THE AI ERA INFRASTRUCTURE

INDUSTRY TRANSFORMATION & INNOVATION OPPORTUNITIES

Meena Arunachalam

Principal Engineer, Intel Architecture and Graphics Software

DISCLAIMERS:

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

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 computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

No computer system can be absolutely secure.

Intel, the Intel logo, Xeon, Mobileye, DL Boost,Nervana, Atom, Arria, Stratix and Movidius are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

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

INDUSTRIAL REVOLUTION -- FROM PHYSICAL TO DIGITAL

Fourth Industrial Revolution represents entirely new ways in which technology becomes embedded within industries, societies and even our human bodies

4th

1st

Mechanization
Steam power
Water Power

18th Century

2nd

Mass production
Assembly line
Electricity

19th Century

3rd

Electronics
Internet & IT
Automation

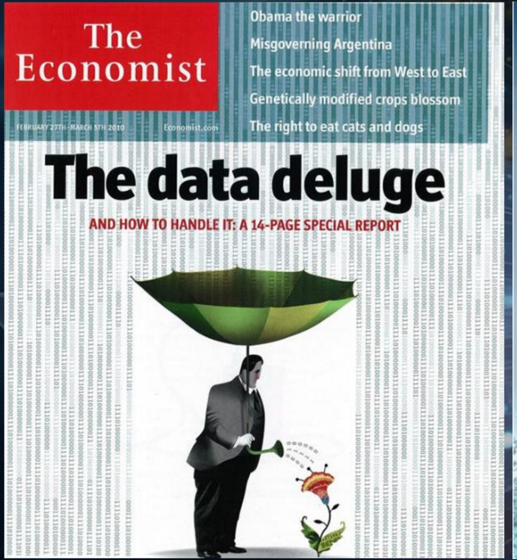
20th Century

Cyber Physical
Systems

21st Century

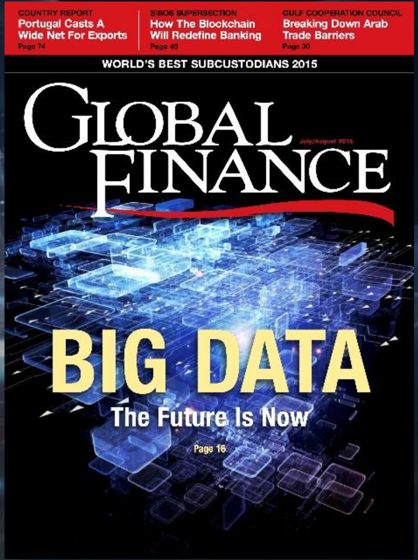
WE ARE ~~WITNESSING~~ CREATING THIS DIGITAL TRANSFORMATION

DATA DEFINES THE FUTURE


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
FEBRUARY 27TH–MARCH 5TH 2010
Economist.com
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

The Economist
Crunch time in France
Ten years on: banking after the crisis
South Korea's unfinished revolution
Biology, but without the cells
MAY 6TH–12TH 2017
The world's most valuable resource
amazon
UBER
Microsoft
Google
TESLA
Data and the new rules of competition


COUNTRY REPORT
Portugal Casts A Wide Net For Exports
Page 74
SIBOS SUPERSECTION
How The Blockchain Will Redefine Banking
Page 40
GULF COOPERATION COUNCIL
Breaking Down Arab Trade Barriers
Page 30
WORLD'S BEST SUBCUSTODIANS 2015
GLOBAL FINANCE
July/August 2015
BIG DATA
The Future Is Now
Page 16

COMPETITIVENESS AND BUSINESS GROWTH ARE INCREASINGLY DETERMINED BY THE POWER OF DATA.

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

INNOVATION ACROSS ALL INDUSTRIES

AI / ML

Processing Power

"DIGITAL FUSION"

Blending of Traditional & Digital Business Models

Smart Cities & Surveillance

Retail: Real-Time Pricing & Inventory

Enterprise/Consumer Analytics

Precision Medicine & Genomic Analytics

Autonomous Cars & 5G Connectivity

Virtual/Augmented Reality

Data

AI IS DRIVING ADVANCED ANALYTICS

DATA DELUGE (2019)

