AI & Vision update Microsoft

Q3 2019

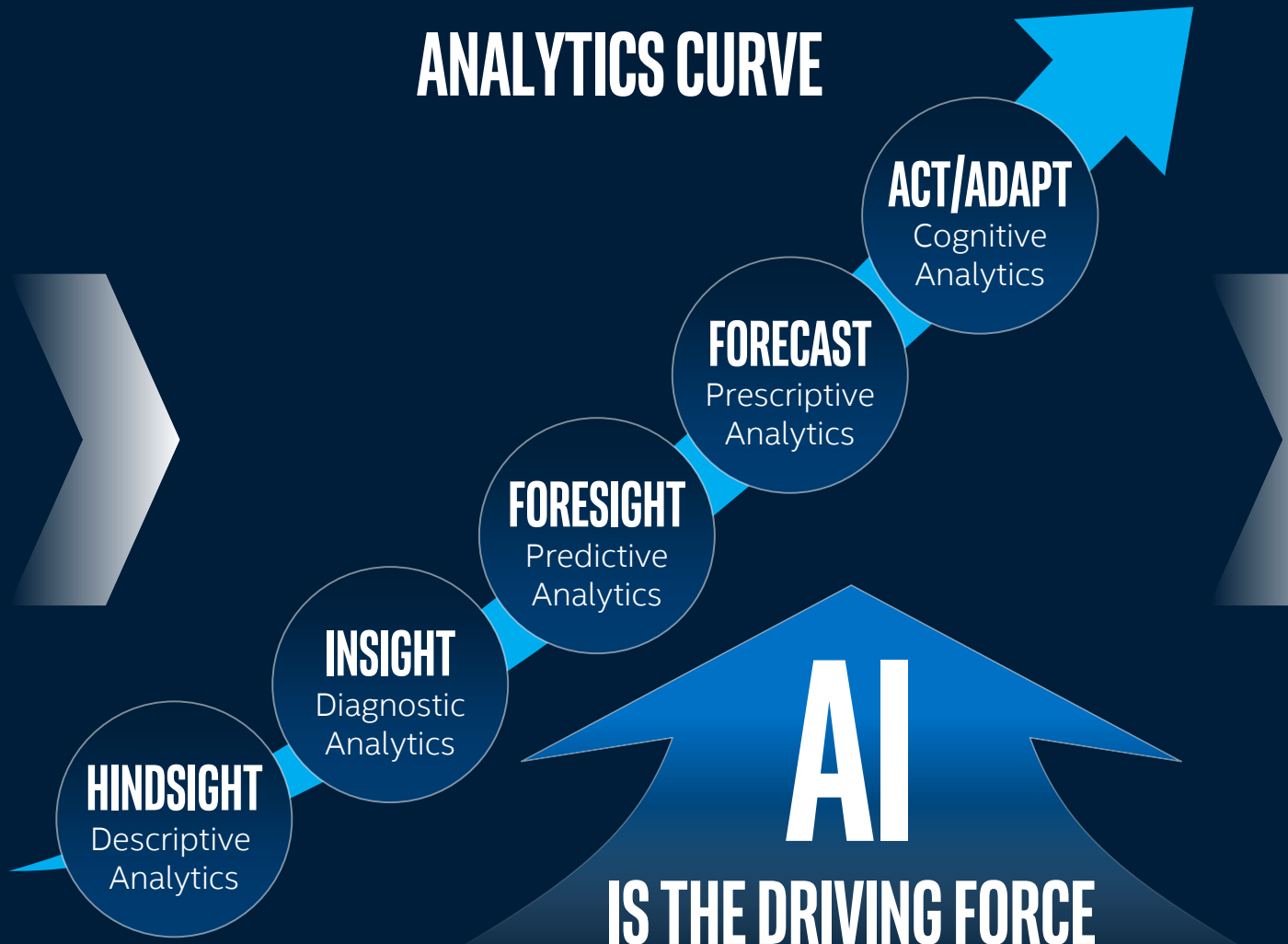
Stefani Eisele , IoT Sales Director EMEA (D/F)
Julian Fischer, BDM AI Sales Germany

WHY AI NOW?

DATA DELUGE (2019)

- **25 GB** [1] per month — INTERNET USER
- **50 GB** [2] per day — SMART CAR
- **3 TB** [2] per day — SMART HOSPITAL
- **40 TB** [2] per day — AIRPLANE DATA
- **1 PB** [2] per day — SMART FACTORY
- **50 PB** [2] per day — CITY SAFETY

ANALYTICS CURVE

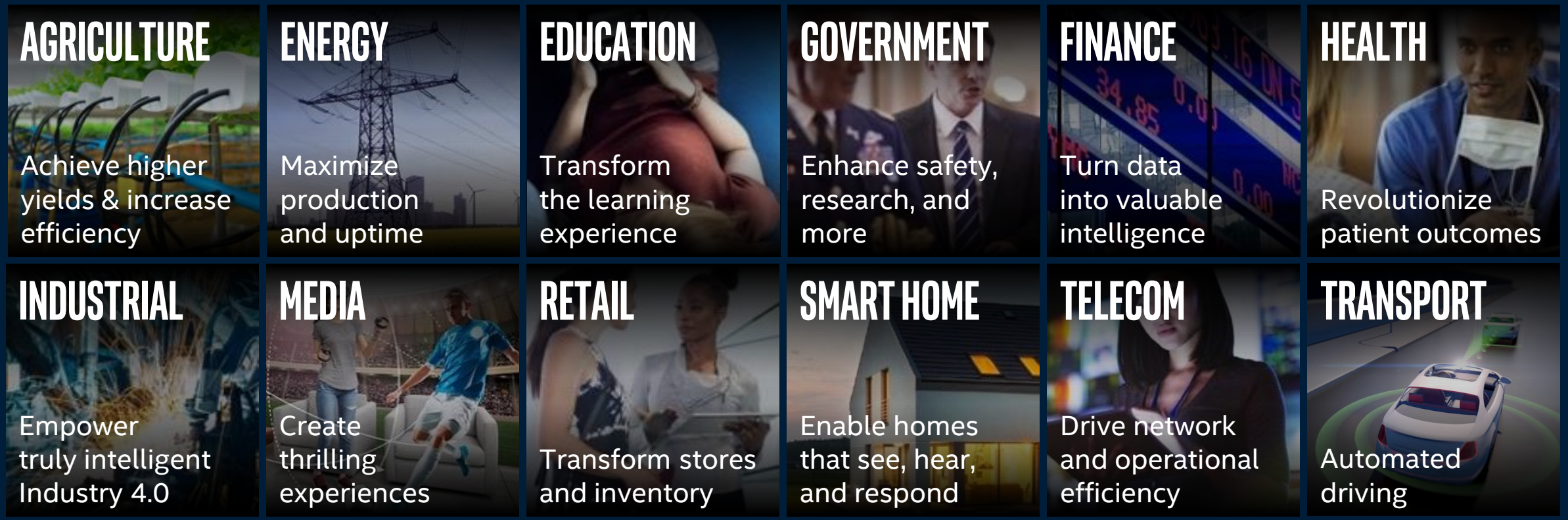


INSIGHTS

- BUSINESS
- OPERATIONAL
- SECURITY

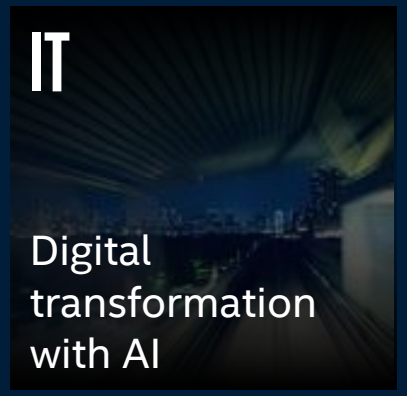
1. Source: http://www.cisco.com/c/en/us/solutions/service-provider/vni-network-traffic-forecast/infographic.html
2. Source: https://www.cisco.com/c/dam/m/en_us/service-provider/ciscoknowledgenetwork/files/547_11_10-15-DocumentsCisco_GCI_Deck_2014-2019_for_CKN__10NOV2015_.pdf

