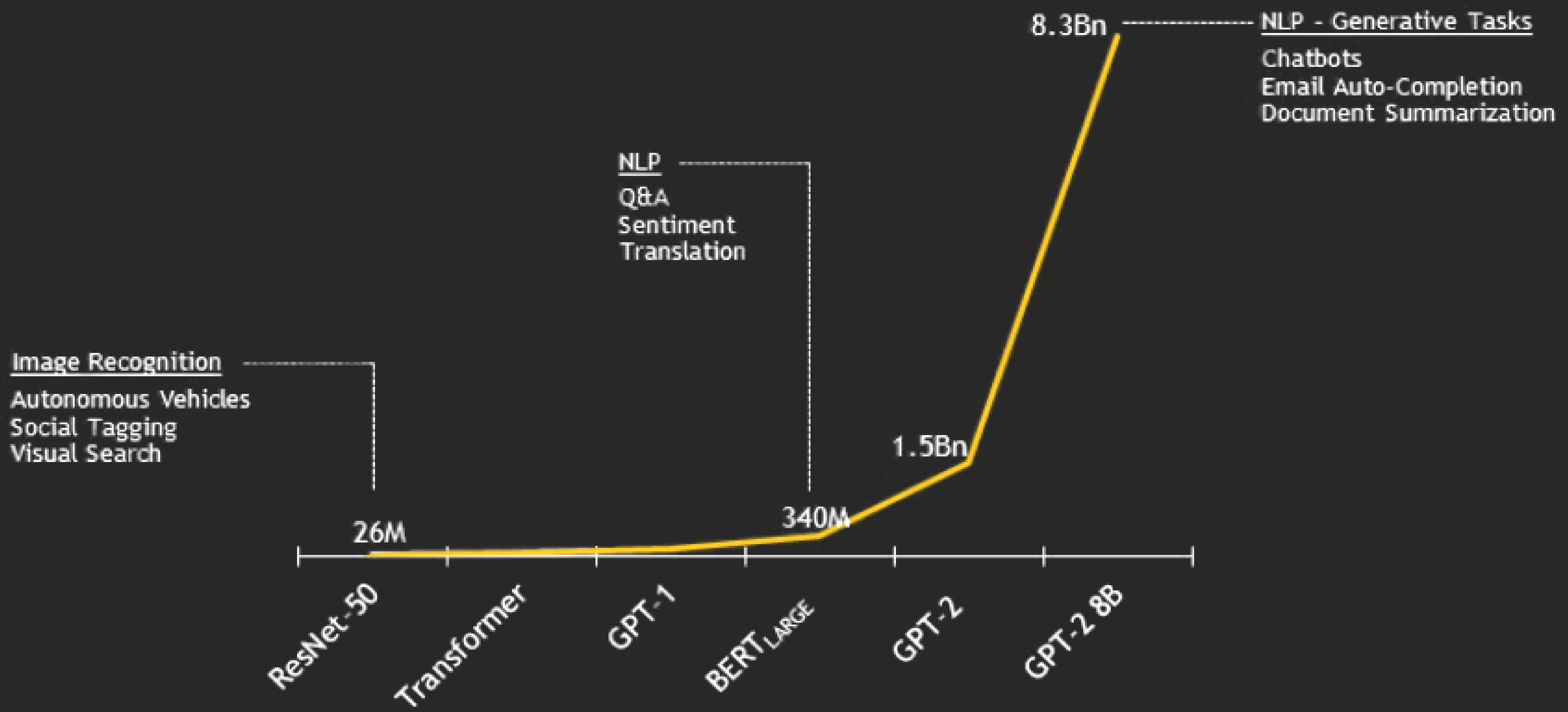


Learning at scale



Exploding model complexity

Number of parameters by network



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

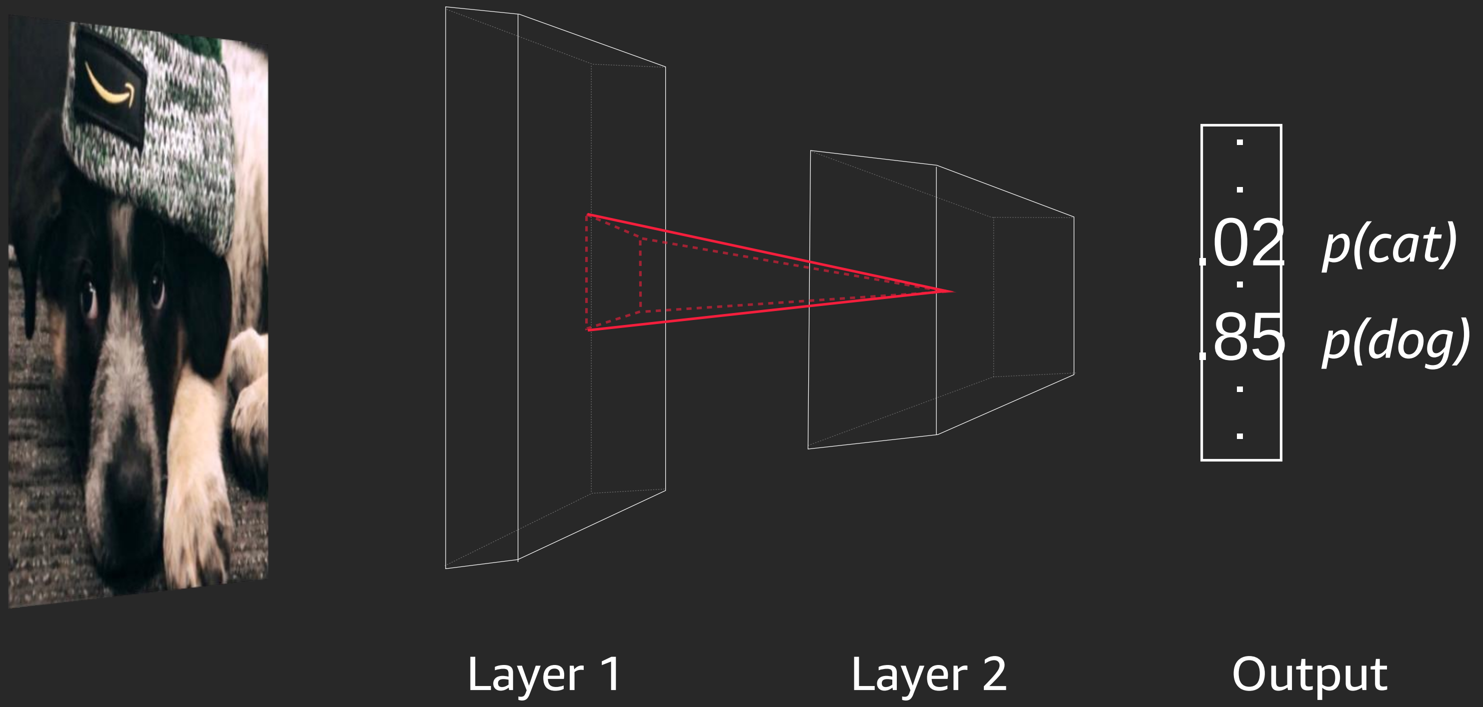
NLU scenarios

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

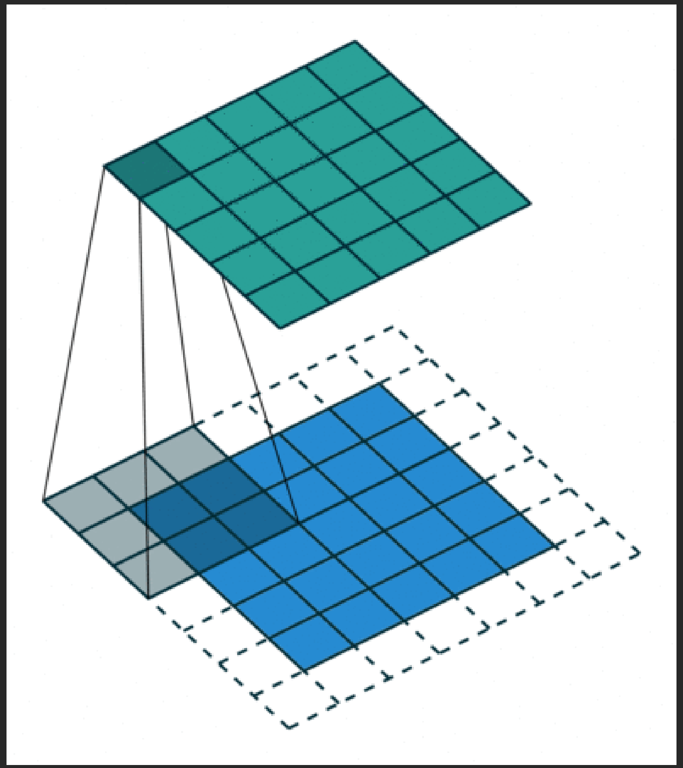
Convolution neural network

Deep learning in computer vision

Explore spatial information with convolution layers



Convolutional neural network



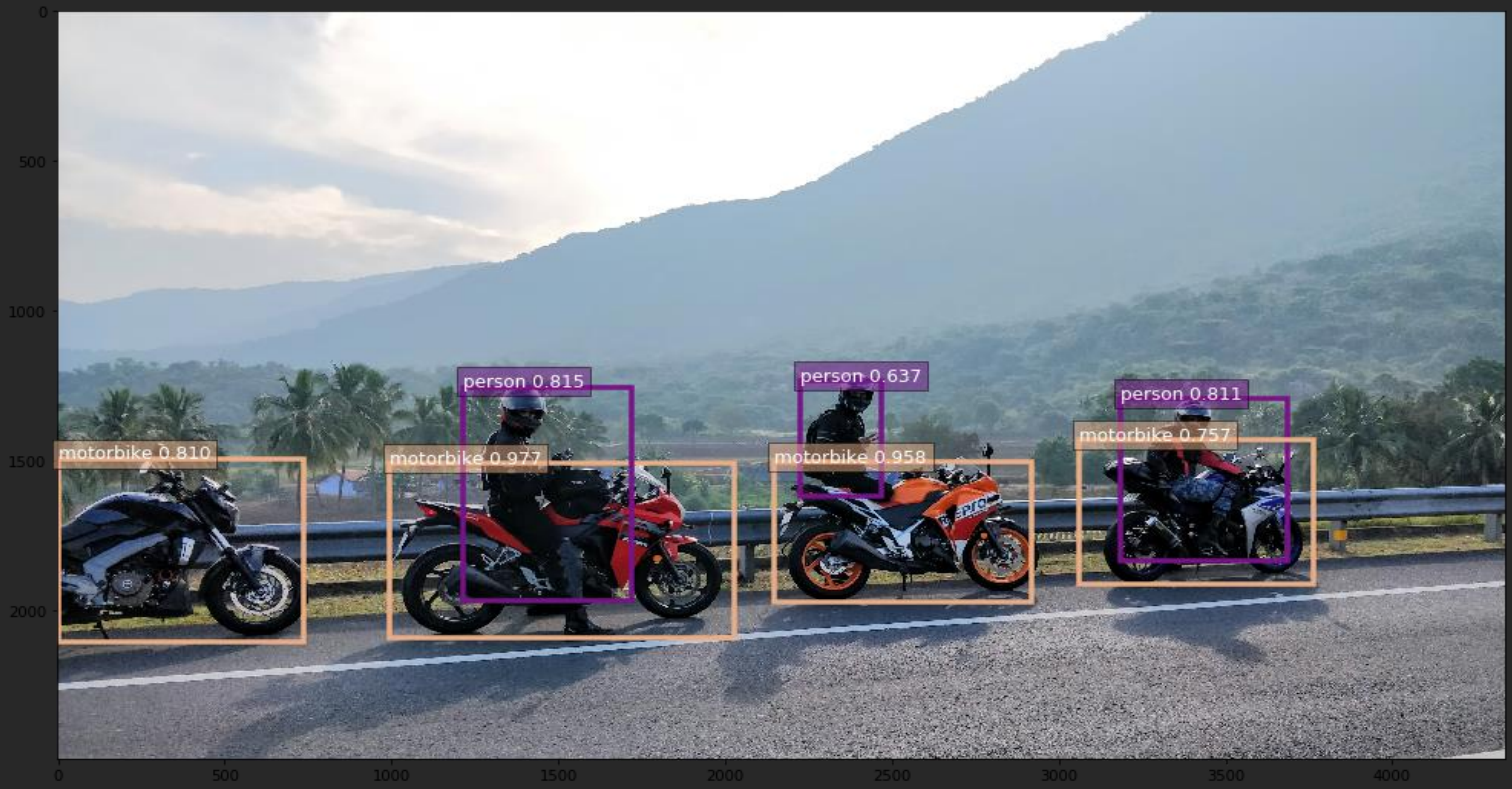
https://leonardoaraujosantos.gitbooks.io/artificial-inteligence

