DEEP LEARNING INFERENCE USING INTEL® OPENVINO™ TOOLKIT

Elroy Ashtian, Jr.

11/7/2018

Core and Visual Computing Group

OpenVINO toolkit overview video: Innovate Vision Solutions with OpenVINO™ toolkit

0:00 / 2:53

EMERGENCY RESPONSE

FINANCIAL SERVICES

MACHINE VISION

CITIES/TRANSPORTATION

VIDEO: THE "EYE OF IOT"
USE OF VIDEO, COMPUTER VISION AND DEEP LEARNING IS GROWING RAPIDLY

AUTONOMOUS VEHICLES

RESPONSIVE RETAIL

MANUFACTURING

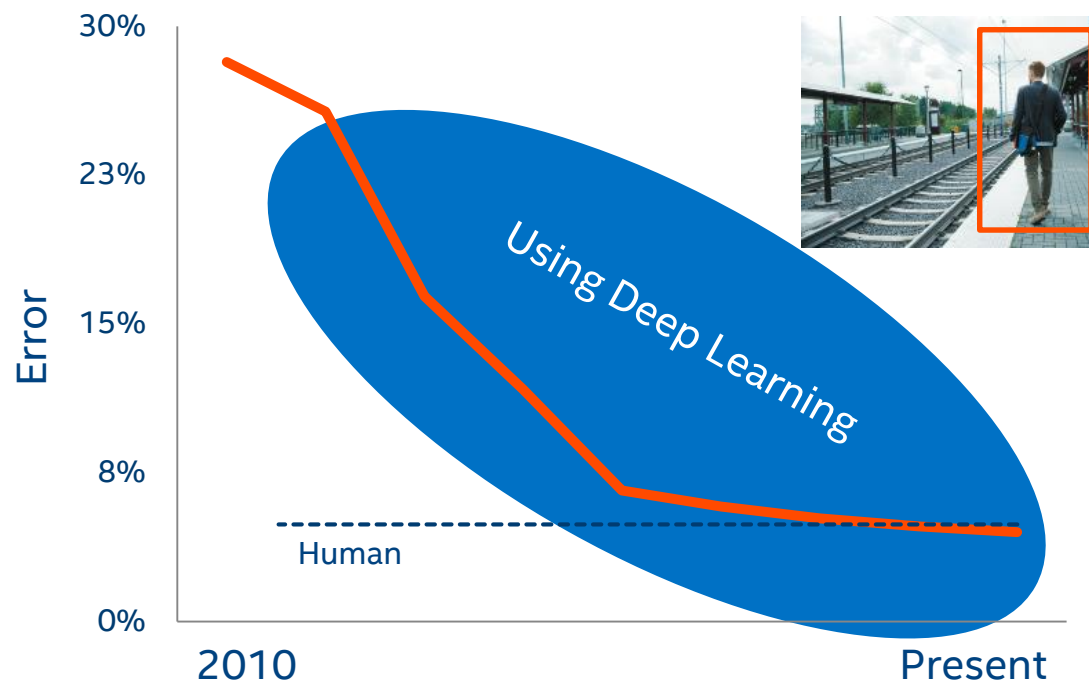PUBLIC SECTOR

Deep Learning Usage Is Increasing

Deep learning revenue is estimated to grow from $655M in 2016 to **$35B** by 2025[1].

