1224-CMP

Developing Deep Learning Models for Computer Vision with Amazon EC2 P3 Instances

Wenming Ye
Sr. Solutions Architect AI/ML
Amazon Web Services

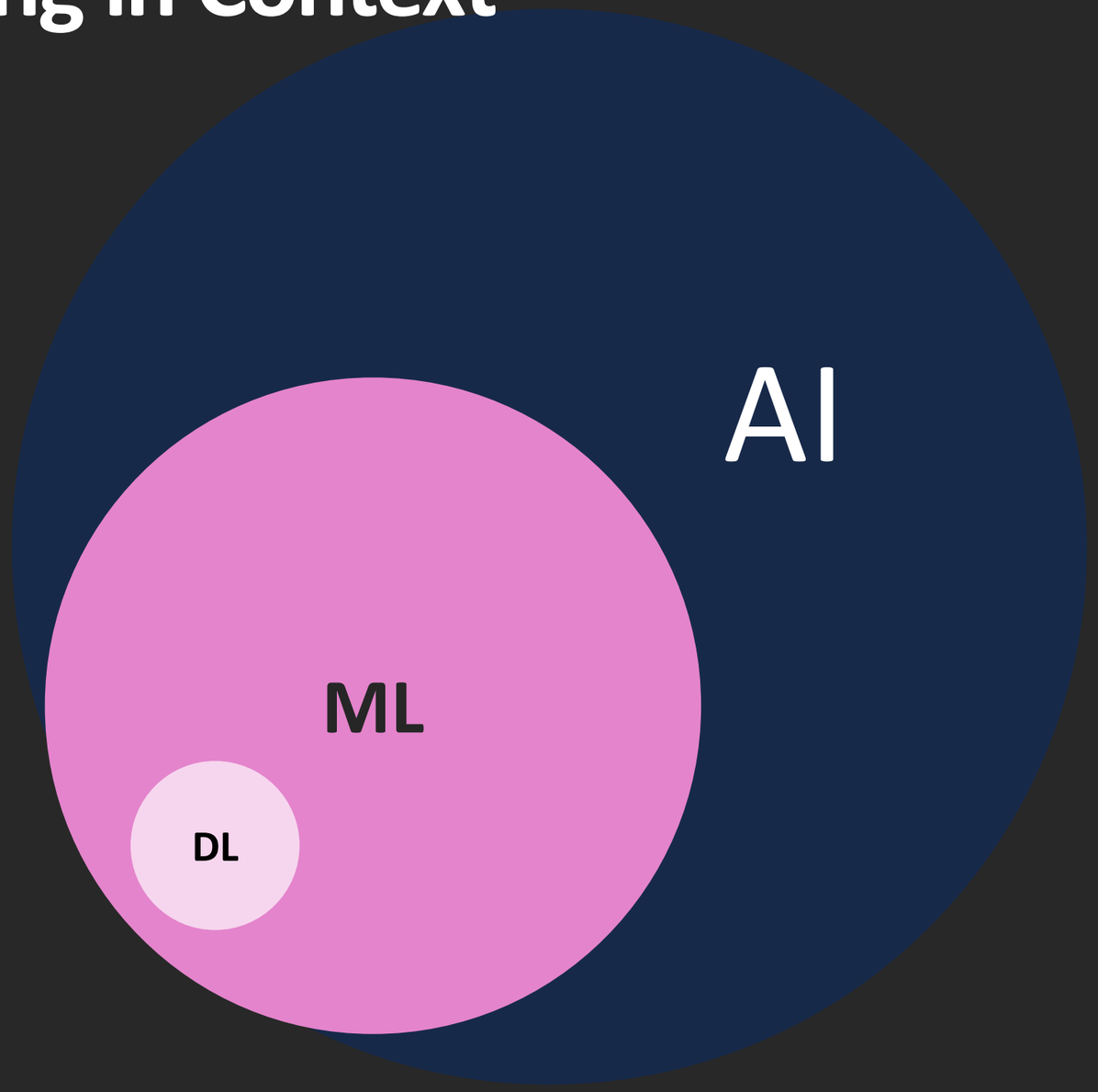
AWS re:Invent

aws

Agenda

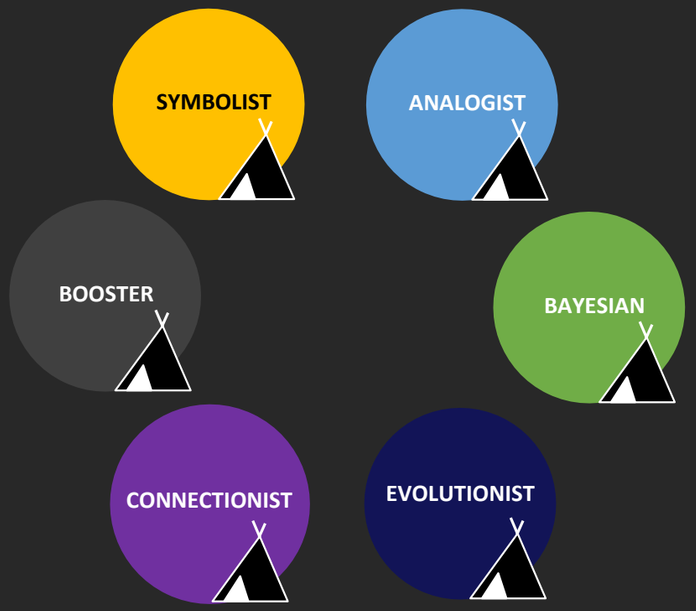
- Deep Learning
- Computer Vision Scenarios and DL Architecture
- P3 Instance Details
- Learning Resources