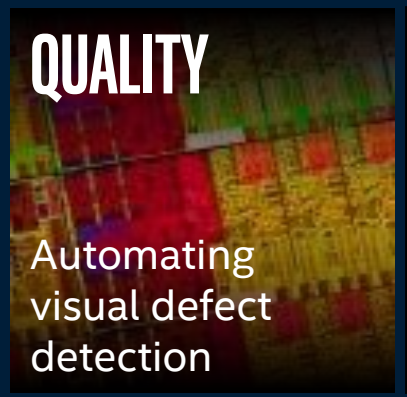
AI SOLUTIONS IN EVERY MARKET

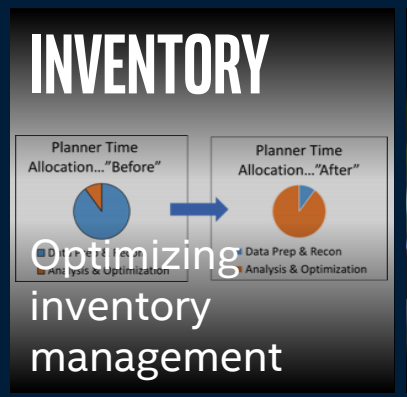


OUR PARTNERS ARE DRIVING REAL-WORLD VALUE WITH INTEL AI

AI INSIDE INTEL







INTEL® IS INFUSING AI INTO EVERYTHING WE DO

THE AI DEVELOPMENT PROCESS

Data cleansing represents the vast majority of the development process

PROOF OF CONCEPT

Label Data | Load Data | Augment Data | Experiment with Topologies | Tune Hyper-parameters | Support Inference | Share Results

15% | 15% | 23% | 15% | 15% | 8% | 8%

Training

Labor-intensive | Compute-intensive | Labor-intensive

BUILD, DEPLOY & SCALE

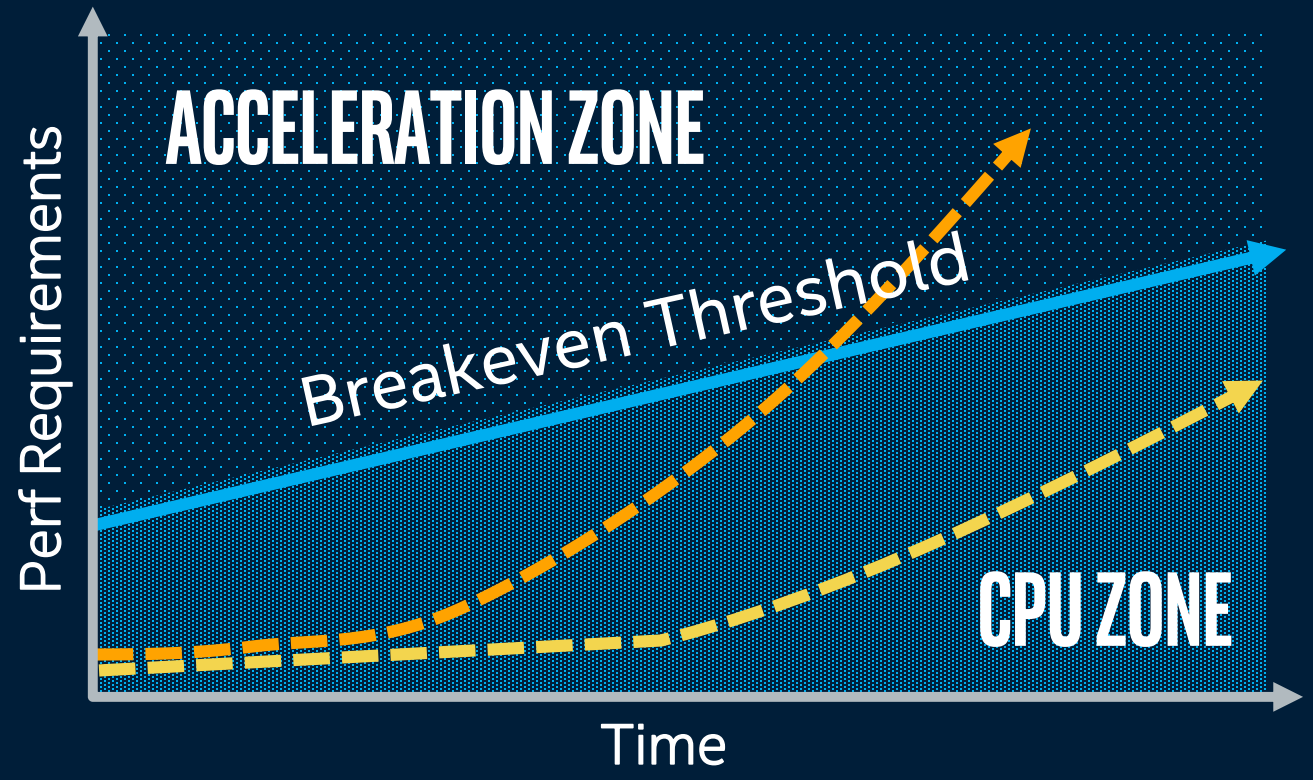
Source Data | Dev Cycle | Scale & Deploy Inference | Scale & Deploy inference within broader application

15% | 15% | 23% | 15% | 15% | 8% | 8%

TIME-TO-SOLUTION

Opportunity → Hypotheses → Data → Modeling → Deployment → Iteration → Evaluation

CPU'S POWER THE MAJORITY OF AI APPLICATIONS TODAY



"A GPU is required for deep learning..."

FALSE

- Most enterprises (---) use CPU for machine & deep learning needs
- Some early adopters (---) may reach a deep learning tipping point when acceleration is needed[1]

1"Most" of enterprise customers based on survey of Intel direct engagements and internal market segment analysis

INTEL AI STRATEGY

SEED & DRIVE THE ECOSYSTEM

- Seed emerging use cases
- Attract & develop top talent
- Pioneer leading-edge AI

SHAPE & WIN INDUSTRY OPEN SOFTWARE STACKS

- Optimize customer software
- Build a unified API
- Evangelize to developers

DELIVER THE BEST AI PLATFORMS

- Extend the CPU
- Most complete portfolio
- Best integrated platforms

INTEL SOFTWARE ECOSYSTEM

USING OPEN AI SOFTWARE

Visit: www.intel.ai/technology

MACHINE LEARNING ⟷ **DEEP LEARNING**

TOOLKITS
App developers

Analytics Zoo
Open source platform for building E2E Analytics & AI applications on Apache Spark* with distributed TensorFlow*, Keras*, BigDL

OpenVINO™
Deep learning inference deployment on CPU/GPU/FPGA/VPU for Caffe*, TensorFlow*, MXNet*, ONNX*, Kaldi*

NAUTA
Open source, scalable, and extensible distributed deep learning platform built on Kubernetes (BETA)

LIBRARIES
Data scientists

Python
- Scikit-learn
- Pandas
- NumPy

R
- Cart
- Random Forest
- e1071

Distributed
- MlLib (on Spark)
- Mahout

Intel-optimized Frameworks

TensorFlow* · Caffe2* · ONNX* · mxnet* · PyTorch* · BigDL

And more framework optimizations underway including PaddlePaddle*, Chainer*, CNTK* & others

KERNELS
Library developers

Intel® Distribution for Python*
Intel distribution optimized for machine learning

Intel® Data Analytics Acceleration Library (DAAL)
High performance machine learning & data analytics library