25 GB[1] per month
INTERNET USER

50 GB[2] per day
SMART CAR

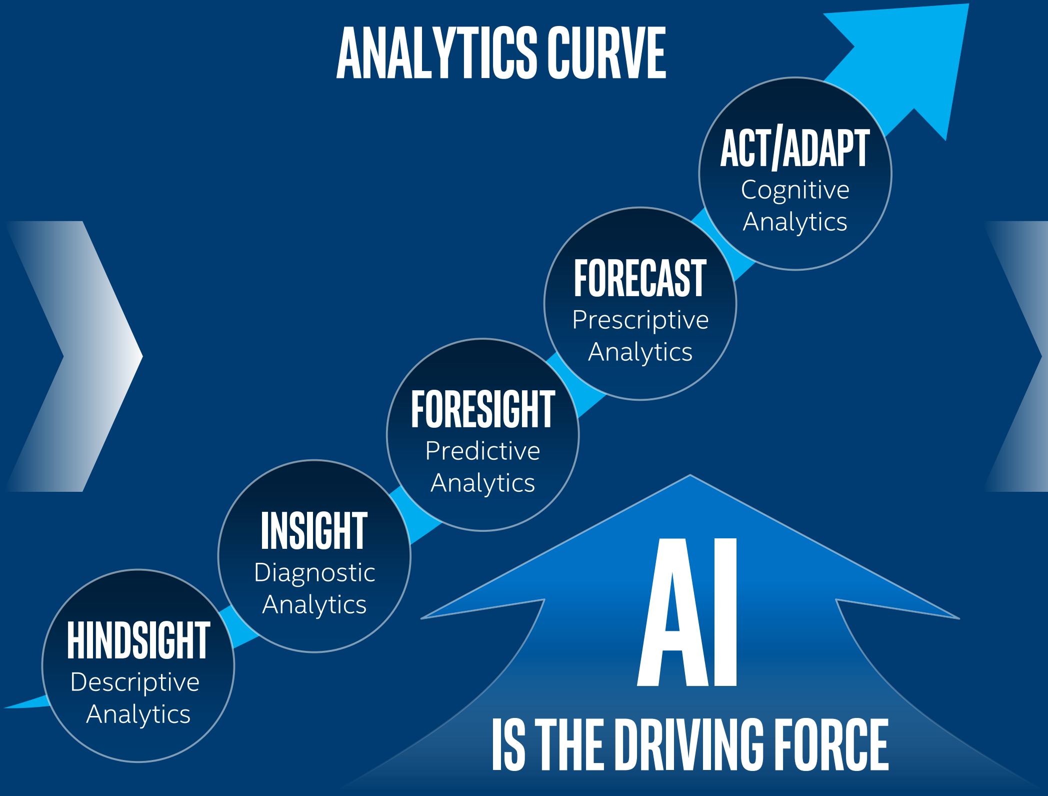
3 TB[2] per day
SMART HOSPITAL

40 TB[2] per day
AIRPLANE DATA

1 PB[2] per day
SMART FACTORY

50 PB[2] per day
CITY SAFETY

ANALYTICS CURVE



HINDSIGHT
Descriptive Analytics

INSIGHT
Diagnostic Analytics

FORESIGHT
Predictive Analytics

FORECAST
Prescriptive Analytics

ACT/ADAPT
Cognitive Analytics

AI
IS THE DRIVING FORCE

INSIGHTS

BUSINESS

OPERATIONAL

SECURITY

AI INSIDE INTEL

REGULATORY

Audit & compliance automation

HR

Diversity, recruiting & retention

IT

Digital transformation with AI

LOGISTICS

Supply chain optimization

SALES

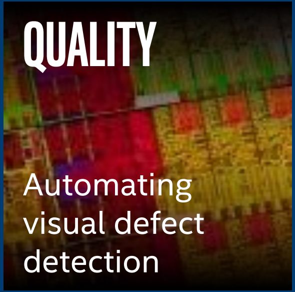
Info processing to improve efficiency

HEALTH

Pharmaceutical analytics platform

PRODUCTION

Factory process automation

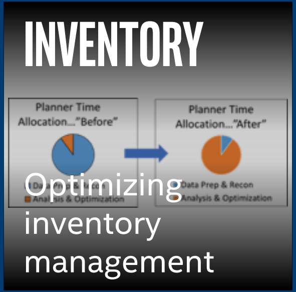


QUALITY

Automating visual defect detection

RELIABILITY

Accelerating product validation



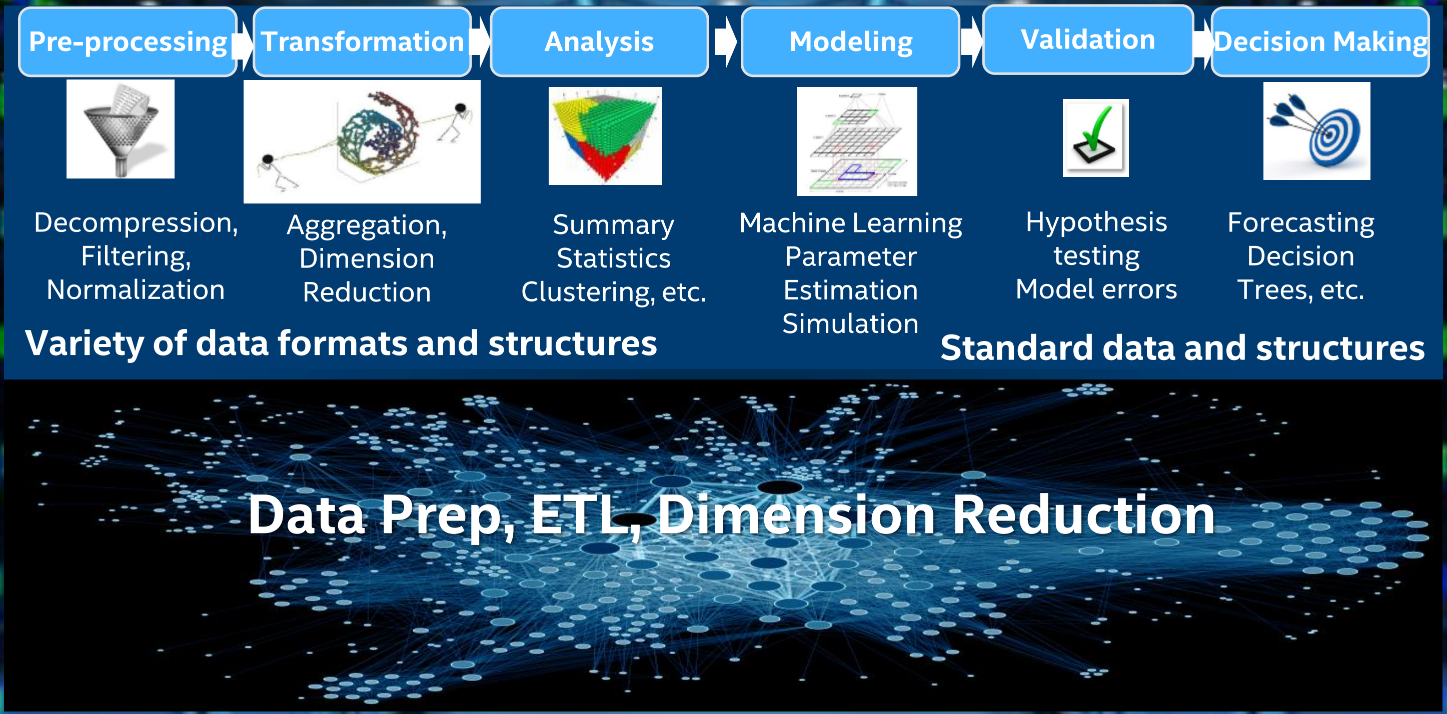
INVENTORY

Optimizing inventory management

AND MORE...

INTEL® IS INFUSING AI INTO EVERYTHING WE DO

End to End Data Analytics Flow



Pre-processing → Transformation → Analysis → Modeling → Validation → Decision Making

Decompression, Filtering, Normalization

Aggregation, Dimension Reduction

Summary Statistics Clustering, etc.

Machine Learning Parameter Estimation Simulation

Hypothesis testing Model errors

Forecasting Decision Trees, etc.