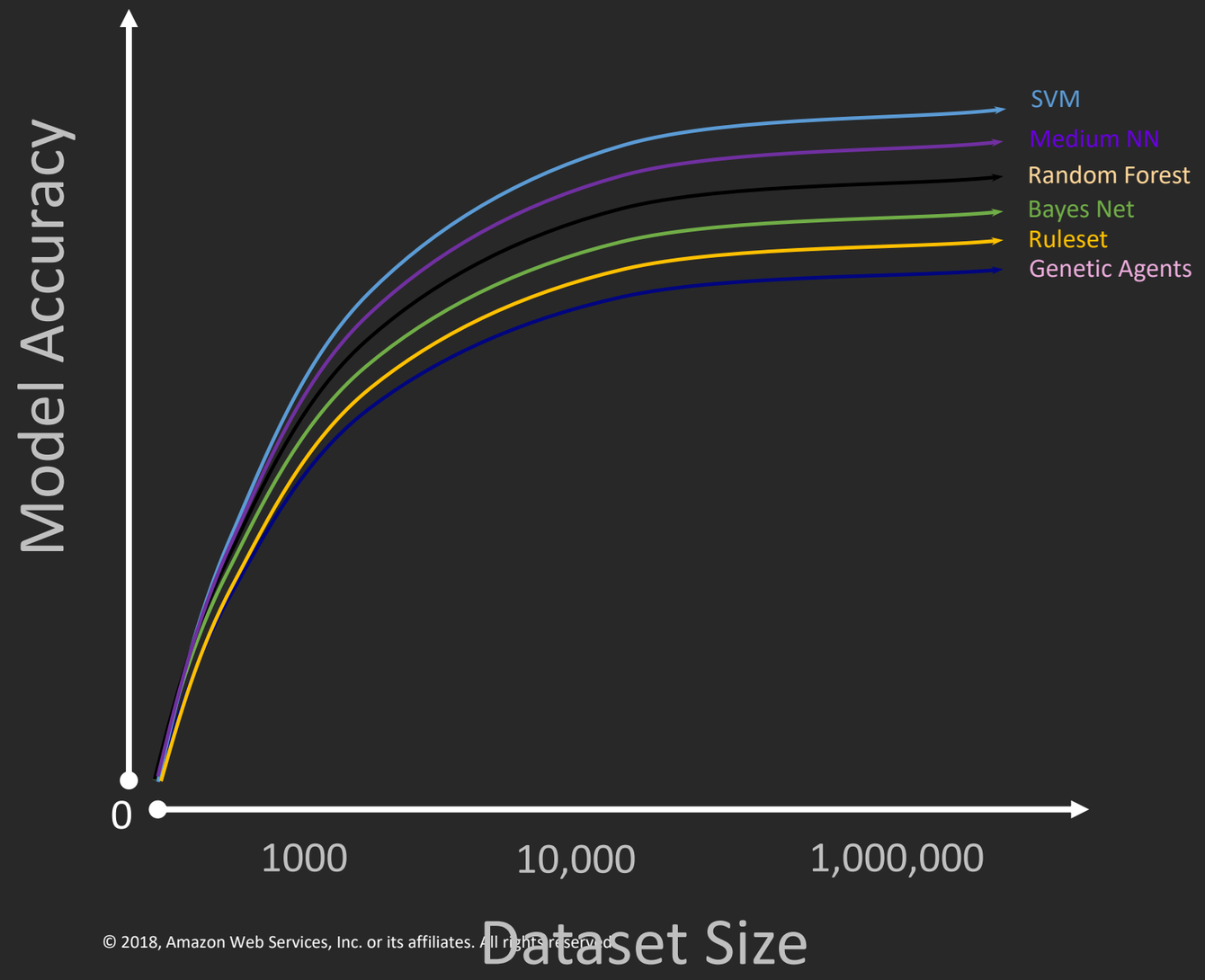
Deep Learning in Context

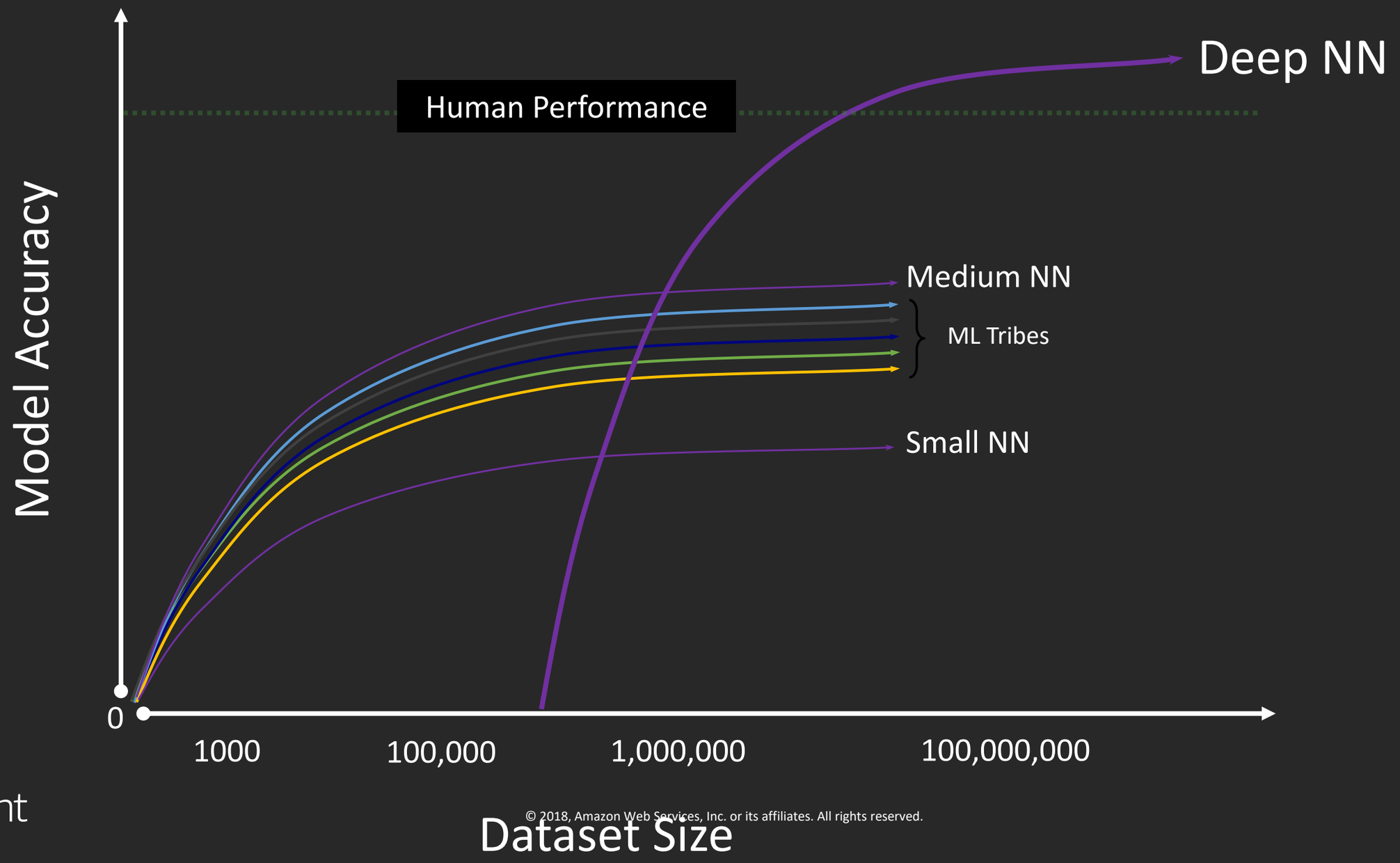


Generalization Performance

rank ~= human effort







Scenarios and DL Architecture

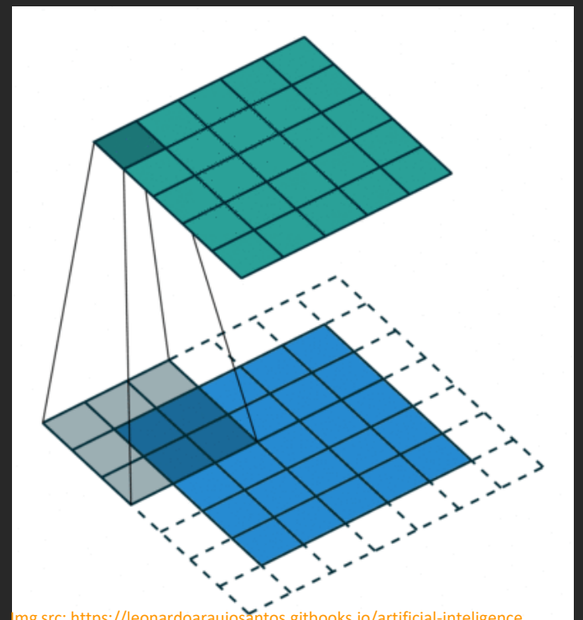
- Architecture
- Convolutional Neural Network (CNN)

- Scenarios
- Image Classification
- Object Detection
- Image Segmentation
- Visual Search
- GANs for Item Generation