Demo: Convolution neural network

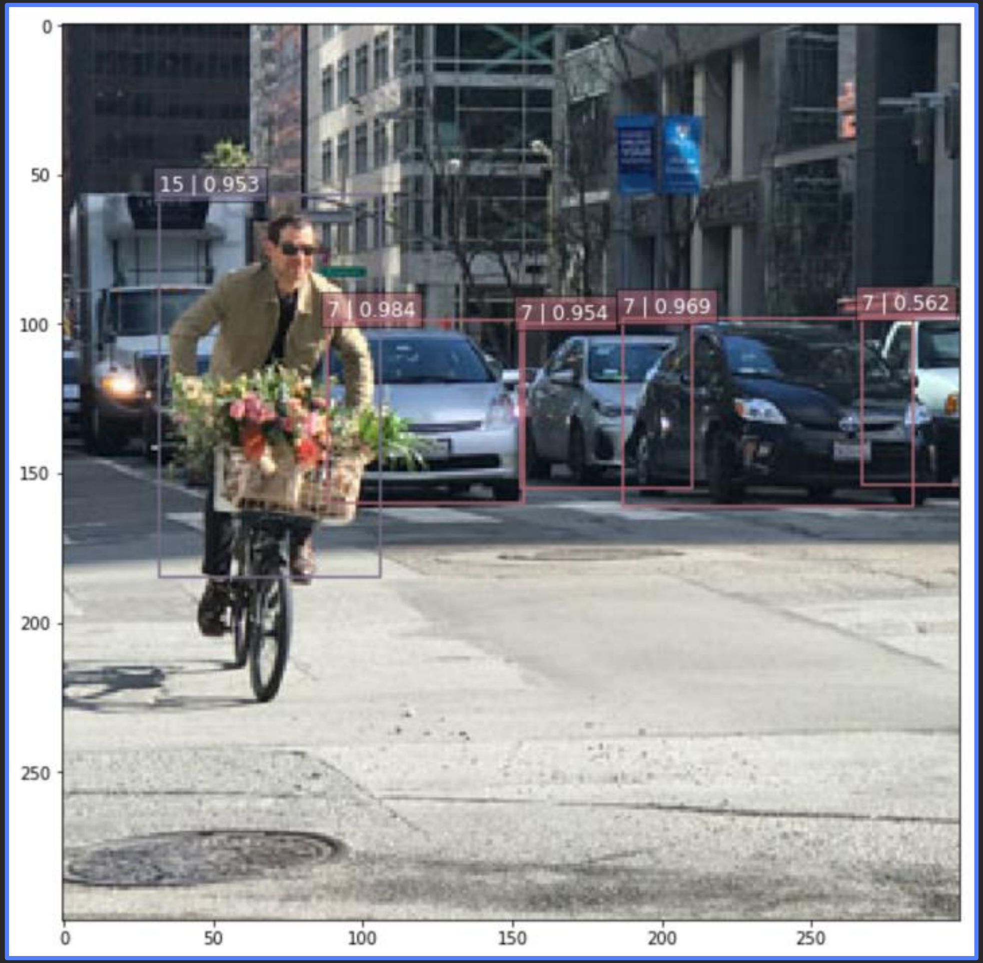
http://scs.ryerson.ca/~aharley/vis/conv

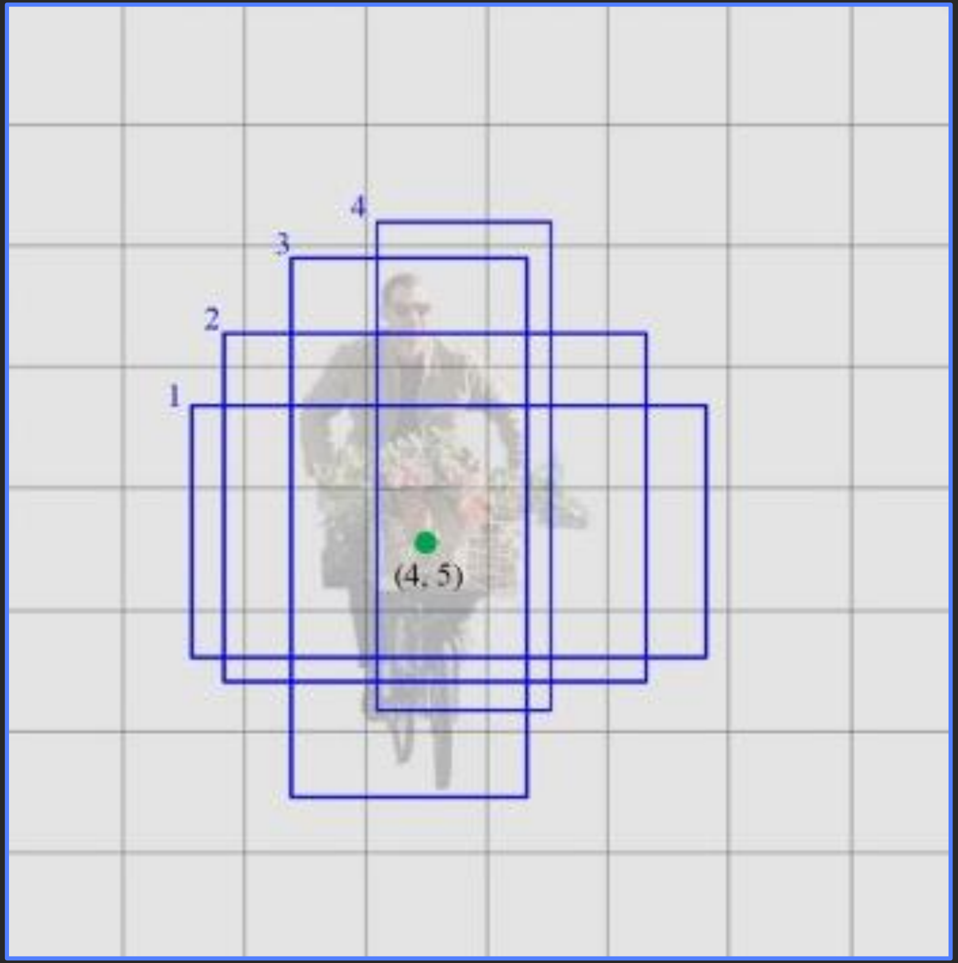
Demo: Object detection

Object detection



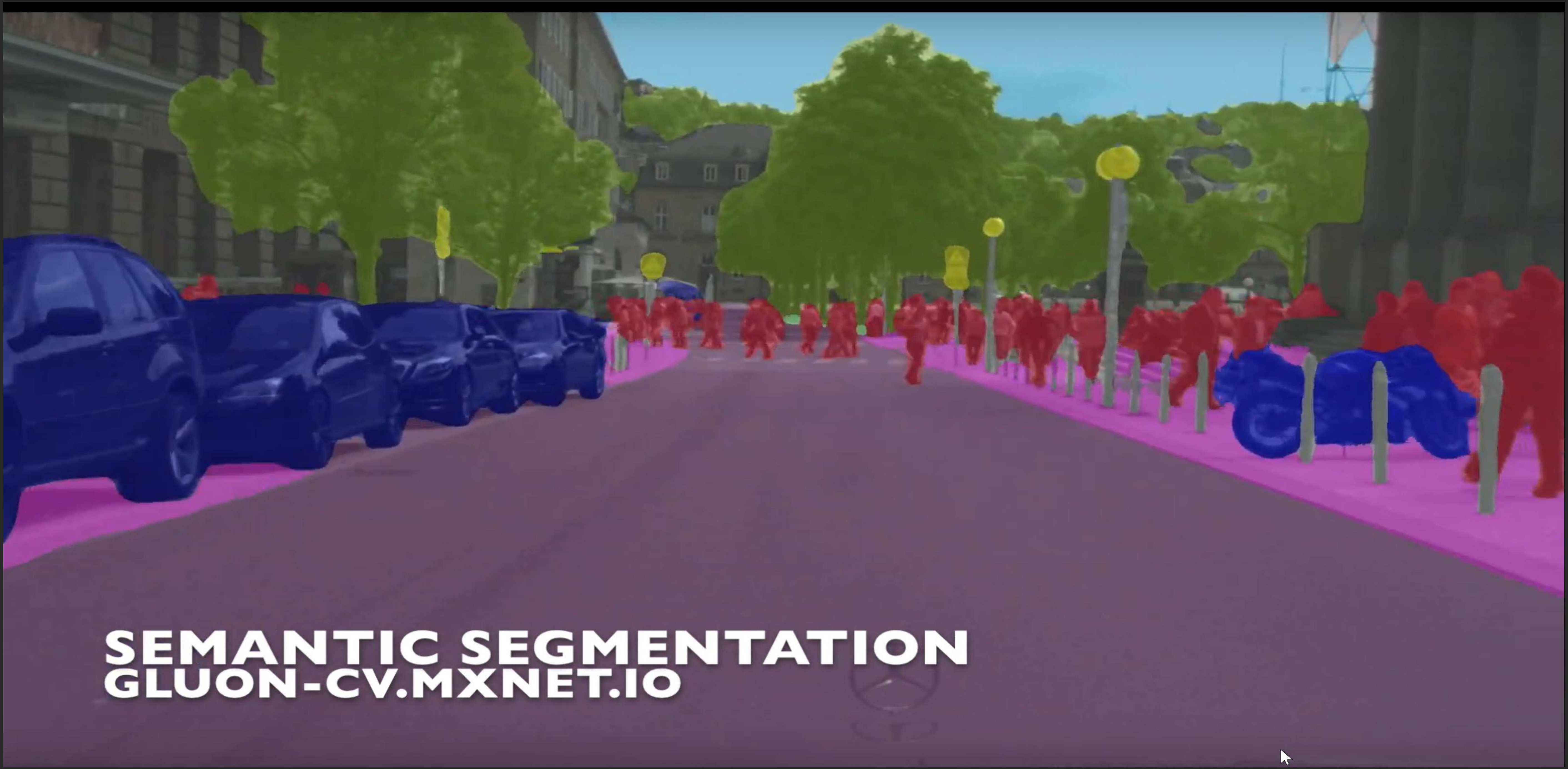
Single shot detector



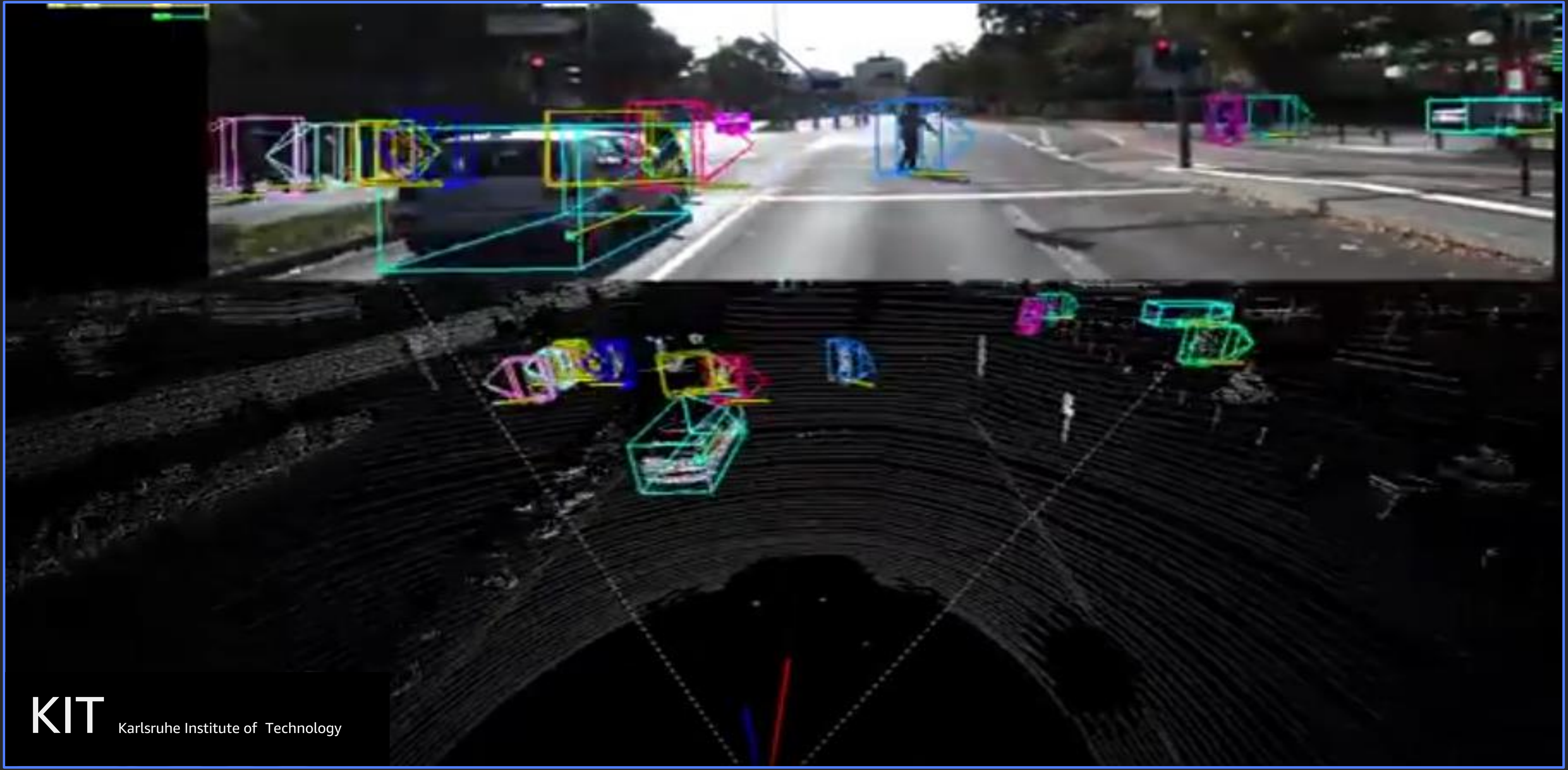


Demo: Image segmentation