Image Recognition

Using Deep Learning

Error

30%
23%
15%
8%
0%

Human

2010 — Present

Traditional Computer Vision
Object Detection

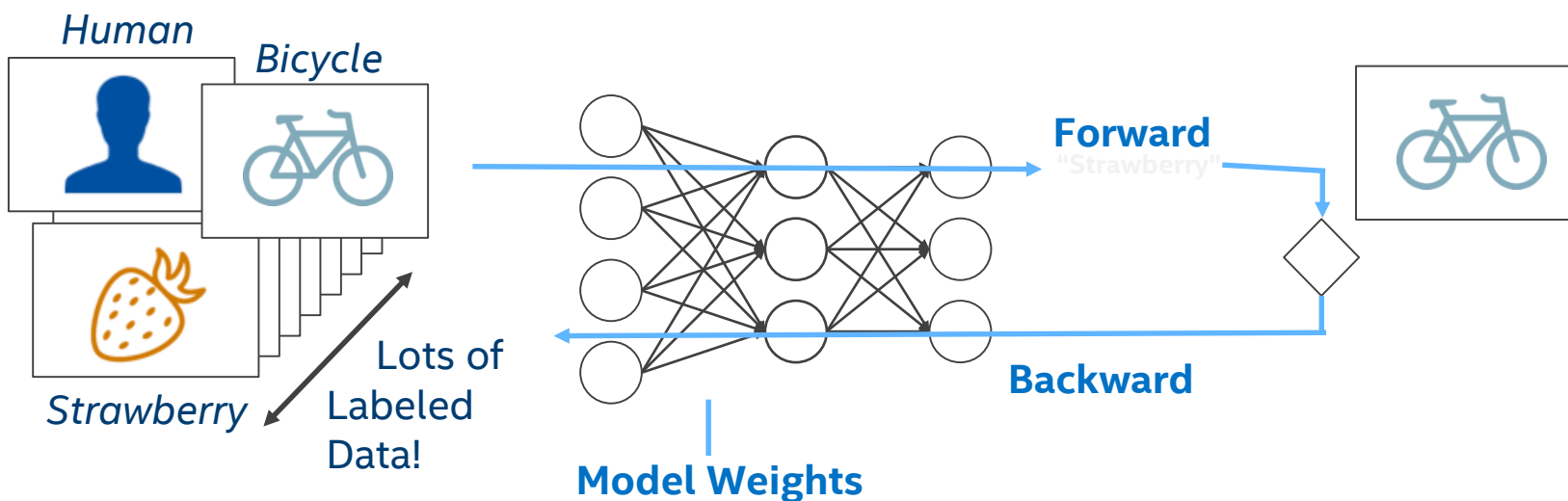Deep Learning Computer Vision
Person Recognition

person

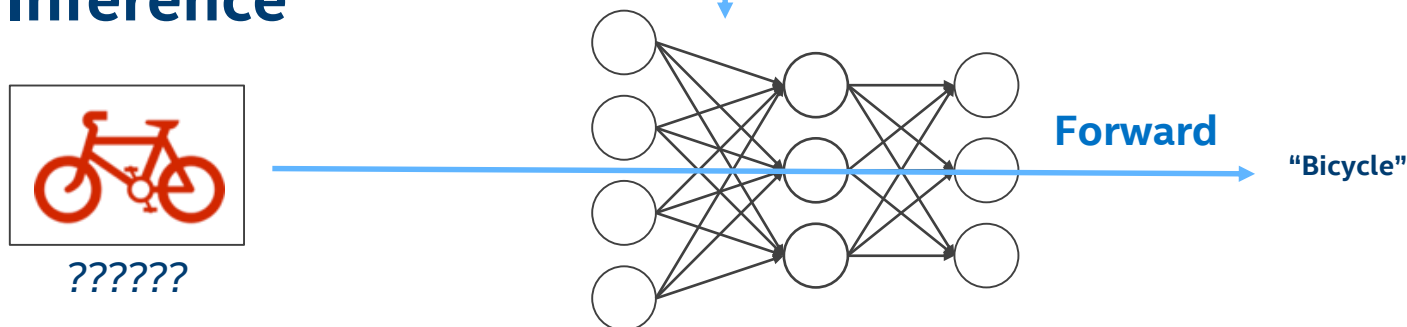Market Opportunities + Advanced Technologies Have Accelerated Deep Learning Adoption

[1]Tractica* 2Q 2017

Deep Learning: Training vs. Inference

Training

Human
Bicycle
Strawberry

Lots of Labeled Data!

Forward
"Strawberry"

Backward

Model Weights

Inference

??????