Variety of data formats and structures

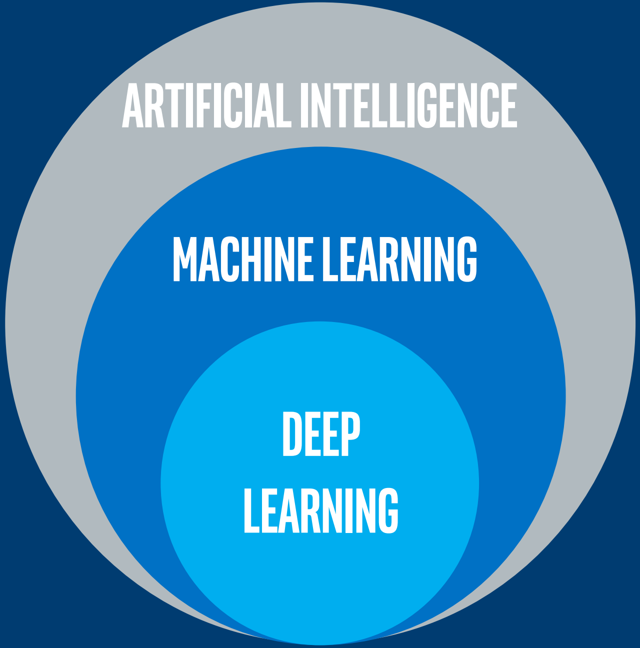
Standard data and structures

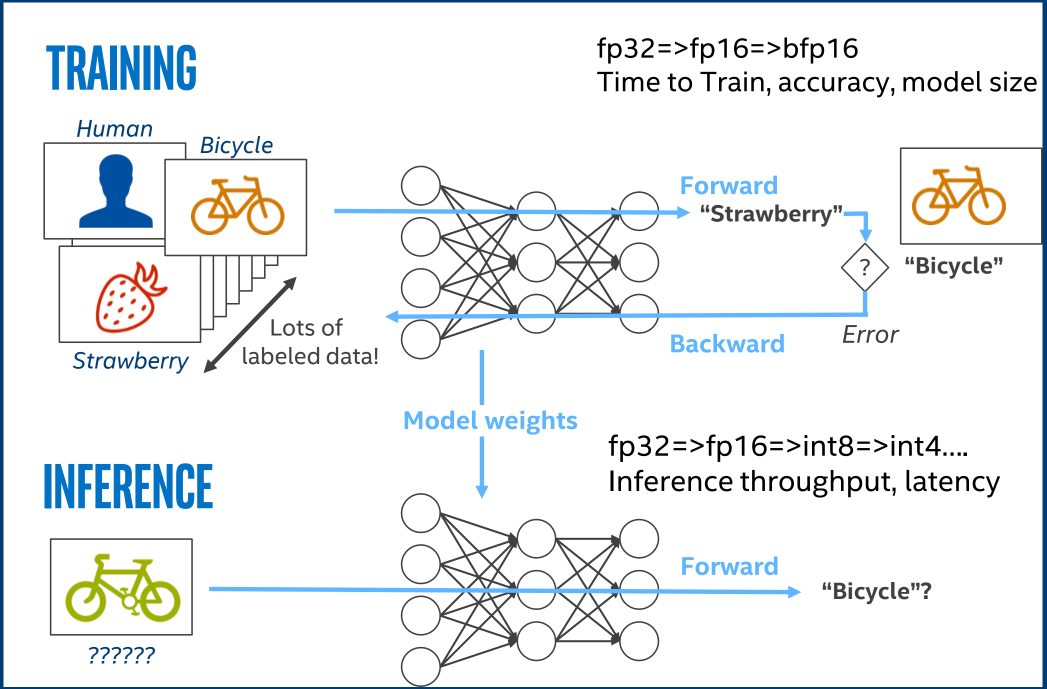
Data Prep, ETL, Dimension Reduction

Start with Data

ETL → Feature Engr, Classical ML → Data Prep

DEEP LEARNING FLOW





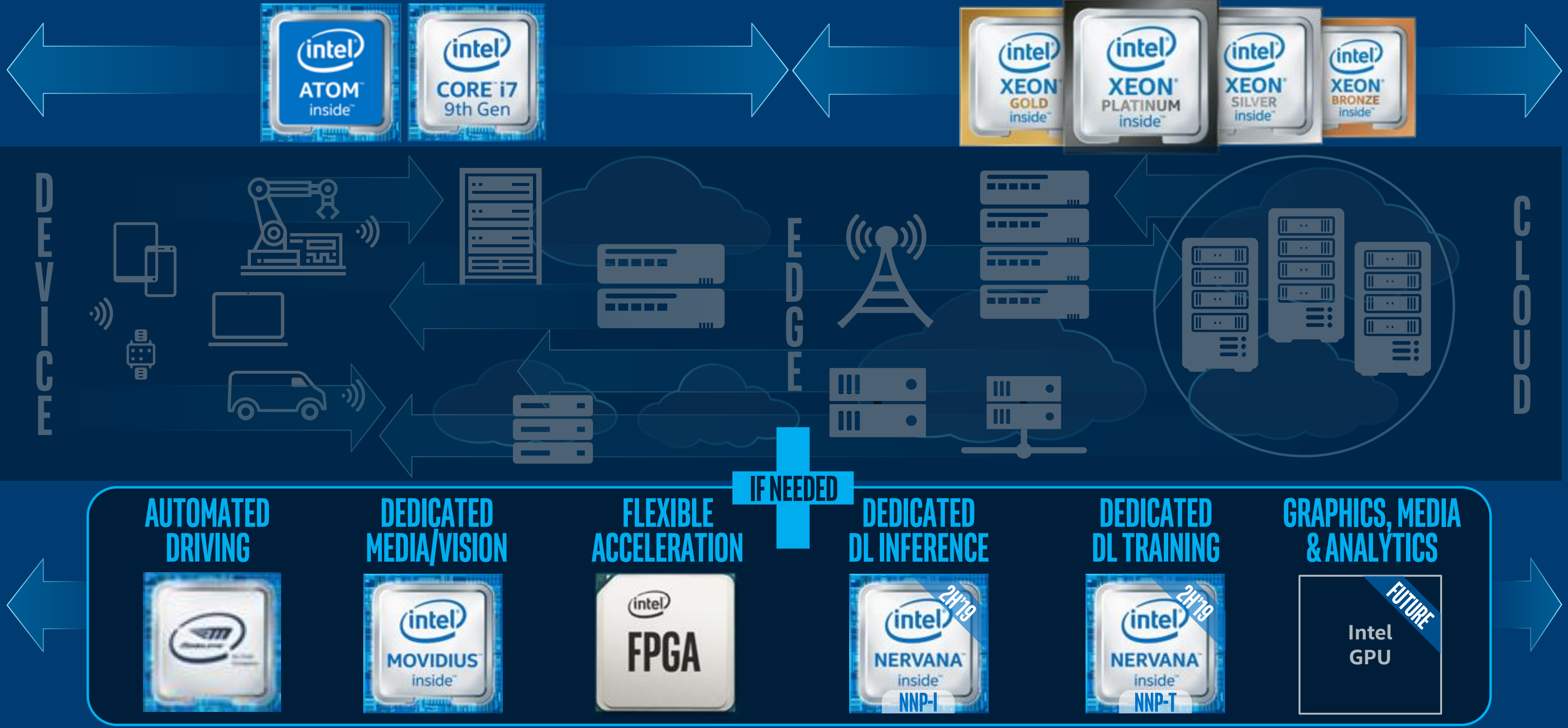
fp32=>fp16=>bfp16
Time to Train, accuracy, model size

fp32=>fp16=>int8=>int4….
Inference throughput, latency

Visit: www.intel.ai/technology

AI IS EXPANDING

Deploy AI anywhere
with unprecedented hardware choice



AUTOMATED DRIVING

DEDICATED MEDIA/VISION

FLEXIBLE ACCELERATION

DEDICATED DL INFERENCE

DEDICATED DL TRAINING

GRAPHICS, MEDIA & ANALYTICS

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

CPU FOUNDATION FOR ARTIFICIAL INTELLIGENCE

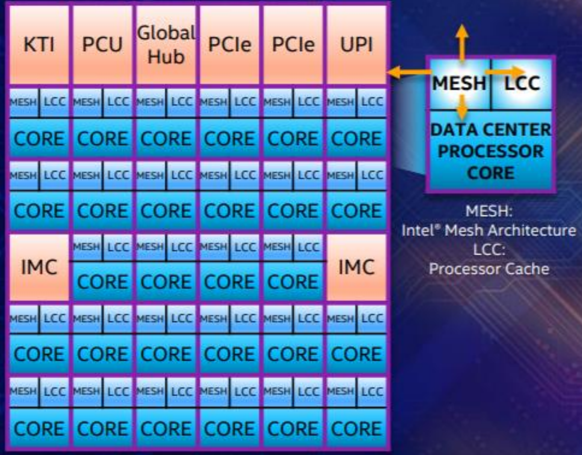
MATRIX OPERATIONS
Intel® Advanced Vector Extensions

LOWER & MIXED PRECISION
Intel® Deep Learning Boost

LARGER CACHES, MEMORY LATENCY & BANDWIDTH

OPTIMAL DATA MOVEMENT & TRANSFORMATIONS

OPTIMIZED LIBRARIES AND FRAMEWORKS



KTI	PCU	Global Hub	PCIe	PCIe	UPI
MESH LCC	MESH LCC	MESH LCC	MESH LCC	MESH LCC	MESH LCC
CORE	CORE	CORE	CORE	CORE	CORE
MESH LCC	MESH LCC	MESH LCC	MESH LCC	MESH LCC	MESH LCC
CORE	CORE	CORE	CORE	CORE	CORE
IMC	MESH LCC	MESH LCC	MESH LCC	MESH LCC	IMC
	CORE	CORE	CORE	CORE	
MESH LCC	MESH LCC	MESH LCC	MESH LCC	MESH LCC	MESH LCC
CORE	CORE	CORE	CORE	CORE	CORE
MESH LCC	MESH LCC	MESH LCC	MESH LCC	MESH LCC	MESH LCC
CORE	CORE	CORE	CORE	CORE	CORE

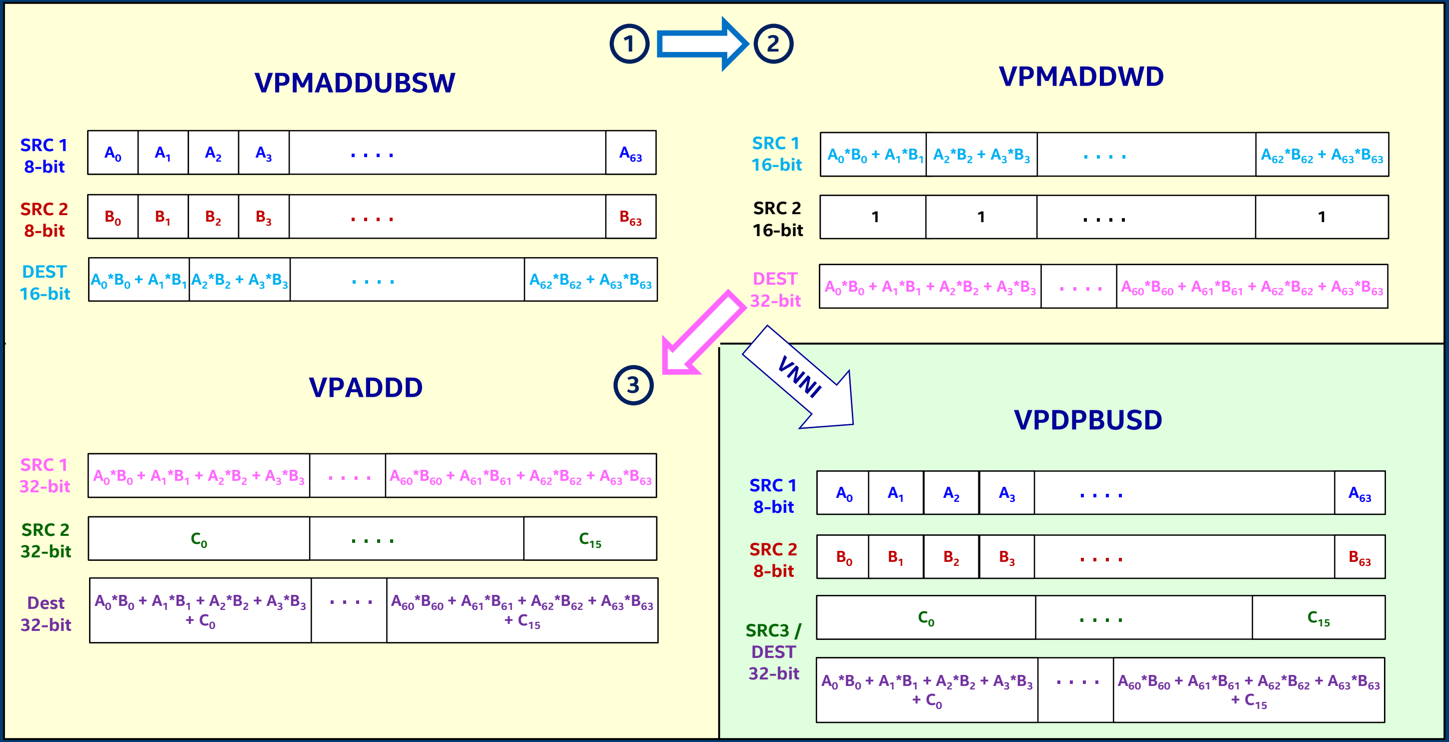
MESH:
Intel® Mesh Architecture
LCC:
Processor Cache