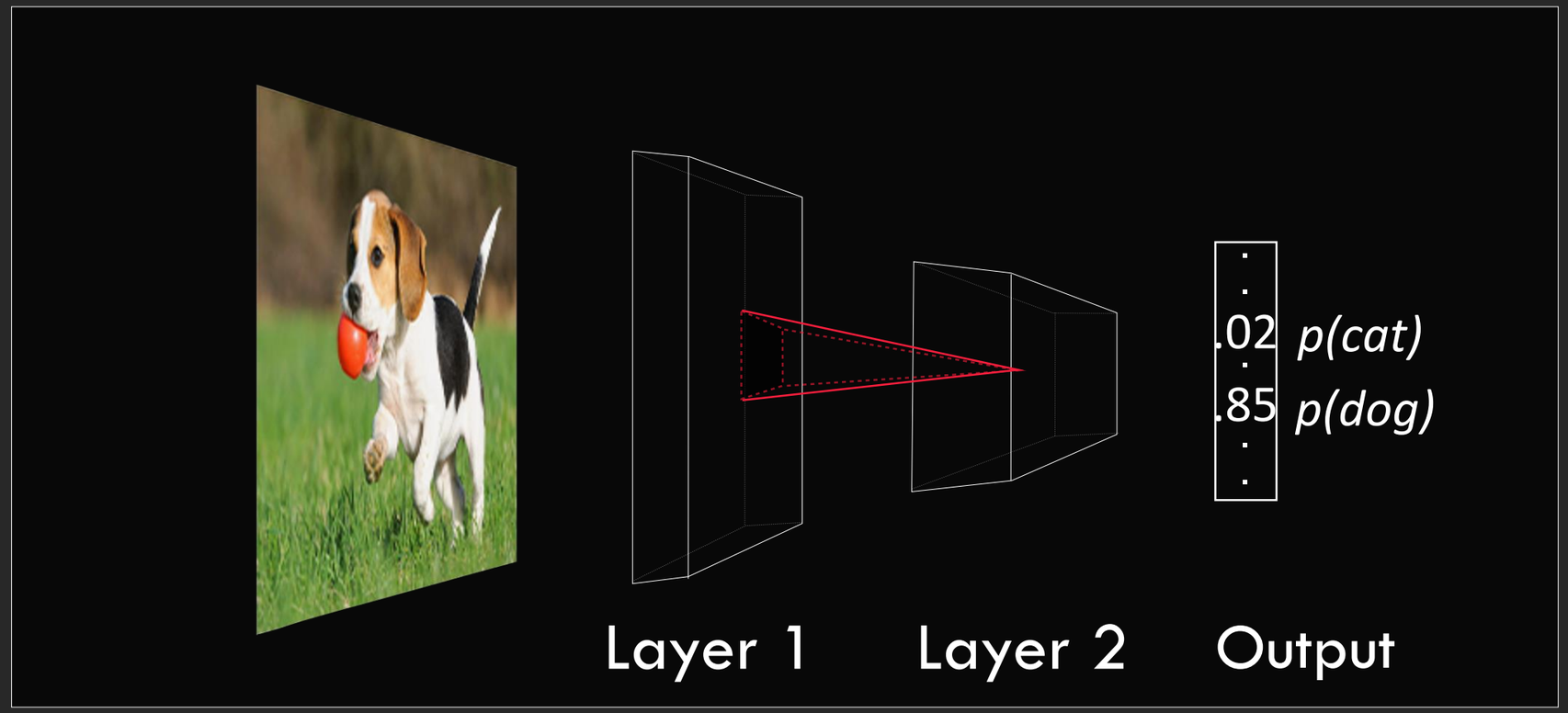
Convolution Neural Network

Deep Learning in Computer Vision

Convolutional neural network



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



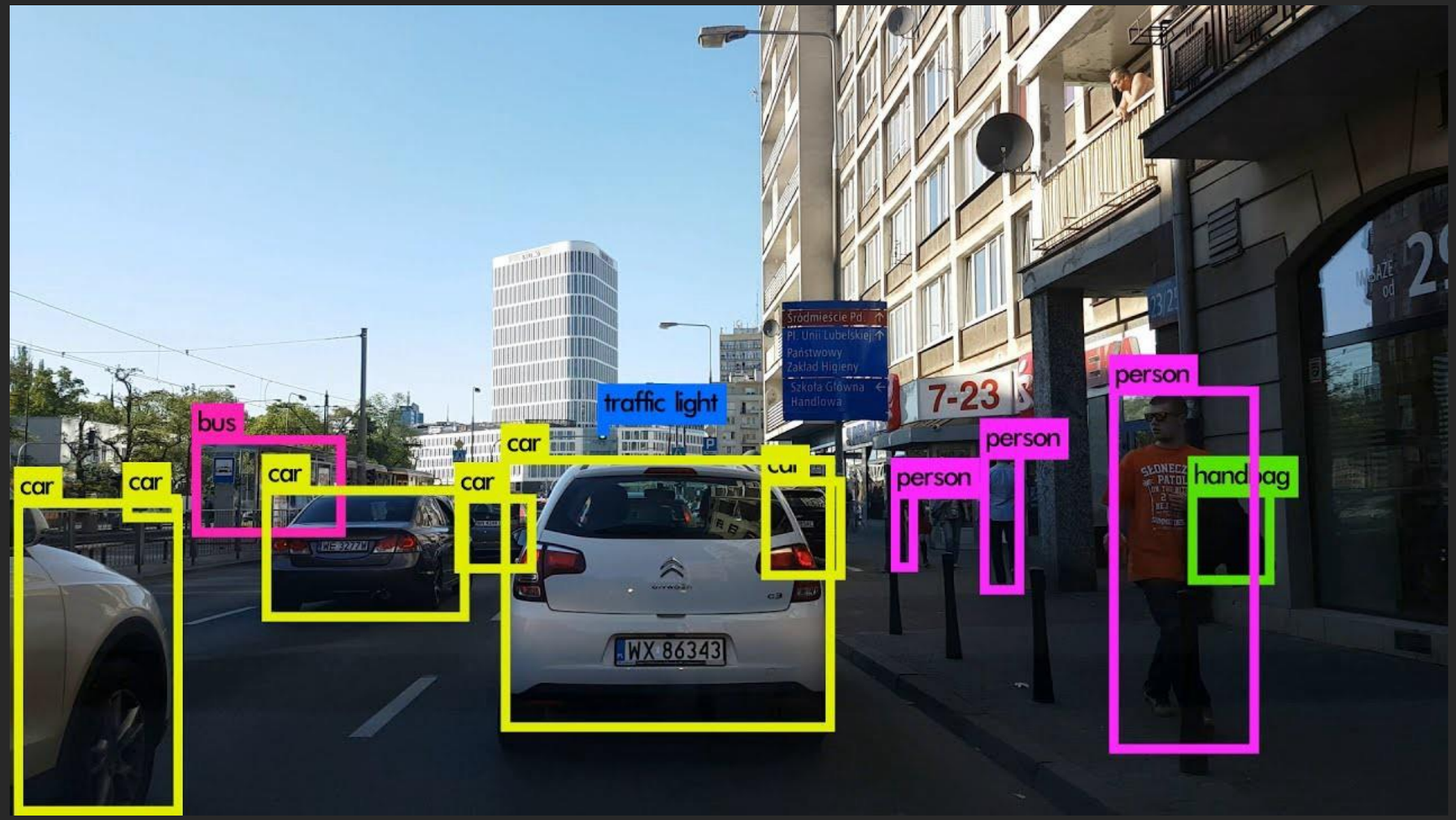
Explore spatial information with convolution layers