Image segmentation

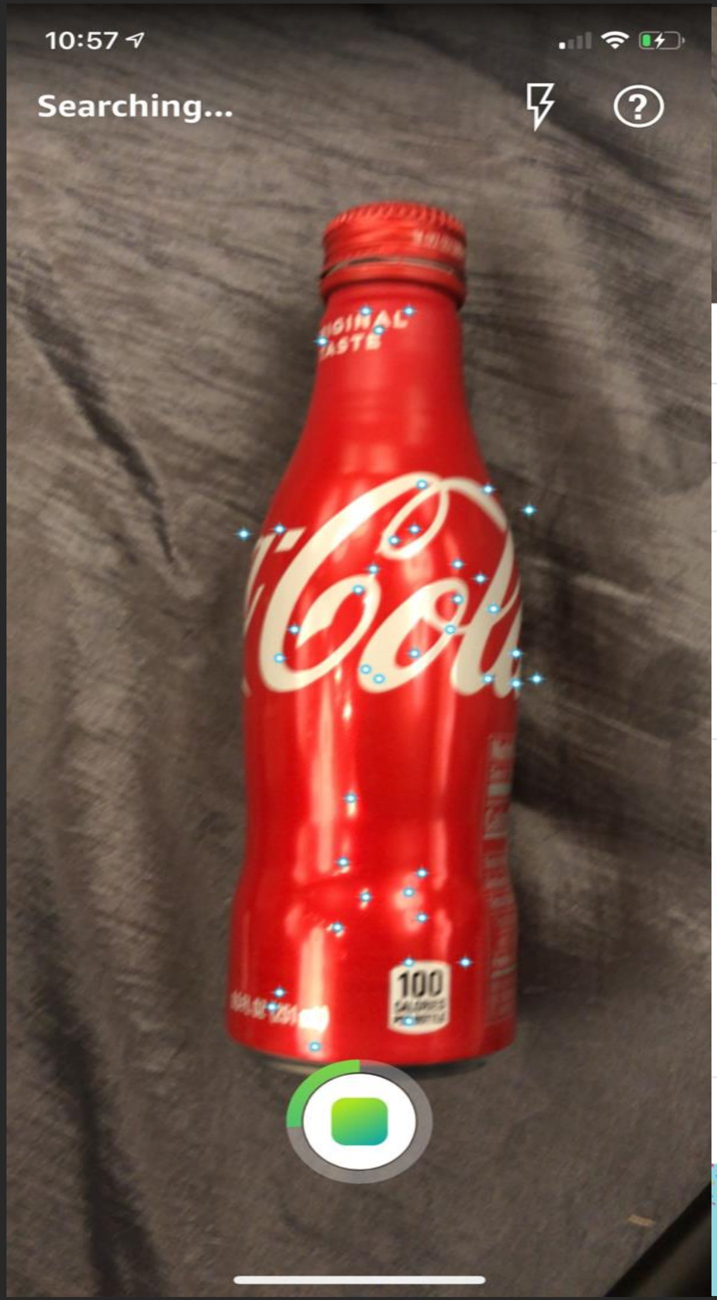


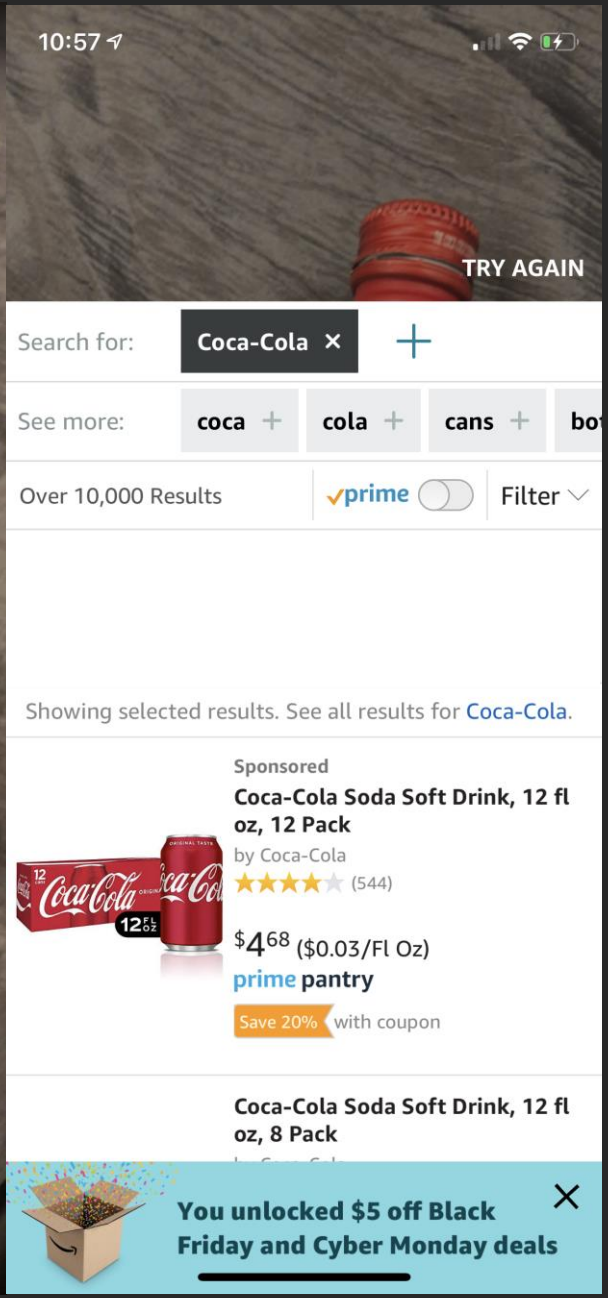
KITTI



Visual search

Visual search





Search for: Coca-Cola × +

See more: coca + cola + cans + bo

Over 10,000 Results ✓prime Filter

Showing selected results. See all results for Coca-Cola.

Sponsored
Coca-Cola Soda Soft Drink, 12 fl oz, 12 Pack