Forward

"Bicycle"

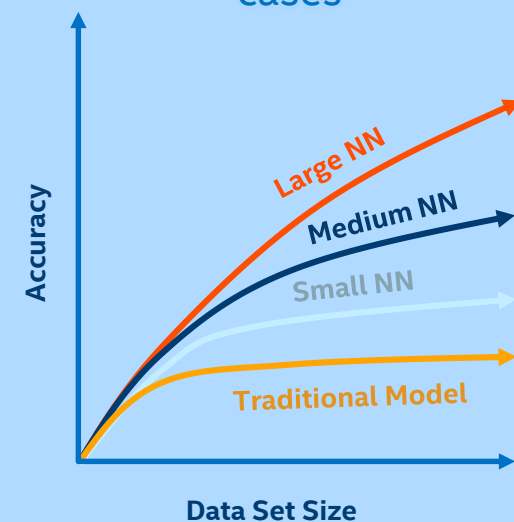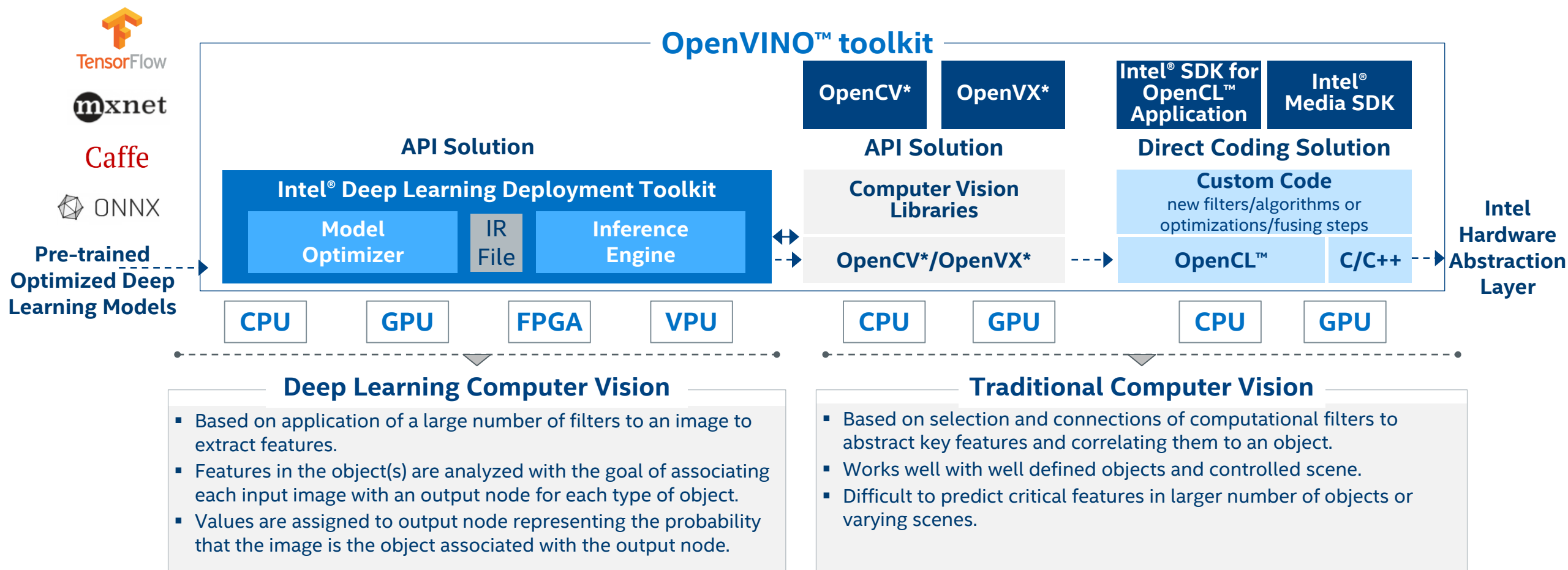**Did You Know?**

Training requires a very large data set and deep neural network (many layers) to achieve the highest accuracy in most cases

