"AI BEYOND THE BUZZWORD: DO IT WELL OR DO IT TWICE"

Walter Riviera
AI – Technical Solution Specialist
EMEA

NOTICES AND DISCLAIMERS

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

No product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .

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/benchmarks .

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

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.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

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.

Arria, Celeron, Intel, the Intel logo, Intel Atom, Intel Core, Intel Nervana, Intel Optane, Intel Xeon, Iris, Movidius, OpenVINO, Pentium, Stratix and the Stratix logo and are trademarks of Intel Corporation in the U.S. and/or other countries.

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

INTEL® AI TRAILER

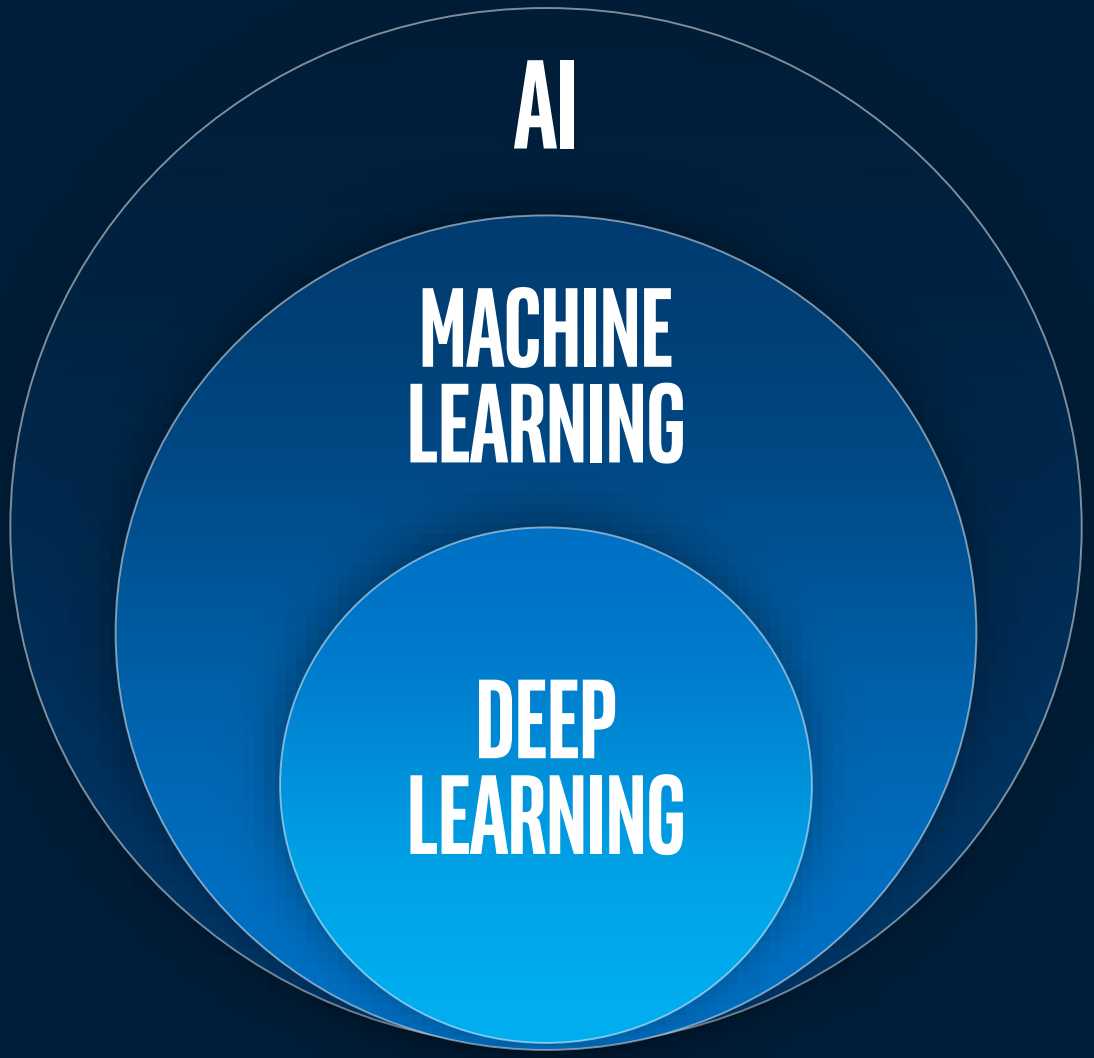
https://digitallibrary.intel.com/content/digital-library/us/en/assetdetail.html/content/dam/intel-digital-library/video-and-animation/intel-corporate-narrative-artificial-intelligence-video.mp4

INTEL® AI TODAY

THE JOURNEYS

INTEL® AI TOMORROW

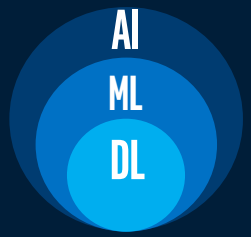
WHAT IS AI?



MACHINE LEARNING VS DEEP LEARNING?

MACHINE LEARNING

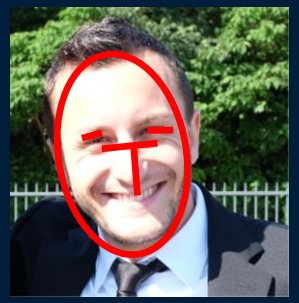
How do you engineer the best features?



DEEP LEARNING

Find the best parameters

$N \times M$



$(f_1, f_2, \ldots, f_K)$

Roundness of face
Dist between eyes
Nose width
Eye socket depth
Cheek bone structure
Jaw line length
...etc.

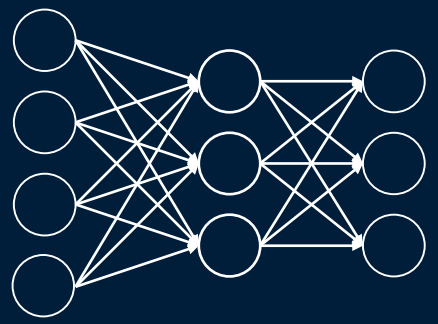
CLASSIFIER ALGORITHM

SVM
Random Forest
Naïve Bayes
Decision Trees
Logistic Regression
Ensemble methods

Walter

$N \times M$

NEURAL NETWORK



Walter

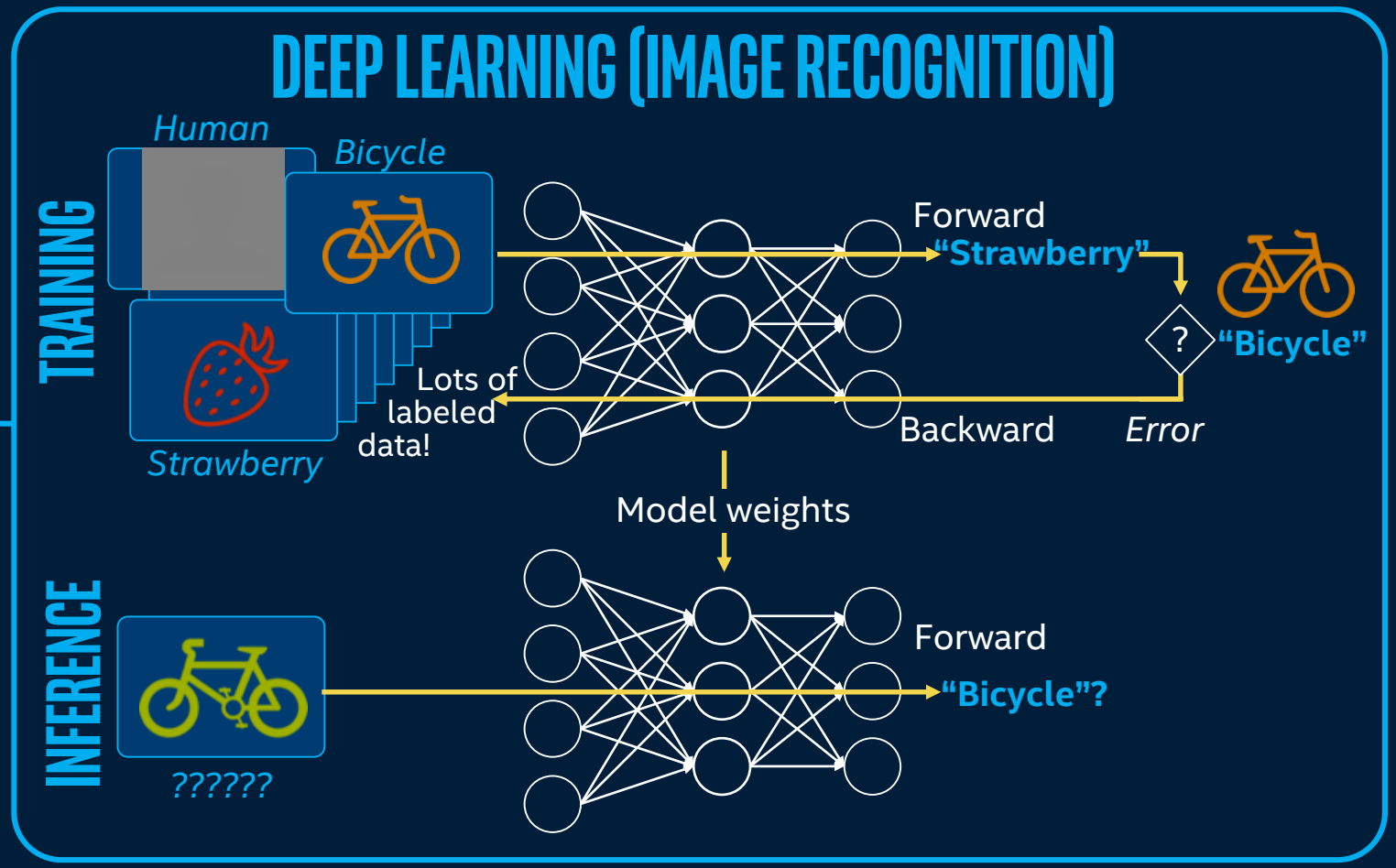
HOW DOES DEEP LEARNING WORK?