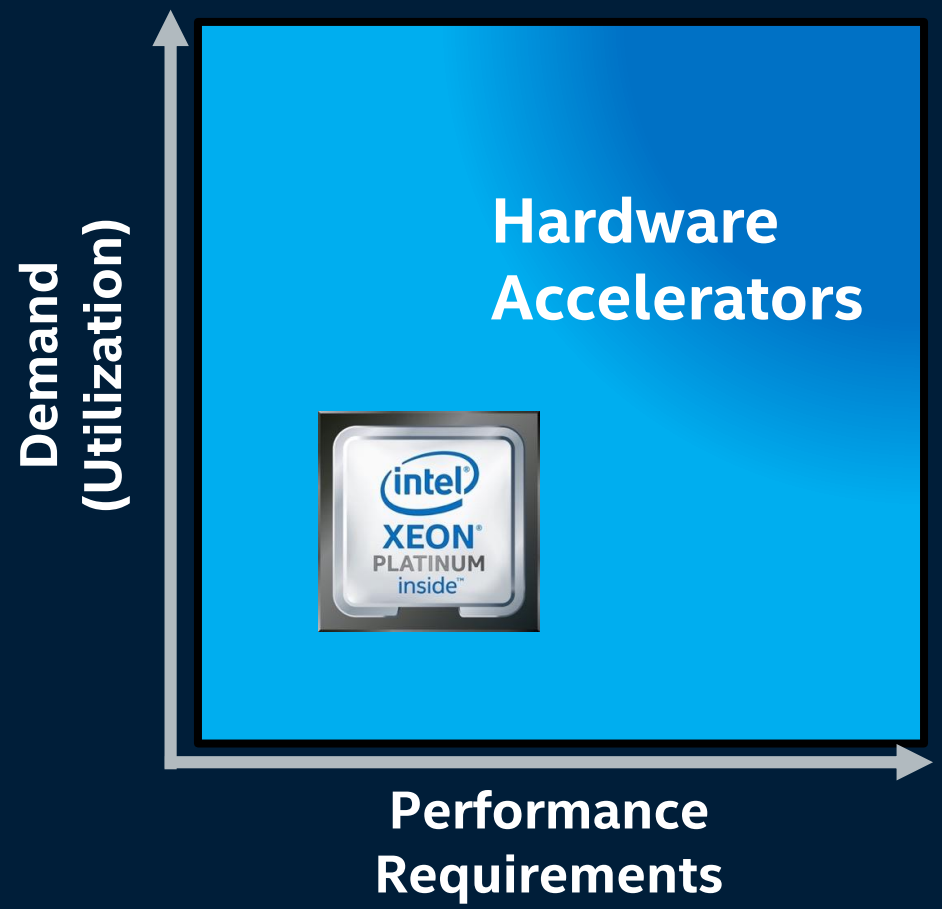
Intel® Math Kernel Library for Deep Neural Networks (MKL-DNN)
Open source DNN functions for CPU / integrated graphics

nGraph
Open source compiler for deep learning model computations optimized for multiple devices (CPU, GPU, NNP) from multiple frameworks (TF, MXNet, ONNX)

[1] An open source version is available at: 01.org/openvinotoolkit
Developer personas show above represent the primary user base for each row, but are not mutually-exclusive
All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

*Other names and brands may be claimed as the property of others.

INTEL XEON PROCESSOR



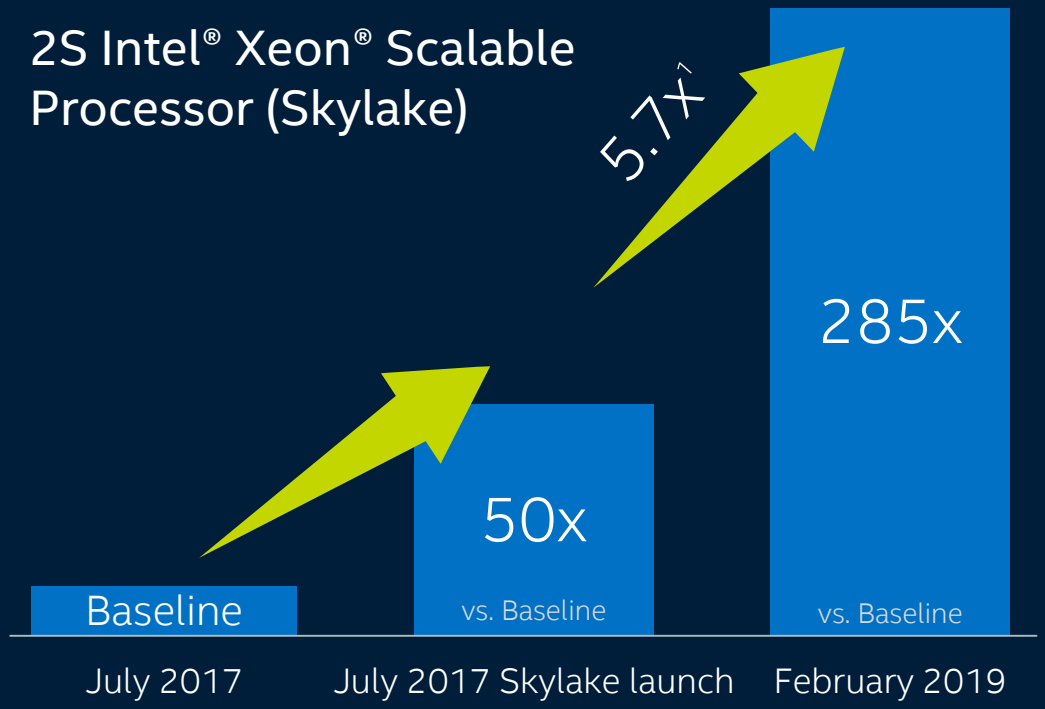
DEEP LEARNING FRAMEWORKS

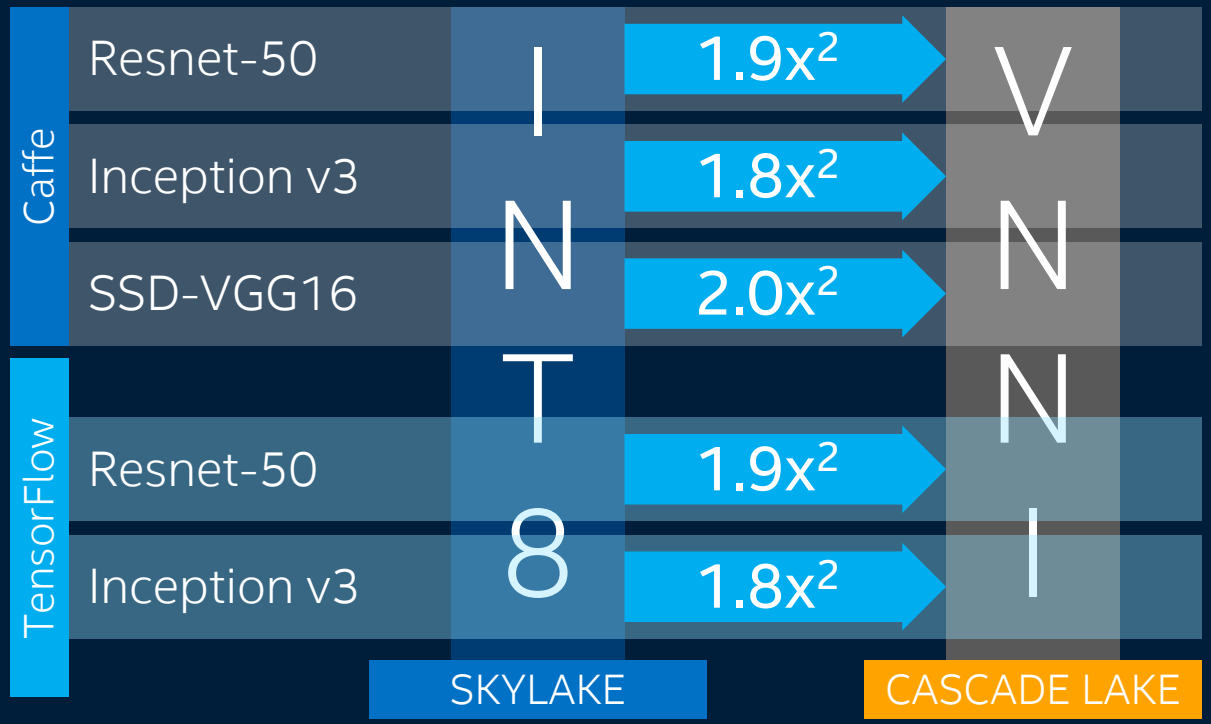
POWER THE VAST MAJORITY OF AI APPLICATIONS TODAY