Demo: Convolution Neural Network

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

Object Detection

Object Detection



Single Shot Detector

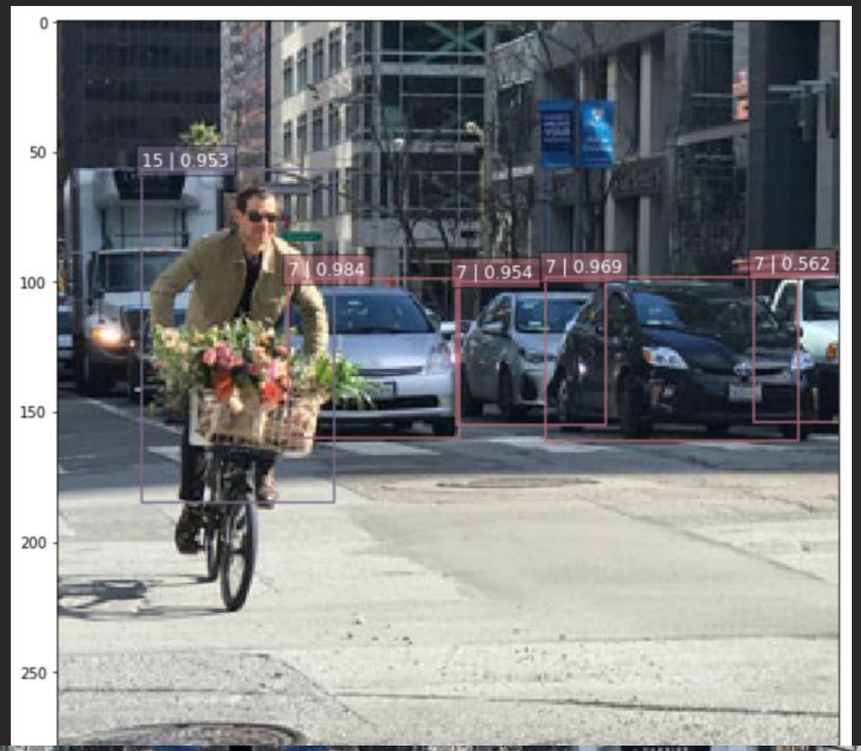
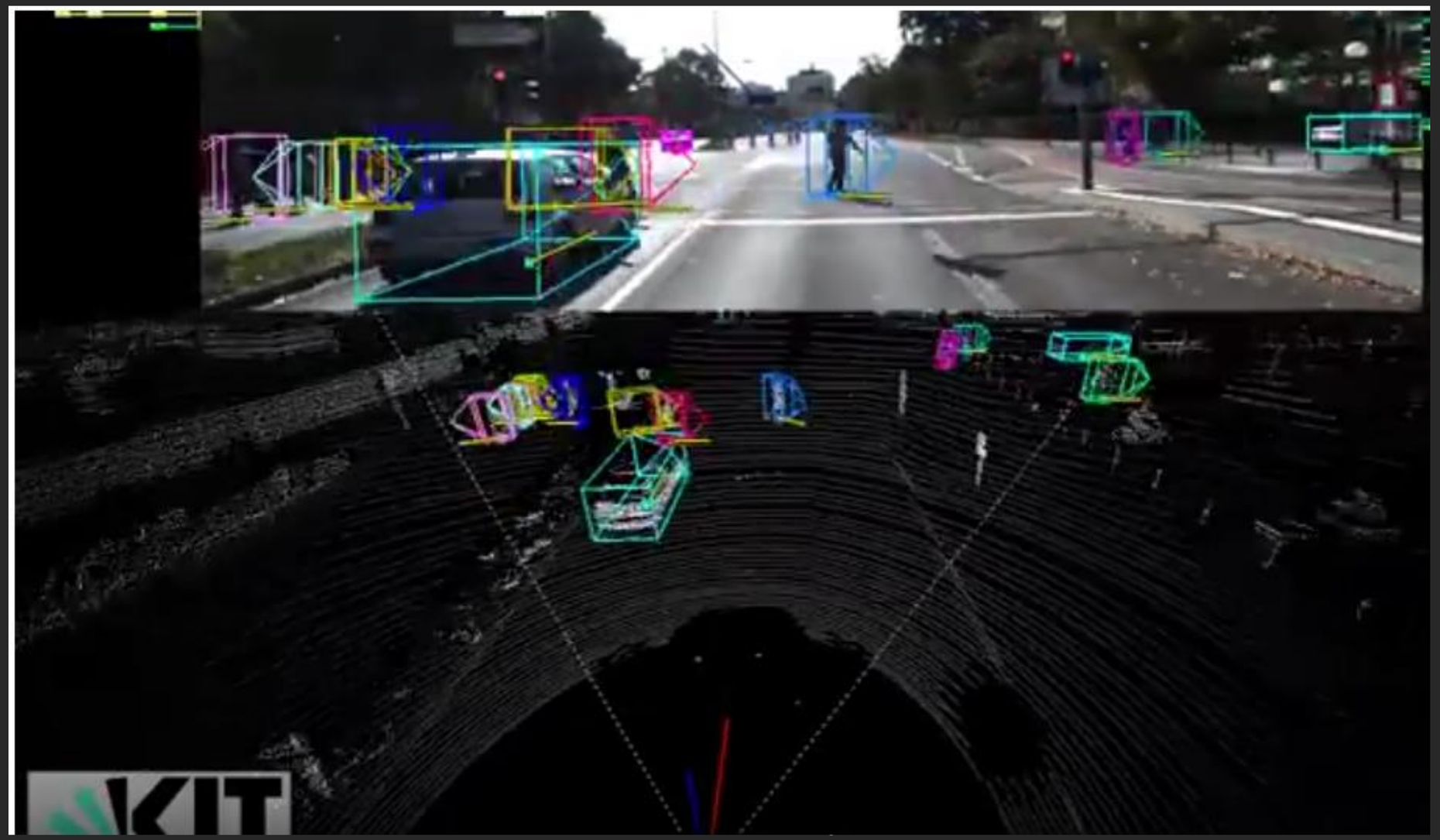