MACHINE LEARNING

- Regression
- Classification
- Clustering
- Decision Trees
- Data Generation

DEEP LEARNING

- Image Processing
- Speech Processing
- Natural Language Processing
- Recommender Systems
- Adversarial Networks
- Reinforcement Learning



CHOOSE THE BEST AI APPROACH FOR YOUR CHALLENGE

SPEED UP DEVELOPMENT
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

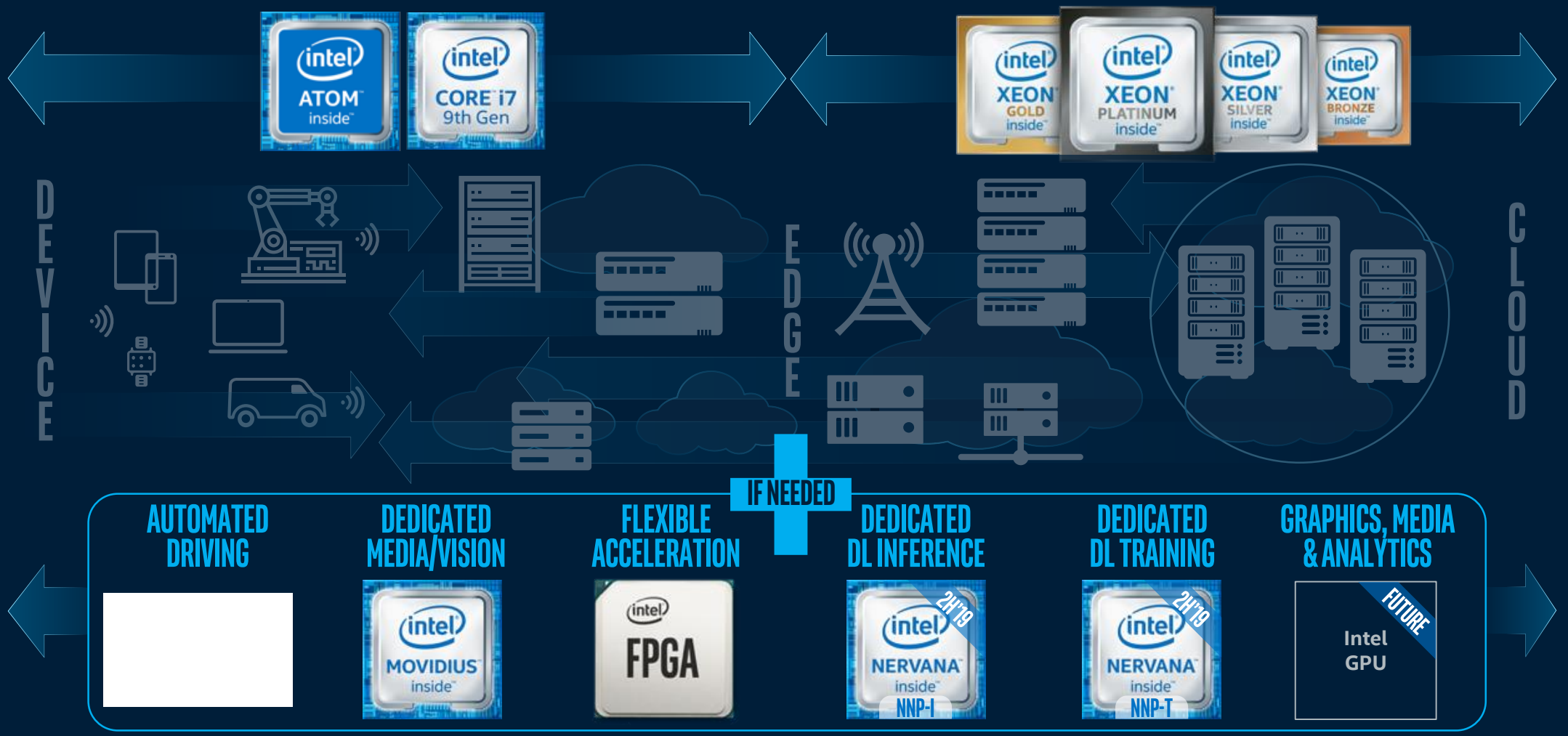
nGraph
Open source compiler for deep learning model computations optimized for multiple devices (CPU, GPU, NNP) from multiple frameworks (TF, MXNet, ONNX)

[1] An open source version is available at: 01.org/openvinotoolkit *Other names and brands may be claimed as the property of others.
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.

DEPLOY AI ANYWHERE

WITH UNPRECEDENTED HARDWARE CHOICE

Visit: www.intel.ai/technology



BREAKING BARRIERS BETWEEN AI THEORY AND REALITY

PARTNER WITH INTEL® TO ACCELERATE YOUR AI JOURNEY

CHOOSE ANY APPROACH
from machine to deep learning

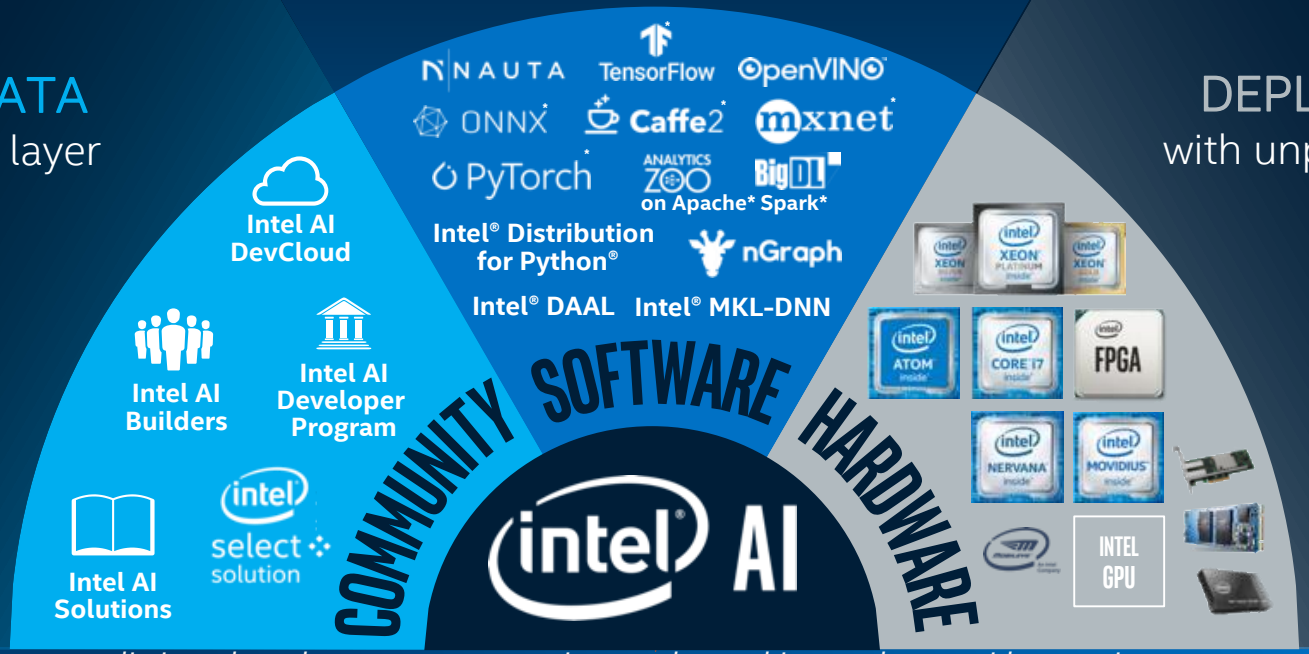
SPEED UP DEVELOPMENT
using open AI software