[1]*Percentage of enterprise customers based on Intel direct engagements*

DEEP LEARNING PERFORMANCE ON CPU

HARDWARE + SOFTWARE IMPROVEMENTS FOR INTEL® XEON® PROCESSORS





1 5.7x inference throughput improvement with Intel® Optimizations for Caffe ResNet-50 on Intel® Xeon® Platinum 8180 Processor in Feb 2019 compared to performance at launch in July 2017. See configuration details on Config 1
Performance results are based on testing as of dates shown in configuration and may not reflect all publicly available security updates.
28/24/2018) Results have been estimated using internal Intel analysis or architecture simulation or modeling and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. No product can be absolutely secure. See configuration disclosure for details. Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit: http://www.intel.com/performance

HARDWARE

Multi-purpose to purpose-built
AI compute from cloud to device

intel AI

INTENSIVE

TRAINING

+

NNP-I

NNP-L

DEEP LEARNING

INFERENCE

+

OpenVINO™

AI AT THE EDGE

REQUIRES THE RIGHT MIX OF PERFORMANCE, POWER & PRICE

Transportation • Retail
Public Sector • Logistics • Smart Cities

Video • Healthcare • Manufacturing
Smart Buildings • Energy

EDGE REQUIREMENTS

LATENCY
BANDWIDTH
SECURITY
CONNECTIVITY

DEVICES • THINGS

EDGE COMPUTE NODE

NETWORK HUB OR REGIONAL DATA CENTER

CORE NETWORK

CLOUD DATA CENTER

VIDEO END CUSTOMERS & USE CASES – SECURITY & BEYOND

CITIES, STATE & FEDERAL

Public Safety & Surveilance
Traffic, Parking and LPR
Emergency Response

FINANCE/BANKING

People Counting
Customer (i.e. Gender, Wait Time)
ATM Facial Recognition

INDUSTRIAL

Machine Vision
Asset Inspection (i.e. Pipeline)
Augmented Reality

CASINO GAMING

Public Safety & Surveillance
Facial Recognition
(see Retail)

TRANSPORTATION

Autonomous Vehicles
Public Safety (i.e. Bus/Rail)
Traffic & People Counting

HOME, RETAIL & SURVEILLANCE

Security & Surveillance
Responsive Retail Advertising
Digital Home Assitant

ROBOTICS

Manufacturing Automation
Industrial (i.e. Pipeline Welding)
Quality Control

DRONES

Emergency Response
Asset Inspection (i.e. Windmill)
Inventory Counting

AI IN PRODUCTION:
FACTORY YIELD IMPROVEMENTS

Aluminum alloy die-casting factories improved defect detection accuracy 5X from manual detection to automatic detection with OpenVINO™.

*Other names and brands names may be claimed as the property of others. 5x accuracy improvement as reported by Alibaba at AIDC on 11/14/18 For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

INTEL AI VISION PORTFOLIO

END POINT
IOT SENSORS

Intel Movidius inside

Vision & Inference

Intel Arria 10 inside

EDGE
GATEWAYS

Intel® Vision Accelerators

Intel Atom inside — Basic Inference, Media & Vision

Intel Core i7 8th Gen

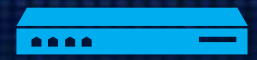
Intel Xeon inside — Video Storage, Analytics, DL Servers

+

Intel Movidius inside — Best Efficiency, Lowest Power Mid/Small Memory Footprint

Intel Arria 10 inside — High Perf, Large Memory Custom/New HW Architecture

DATA CENTER
SERVERS & APPLIANCES



Intel Xeon Platinum inside

+

Most Use Cases

Intel Nervana inside (COMING SOON) — ***NNP-L / NNP-I*** Most Intensive Use Cases

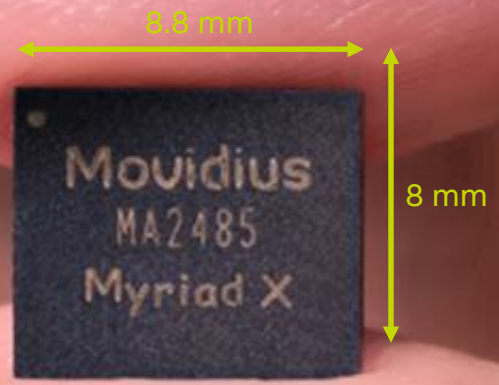
Intel Stratix 10 inside — Flexible & Memory Bandwidth-Bound Use Cases

OpenVINO™ Toolkit
Deploy across Intel® CPU, GPU, VPU, FPGA; Leverage common algorithms

OPENVINO DRIVES AI SCALE AND VALUE AT THE EDGE

INTEL® MOVIDIUS™ MYRIAD™ VPU TECHNOLOGY

Dedicated Imaging, Vision, and Deep Neural Networks at the Edge



- 8 x 8 mm
- 4 TOPS
- < 2 Watts

INTEL® NEURAL COMPUTE STICK 2

MORE CORES. MORE AI INFERENCE.

- Start quickly with plug-and-play simplicity
- Develop on common frameworks and out-of-box sample applications
- Prototype on any platform with a USB port
- Operate without cloud compute dependence

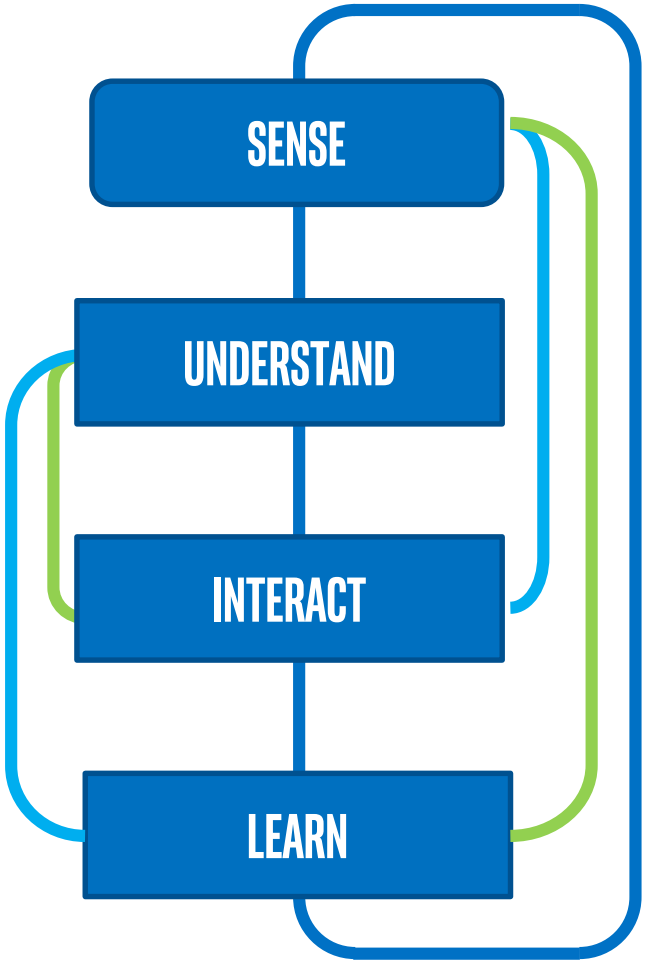
Intel® Movidius™ Myriad™ X VPU delivers industry leading performance