Illustration: Intel® Xeon® Scalable Processor

INTEL® XEON® SCALABLE PROCESSOR: ENABLES INFRASTRUCTURE-WIDE AI READINESS

INTEL DLBOOST – VNNI EXAMPLE

8-bit Convolution Inner loop



VPMADDUBSW

SRC 1 8-bit	A_0	A_1	A_2	A_3	. . . .	A_{63}
SRC 2 8-bit	B_0	B_1	B_2	B_3	. . . .	B_{63}
DEST 16-bit	$A_0*B_0 + A_1*B_1$		$A_2*B_2 + A_3*B_3$		. . . .	$A_{62}*B_{62} + A_{63}*B_{63}$

VPMADDWD

SRC 1 16-bit	$A_0*B_0 + A_1*B_1$	$A_2*B_2 + A_3*B_3$	. . . .	$A_{62}*B_{62} + A_{63}*B_{63}$
SRC 2 16-bit	1	1	. . . .	1
DEST 32-bit	$A_0*B_0 + A_1*B_1 + A_2*B_2 + A_3*B_3$		. . . .	$A_{60}*B_{60} + A_{61}*B_{61} + A_{62}*B_{62} + A_{63}*B_{63}$

VPADDD

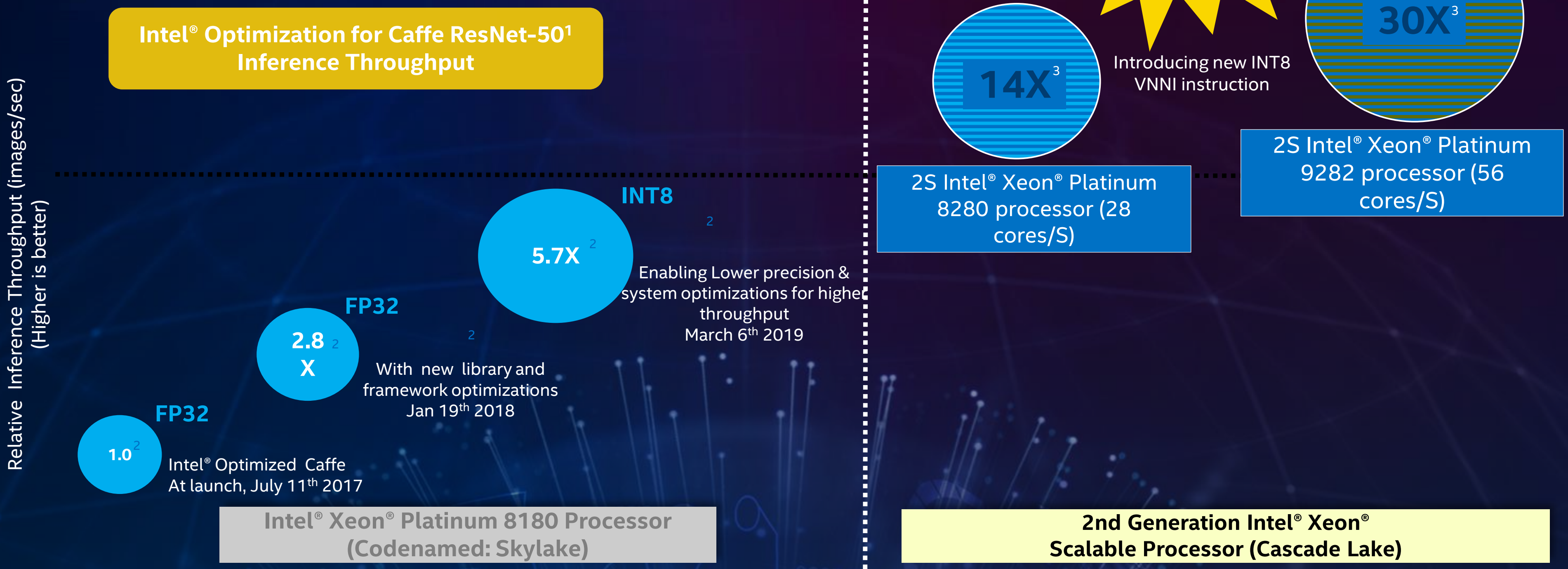
SRC 1 32-bit	$A_0*B_0 + A_1*B_1 + A_2*B_2 + A_3*B_3$	. . . .	$A_{60}*B_{60} + A_{61}*B_{61} + A_{62}*B_{62} + A_{63}*B_{63}$
SRC 2 32-bit	C_0	. . . .	C_{15}
Dest 32-bit	$A_0*B_0 + A_1*B_1 + A_2*B_2 + A_3*B_3 + C_0$	. . . .	$A_{60}*B_{60} + A_{61}*B_{61} + A_{62}*B_{62} + A_{63}*B_{63} + C_{15}$

VPDPBUSD

SRC 1 8-bit	A_0	A_1	A_2	A_3	. . . .	A_{63}
SRC 2 8-bit	B_0	B_1	B_2	B_3	. . . .	B_{63}
SRC3 / DEST 32-bit	C_0				. . . .	C_{15}
	$A_0*B_0 + A_1*B_1 + A_2*B_2 + A_3*B_3 + C_0$				. . . .	$A_{60}*B_{60} + A_{61}*B_{61} + A_{62}*B_{62} + A_{63}*B_{63} + C_{15}$

3X peak operations providing significant improvement in inferencing performance

REINVENTING XEON FOR AI



Intel® Optimization for Caffe ResNet-50[1] Inference Throughput

Intel® Optimized Caffe
At launch, July 11th 2017

With new library and framework optimizations
Jan 19th 2018

Enabling Lower precision & system optimizations for higher throughput
March 6th 2019

Intel® Deep Learning Boost

Introducing new INT8 VNNI instruction

2S Intel® Xeon® Platinum 8280 processor (28 cores/S)

2S Intel® Xeon® Platinum 9282 processor (56 cores/S)

Intel® Xeon® Platinum 8180 Processor (Codenamed: Skylake)

2nd Generation Intel® Xeon® Scalable Processor (Cascade Lake)

[1] Intel® Optimization for Caffe Resnet-50 performance does not necessarily represent other Framework performance.