Accuracy

Large NN
Medium NN
Small NN
Traditional Model

Data Set Size

Open Visual Inference & Neural network Optimization (OpenVINO™) toolkit

TensorFlow

mxnet

Caffe

ONNX

Pre-trained Optimized Deep Learning Models

OpenVINO™ toolkit

| OpenCV* | OpenVX* | Intel® SDK for OpenCL™ Application | Intel® Media SDK |

API Solution

API Solution

Direct Coding Solution

Intel® Deep Learning Deployment Toolkit
- **Model Optimizer**
- IR File
- **Inference Engine**

Computer Vision Libraries

Custom Code
new filters/algorithms or optimizations/fusing steps

OpenCV*/OpenVX*

| OpenCL™ | C/C++ |

Intel Hardware Abstraction Layer

| CPU | GPU | FPGA | VPU |

| CPU | GPU |

| CPU | GPU |

Deep Learning Computer Vision
- Based on application of a large number of filters to an image to extract features.
- Features in the object(s) are analyzed with the goal of associating each input image with an output node for each type of object.
- Values are assigned to output node representing the probability that the image is the object associated with the output node.

Traditional Computer Vision
- Based on selection and connections of computational filters to abstract key features and correlating them to an object.
- Works well with well defined objects and controlled scene.
- Difficult to predict critical features in larger number of objects or varying scenes.

IR = Intermediate Representation File
GPU = Intel CPU with integrated graphics processing unit/Intel® Processor Graphics
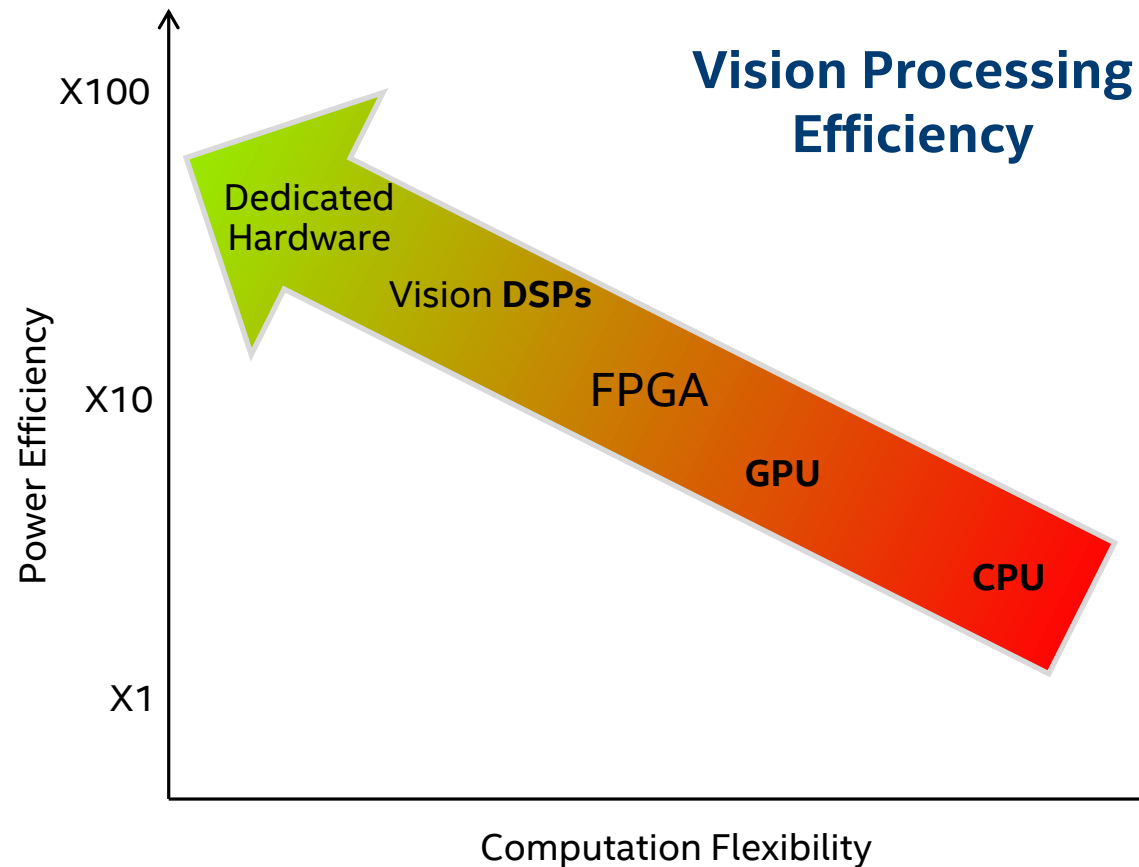VPU = Intel® Movidius™ Vision Processing Unit