TAME YOUR DATA
with a robust data layer

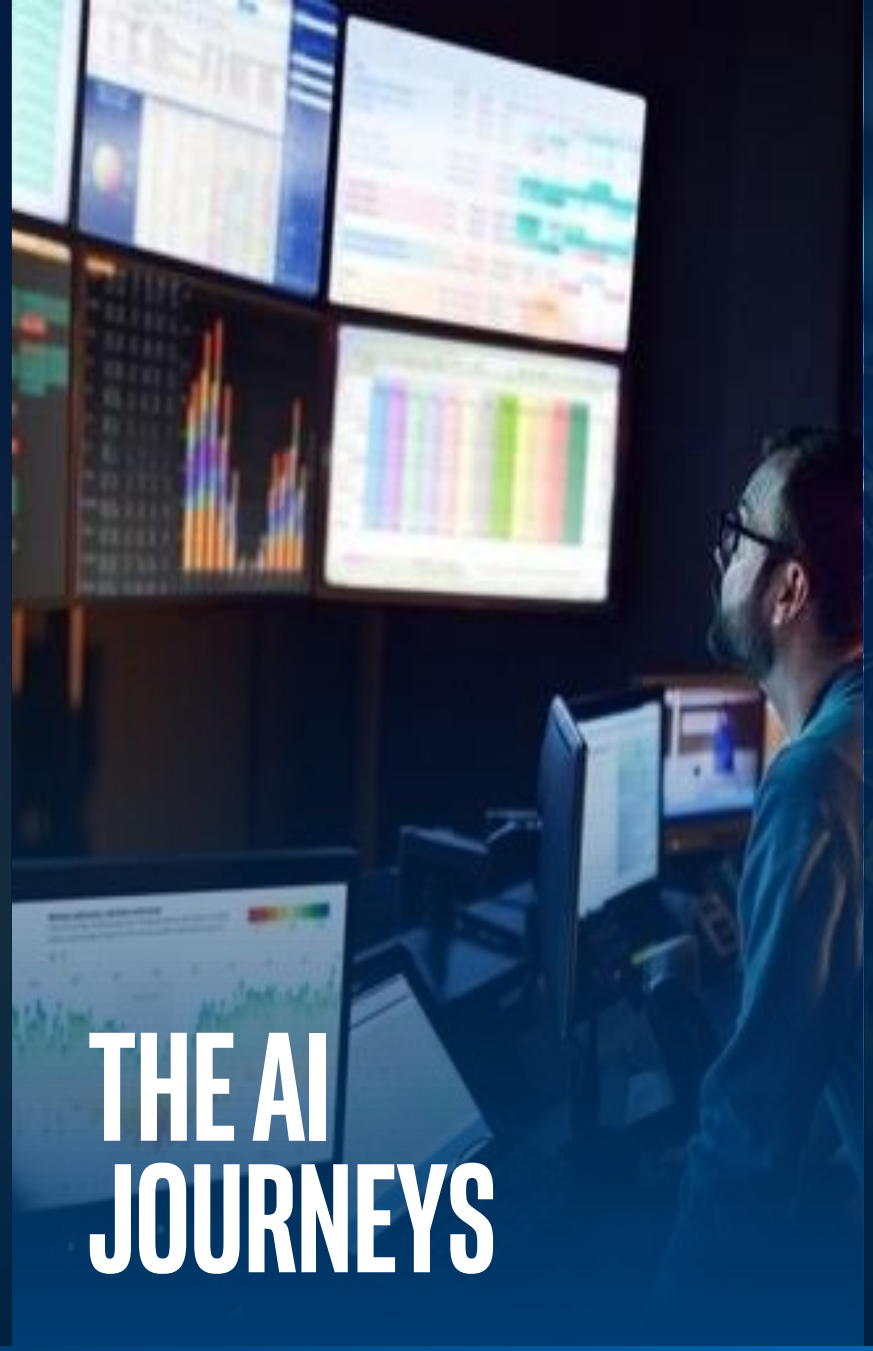
DEPLOY AI ANYWHERE
with unprecedented HW choice

SIMPLIFY AI
using community solutions

SCALE WITH CONFIDENCE
on the platform for IT & cloud



BUSINESS IMPERATIVE



THE AI JOURNEYS

INTEL® AI TOMORROW

RESONATE

50 million customer journeys,
2,000 services per day,
1300 km of track,
82 Stations

- **On time arrivals UP by 9%,**
- **Currently @ 92% of trains on time**
- **Increased automation: additional 6,000 short term schedules**
- **Improved station planning**
- **Over 50,000 schedule adjustments**
- **Time saved in all sectors: Long distance, Express airport services, Commuter, Freight**

https://www.intel.co.uk/content/www/uk/en/analytics/artificial-intelligence/resonate-success-story.html

DATA PROBLEM

INFO 1	INFO 2	INFO 3	Value to Predict
City/town	District	County	Postcode

DATA PROBLEM

INFO 1	INFO 2	INFO 3	Value to Predict
City/town	District	County	Postcode
City/town	District	Old/new	Price

DATA PROBLEM

INFO 1	INFO 2	INFO 3	Value to Predict
City/town	District	County	Postcode
City/town	District	Old/new	Price
Postcode	Property type	Old/new	Price

DATA DRIVES DECISIONS

FRAUD DETECTION

China UnionPay

RESULT

"The new sandwich-structured fraud detection model has performed up to expectations in various assessments by the National Engineering Laboratory for E-commerce and E-payment and ZhongAn Technology."

Intel® MKL-DNN
Intel® Python with DAAL

Client: China UnionPay*, which specializes in banking services and payment systems. It is the 3rd largest payment network in the world.

Challenge: No single ML or DL algorithm delivers a high-enough accuracy when trying to detect online fraudulent transactions using fake/clone cards through POS, ATM, or online payments. Real time detection is required that can service hundreds of millions of transactions per day.

Solution: Intel collaborated with UnionPay* on a GBDT→GRU→RF "sandwich" architecture.[1] The entire solution was based on the Intel® Xeon® Scalable processor, and each layer used Intel-optimized software: GBDT (Apache Spark with BigDL), GRU (TensorFlow* optimized for Intel architecture), and RF (Intel® Python with DAAL).

[1] GBDT = "Gradient Boosting Decision Tree." GRU ="Gated Recurrent Unit." RF = "Random Forest."

https://ai.intel.com/nervana/wp-content/uploads/sites/53/2018/06/Intel-White-Paper-Union-Pay_2_hir-res_Keep-the-Size-of-Figure-6.pdf

Fraud detection – Challenges

Data science challenges:

- DT (Decision Tree), RF (Random Forest), GBDT not effective enough.
- RNNs not autonomous enough

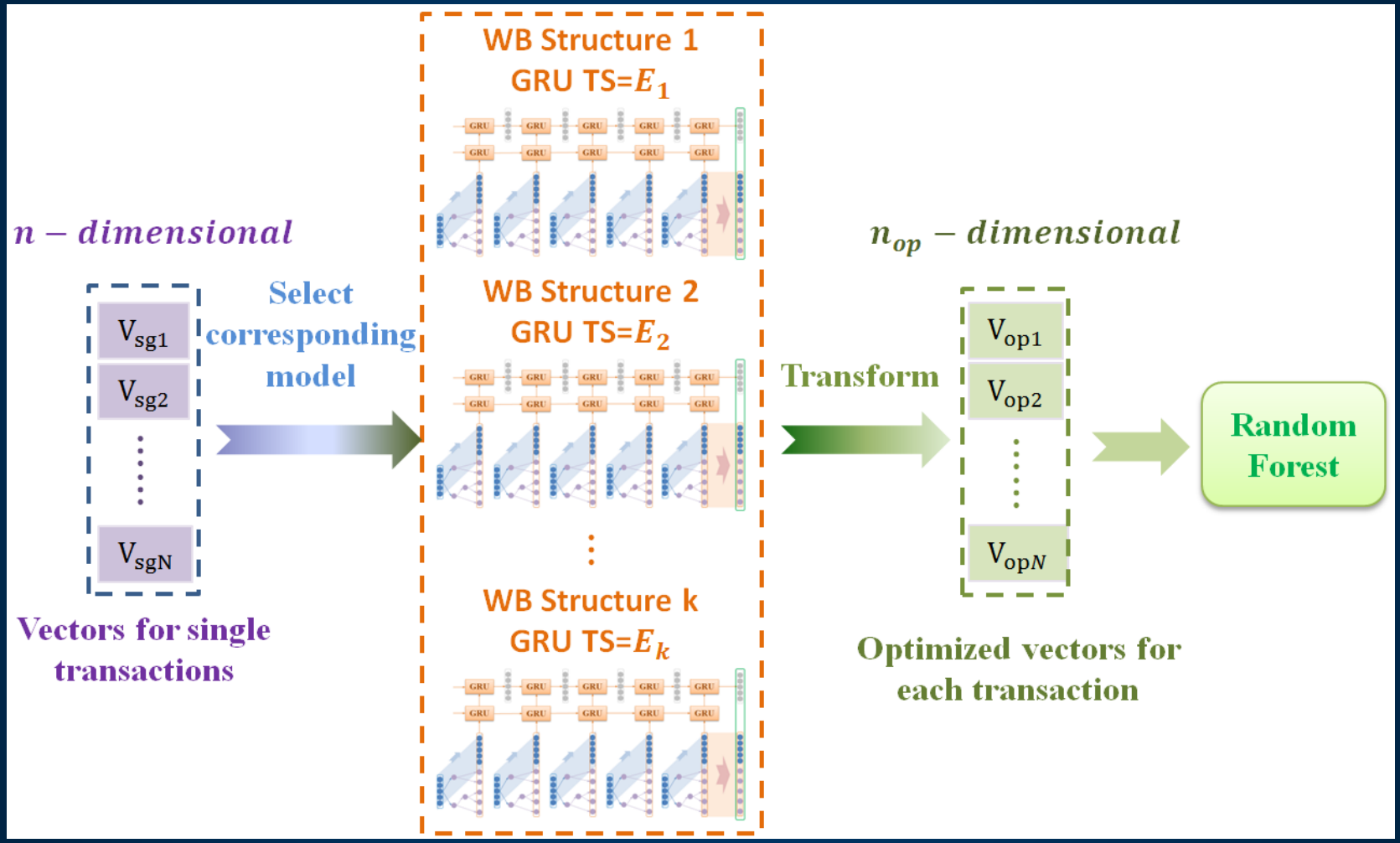
→ Combined approach: **the “Sandwich” architecture**