[2] Based on Intel internal testing: 1X,2.8x,5.7x,14x and 30x performance improvement based on Intel® Optimization for Café ResNet-50 inference throughput performance on Intel® Xeon® Scalable Processor. See Configuration Details 3
Performance results are based on testing as of 7/11/2017(1x) ,11/8/2018 (5.7x), 2/20/2019 (14x) and 2/26/2019 (30x) and may not reflect all publically 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

SOFTWARE IS ESSENTIAL

TOOLKITS
Application Developers

OPENVINO™ TOOLKIT

INTEL® MOVIDIUS™ SDK

LIBRARIES
Data Scientists

MACHINE LEARNING LIBRARIES

Scikit-Learn NumPy MLlib

DEEP LEARNING FRAMEWORKS

TensorFlow mxnet Caffe BigDL FOR Spark

Caffe2 PYTORCH Microsoft CNTK PaddlePaddle

FOUNDATION
Library Developers

ANALYTICS, MACHINE & DEEP LEARNING PRIMITIVES

MKL-DNN clDNN MLSL Python DAAL

DEEP LEARNING GRAPH COMPILER

Intel® nGraph™ Compiler

intel ATOM x inside · intel Iris Graphics · intel CORE i7 8th Gen · intel XEON inside · intel NERVANA inside · intel STRATIX 10 inside · intel ARRIA 10 inside · intel MOVIDIUS inside · INTEL GNA (IP) · MOBILEYE An Intel Company

HARDWARE

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

intel AI

INTEL® DISTRIBUTION FOR PYTHON*

python™ software.intel.com/intel-distribution-for-python

FOR DEVELOPERS USING THE MOST POPULAR AND FASTEST GROWING PROGRAMMING LANGUAGE FOR AI

EASY, OUT-OF-THE-BOX ACCESS TO HIGH PERFORMANCE PYTHON

- Prebuilt, optimized for numerical computing, data analytics, HPC
- Drop in replacement for your existing Python (no code changes required)

DRIVE PERFORMANCE WITH MULTIPLE OPTIMIZATION TECHNIQUES

- Accelerated NumPy/SciPy/Scikit-Learn with Intel® MKL
- Data analytics with pyDAAL, enhanced thread scheduling with TBB, Jupyter* Notebook interface, Numba, Cython
- Scale easily with optimized MPI4Py and Jupyter notebooks

FASTER ACCESS TO LATEST OPTIMIZATIONS FOR INTEL® ARCHITECTURE

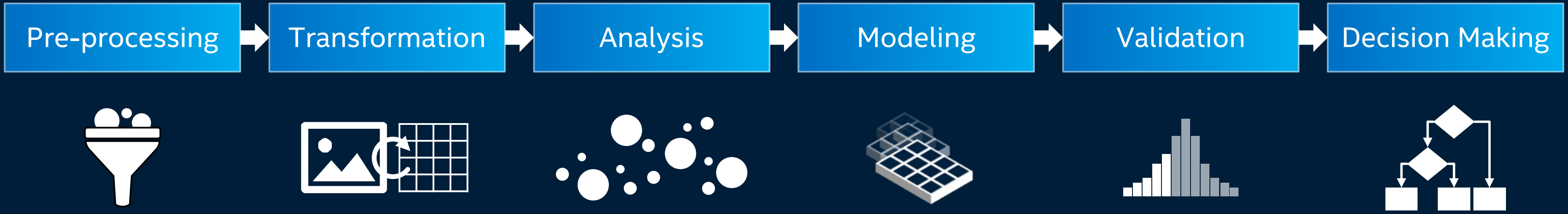
- Distribution and individual optimized packages available through conda and Anaconda Cloud
- Optimizations upstreamed back to main Python trunk

ADVANCING PYTHON* PERFORMANCE CLOSER TO NATIVE SPEEDS

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® DATA ANALYTICS ACCELERATION LIBRARY (INTEL® DAAL)

BUILDING BLOCKS FOR ALL DATA ANALYTICS STAGES, INCLUDING DATA PREPARATION, DATA MINING & MACHINE LEARNING


Pre-processing → Transformation → Analysis → Modeling → Validation → Decision Making

Open Source | Apache* 2.0 License

Common Python, Java and C++ APIs across all Intel hardware

Optimized for large data sets including streaming and distributed processing

Flexible interfaces to leading big data platforms including Spark* and range of data formats (CSV, SQL, etc.)

HIGH PERFORMANCE MACHINE LEARNING AND DATA ANALYTICS LIBRARY

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.

Optimization Notice

INTEL® MATH KERNEL FOR DEEP LEARNING NEURAL NETWORKS (INTEL® MKL-DNN)

FOR DEVELOPERS OF DEEP LEARNING FRAMEWORKS FEATURING OPTIMIZED PERFORMANCE ON INTEL HARDWARE

DISTRIBUTION DETAILS

- Open Source
- Apache* 2.0 License
- Common DNN APIs across all Intel hardware.
- Rapid release cycles, iterated with the DL community, to best support industry framework integration.
- Highly vectorized & threaded for maximal performance, based on the popular Intel® MKL library.

github.com/01org/mkl-dnn

EXAMPLES:

- Direct 2D Convolution
- Local response normalization (LRN)
- Rectified linear unit neuron activation (ReLU)
- Maximum pooling
- Inner product

Accelerate Performance of Deep Learning Models

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® DISTRIBUTION OF OPENVINO TOOLKIT

OPTIMIZE EXISTING MODELS, RUN INFERENCE WHERE YOU NEED IT

Build high performance deep learning inference and computer vision

OpenVINO™

A toolkit to accelerate development of **high performance computer vision** & **deep learning inference into vision/AI applications** from edge to cloud. It enables deep learning on hardware accelerators and easy deployment across multiple types of Intel® platforms (CPU, GPU/Intel® Processor Graphics, FPGA, VPU).

Who needs it?

- Computer vision, deep learning developers
- Data scientists
- OEMs, ISVs, system integrators

Usages

Security surveillance, robotics, retail, healthcare, AI, office automation, transportation, non-vision use cases (speech, text) & more.

HIGH PERFORMANCE AI EDGE TO CLOUD

STREAMLINED & OPTIMIZED DEEP LEARNING INFERENCE

HETEROGENEOUS, CROSS-PLATFORM FLEXIBILITY

Free Download ▶ software.intel.com/openvino-toolkit

Open Source version ▶ 01.org/openvinotoolkit

USE CASE

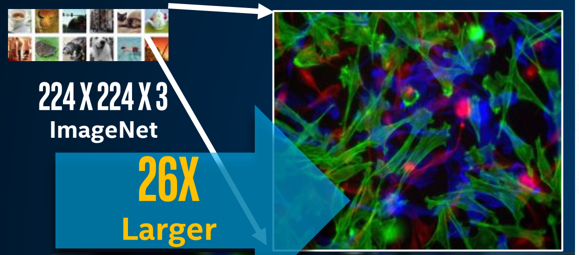
HPC← →AI: IMAGE ANALYSIS FOR DRUG DISCOVERY

NOVARTIS

Joint Intel & Novartis collaboration

RESULTS

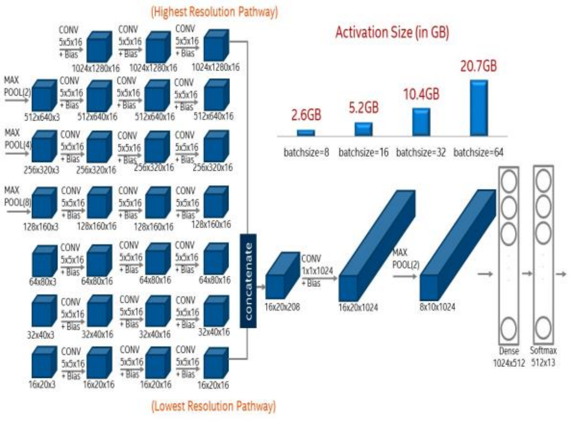
Processing 1024x1280 large image dataset, reducing the training time to 31 minutes to >99% accuracy on 2S Intel® Xeon® processor based cluster.