Choosing the "Right" Hardware

Considerations
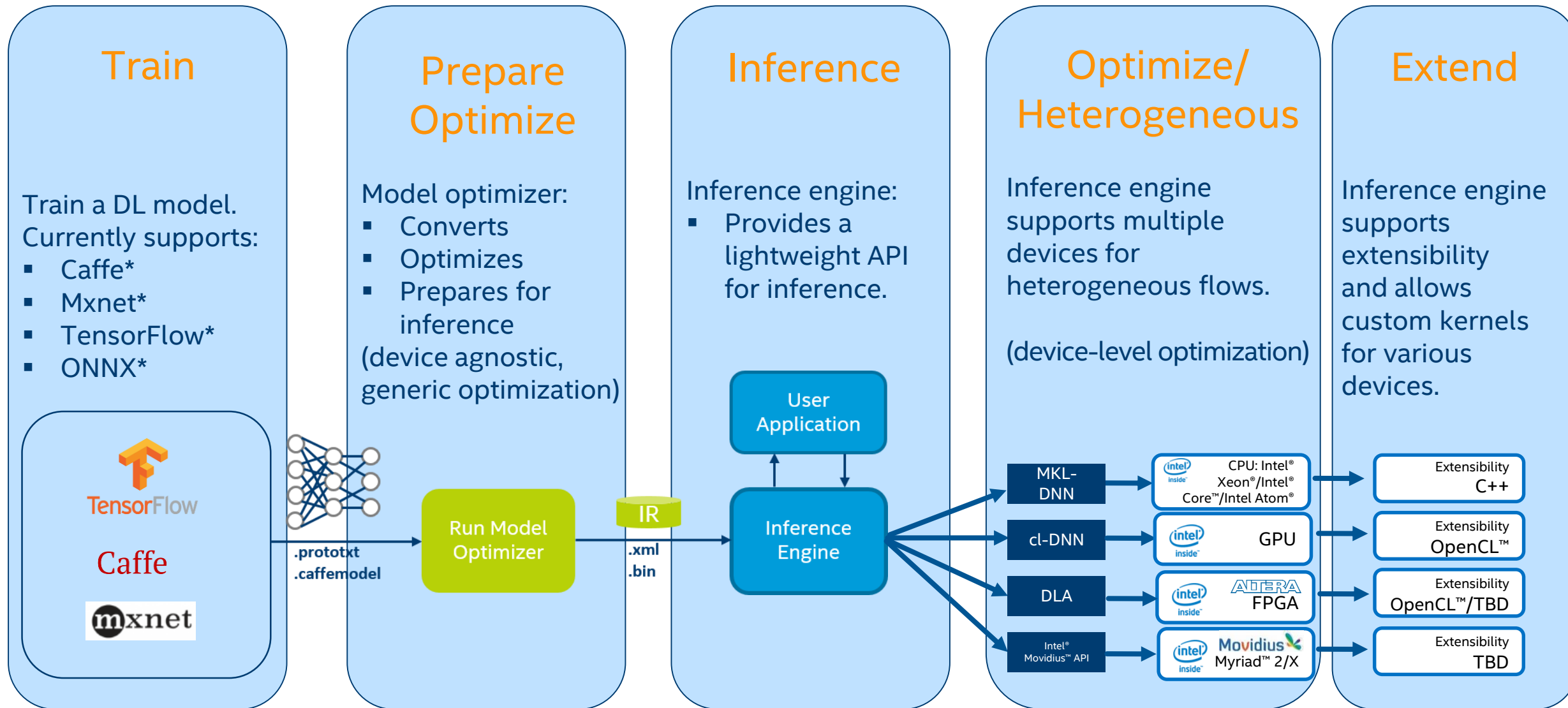
- Workload
- Hardware acceleration
- Heterogeneous computing