IT challenges:

- Leverage the existing Data infrastructure with Hadoop storage system
- Enable Tensorflow

→ **Intel BigDL**

Fraud detection



AI ON APACHE Spark™

BigDL

HIGH-PERFORMANCE DEEP LEARNING FRAMEWORK FOR APACHE SPARK

software.intel.com/bigdl

ANALYTICS ZOO

UNIFIED ANALYTICS + AI PLATFORM

DISTRIBUTED TENSORFLOW, KERAS AND BIGDL ON APACHE SPARK

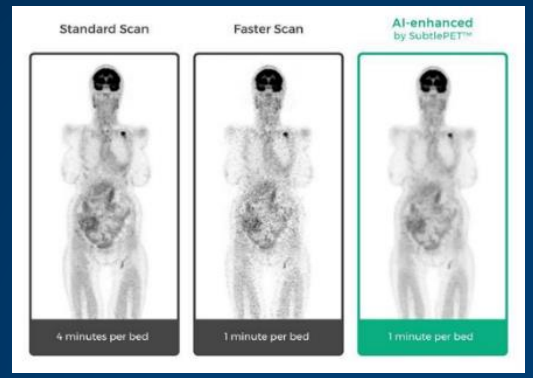
Reference Use Cases, AI Models,
High-level APIs, Feature Engineering, etc.

https://github.com/intel-analytics/analytics-zoo

UNIFYING ANALYTICS + AI ON APACHE SPARK

Case Study : **SubtlePET** – Enabling faster PET Scans

Subtle Medical



RESULTS

2.1X FASTER

Inference performance

Customer:

SubtlePET enhances the quality of the DICOM images produced by PET scanners using faster scan protocols to maintain clinically-equivalent results.

Challenge:

SubtlePET, for some customers, has to be deployed to computing devices that are already in the hospital. In most cases, this is an Intel® CPU. Improving inference speed on existing clinical infrastructure is paramount to accelerating the adoption of SubtlePET into the clinical workflow.

Solution:

SubtlePET incorporated the Intel® OpenVINO toolkit into their C++ solution which Improved image quality and enabled faster PET scans by increasing signal-to-noise ratio. The integration of Intel® distribution of OpenVINO™ toolkit increased the inference speed of their typical PET/CT whole body by 2.1x on Intel® Xeon E5 processor

Other Marketing Assets: https://builders.intel.com/ai/blog/subtle-medical-subtlepet-openvino-toolkit

*Other names and brands may be claimed as the property of others.
Configuration: AWS p3.2xlarge instance, Intel® Xeon® E5-2686 v4 @ 2.30 GHz, 61 GB RAM, Intel® OpenVINO™ Toolkit
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

Performance results are based on testing as of dates shown in configuration and may not reflect all publicly available security updates. 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

Intel® Distribution of OpenVINO™ toolkit

DEEP LEARNING

Caffe TensorFlow ONNX mxnet KALDI

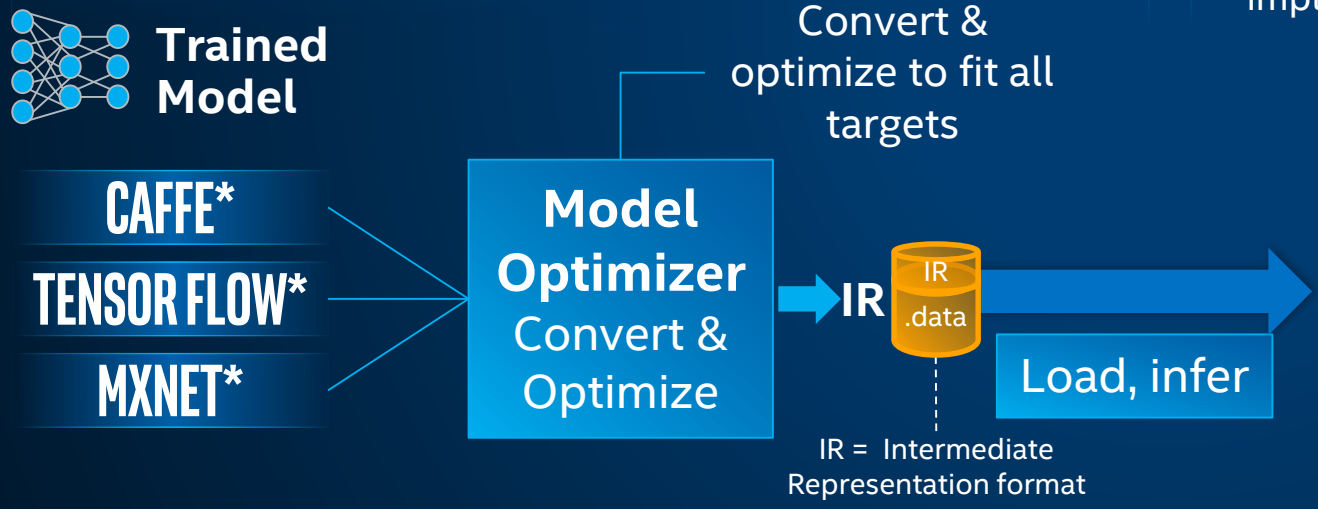
OpenCV

COMPUTER VISION

OpenVX OpenCL™

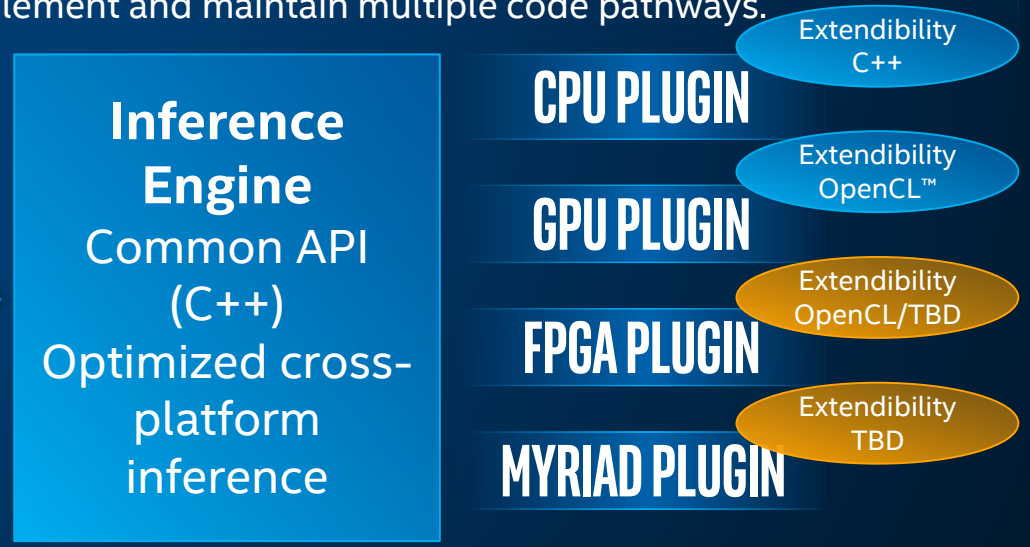
Model Optimizer

- What it is: Preparation step -> imports trained models
- Why important: Optimizes for performance/space with conservative topology transformations; biggest boost is from conversion to data types matching hardware.



Inference Engine

- What it is: High-level inference API
- Why important: Interface is implemented as dynamically loaded plugins for each hardware type. Delivers best performance for each type without requiring users to implement and maintain multiple code pathways.



GPU = Intel CPU with integrated graphics processing unit/Intel® Processor Graphics
OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos
*Other names and brands may be claimed as the property of others.