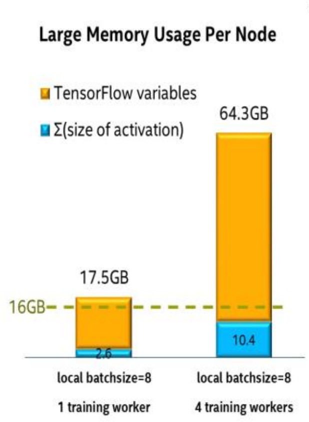


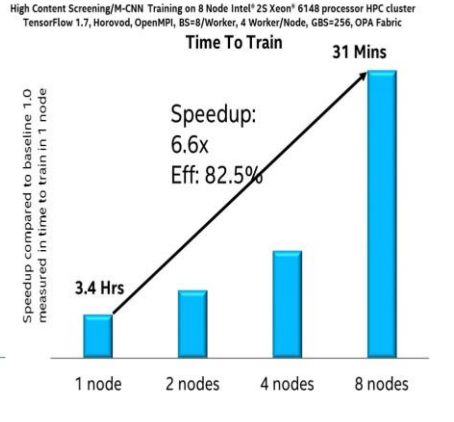
224 X 224 X 3
ImageNet

26X
Larger

1024 X 1280 X 3
Microscopic
Images






High Content Screening/M-CNN Training on 8 Node Intel® 2S Xeon® 6148 processor HPC cluster
TensorFlow 1.7, Horovod, OpenMPI, BS=8/Worker, 4 Worker/Node, GBS=256, OPA Fabric

Customer:
Novartis Inst. of Biomedical Research (Switzerland) is one of the largest pharmaceutical companies in the world

Challenge: High content screening of cellular phenotypes is a fundamental tool supporting early stage drug discovery. While analyzing whole microscopic images are desirable, these images are 26X larger than benchmark dataset such as ImageNet*-1K. As a result, the high computational workload with high memory requirement would be prohibitive for deep learning model training

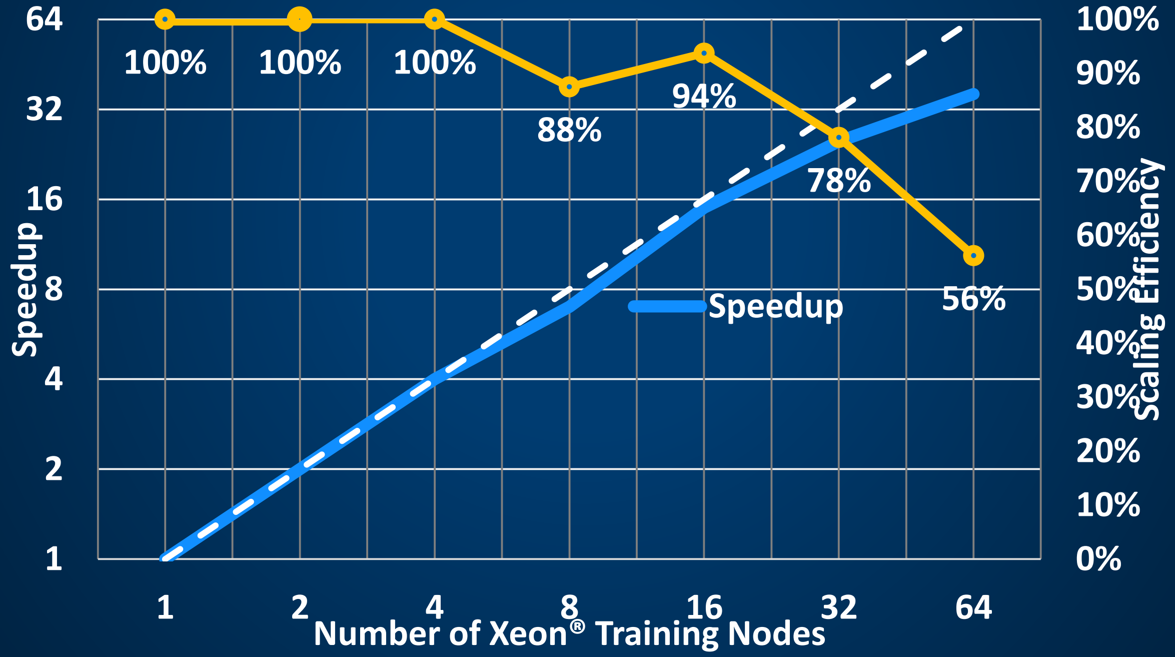
Solution: Intel and Novartis teams were able to scale and train the model with 32 TensorFlow* workers in 31 minutes.

http://aidc.gallery.video/detail/video/5790618241001/deep-learning-based-classification-of-high-content-cellular-images-on-intel-architecture?autoStart=true&q=Datta

Performance results are based on testing as of May 17, 2018 and may not reflect all publicly available security update. See configuration disclosure for details. No product can be absolutely secure.
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

HPC ←→ AI: IMAGE ANALYSIS FOR DRUG DISCOVERY

High Content Screening Training with 313K Images on 64-Node Intel® 2S Xeon® Scalable processor 6148, TensorFlow*, “horovod*”, OpenMPI*, Batch Size=32/Node, Intel® Omni-Path™ Fabric



Performance results are based on testing as of April, 2019 and may not reflect all publicly available security update. See configuration disclosure for details. No product can be absolutely secure.
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

HPC←→AI: DIS/REPLACING MONTE CARLO SIM.

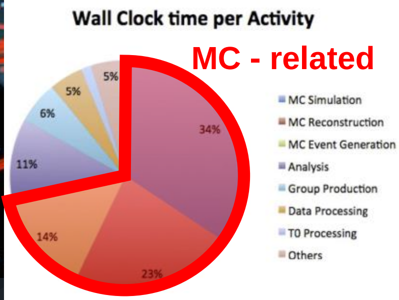
CERN HIGH ENERGY PHYSICS

JOINT COLLABORATION WITH INTEL AND SURFSARA

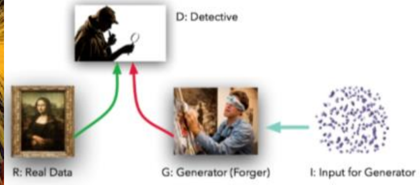
USE CASE

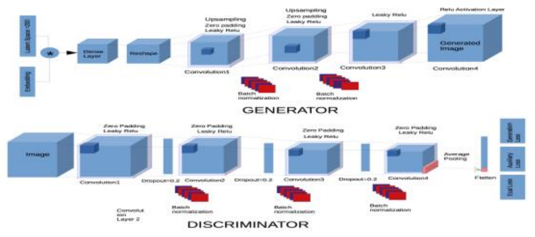
RESULT

94% scaling efficiency up to 128 nodes, with a significant reduction in training time per epoch for 3D-GANs & >2500X Inference



WLCG Wall Clock time for the ATLAS experiment





3D-Generative Adversarial Networks(GANs)

Time to create an electron shower		
Method	Machine	Time/Shower (msec)
Full Simulation (geant4)	2S Intel® Xeon® Platinum 8180	17000
3D GAN (batch size 128)	**2S Intel® Xeon® Platinum 8180**	**7**

Inference Perf: >2500X

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 then scaled using TensorFlow* optimized with Intel® MKL-DNN, using 4 workers/node and an optimized number of convolutional filters.

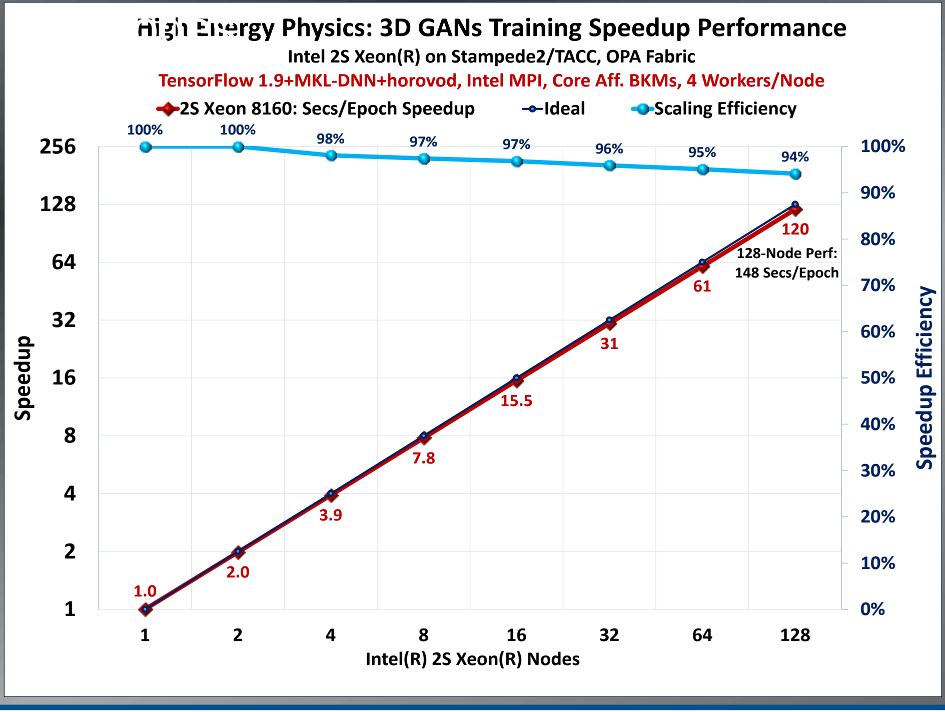
https://www.rdmag.com/article/2018/11/imagining-unthinkable-simulations-without-classical-monte-carlo