by Coca-Cola
★★★★☆ (544)
$4.68 ($0.03/Fl Oz)
prime pantry
Save 20% with coupon

Coca-Cola Soda Soft Drink, 12 fl oz, 8 Pack

You unlocked $5 off Black Friday and Cyber Monday deals

Pipeline stages

Image query processing

Data normalization/augmentation

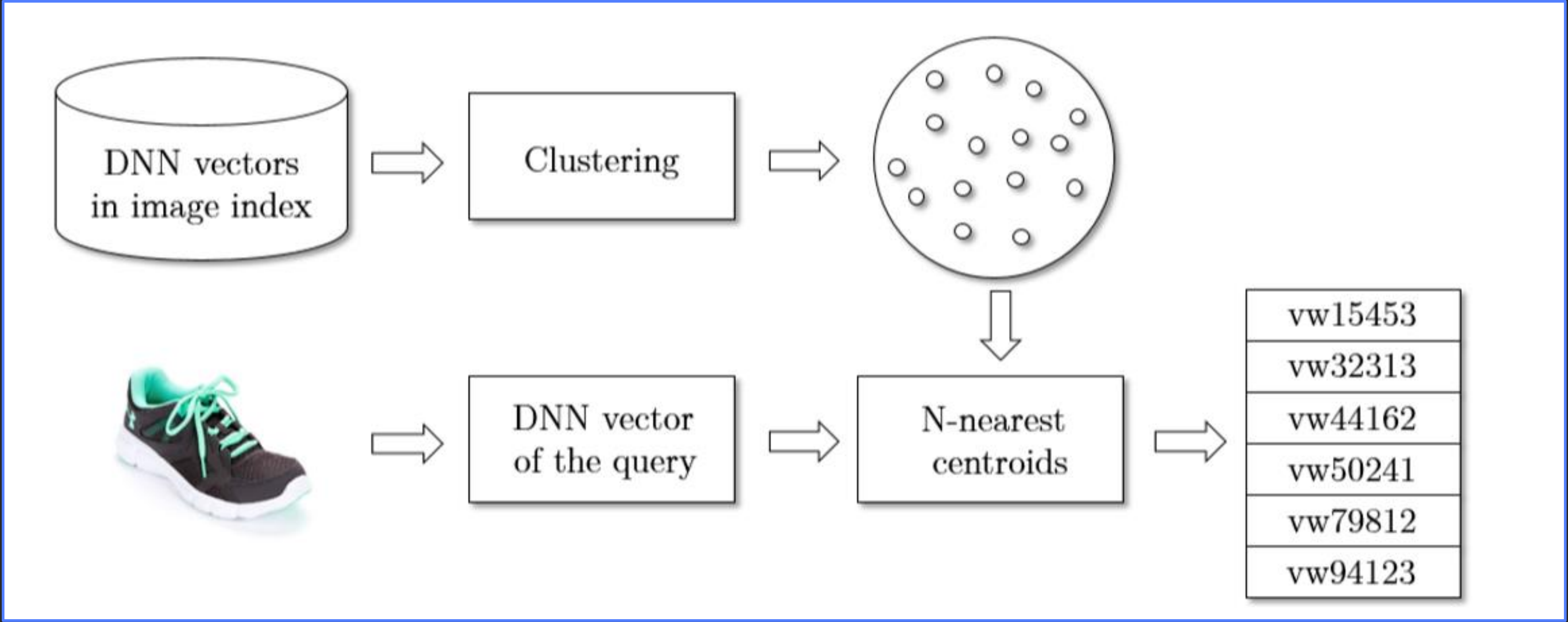
Embedding

DNN model(s)