Image Segmentation

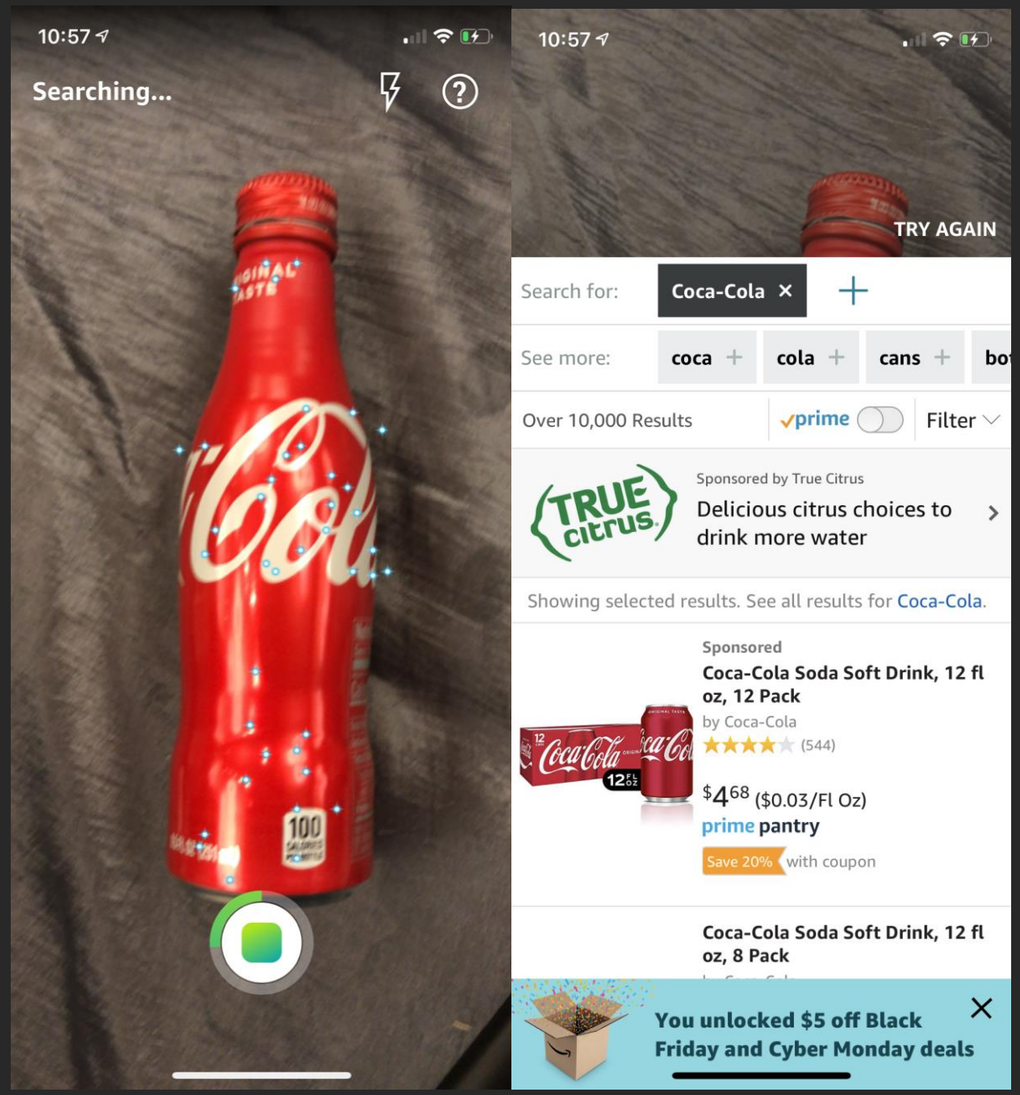
Image Segmentation

KITTI

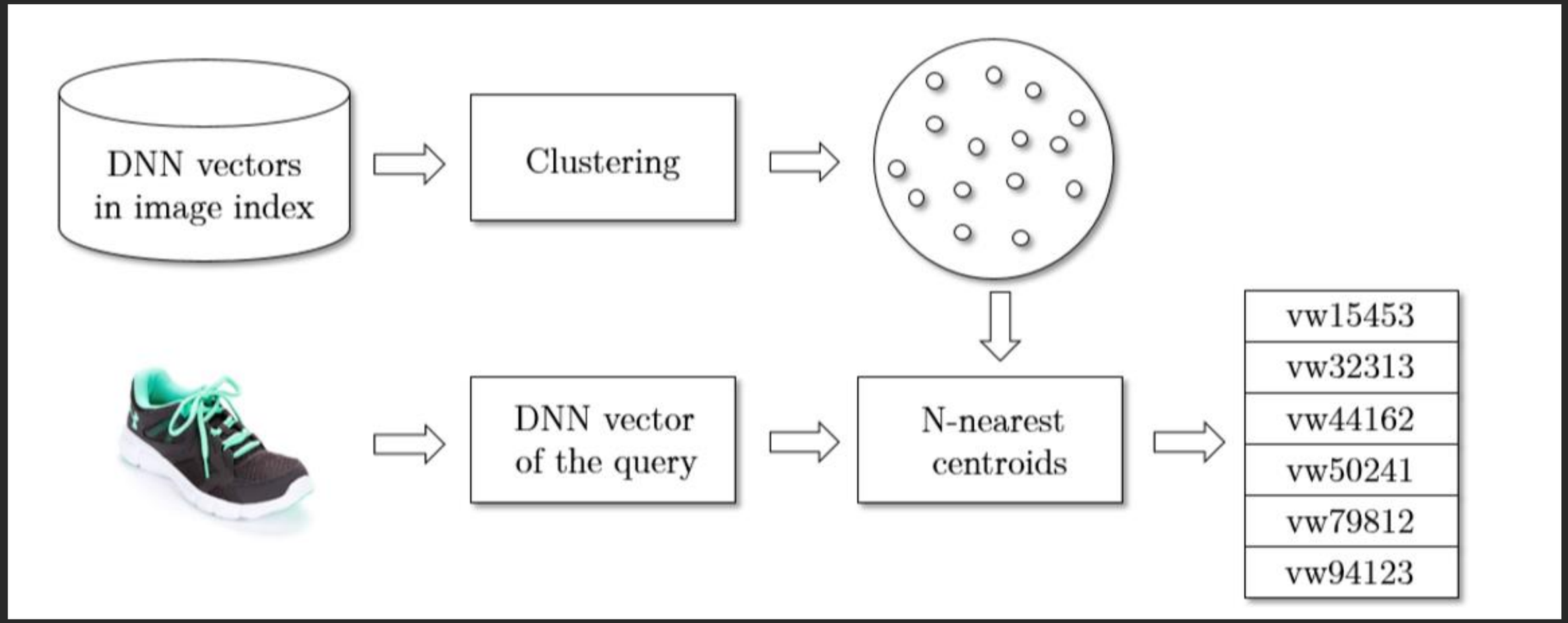


Visual Search

Visual Search



Architecture



Pipeline Stages

Image Query Processing

- data normalization / augmentation

Embedding

- DNN model (s)