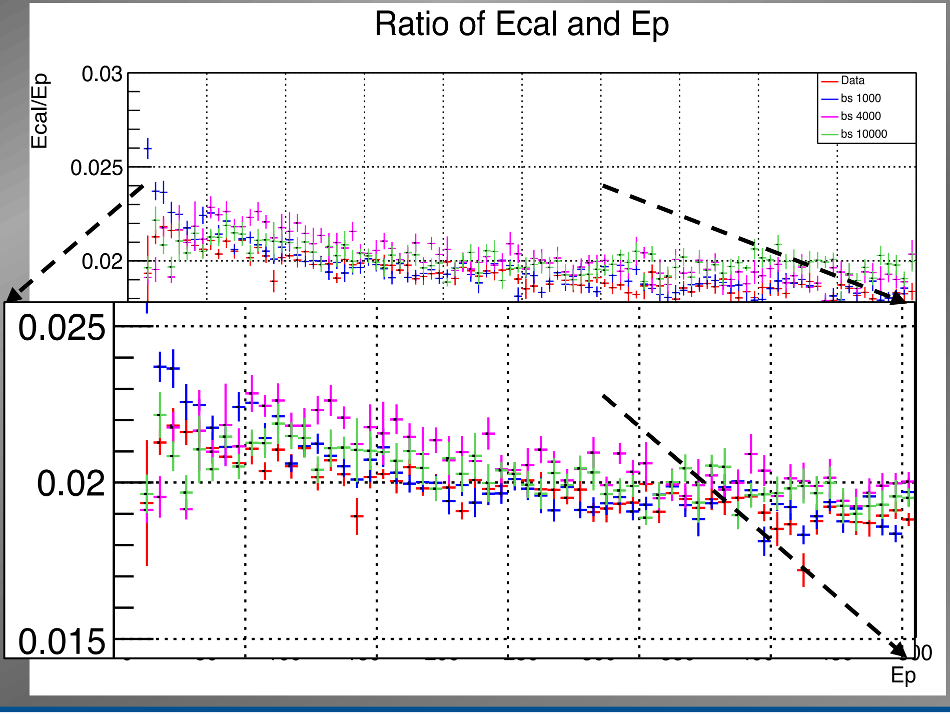
Performance results are based on testing as of May 17, 2018 and may not reflect all publicly available security update. See configuration disclosure for details. No product can be absolutely secure.
Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors 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

Multi-Node Training Performance & Accuracy (2018)

Distributed training using data parallelism

94% Scaling efficiency up to 128





HPC ← → AI: CHEST X-RAY IMAGE CLASSIFICATION

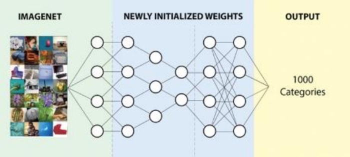
DellEMC

Joint collaboration with SURFsara, & DellEMC

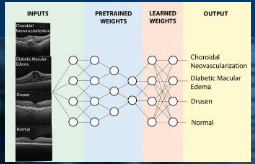
Identifying thoracic pathologies from the NIH ChestXray14 dataset

Emphysema affects more than: 3 Mil U.S & 65 Mil Worldwide

Pneumonia affects more than: 1 Mil US & 450 Mil Worldwide



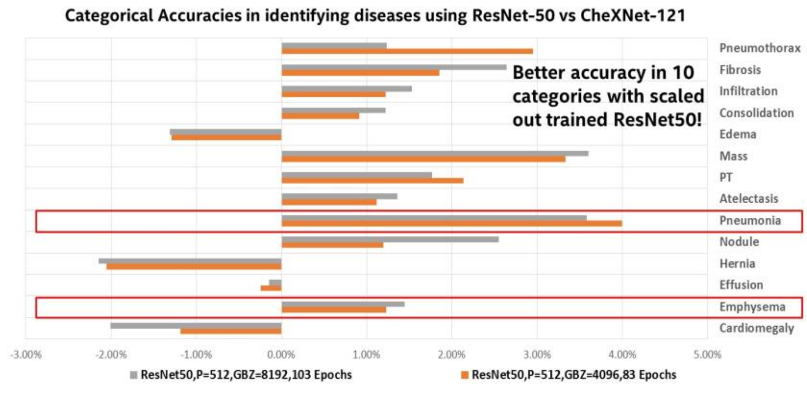
Transfer Learning →



USE CASE

RESULT

Training time reduced to 11 mins while increasing the accuracy across 10 categories & 4% (>90%) better relative to the existing DenseNet-121 model


Categorical Accuracies in identifying diseases using ResNet-50 vs CheXNet-121

Better accuracy in 10 categories with scaled out trained ResNet50!

Customer: DellEMC* Research on AI applications on Intel® Xeon® CPUs: Medical, Cloud, HPC, etc.

Challenge: Train a chest X-ray model that delivers highly-efficient scaling performance on Intel® Xeon® processor nodes, while also delivering higher accuracy than the existing ChexNet-121 model

Solution: 256-node cluster consisting of dual Intel® Xeon® Gold 6148 processor, Intel® Omni-Path fabric, and ResNet-50 topology. ResNet50 tests performed with TensorFlow* and Horovod*.

https://ai.intel.com/diagnosing-lung-disease-using-deep-learning/

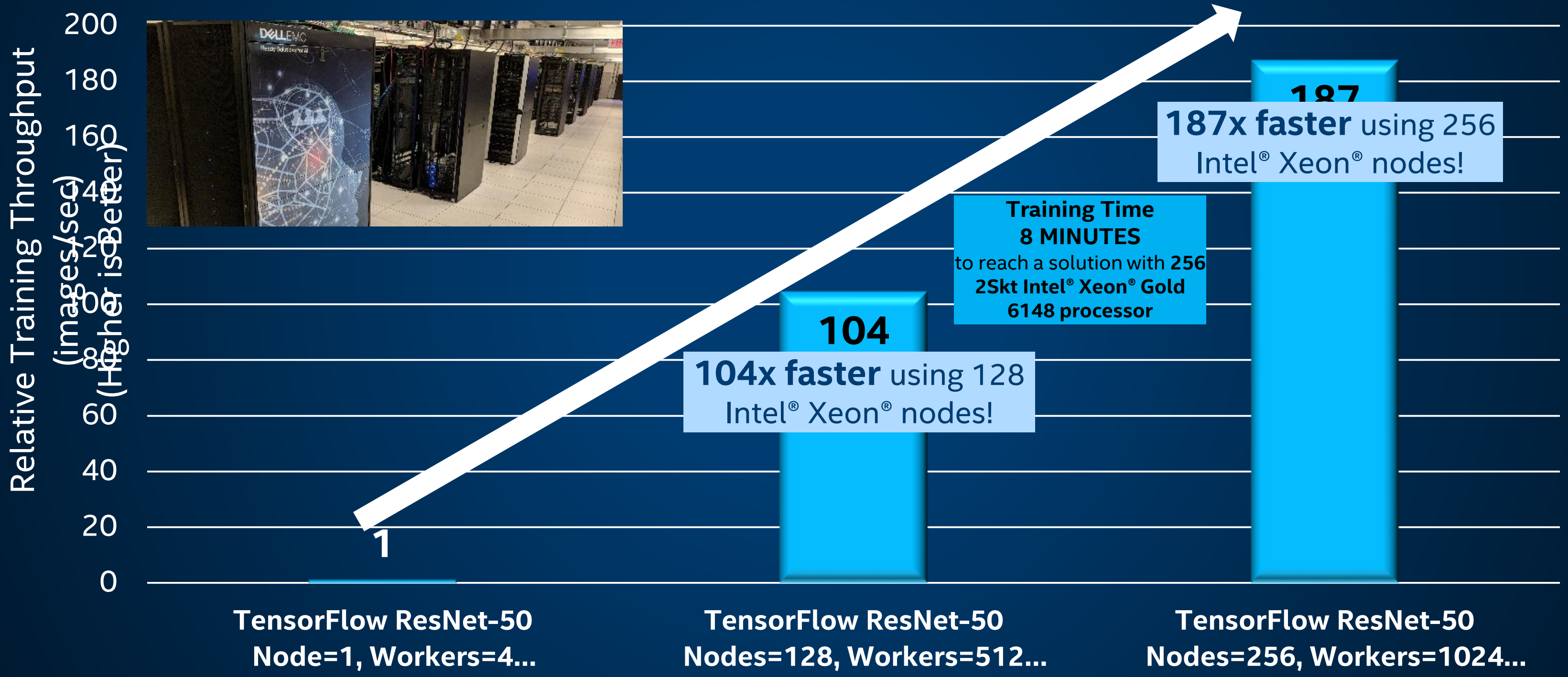
Performance results are based on testing as of May 17, 2018 and may not reflect all publicly available security update. See configuration disclosure for details. No product can be absolutely secure.
Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors 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