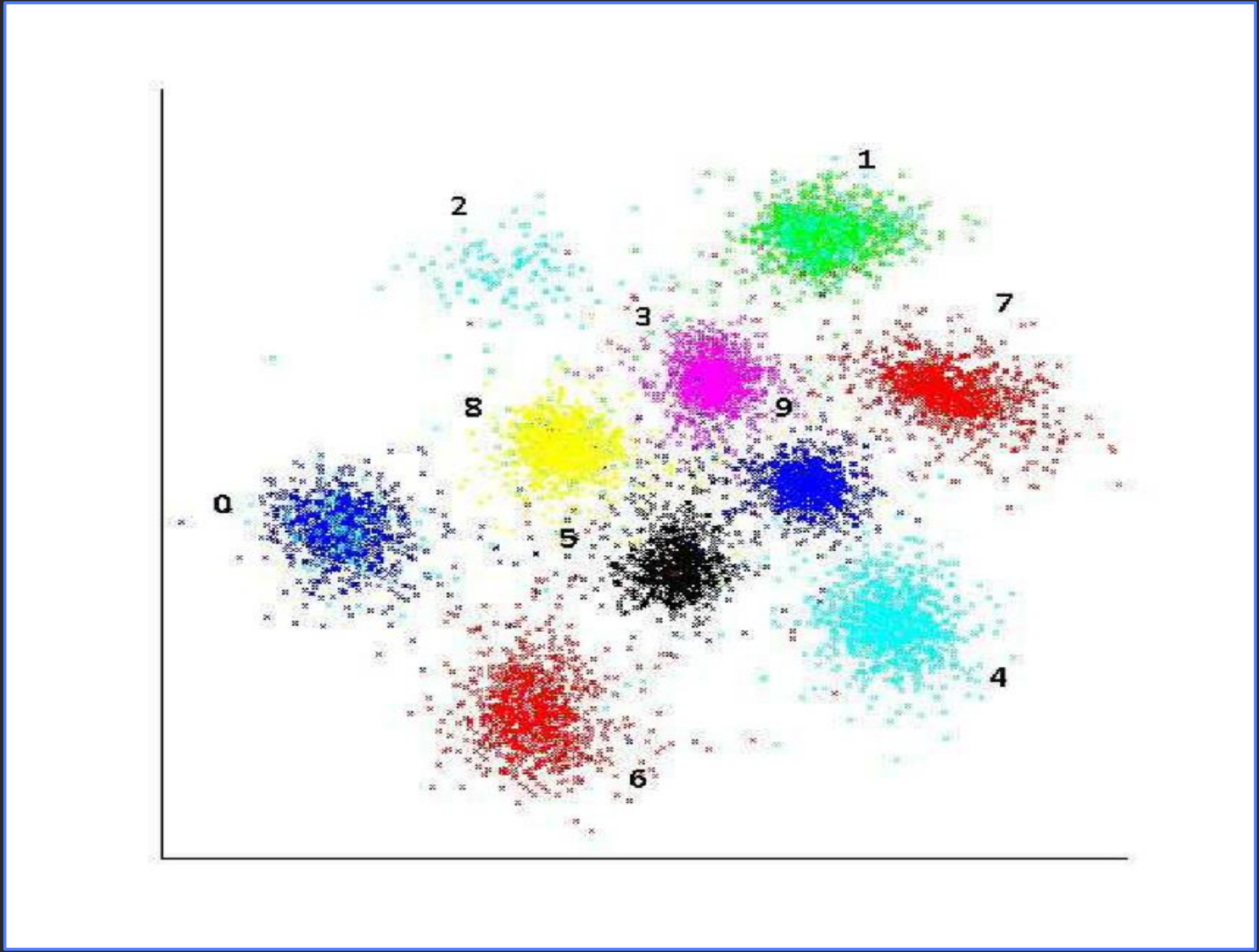
kNN + ranking

Post-processing, de-dup

Architecture



Embedding = learned representation space



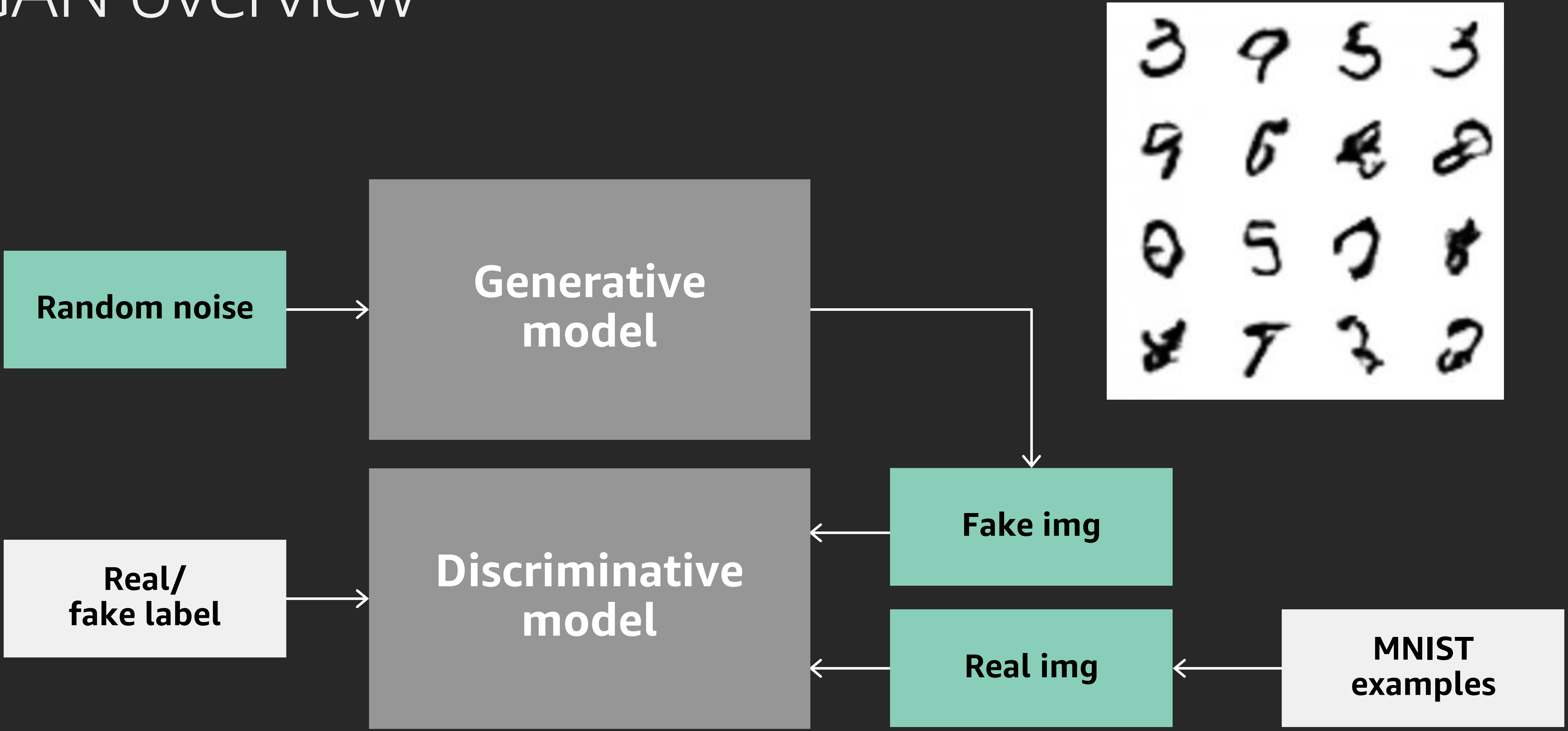
Demo: Image embedding

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 nonadult 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)