+

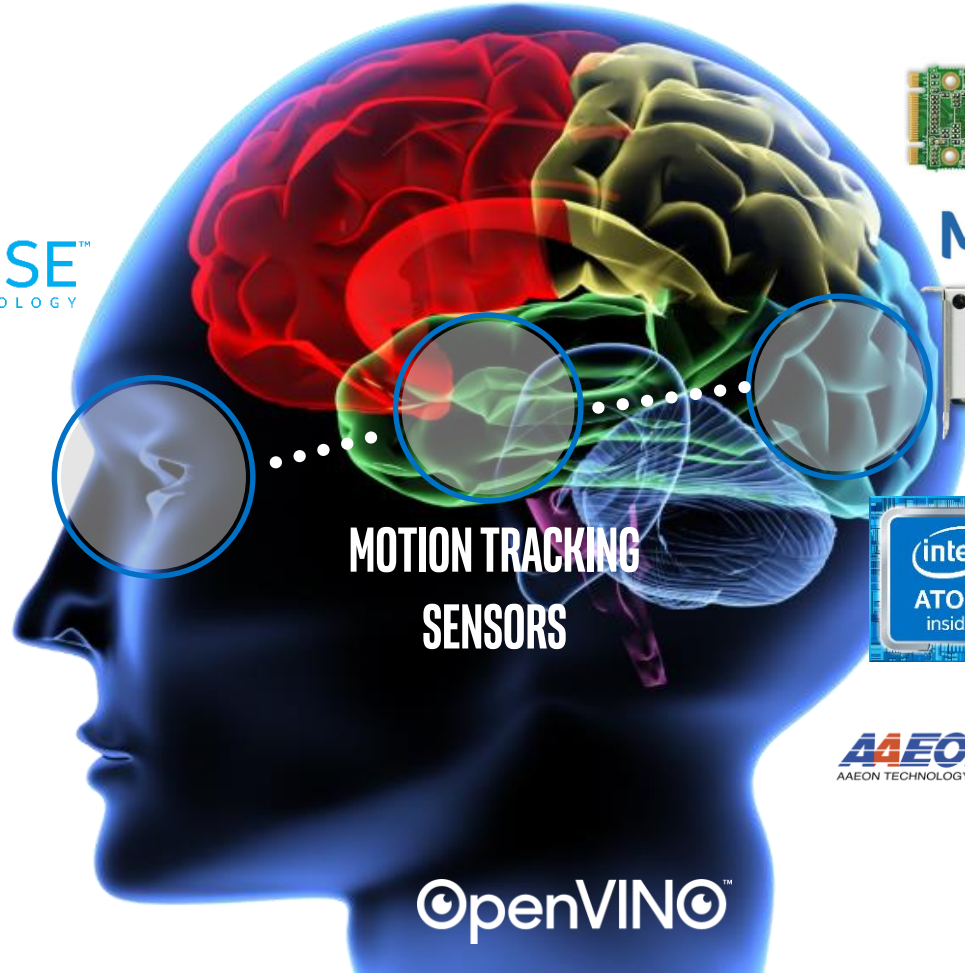
OpenVINO™

Intel® Distribution of OpenVINO™ toolkit accelerates solution development and streamlines deployment

"MIMICKING" THE HUMAN PERCEPTUAL SYSTEM



intel REALSENSE TECHNOLOGY



intel XEON PLATINUM inside

intel NERVANA inside COMING SOON

Movidius

intel ATOM inside

intel CORE i7 8th Gen

AAEON TECHNOLOGY INC.

IEI Integration Corp.

* Other names and brands may be claimed as the property of others

SIMPLIFY AI

USING COMMUNITY SOLUTIONS

SOLVE

***Solve** your challenge using one of 70+ AI solutions in the Intel AI Builders program*

Visit: builders.intel.com/ai

OR

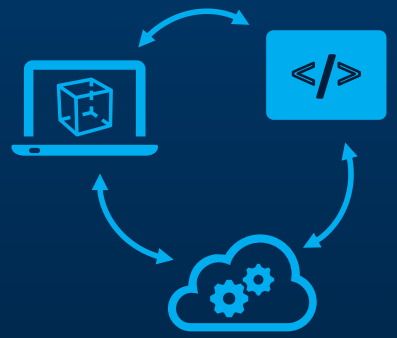
DEPLOY

***Deploy** AI-optimized systems including Intel® Select Solutions and ecosystem Partner solutions*

Visit: builders.intel.com/ai

OR

DEVELOP



***Develop** your own AI solutions using Intel's FREE¥ developer courses, tools and cloud access*

Visit: software.intel.com/ai

¥Free = available to download/access at no cost to qualified developers who are enrolled in the program
*Other names and brands may be claimed as the property of others.

PARTNER WITH INTEL TO ACCELERATE YOUR AI JOURNEY

WHY INTEL® AI?

SIMPLIFY AI
using community solutions

TAME YOUR DATA
with a robust data layer

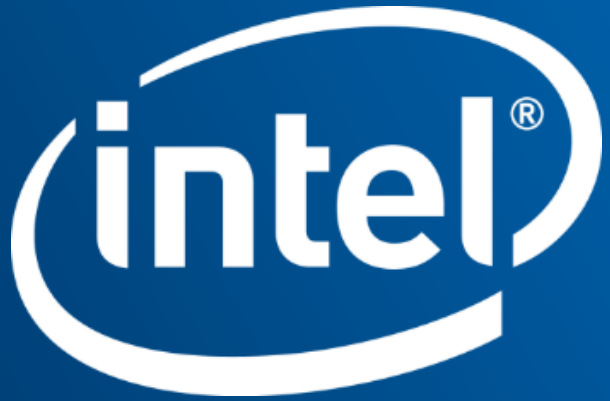
CHOOSE ANY APPROACH
from machine to deep learning

SPEED UP DEVELOPMENT
with open AI software

DEPLOY AI ANYWHERE
with unprecedented HW choice

SCALE WITH CONFIDENCE
on the engine for IT & cloud

LEARN MORE AT WWW.INTEL.AI


intel


intel
CV, DL & AI
LIVING AT THE EDGE
CO-PRESENTING DR. LYDIA NEMEC,
DATA SCIENTIST CARL ZEISS
SEP, 2019
OpenVINO™

HISTORY

- Intel acquired Itseez in July, 2016 and Itseez joined Intel IOT
- Itseez is widely known as OpenCV developer/maintainer
- Team focus:
- CV/DL ecosystem enablement: OpenCV, OpenVINO™
- CV/DL models/algorithms/applications

OpenVINO™

VISION/VIDEO IS DRIVING ARTIFICIAL INTELLIGENCE TO THE EDGE NOW

2500 PETABYTES VIDEO SURVEILLANCE DATA GENERATED DAILY IN 2019

- Bandwidth & Storage Costs
- Latency & Response Predictability
- Privacy & Security Limitations