kNN + Ranking

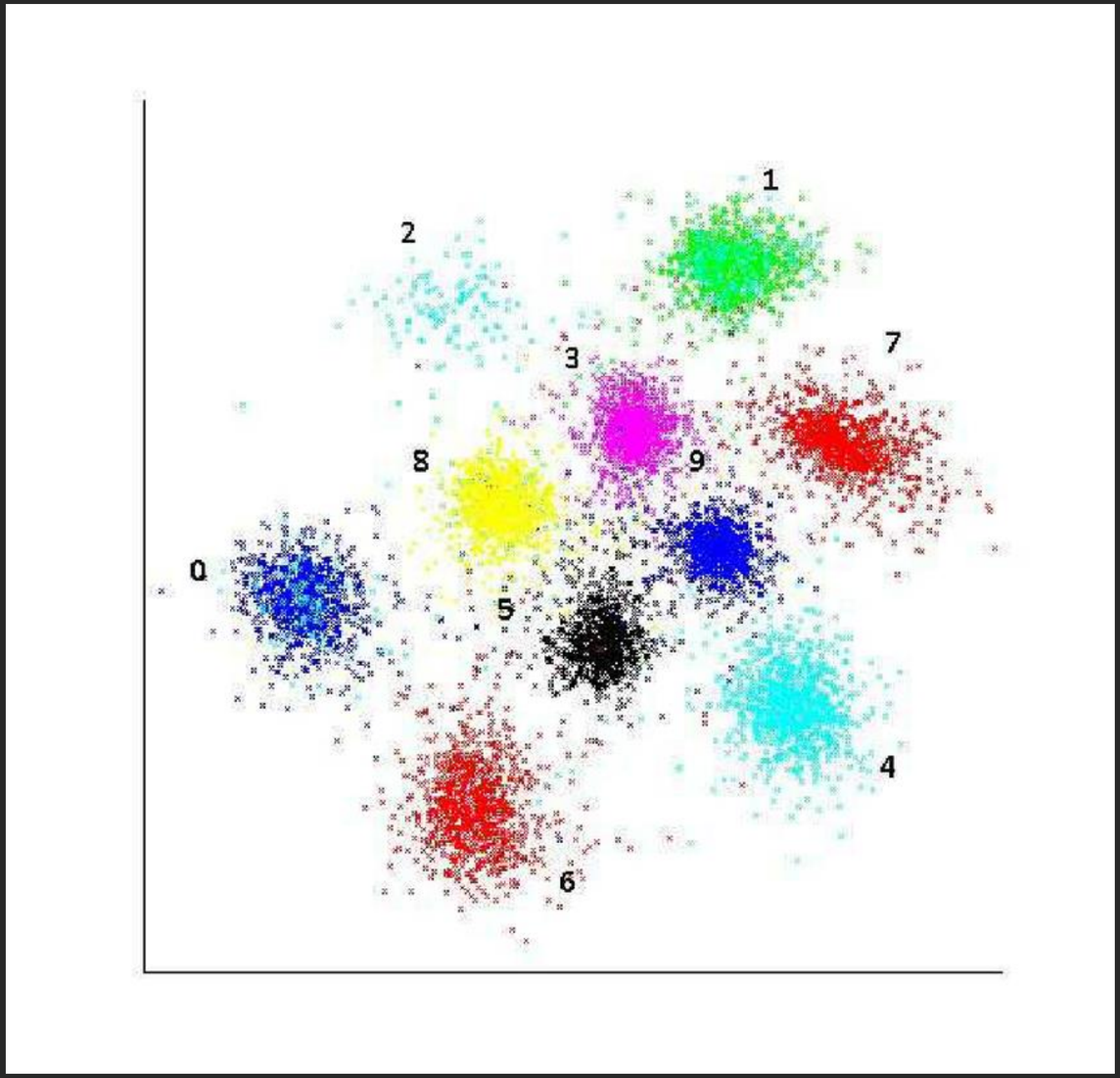
- post-processing, de-dup

Demo: Amazon Shopping

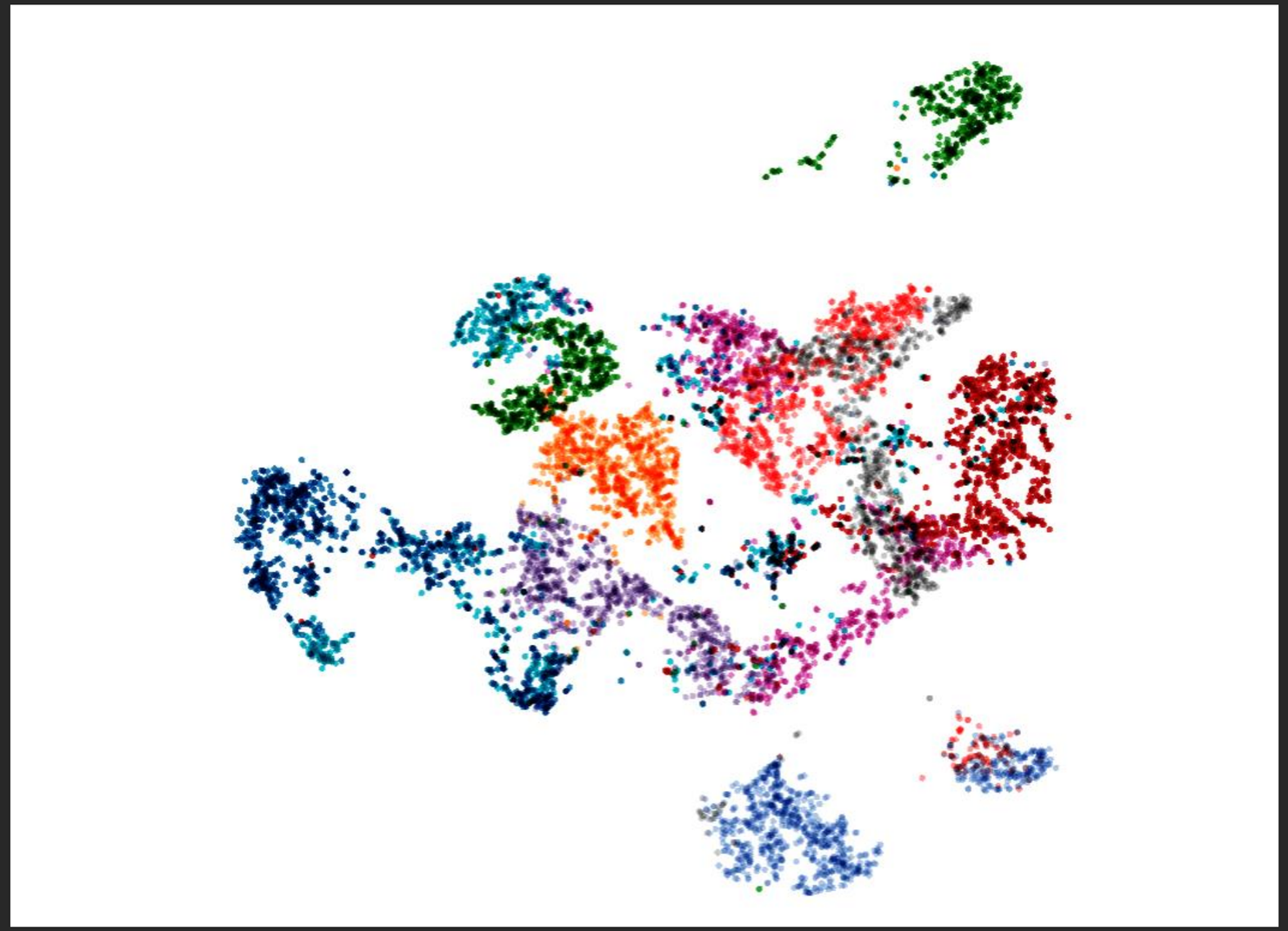
Domains

Domain	Purpose
Generic	Optimized 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.

Embedding = Learned Representation Space



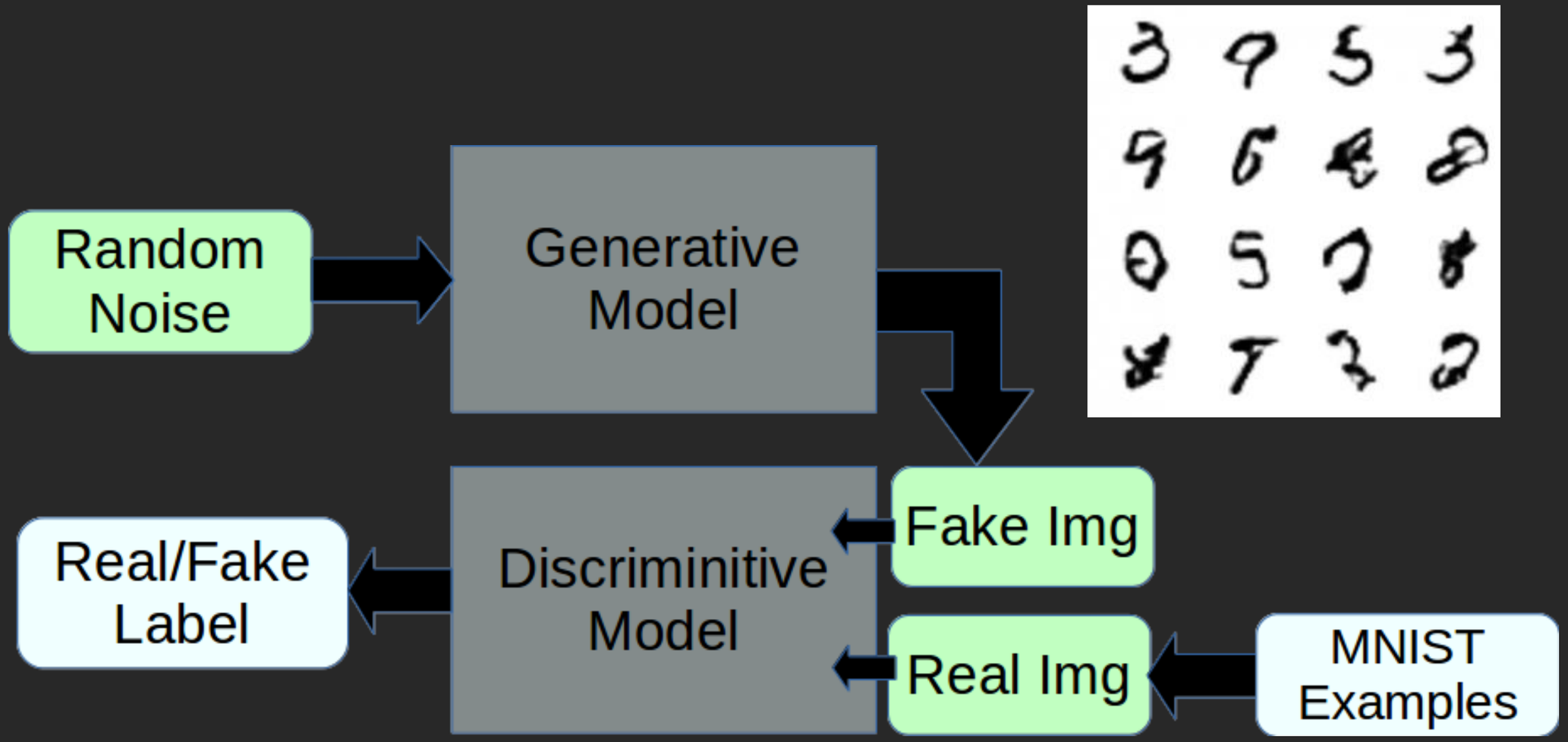
Demo Notebook + Tensorboard



https://github.com/miroenev/teach_DL/blob/master/data_buckets_fashion_mnist_cnn_ae.ipynb