The “DeepAd” Project with Cloudian* & Dentsu*

Targeted Advertisement

RESULT

“Multi-core Intel® Xeon® Processor E5 family [...] were critical to enabling the speed at which real-time detection, control, and decision making were possible [and the] the Intel NUC kept the technology footprint small in the field...”

Client: The “DeepAd project” comprised primarily of Cloudian*, Dentsu*, and Intel testing innovative digital billboard solutions with dynamic content

Challenge: Utilizing machine learning and deep learning to increase the effectiveness of OOH (out of home) signage along busy freeways for targeted advertising, detecting and tracking different type of automobiles and displaying content based on the recognized car.

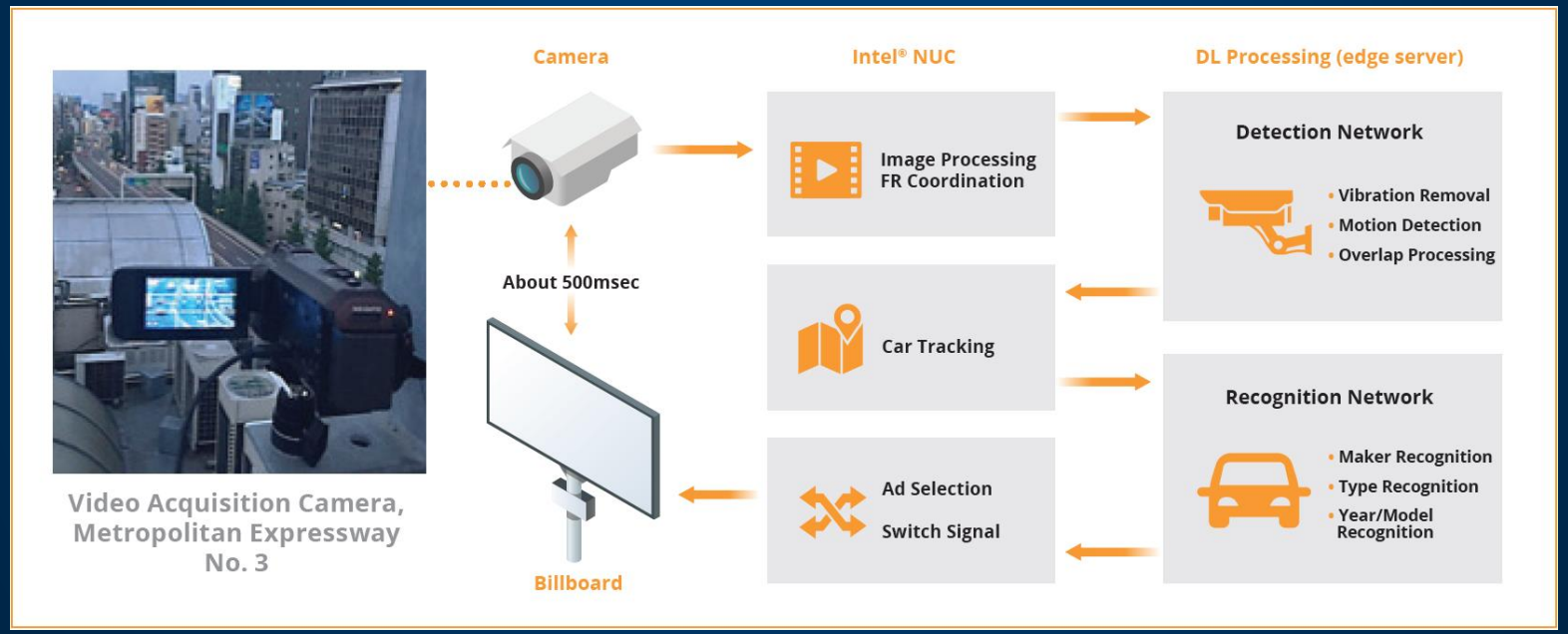
Solution: Intel® Xeon® processor E5 family servers used to train a models consisting of over 5000 vehicles, with Intel® NUCs in the field which delivered the speed needed to help accurately detect and track cars – all within less than a second!

https://builders.intel.com/docs/aibuilders/deep_learning_enables_intelligent_billboard_for_dynamic_targeted_advertising_on_tokyo_expressway-ai.pdf

DEEP-AD - Challenges

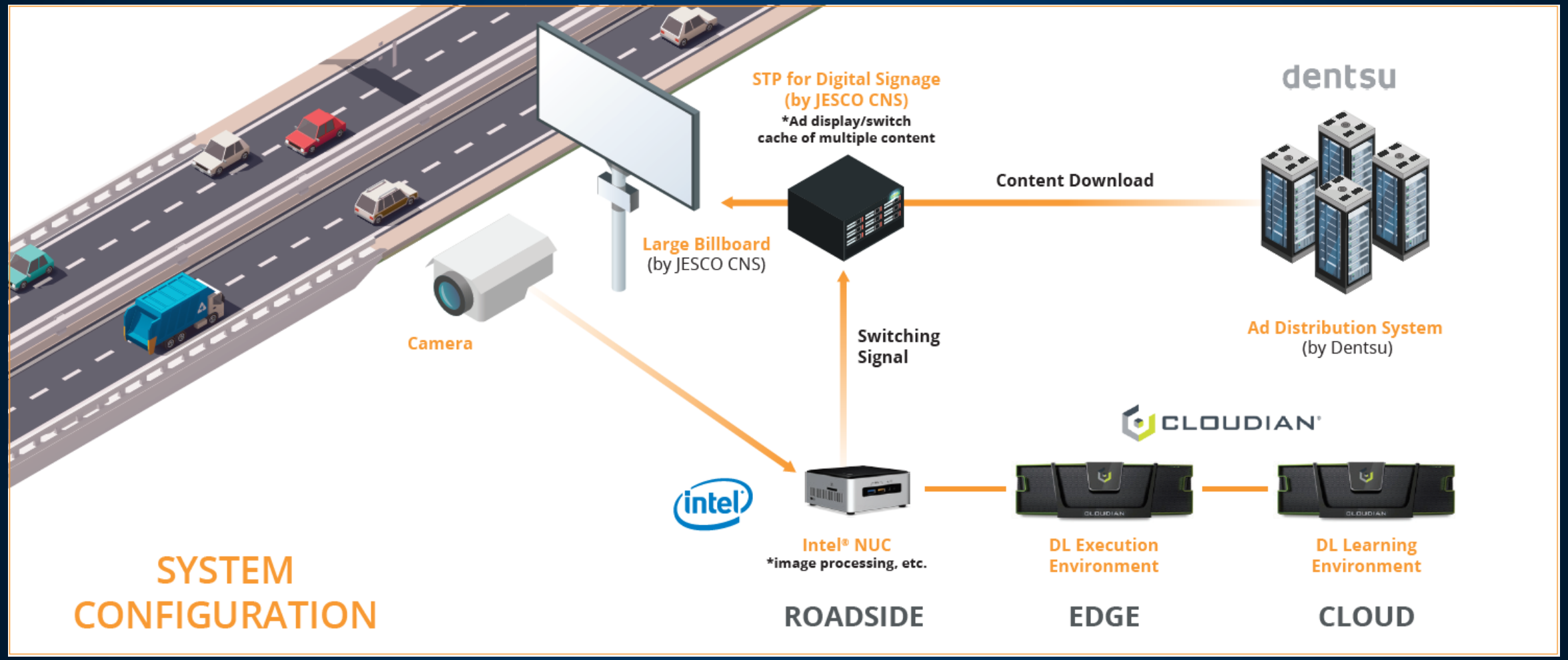
IT challenges:

- Build an end-to-end solution Datacentre - Edge
- Fast decision making happening at the Edge



→ Intel portfolio of AI solutions (Widest!)