GAN overview



Random noise

Generative model

Fake img

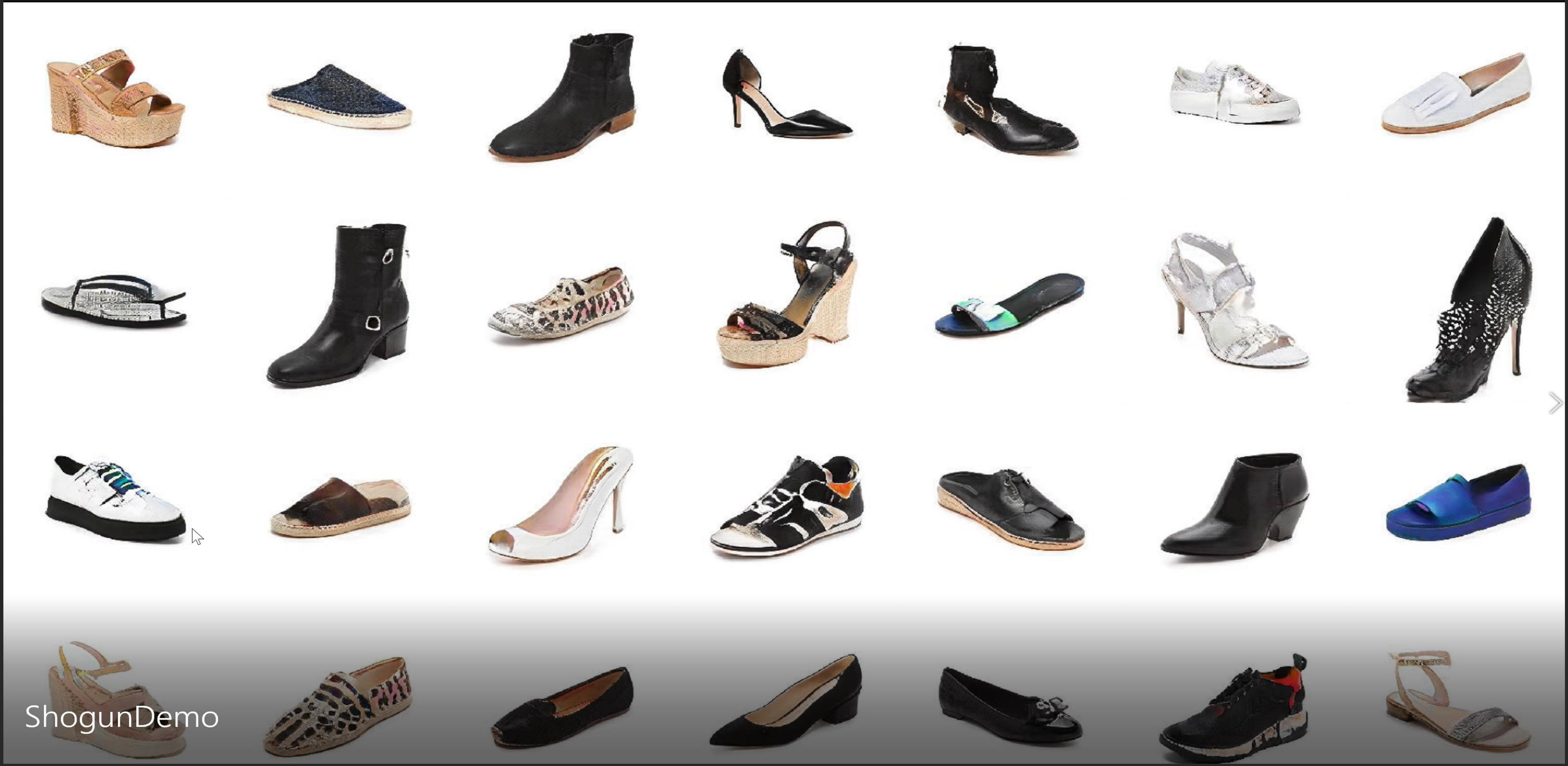
Real/ fake label

Discriminative model

Real img

MNIST examples

Helping ShopBop to Look at AI Shoe Designs



Video: Generative adversarial networks (GANs)

BERT: SOTA for language modeling

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 builds 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

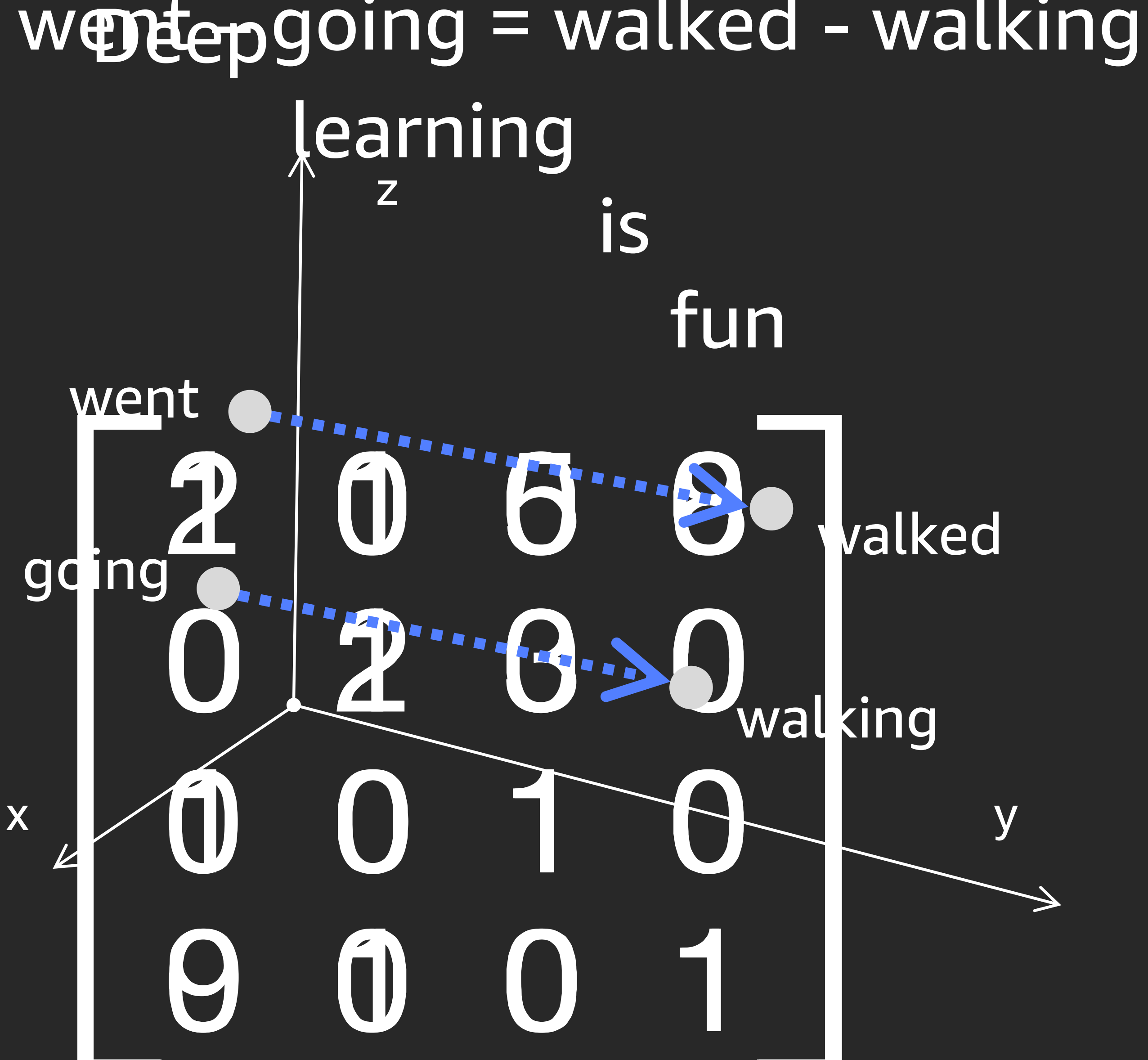
Deep **learning** is fun!

P(learning | deep, is, fun)

Prediction

- Inferences via vector lookups

went - going = walked - walking



Word analogy in Word2Vec

Representation learning with BERT

Word embeddings

Vector representations of words

Word2Vec (shallow)

Amazon is on fire...

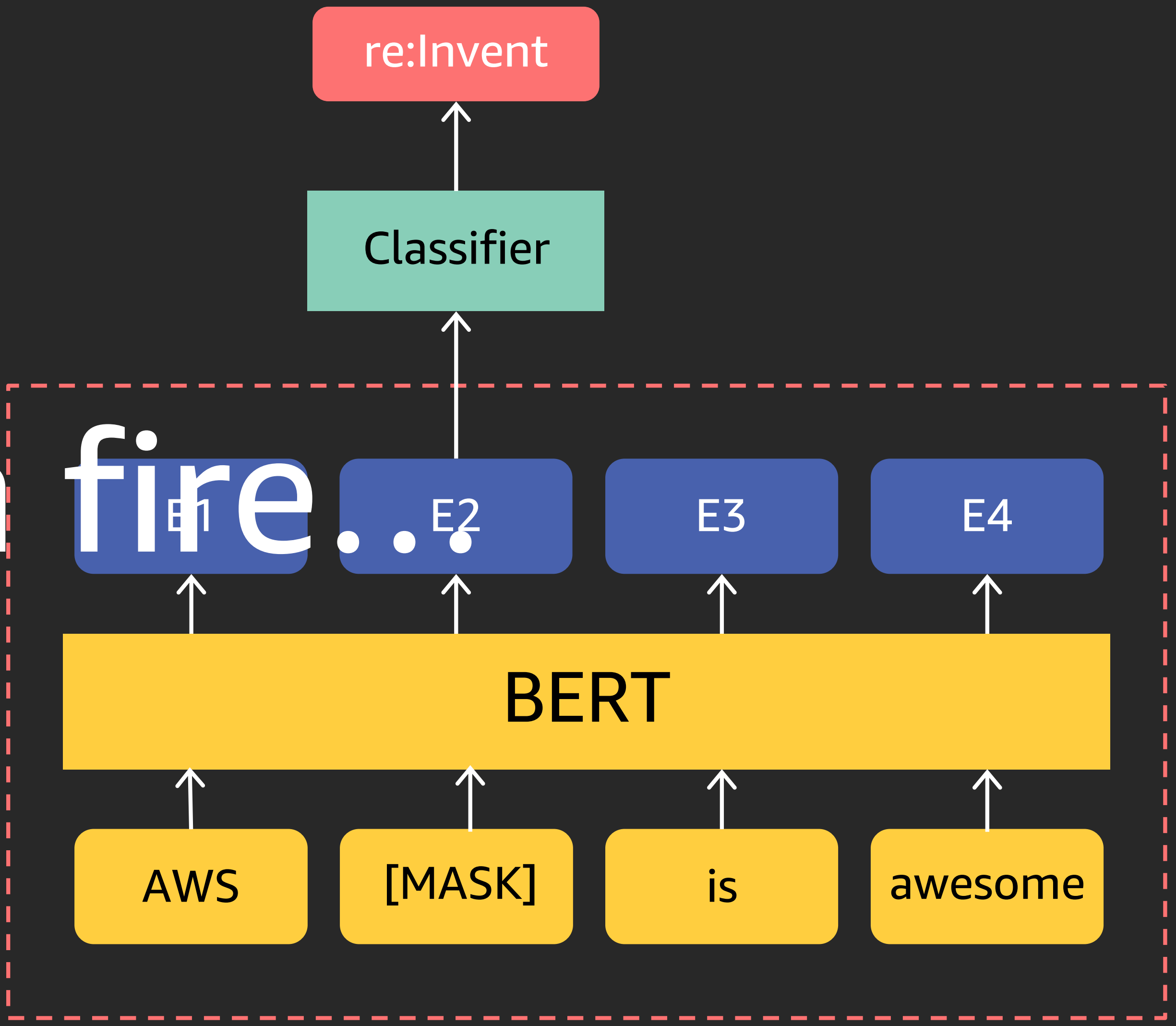
BERT (deep)