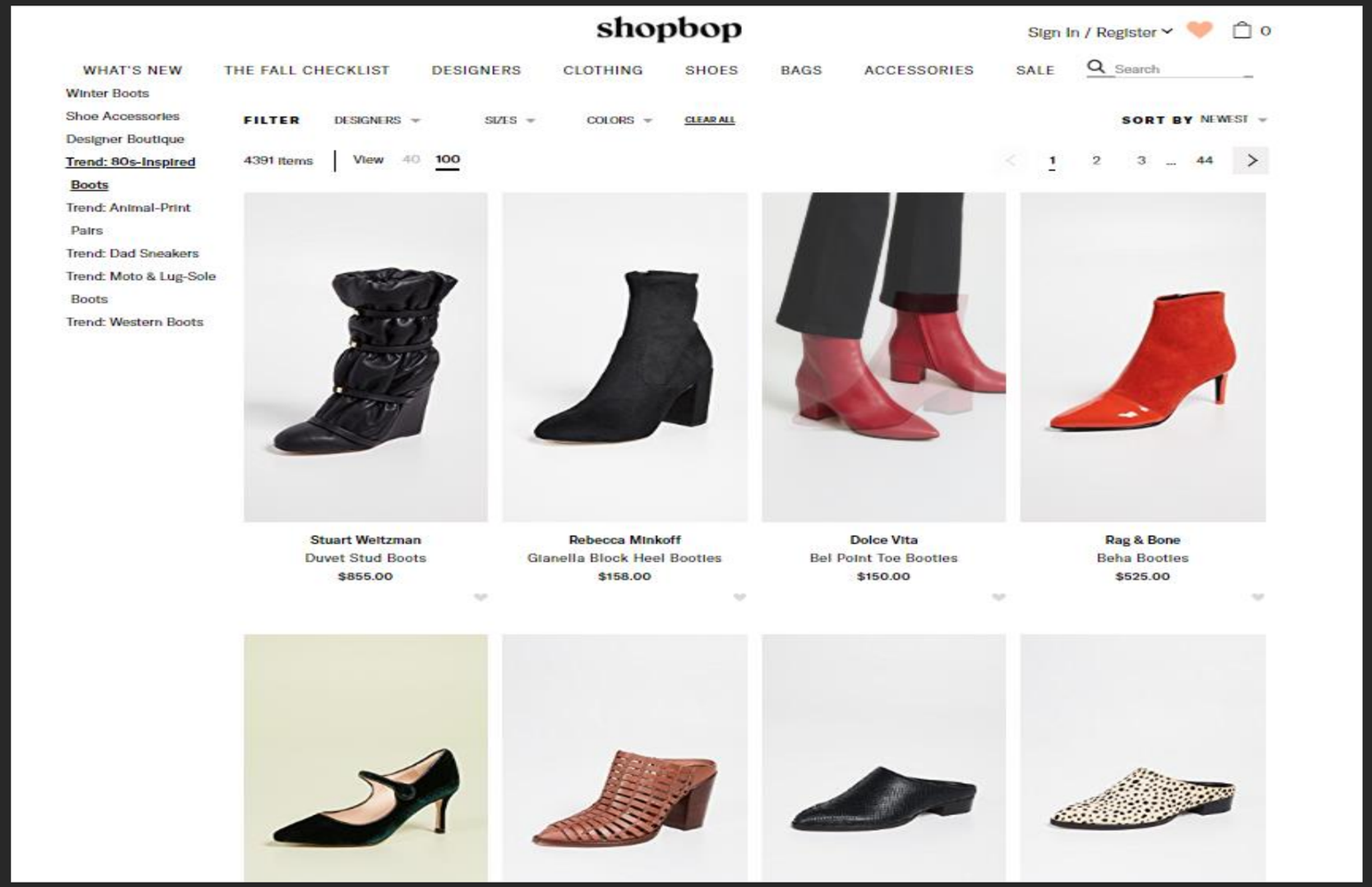
Generative Adversarial Networks (GANs)

GAN Overview



GANs @ Celebrity Faces

Helping ShopBop to design new shoes using Progressive GANs



Video: ShopBop Progressive GAN Demo

Amazon EC2 P3 and P3dn Instances

One of the fastest, most powerful GPU instances in the cloud

- 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*
- 16GB GPU memory with 900 GB/sec peak GPU memory bandwidth

airbnb
salesforce
TOYOTA RESEARCH INSTITUTE
Western Digital.
OpenAI
SCHRÖDINGER.

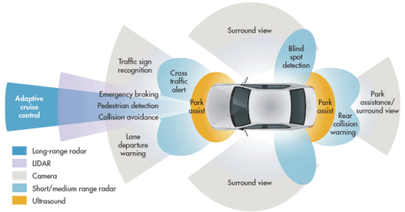
Use-Cases for P3 Instances

Machine Learning/AI

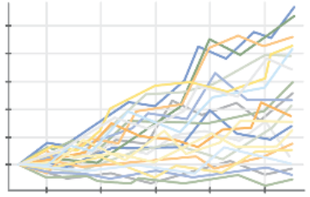
Natural Language Processing

Image and Video recognition

Autonomous vehicle systems

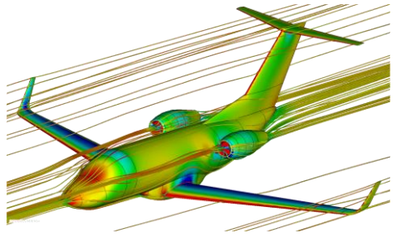


Recommendation Systems



High Performance Computing

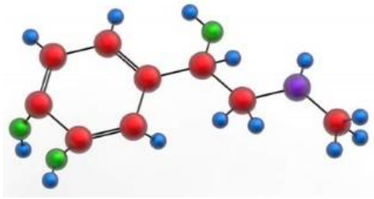
Computational Fluid Dynamics



Financial and Data Analytics

Weather Simulation

Computational Chemistry



P3 Instances Details

Instance Size	GPUs	GPU Peer to Peer	vCPUs	Memory (GB)	Network Bandwidth	EBS Bandwidth	On-Demand Price/hr*	1-yr RI Effective Hourly*	3-yr RI Effective Hourly*
P3.2xlarge	1	No	8	61	Up to 10Gbps	1.7Gbps	$3.06	$1.99 (35% Disc.)	$1.23 (60% Disc.)
P3.8xlarge	4	NVLink	32	244	10Gbps	7Gbps	$12.24	$7.96 (35% Disc.)	$4.93 (60% Disc.)
P3.16xlarge	8	NVLink	64	488	25Gbps	14Gbps	$24.48	$15.91 (35% Disc.)	$9.87 (60% Disc.)

Note: P3 instances are also supported via Spot Instances with up to 70% discount

Regional Availability

P3 instances are generally available in AWS US East (Northern Virginia), US East (Ohio), US West (Oregon), EU (Ireland), Asia Pacific (Seoul), Asia Pacific (Tokyo), AWS GovCloud (US) and China (Beijing) Regions.

Framework Support

P3 instances and their V100 GPUs supported across all major frameworks (such as TensorFlow, MXNet, PyTorch, Caffe2 and CNTK)

New larger P3 size – P3dn.24xlarge

Optimized for distributed ML training

- One of the most powerful GPU instances available in the cloud
- 100 Gbps of networking throughput
- 96 vCPU using AWS customer Skylake CPUs and 768 GB of system memory
- Based on NVIDIA's latest GPU Tesla V100 with 32 GB of memory

Instance size	GPUs	GPU memory	GPU peer to peer	vCPUs	CPU type	Memory (GB)	Network Bandwidth	Amazon EBS Bandwidth	Local instance storage
P3.2xlarge	1 x V100	16 GB/GPU	No	8	Broadwell	61	Up to 10 Gbps	1.7 Gbps	NA
P3.8xlarge	4 x V100	16 GB/GPU	NVLink	32	Broadwell	244	10 Gbps	7 Gbps	NA
P3.16xlarge	8 x V100	16 GB/GPU	NVLink	64	Broadwell	488	25 Gbps	14 Gbps	NA
NEW! P3dn.24xlarge	8 x V100	32 GB/GPU	NVLink	96	Skylake	768	100 Gbps	14 Gbps	2 TB NVMe