Thanks to DellEMC for Access to Internal Clustetr

TRAINING PERFORMANCE: RESNET-50 ON CHESTXRAY14

INTEL® 2S XEON® GOLD 6148F PROCESSOR BASED DELLEMC* POWEREDGE C6420 ZENITH* CLUSTER ON OPA™ FABRIC

TENSORFLOW* + “HOROVOD*”, IMPI



104x faster using 128 Intel® Xeon® nodes!

187x faster using 256 Intel® Xeon® nodes!

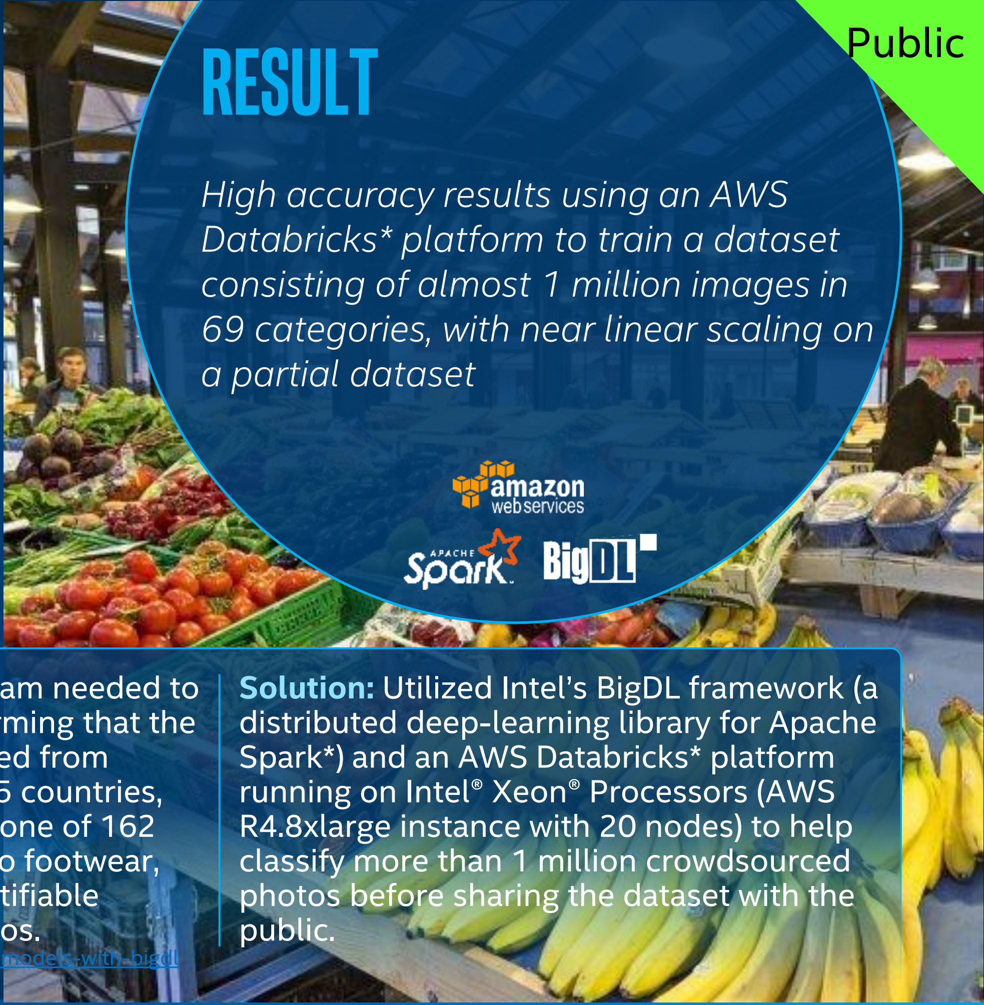
Training Time
8 MINUTES
to reach a solution with **256 2Skt Intel® Xeon® Gold 6148 processor**

Performance results are based on testing as of May 17, 2018 and may not reflect all publicly available security update. See configuration disclosure for details. No product can be absolutely secure.
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

Public

Case Study: Image Recognition

World Bank



RESULT

High accuracy results using an AWS Databricks platform to train a dataset consisting of almost 1 million images in 69 categories, with near linear scaling on a partial dataset*

amazon web services

Apache Spark BigDL

Client: The International Comparison Program (ICP) in the World Bank Development Data Group

Challenge: The World Bank team needed to automate the process of confirming that the crowd-sourced photos, gathered from cellphone contributors from 15 countries, were accurately classified into one of 162 categories ranging from food to footwear, and to remove personally identifiable information (PII) from the photos.

Solution: Utilized Intel's BigDL framework (a distributed deep-learning library for Apache Spark*) and an AWS Databricks* platform running on Intel® Xeon® Processors (AWS R4.8xlarge instance with 20 nodes) to help classify more than 1 million crowdsourced photos before sharing the dataset with the public.

https://databricks.com/session/using-crowdsourced-images-to-create-image-recognition-models-with-bigdl
https://itpeernetwork.intel.com/artificial-intelligence-world-bank-image-recognition/

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.

CENTER FOR DIGITAL HEALTH INNOVATION (CDHI) AT UCSF*

RESULT

"Dataset, model development and training […] implementing 3D CNN in BigDL to analyze MRI scans and classify OA (osteoarthritis) […] provides rich 3D imaging support […] on the same cluster where data is stored"

Client: Center for Digital Health Innovation (CDHI) at UCSF, leveraging new digital health technologies to transform healthcare.

Challenge: Projected by 2040 – 78M adults with doctor-diagnosed OA & 35M with arthritis-attributable activity limitations. Need automated system that classifies menisci based on presence/absence of lesions, provides immediate objective results at MRI scan, & eliminates intra-user variability.

Solution: Apache Spark* with BigDL on CDH 5.9*, on Intel® Xeon® servers from Dell*. With 3D image convolution in BigDL, the CDHI team built a MRI classification system & deployed it on their CDH Dell cluster.

https://cdn.oreillystatic.com/en/assets/1/event/269/Automatic%203D%20MRI%20knee%20damage%20classification%20with%203D%20CNN%20using%20BigDL%20on%20Spark%20Presentation.pdf

UCSF = University of California – San Francisco

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

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.

Artificial Intelligence will empower

TRANSFORMATIVE INNOVATIONS

WE ARE ~~WITNESSING~~ CREATING THIS DIGITAL TRANSFORMATION intel® AI