Bidirectional, “contextual,” deep

Masked language modeling

AWS [MASK] is awesome

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



BERT pre-training

BERT fine-tuning

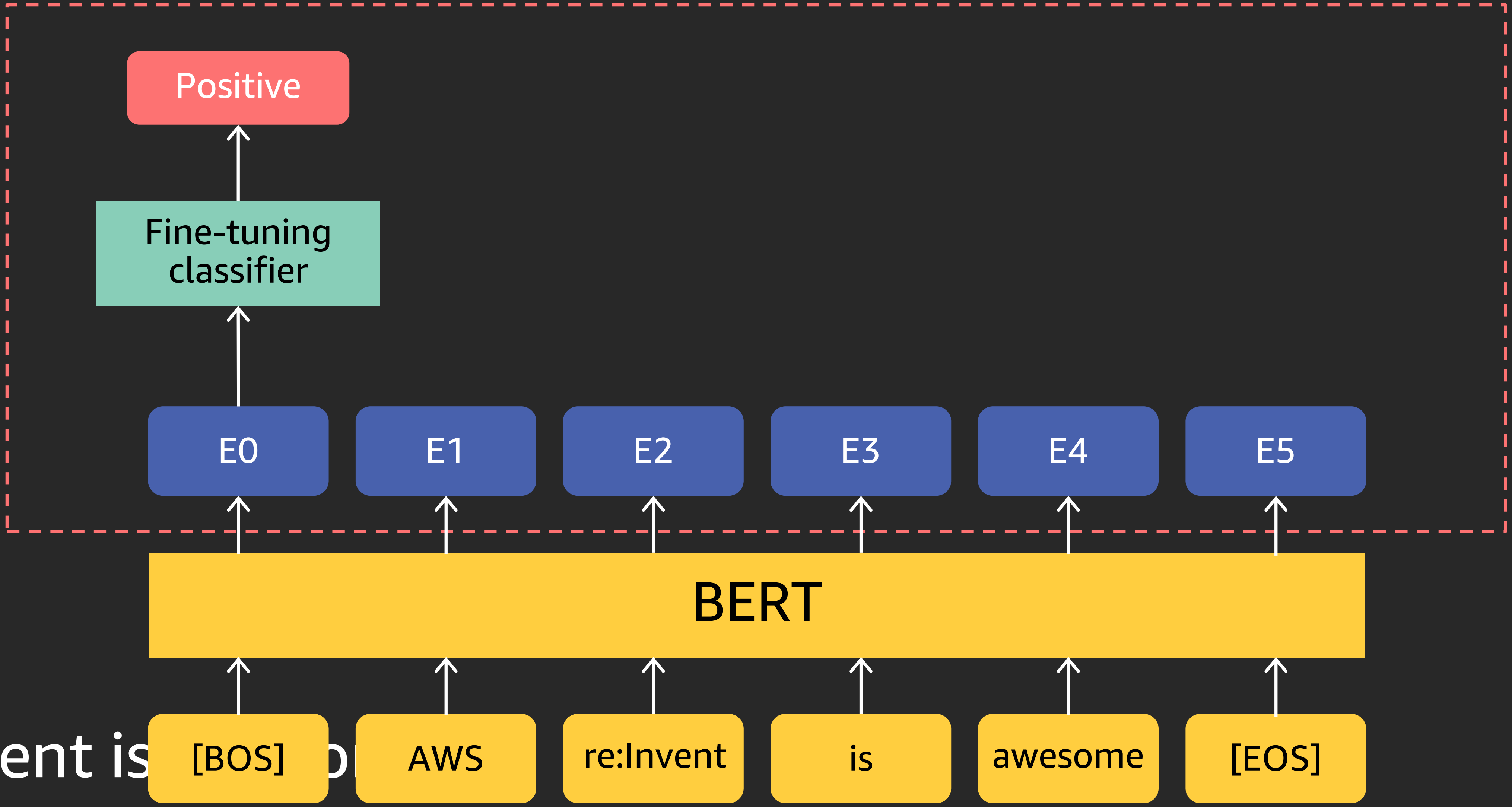
Sentiment analysis

Output: positive

Embedding:

Input: AWS re:Invent is a

BERT fine-tuning (sentiment analysis)



BERT fine-tuning

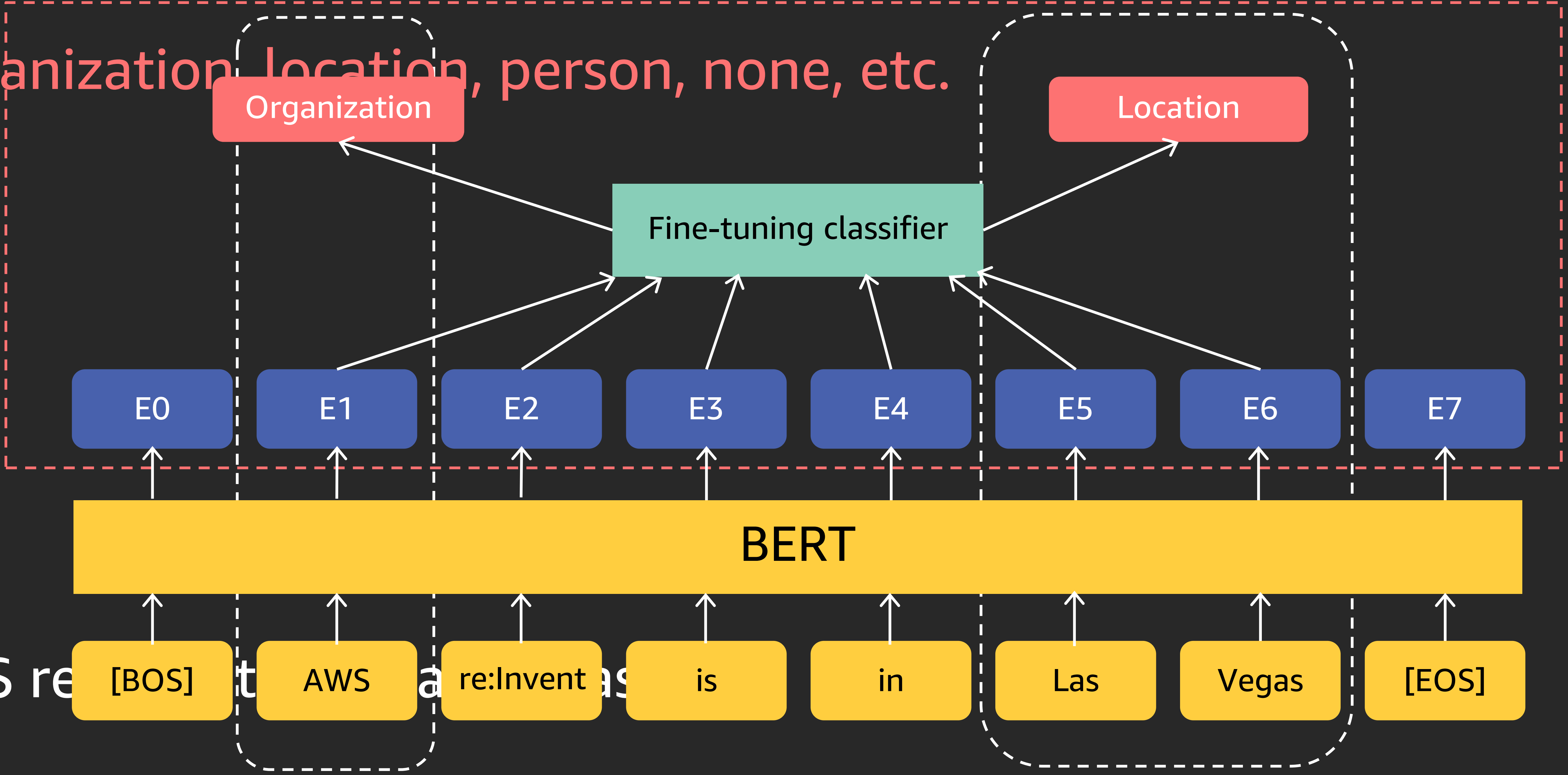
Name entity recognition (NER)

Output: organization, location, person, none, etc.