Tradeoffs

- Power/performance
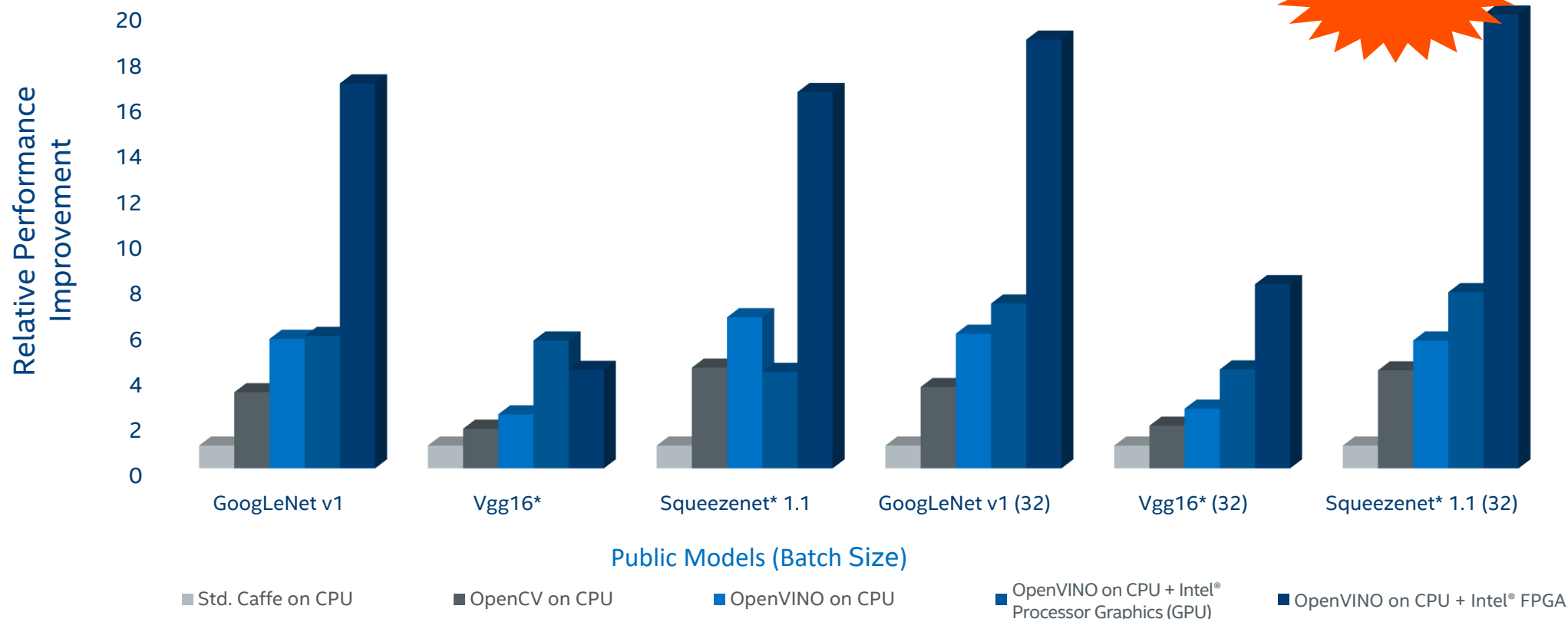- Price
- Software flexibility, portability

Vision Processing Efficiency

Dedicated Hardware

Vision **DSPs**

FPGA

GPU

CPU

Power Efficiency

X100

X10

X1

Computation Flexibility

Application development with OpenVINO™ Toolkit

Train

Train a DL model.
Currently supports:
- Caffe*
- Mxnet*
- TensorFlow*
- ONNX*

TensorFlow

Caffe

mxnet

Prepare Optimize

Model optimizer:
- Converts
- Optimizes
- Prepares for inference

(device agnostic, generic optimization)

.prototxt
.caffemodel

Run Model Optimizer

Inference

Inference engine:
- Provides a lightweight API for inference.