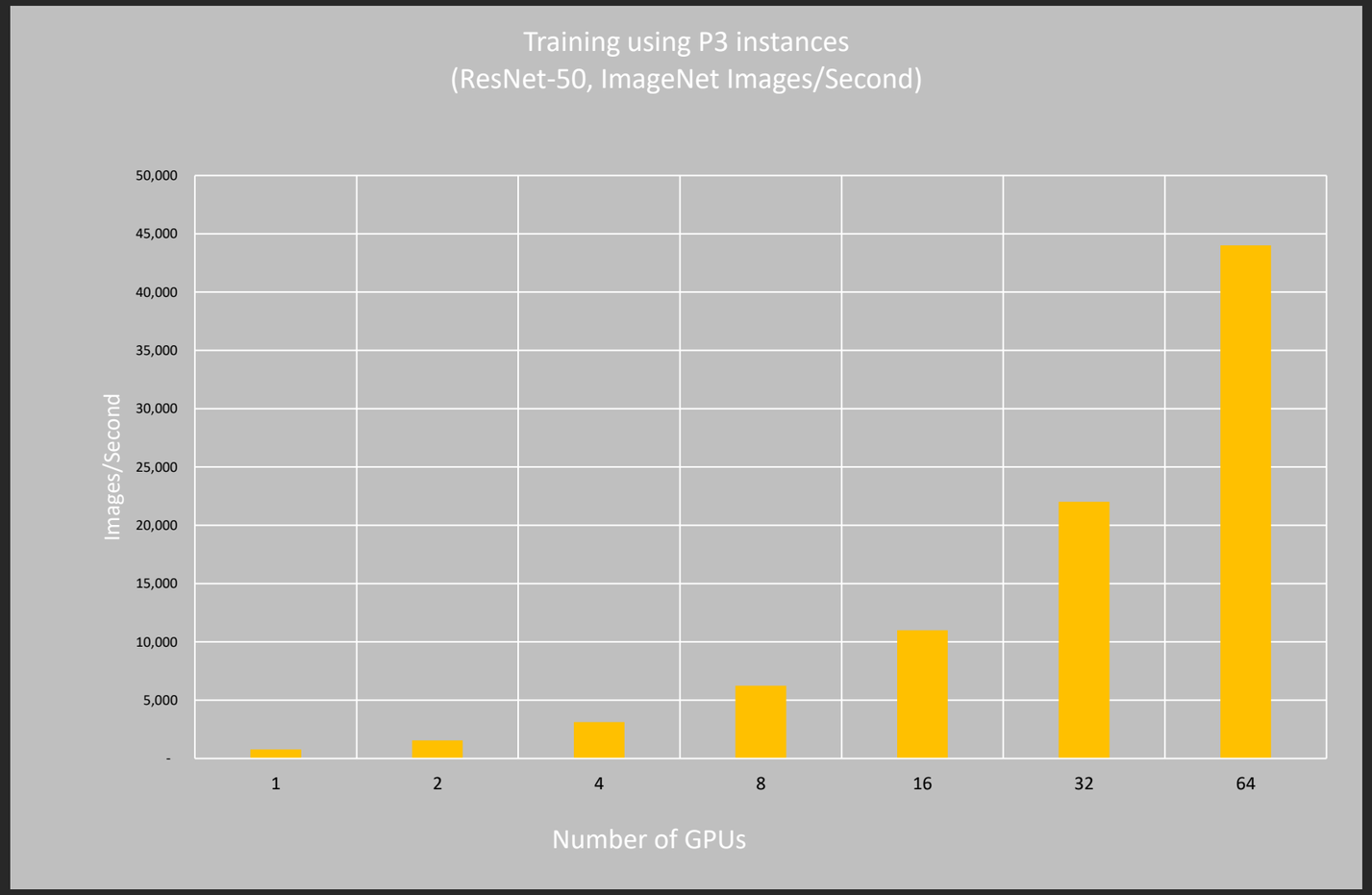
Latest NVIDA V100 GPU with 32 GB memory for large models and higher batch sizes

96 Skylake vCPUs with support for AVX-512 instructions for pre-processing of training data

100 Gbps of networking throughput for large-scale distributed training & fast data access

Scaling Performance using Distributed Training

- Using single P3 instances, with Volta GPUs, customers can cut down training times of their machine learning models from days to a few hours.
- Using distributed training via multiple P3 instances, high performance networking and storage solutions, customers can further cut down their time-to-train from hours in to minutes.
- Example – We been able to train ResNet-50 to Top1 validation accuracy of 76% in 47 mins using 8 P3.16xlarge instances.

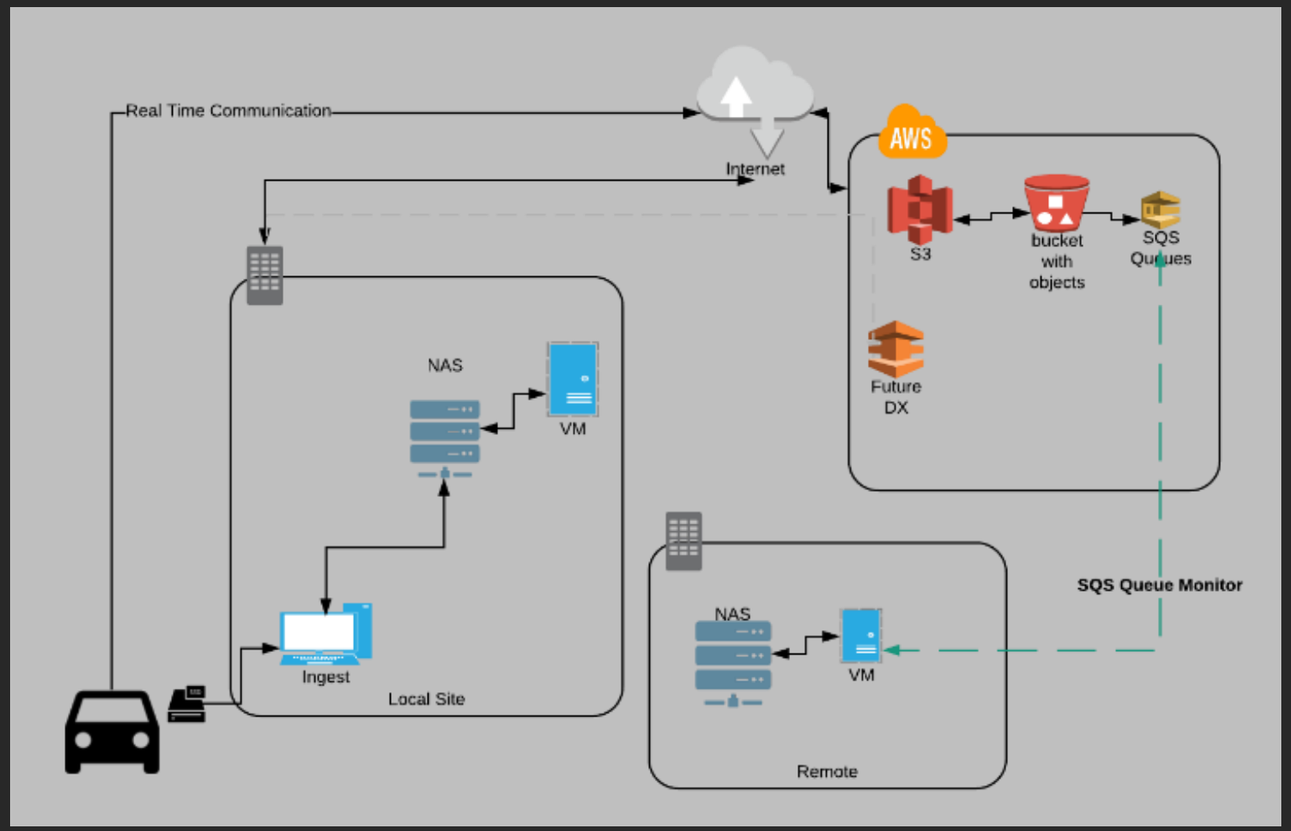


https://aws.amazon.com/blogs/machine-learning/scalable-multi-node-deep-learning-training-using-gpus-in-the-aws-cloud/

Toyota Research Institute

Toyota Research Institute's (TRI) mission is to improve the quality of human life through advances in artificial intelligence, automated driving, and robotics. The core compute capability needed by TRI to accelerate the training of their machine learning models is powered by multiple Amazon EC2 P3 instances.

With P3 instances, TRI is seeing a 4X faster time-to-train than the P2 instances they had used previously. This gives them significant agility to optimize and retrain their models quickly and deploy them in their test cars or simulations environment for further testing. Furthermore, the significant performance improvement in P3s over P2s, coupled with pay-as-you-go model translates to lower operating costs.



P3 Instance Demo

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

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

Amazon SageMaker: Launch Customers

Amazon SageMaker

1

Notebook Instances

2

Algorithms

3

ML Training Service

4

ML Hosting Service

SageMaker Demo

AWS re:Invent

aws

Hands on Labs to Try

- Image classification
- Object detection
- URL: https://bit.ly/2TNdEI0

Where to go from here?

- Other Topics (coming soon)
  - Other DL architectures [Auto Encoder, RNNs, and more]
  - DL Inference
    - SageMaker hosting
    - Elastic Inference
    - Inference at the Edge [Deeplens, Green Grass Inference]
- Learning Resources:
  - New Book: https://en.diveintodeeplearning.org/chapter_deep-learning-computation/index.html

Thank you!

Wenming Ye
twitter: @wenmingye

AWS re:Invent

aws