DEEP-AD - Challenges

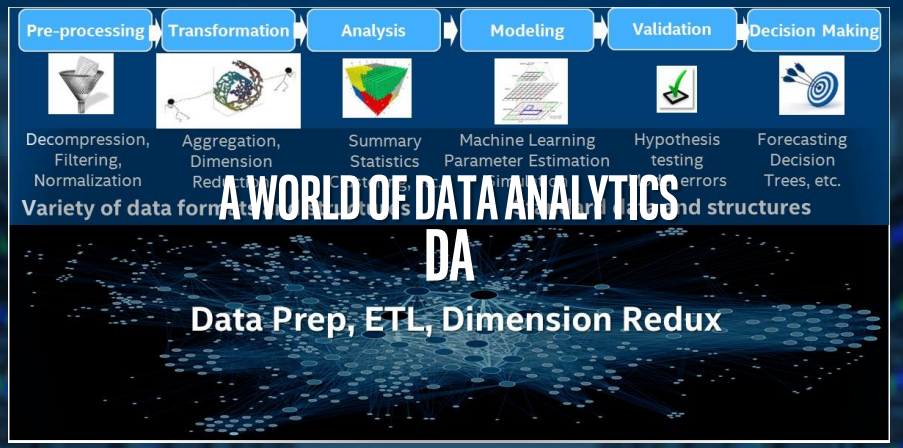


Convergence

HPC ↔ DA ↔ AI

Better Theory Guided Data Science via consistency

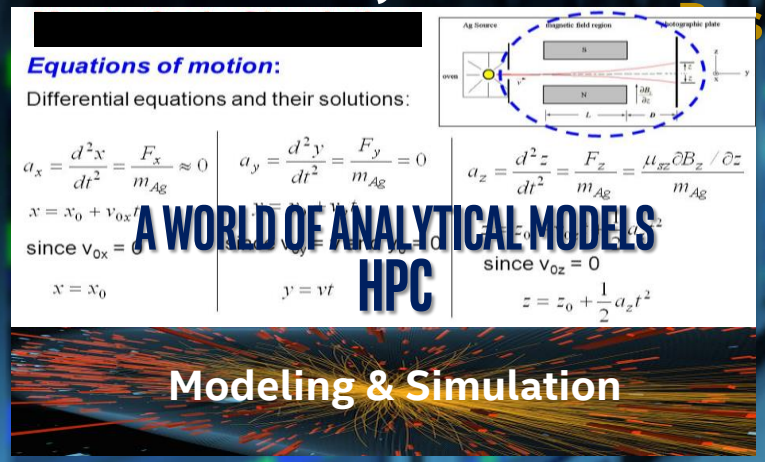
Steering in High Dimensionality space, In-situ Processing



Feature Vectors for Training

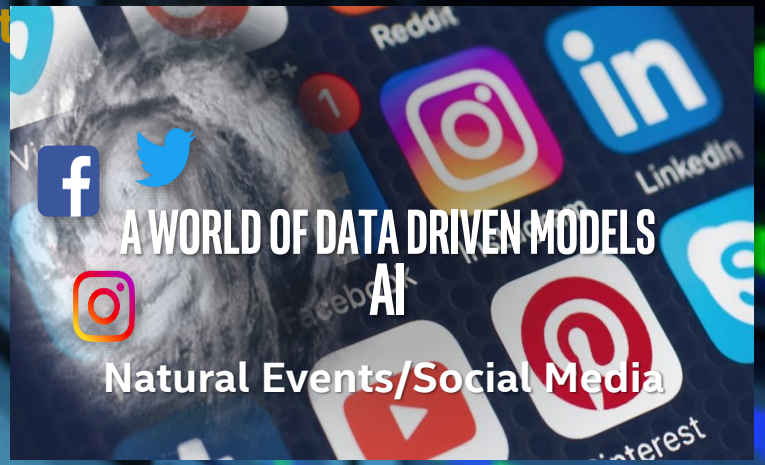
Data Compression via Learning

Convergence Today on Intel® Xeon® Processor Based Supercomputing Infrastructure



Training Data augmenting Real World Data

Displacement of Analytical Models



https://www.google.com/search?rlz=1C1GGRV_enUS751US751&biw=1506&bih=871&tbm=isch&sa=1&ei=91fDW8ClD6jj0gLuio_ICA&q=Monte+Carlo+Simulation+equations&oq=Monte+Carlo+Simulation+equations&gs_l=img.3.0.35i39l2j0i24.130457.130696..132104...0.0..0.80.149.2.....1....1..gws-wiz-img.fEfNvMCo2_c#imgrc=mlfR_gpr22qbqM:

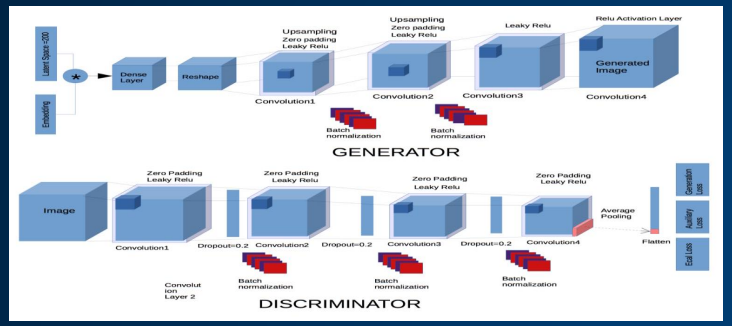
https://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/bdec_pathways.pdf

27

HPC POC: High Energy Physics Simulation

CERN

Joint collaboration with Intel and SURFsara and TACC



RESULT

94% scaling efficiency up to 128 nodes, with a significant reduction in training time per epoch for 3D GANs

Customer: CERN, the European Organization for Nuclear Research, which operates the Large Hadron Collider (LHC), the world's largest and most powerful particle accelerator

Challenge: CERN currently uses Monte Carlo simulations for complex physics and geometry modeling, which is a heavy computational load that consumes up to >50% of the Worldwide LHC (Large Hadron Collider) Computing Grid (WLCG) power for electron shower simulations.

Solution: Distributed training using 128 nodes of the TACC Stampede 2 cluster (Intel® Xeon® Platinum 8160 processor, Intel® OPA) and a 3D Generative Adversarial Network (3D GAN). Performance was first optimized on a single node using TensorFlow* optimized with Intel® MKL-DNN, using 4 workers/node and an optimized number of convolutional filters.

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

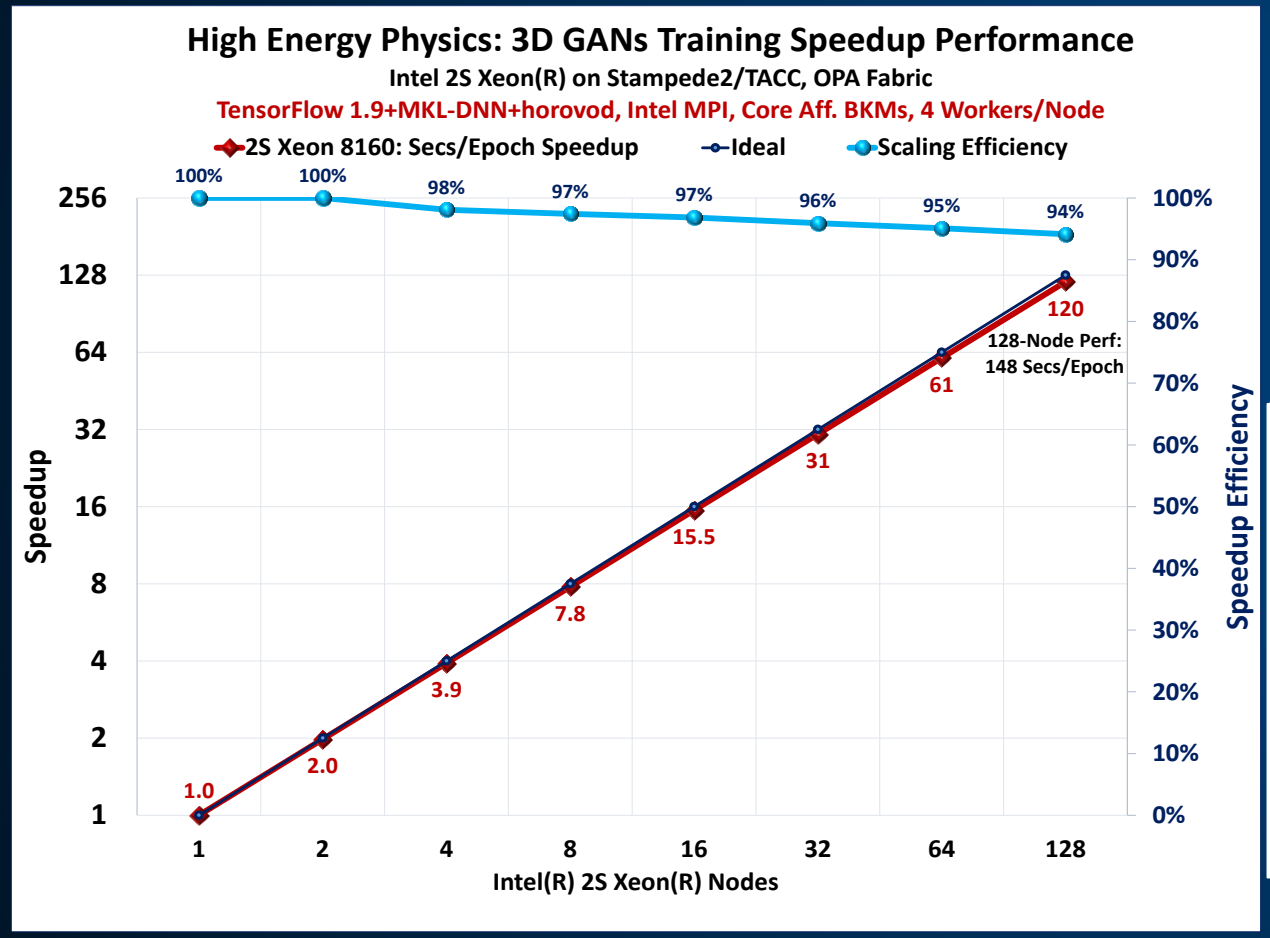
[1] *Stampede2/TACC: https://portal.tacc.utexas.edu/user-guides/stampede2. Compute nodes: 2 sockets Intel® Xeon® Platinum 8160 processor with 24 cores each @ 2.10GHz for a total of 48 cores per node, 2 Threads per core, L1d 32K; L1i cache 32K; L2 cache 1024K; L3 cache 33792K, 96 GB of DDR4, Intel® OmniPath Host Fabric Interface, dual-rail. Software: Intel® MPI Library 2017 Update 4Intel® MPI Library 2019 Technical Preview OFI 1.5.0PSM2 w/ Multi-EP, 10 Gbit Ethernet, 200 GB local SSD, Red Hat* Enterprise Linux 6.7. TensorFlow* 1.6: Built & Installed from source: https://www.tensorflow.org/install/install_sources Model: CERN* 3D GANS from https://github.com/sara-nl/3Dgan/tree/tf Dataset: CERN* 3D GANS from https://github.com/sara-nl/3Dgan/tree/tf Performance measured on 256 NodesPerformance measured on 256 Nodes with: OMP_NUM_THREADS=24 HOROVOD_FUSION_THRESHOLD=134217728 export I_MPI_FABRICS=tmi, export I_MPI_TMI_PROVIDER=psm2 \ mpirun -np 512 -ppn 2 python resnet_main.py --train_batch_size 8 \ --num_intra_threads 24 --num_inter_threads 2 --mkl=True \ --data_dir=/path/to/gans_script.py --kmp_blocktime 1.