IR

.xml
.bin

User Application

Inference Engine

Optimize/ Heterogeneous

Inference engine supports multiple devices for heterogeneous flows.

(device-level optimization)

MKL-DNN		intel inside	CPU: Intel® Xeon®/Intel® Core™/Intel Atom®
cl-DNN		intel inside	GPU
DLA		intel inside	ALTERA FPGA
Intel® Movidius™ API		intel inside	Movidius Myriad™ 2/X

Extend

Inference engine supports extensibility and allows custom kernels for various devices.

| Extensibility C++ |
| Extensibility OpenCL™ |
| Extensibility OpenCL™/TBD |
| Extensibility TBD |

Deep Learning Workload Performance on Public Models using OpenVINO™ toolkit & Intel® Architecture
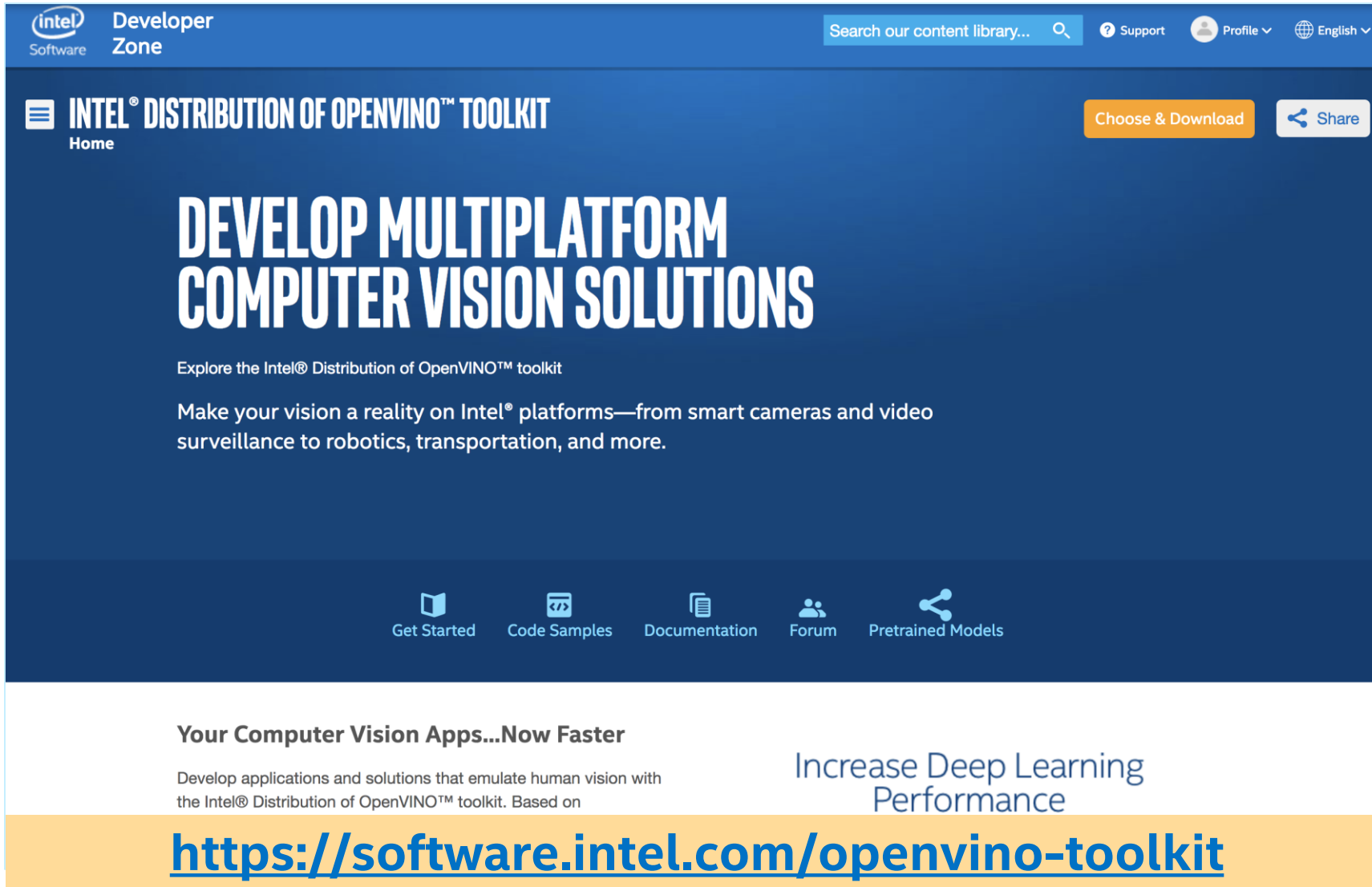
Comparison of Frames per Second (FPS)