Embedding:

Input: AWS re

BERT fine-tuning (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

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

AWS Inf1 instance

- Up to 2000 TOPs with 16x AWS-designed AWS Inferentia accelerators
- Lowest cost per inference in the cloud
- Designed for high throughput and low latency

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

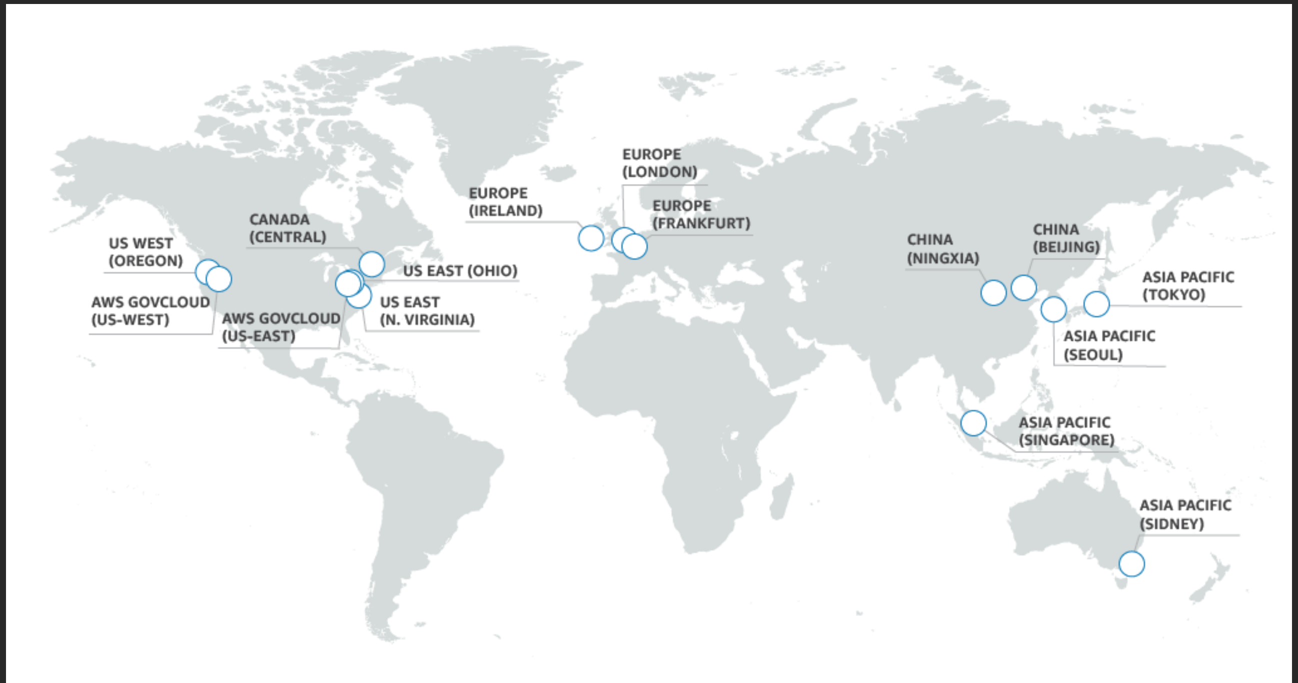
Up to eight NVIDIA Tesla V100 GPUs

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

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

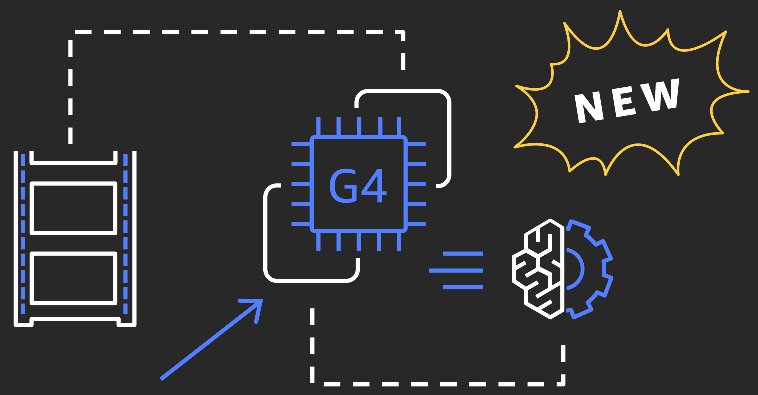
Support all ML frameworks and model types

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

Bringing machine learning to all developers

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
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

End-to-end machine learning platform

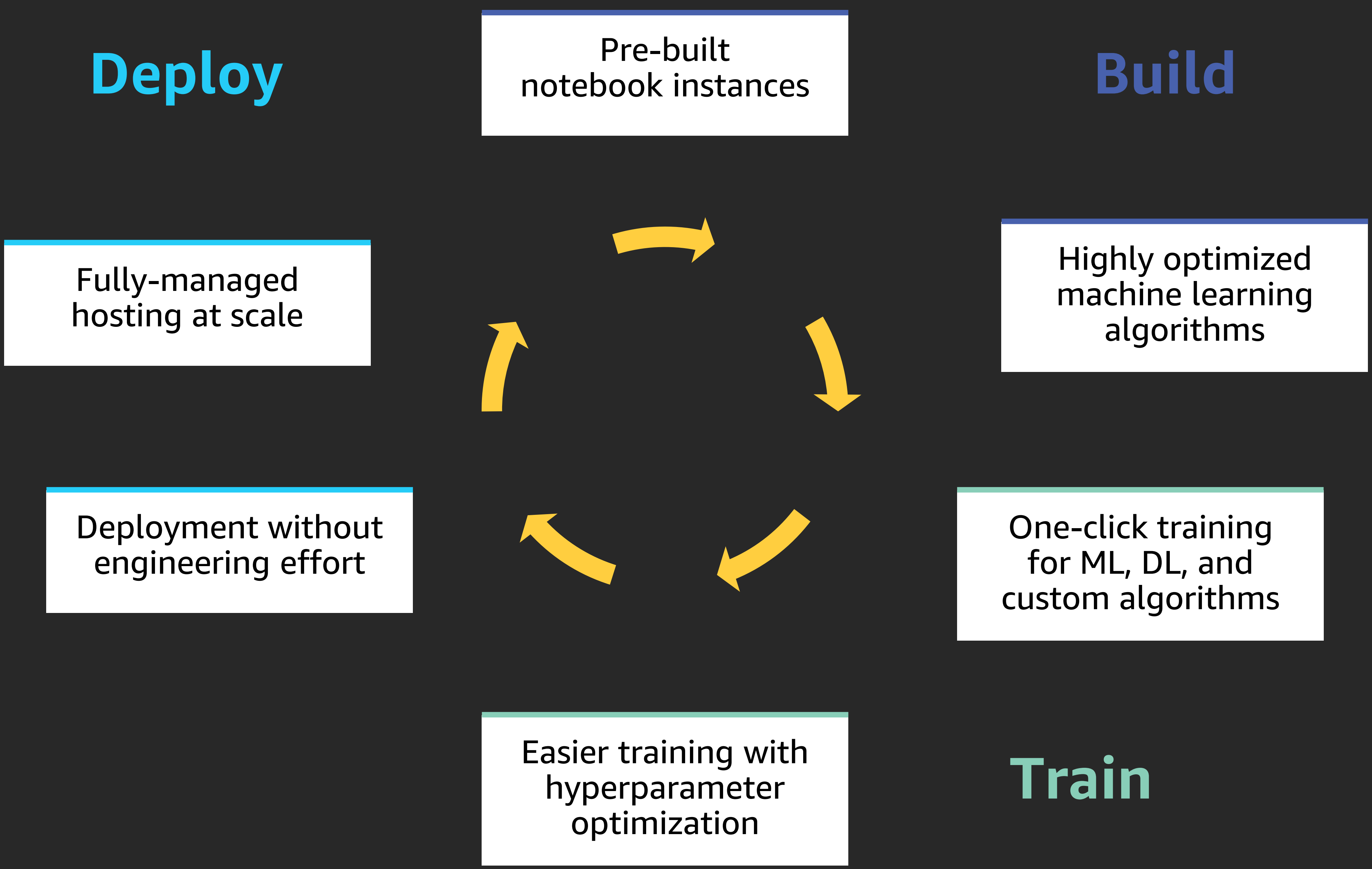
Flexible model training

K mxnet GLUON Chainer

PyTorch TensorFlow HOROVOD

Pay by the second

Amazon SageMaker



Build

Pre-built notebook instances

Highly optimized machine learning algorithms

Train

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

Easier training with hyperparameter optimization

Deploy

Deployment without engineering effort

Fully-managed hosting at scale

scikit learn

GLUON

TensorFlow

PyTorch

mxnet

Hands-on labs

1. Object detection (SSD)

2. Sentiment analysis (BERT)

URL: **https://bit.ly/2sszib8**

Full URL:

https://github.com/awshlabs/reinventGPULab

Resources

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

Gluon:

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

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

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

Dive into Deep Learning Book:

http://d2l.ai/

https://discuss.mxnet.io/

Thank you!

Wen-ming Ye

Twitter: @wenmingye

Please complete the session survey in the mobile app.