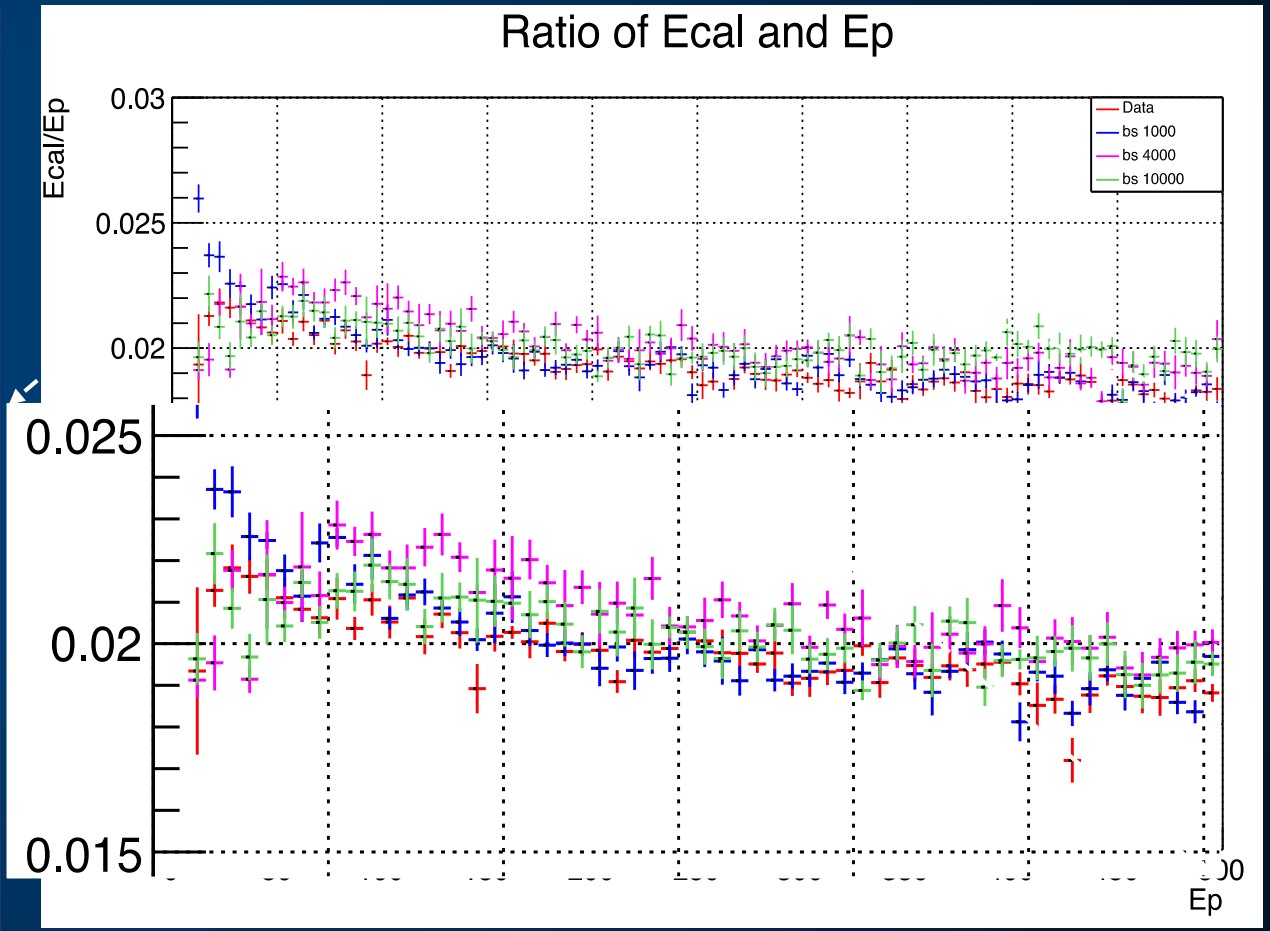
Performance results are based on testing as of (05/17/2018) and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

Multi-Node Training Performance & Accuracy

Distributed training using data parallelism

94% Scaling efficiency up to 128 nodes





High Energy Physics Simulation - CERN

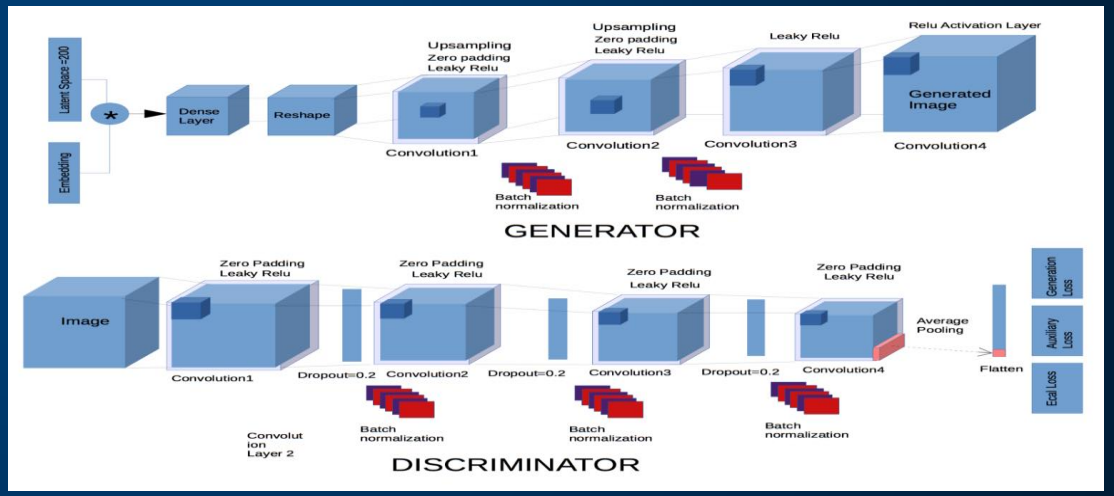
Paper: https://doi.org/10.1007/978-3-030-02465-9_35