19.9x[1]

Relative Performance Improvement (y-axis: 0 to 20)

Public Models (Batch Size): GoogLeNet v1, Vgg16*, Squeezenet* 1.1, GoogLeNet v1 (32), Vgg16* (32), Squeezenet* 1.1 (32)

Legend:
- ■ Std. Caffe on CPU
- ■ OpenCV on CPU
- ■ OpenVINO on CPU
- ■ OpenVINO on CPU + Intel® Processor Graphics (GPU)
- ■ OpenVINO on CPU + Intel® FPGA

Get an even Bigger Performance Boost with Intel® FPGA

[1]Depending on workload, quality/resolution for FP16 may be marginally impacted. A performance/quality tradeoff from FP32 to FP16 can affect accuracy; customers are encouraged to experiment to find what works best for their situation. Performance results are based on testing as of June 13, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. **Configuration:** Testing by Intel as of June 13, 2018. Intel® Core™ i7-6700K CPU @ 2.90GHz fixed, GPU GT2 @ 1.00GHz fixed Internal ONLY testing, Test v3.15.21 – Ubuntu* 16.04, OpenVINO 2018 RC4, Intel® Arria® 10 FPGA 1150GX. Tests were based on various parameters such as model used (these are public), batch size, and other factors. Different models can be accelerated with different Intel hardware solutions, yet use the same Intel software tools.

Intel® OpenVINO™ toolkit on Intel Developer Zone

https://software.intel.com/openvino-toolkit

Inference with Intel® OpenVINO

(a) classification
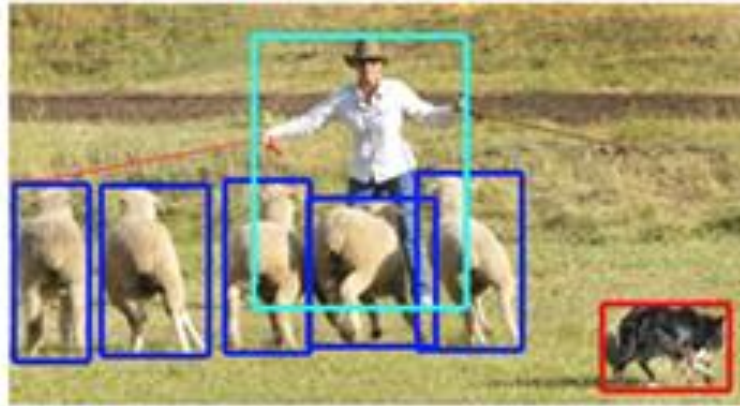
(b) detection

(c) segmentation

https://software.intel.com/openvino-toolkit/documentation/code-samples

Day 2 – November 8th (Hands-on Labs)

09:00 AM - 09:45 AM   Breakfast

10:00 AM - 11:00 AM   Video Performance using Intel® Media SDK

11:00 AM - 12:00 PM   Hardware based Remote Management using Intel® AMT

12:00 PM - 01:00 PM   Lunch

01:00 PM - 05:00 PM   Video Analytics using Intel® OpenVINO™

- Face Detection using OpenVINO™ toolkit lab
- Age and Gender Detection using OpenVINO™ toolkit lab
- Face Data Analysis on the Cloud Lab

Core and Visual Computing Group