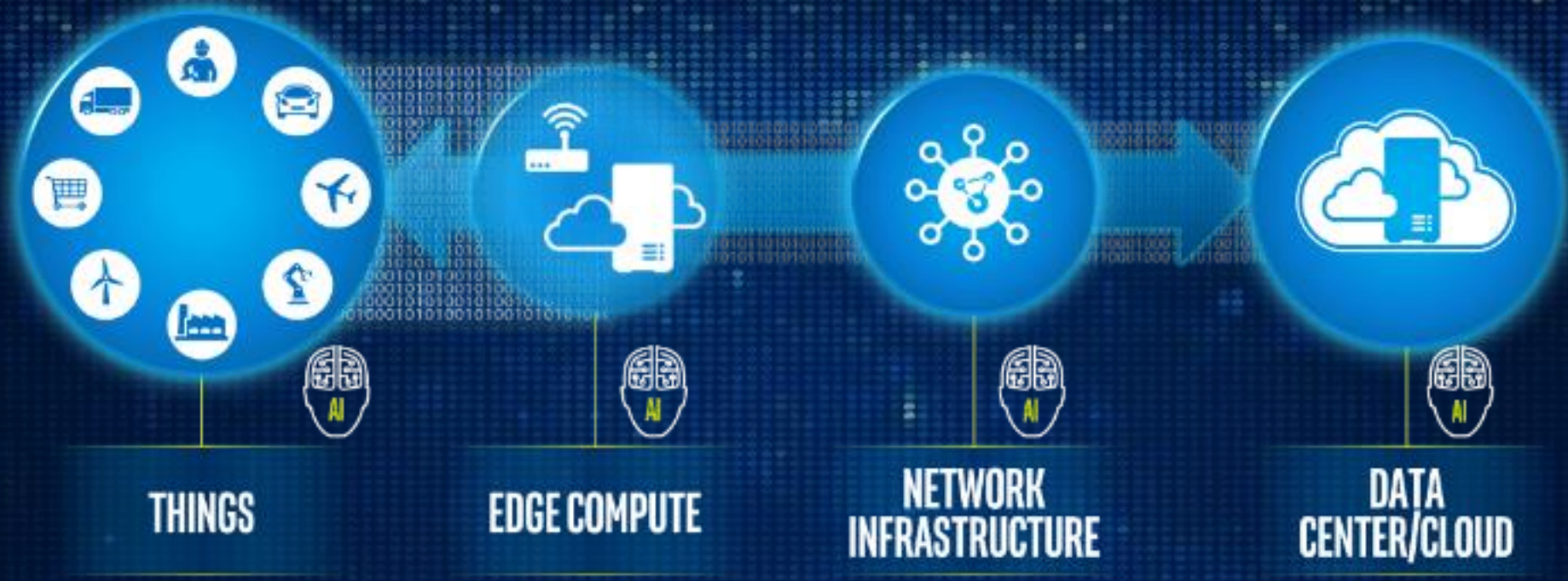
DEEP LEARNING

The overall computer vision market to reach **$17.38 Billion** by 2023[1]

The video analytics market to reach **$11.17 Billion** by 2022 at a CAGR of **21.5%**[3]

Deep learning revenue to grow from **$655 Million** in 2016 to **$35 Billion** by 2025 --a CAGR of **53%**[2]

DRIVERS FOR EDGE: BANDWIDTH, LATENCY, PRIVACY, SECURITY



45% *of data will be stored, analyzed, and acted on at the edge by 2019*[1]

43% *share of AI tasks taking place on edge devices (vs. cloud) in 2023*[2]

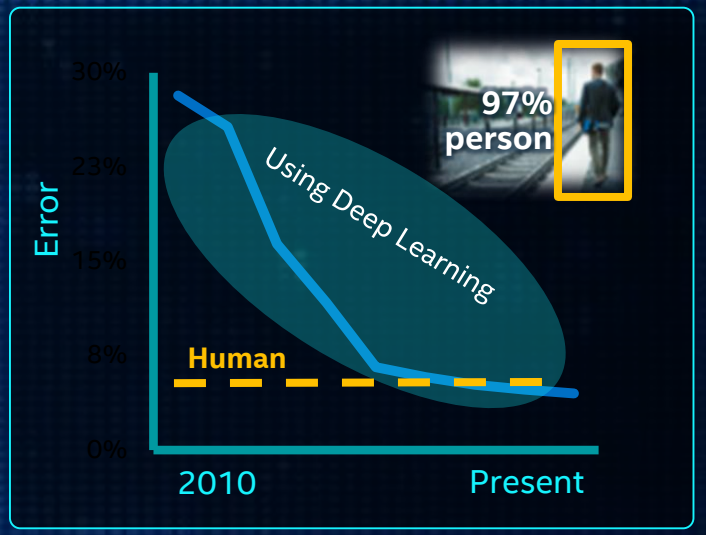
15X *growth in devices with edge AI capabilities by 2023*[2]

1. Source: IDC FutureScape: Worldwide Internet of Things 2017 Predictions
2. ABI Research

WHY EDGE HAS NOT BECOME SMART BEFORE ?

TRADITIONAL COMPUTER VISION

DEEP LEARNING



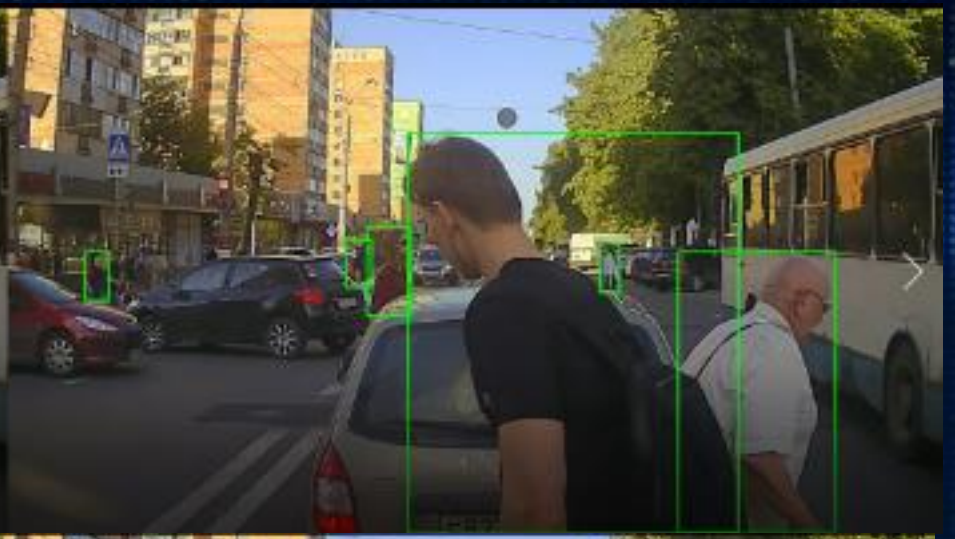
Source: ILSVRC ImageNet winning entry classification error rate each year 2010-2016

- Superior Accuracy
- Forensic to Predictive

MACHINE LEARNING (ML)

DR = **70 %** @ 0.1 % fppi
Dataset size: **10,000**

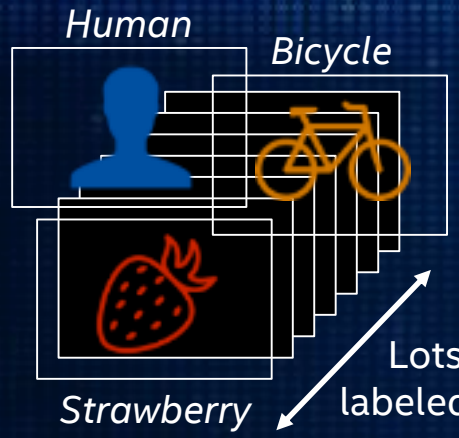
DEEP LEARNING (DL)



DR = **90+ %** @ 0.1 % fppi
Dataset size: **140,000**

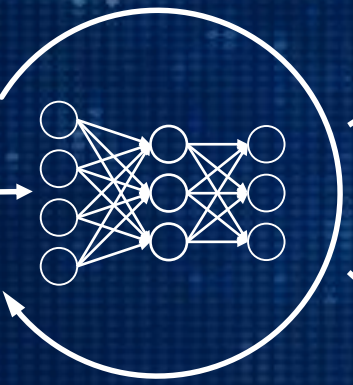
DEEP LEARNING USER WORKFLOW

Data Collection & Annotation



Lots of labeled data

DL Model Selection & Training



DL Inference from Training Framework

HW

HW Optimized Training Framework

DL Inference via OpenVINO™

Model Optimizer (MO)

Inference Engine (IE)

HW

Training

Inference/Deployment

OpenVINO™ FOR CV/DL APPLICATIONS

OpenVINO™:

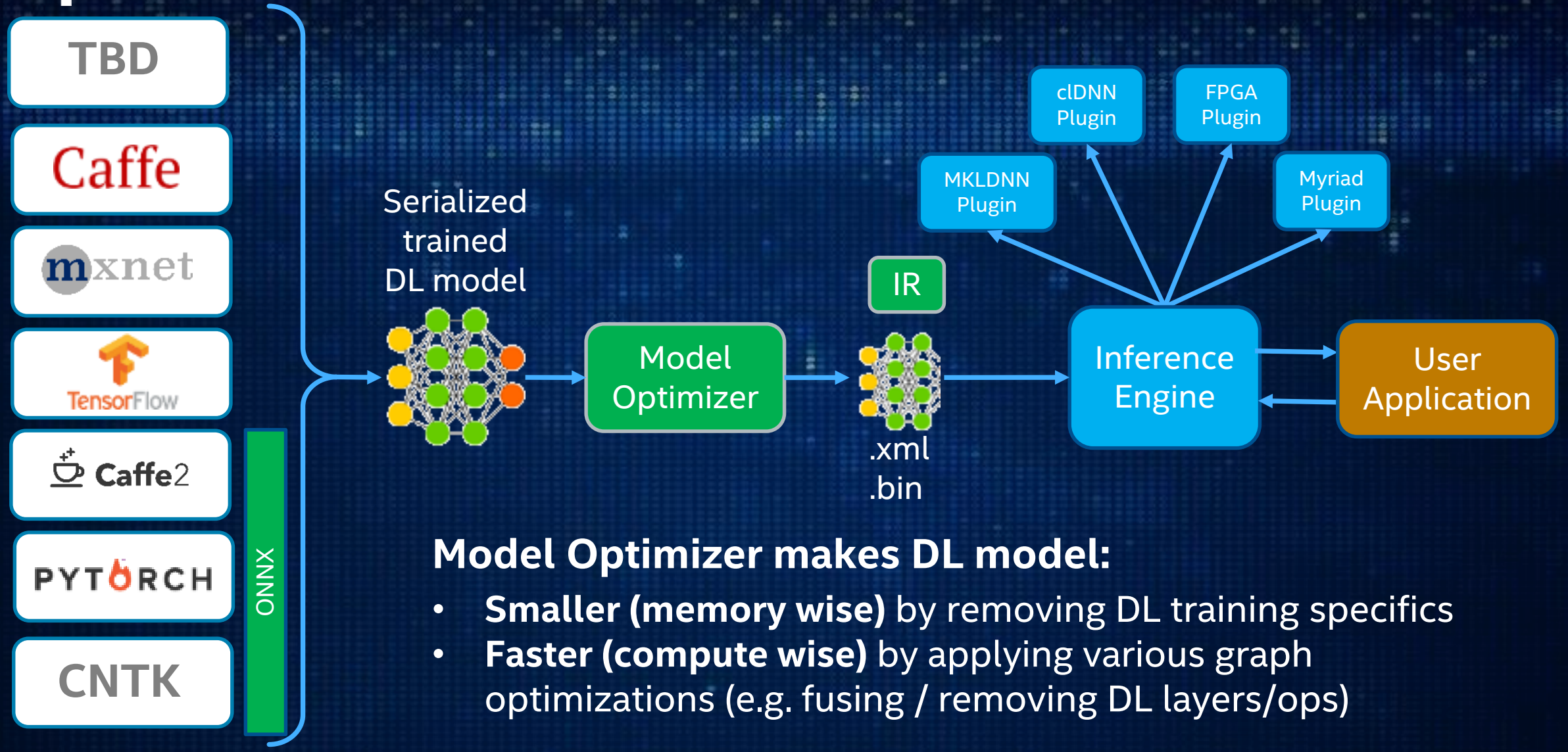
- A fantastic tool for E2E CV/DL application development and deployment
- High CV and DL inference performance on Intel accelerators – **CPU, GPU, Movidius™ VPU and FPGA** - via uniform API and heterogeneous execution model
- No training overhead or specifics, minimal footprint, highly portable code

OpenVINO™ contains:

- **OpenCV** – fastest build optimized for Intel hardware
- **Model Optimizer (MO)** – offline DL model memory/compute optimization tool
- **Inference Engine (IE)** – DL inference solution with ultimate focus on performance +Samples for enablement, rapid prototyping and benchmarking
- **Open Model Zoo** - public and Intel trained DL model sets + demos
- **DL Workbench** - profiler which visualizes key performance metrics such as latency, throughput and performance counters for neural network topologies and its layers.

OpenVINO™ was *open sourced in Oct'18*

OpenVINO™ MODEL OPTIMIZER (MO)



Model Optimizer makes DL model:

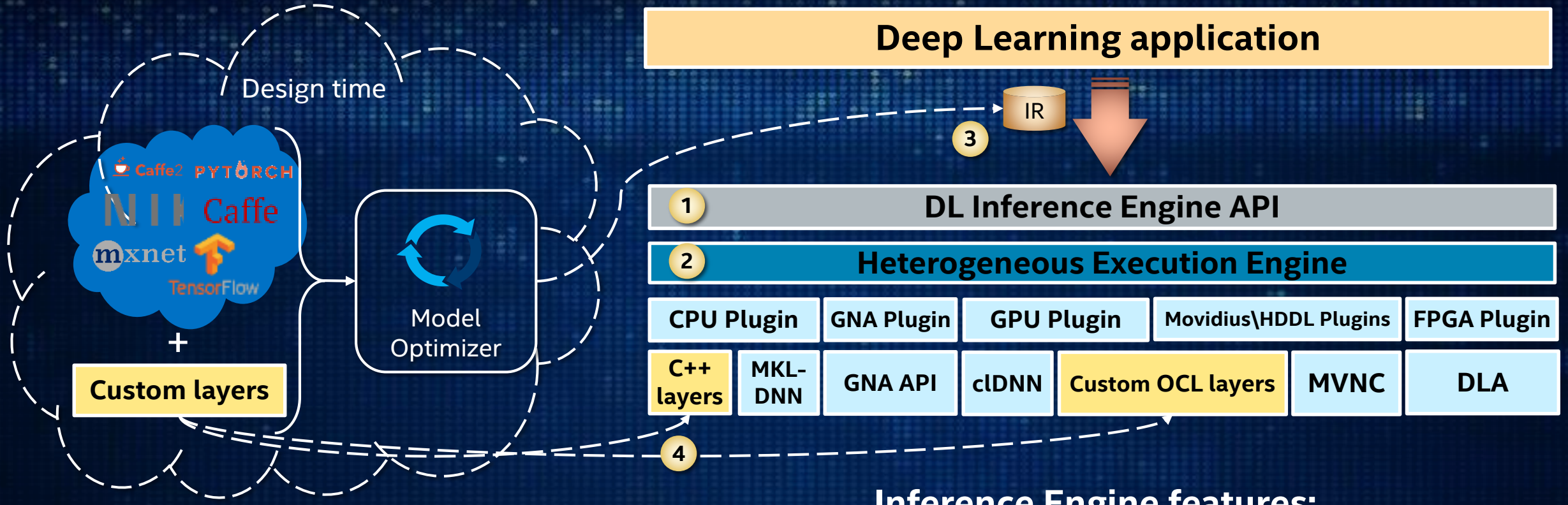
- **Smaller (memory wise)** by removing DL training specifics
- **Faster (compute wise)** by applying various graph optimizations (e.g. fusing / removing DL layers/ops)

FRAMEWORK INFERENCE VS OPENVINO™ INFERENCE

Fast Results on Intel Hardware, even before using Accelerators

[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. The benchmark results reported in this deck may need to be revised as additional testing is conducted. Performance results are based on testing as of April 10, 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 April 10, 2018. Intel® Core™ i7-6700K CPU @ 2.90GHz fixed, GPU GT2 @ 1.00GHz fixed Internal ONLY testing, Test v312.30 – Ubuntu* 16.04, OpenVINO™ 2018 RC4. 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.