Code: https://github.com/svalleco/3Dgan

Distributed Model: https://github.com/svalleco/3Dgan/blob/svalleco/sc18/keras/

The training script is EcalEnergyTrain_hvd.py

Architecture details: EcalEnergyGan.py



Training dataset: https://cernbox.cern.ch/index.php/s/CczhzHSwLrVWD4p

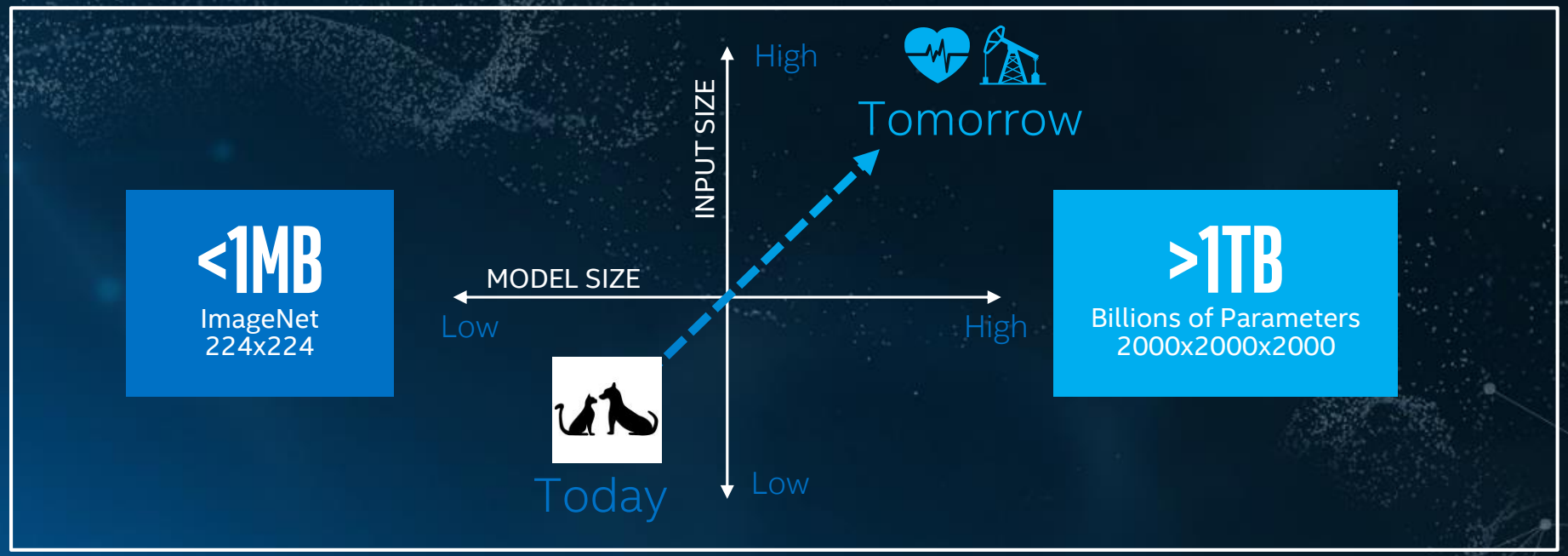
KEY MESSAGEs and TAKEAWAY

1. **Know your goal:** Data drives decision

2. **Max ROI:** turn your HPC/BigData facility into an AI capable infrastructure

3. **Intel AI Portfolio:** 1 solution does not fit all needs. Find your best!

WHAT ABOUT TOMORROW?



BIGGER DATA + BIGGER MODELS = BIGGER COMPUTE

Source: Siemens Healtheneers; https://www.youtube.com/watch?v=o8NIN8dJqdc
Source: https://www.intel.com/content/www/us/en/high-performance-computing/ai-and-hpc-workload-convergence-video.html
Source: https://idealmagazine.co.uk/the-ideal-guide-how-to-order-a-beerpizzataxicoffee-in-15-languages/

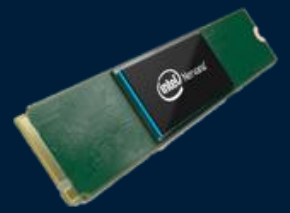
Source: https://openai.com/blog/ai-and-compute/

INTEL® NERVANA™ NEURAL NETWORK PROCESSORS (NNP)¥

NNP-T

DEDICATED
DL TRAINING

Fastest time-to-train with high bandwidth AI server connections for the most persistent, intense usage



NNP-I

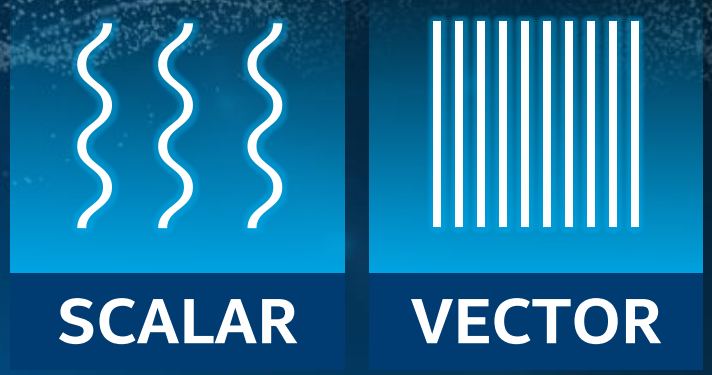
DEDICATED
DL INFERENCE

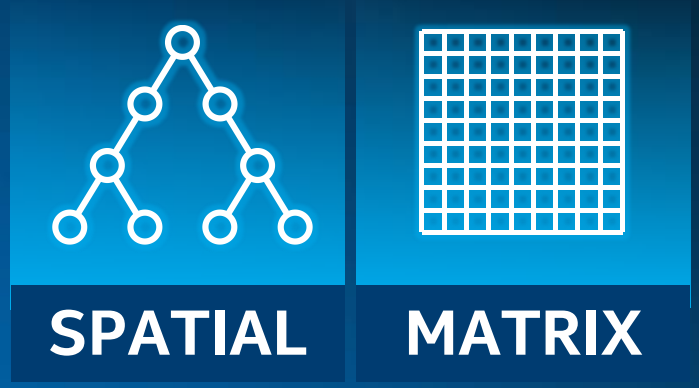
Highly-efficient multi-model inferencing for cloud, data center and intense appliances

https://www.intel.ai/accelerating-for-ai/#gs.8h2ig5

¥ The Intel® Nervana™ Neural Network Processor is a future product that is not broadly available today
All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

RETHINKING ARCHITECTURE TO MEET DEMAND





To quickly process vast, sparse, or complex data for large models within a power budget, AI hardware must deliver a critical balance of:

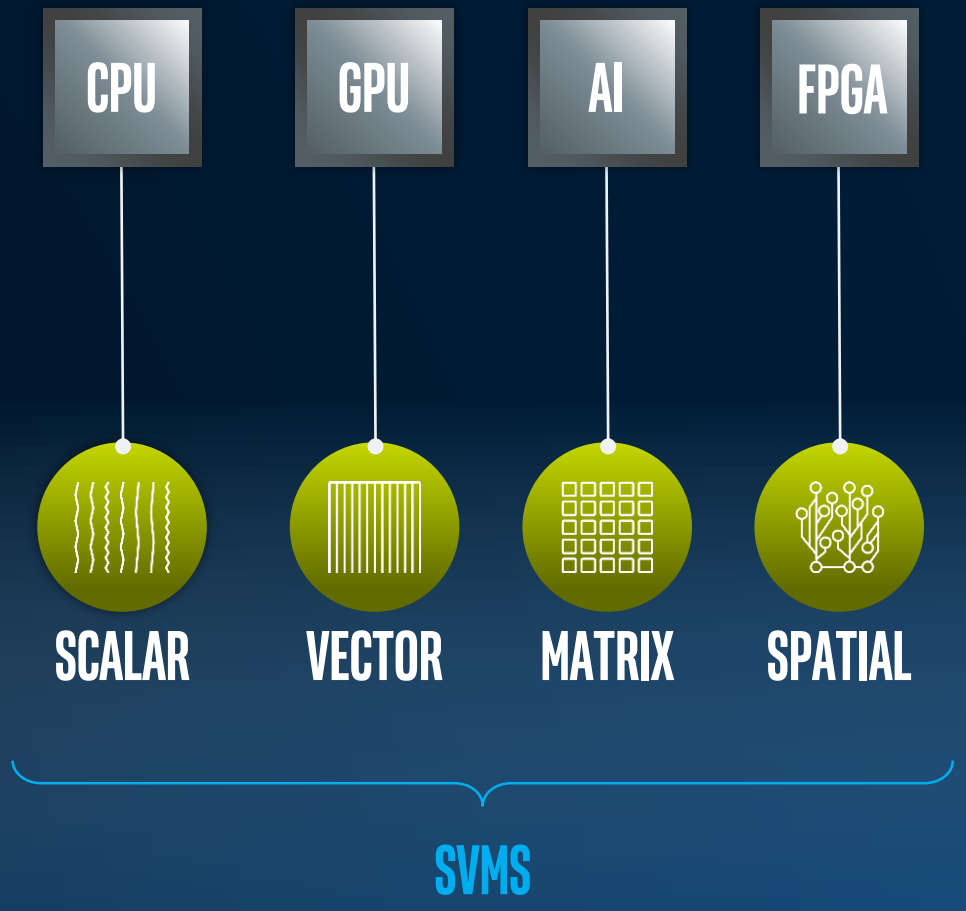
COMPUTE

COMMUNICATION

MEMORY

DIVERSE WORKLOADS REQUIRE DIVERSE ARCHITECTURES

The future is a **diverse** mix of scalar, vector, matrix, and spatial **architectures** deployed in CPU, GPU, AI, FPGA and other accelerators



PROGRAMMING CHALLENGE

Diverse set of data-centric hardware

No common **programming languages or API**

Inconsistent tool support across platform

Each platform require software investment

INTEL'S ONE API CORE CONCEPT

One API is a project to deliver a unified programming model to simplify development across diverse architectures

Common developer experience across Scalar, Vector, Matrix and Spatial (SVMS) architecture

Unified and simplified language and libraries for expressing parallelism

Uncompromised native high-level language performance

Support for CPU, GPU, AI and FPGA

Based on industry standards and open specifications