OpenVINO™ INFERENCE ENGINE (IE)



1. Single API solution across accelerators
2. Heterogeneous DL network execution across accelerators
3. Framework independent lightweight IR
4. Customizations in C++ and OpenCL languages

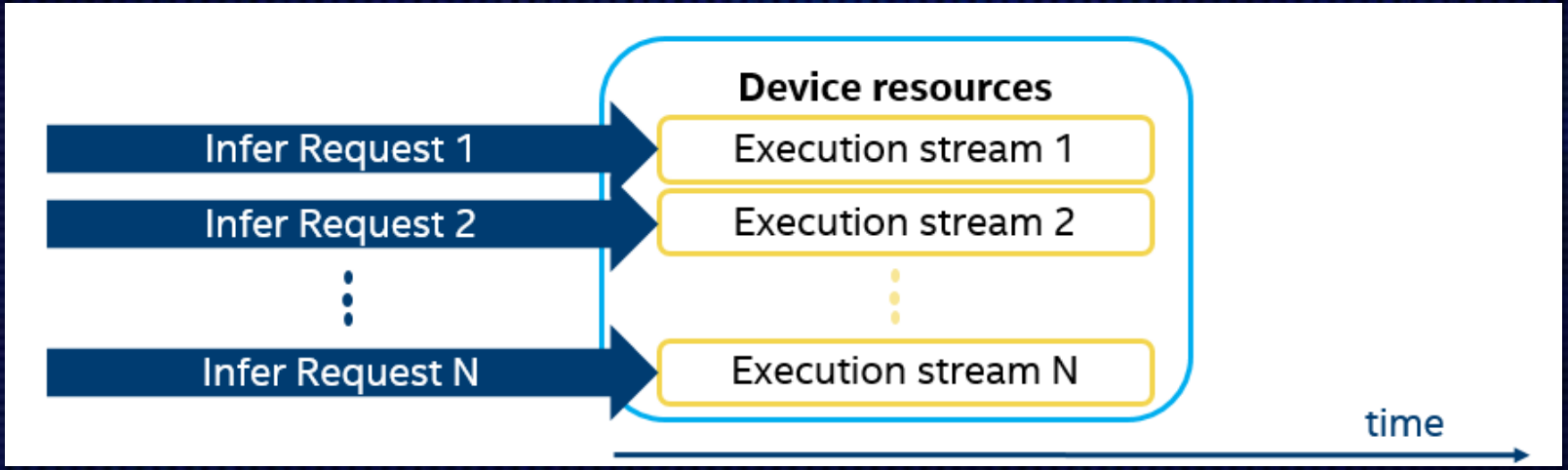
Inference Engine features:

- Superior inference performance on Intel HW
- Minimal memory footprint (model, binary, etc)
- Absolute minimum of dependencies (e.g. no training framework in runtime)

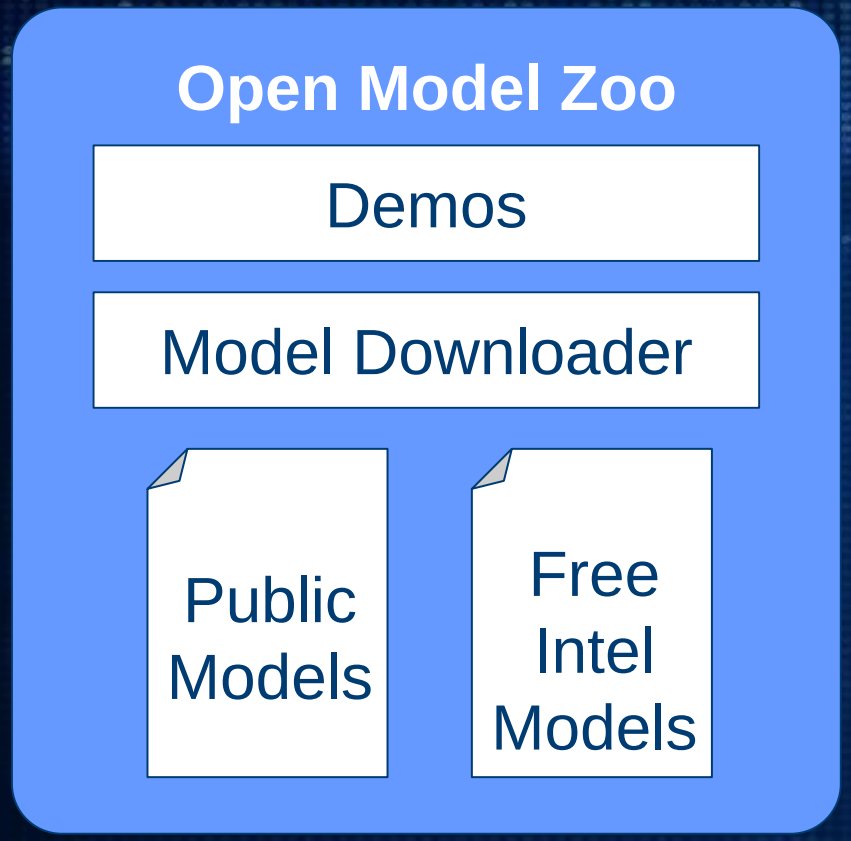
INFERENCE ENGINE (IE) PERFORMANCE FEATURES

Asynchronous API - checks the execution status of inference with the wait, or specify a completion callback

"Throughput" Mode allows the Inference Engine to efficiently run multiple infer requests simultaneously, greatly improving the overall throughput. Device resources are divided into execution "**streams**" – parts which runs infer requests in parallel



OpenVINO™ MODEL ZOO



OpenVINO™ Model Zoo:

- **Public models**: popular and widely used public DL topologies
- **40+ Free Intel models**: low memory/light compute edge/task specific DL models for Object Detection, Object Recognition, Reidentification, Semantic Segmentation, Pose estimation, Image Processing, Text Detection
- **Demos:** 15 C++ and 6 Python

More info:
https://github.com/opencv/open_model_zoo

COMPUTER VISION ANNOTATION TOOL (CVAT)

Object Detection

Image Classification

Semantic Segmentation

Auto annotation Using TF OD

OpenVINO™ WHAT IS NEW IN 2019 R2

feature to provide visualization of ke… …etrics
…ural network topologies and its l… …ncludes easy
…tion, accuracy check, and au… …on of optimal

…ing across av… …o achieve higher throughput when

…to available devices, and provide Query API to get device configuration and metrics to help determin… …ms for deploying deep learning applications. This saves time and eliminates manually loading of ind… …ns.

- **Supports serialized FP16 Intermediate** … (IR) to work uniformly across supported platforms. This helps reduce model size by 2x compared … …ing utilization of available device memory and portability of models. (For CPU, inference remains at FP3…

- **Enables new use ca… translation, n…** with support fo… …non-vision topologi…

- **Provi… distribution methods** - Binar… … Hub*, and .zip/.tgz files through G… toolkit with minimal to no overhead in… …

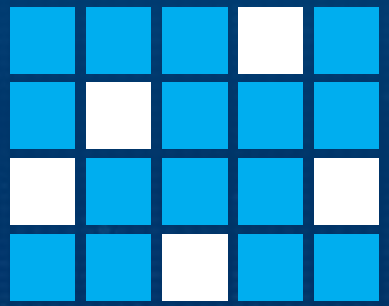
https://iotevents.intel.com/MunichSmartVideo2019/

17. September 2019
and/or
18. September 2019
(1 day, technical and business tracks in parallel

DRIVING INDUSTRY MOMENTUM
IN DEEP LEARNING INFERENCE AT THE EDGE

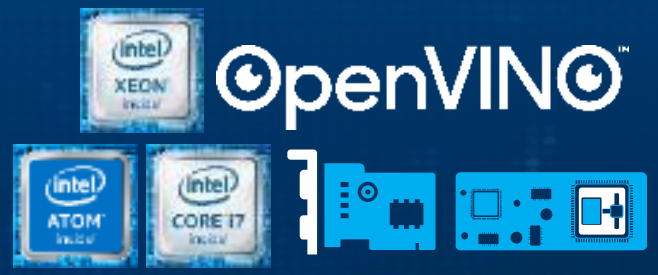
DEVELOPERS

Over **15000** Unique Developers



CUSTOMERS

25+ Customer Products Launched Based on OpenVINO™ toolkit



PRODUCTS

Breadth of vision product portfolio, **and now, introducing...**

STRONG ADOPTION + RAPIDLY EXPANDING CAPABILITY


intel