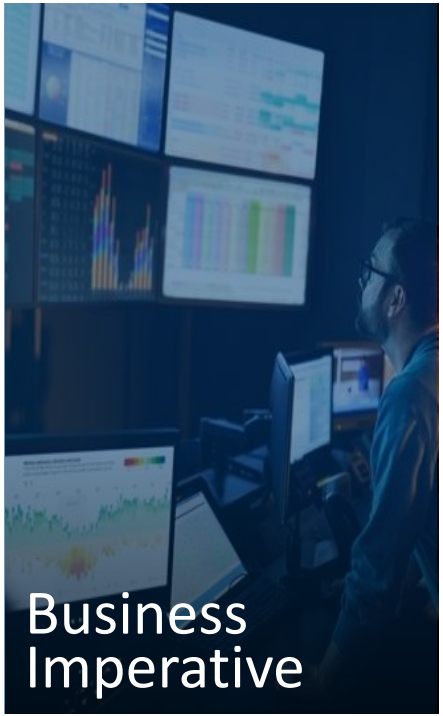
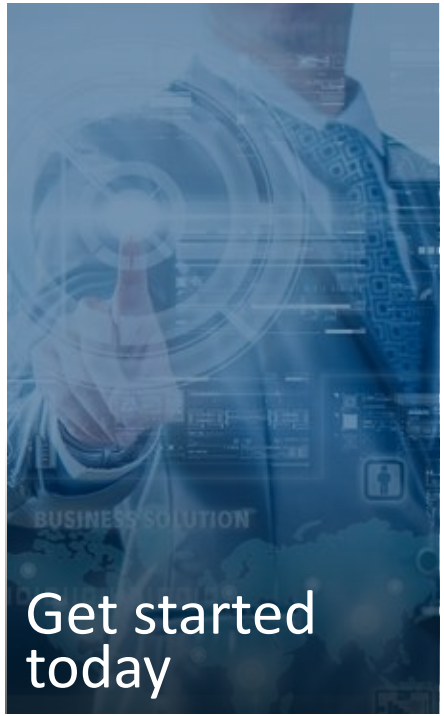
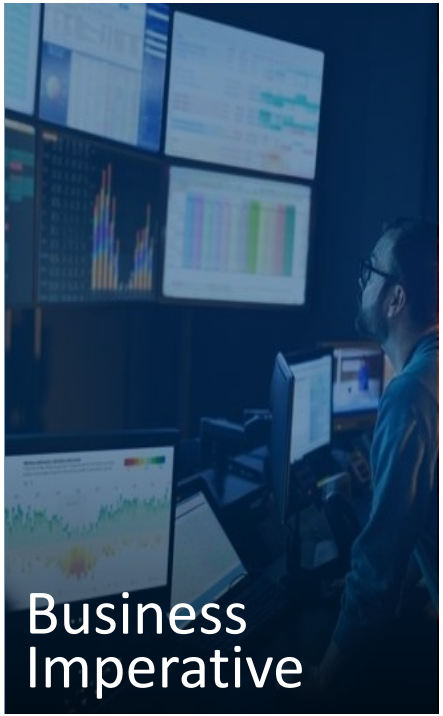

Business
Imperative


What is AI?


AI with Intel


Get started
today


Business
Imperative

What is AI?

AI with Intel

Get started
today

The deluge of data

Daily By 2020

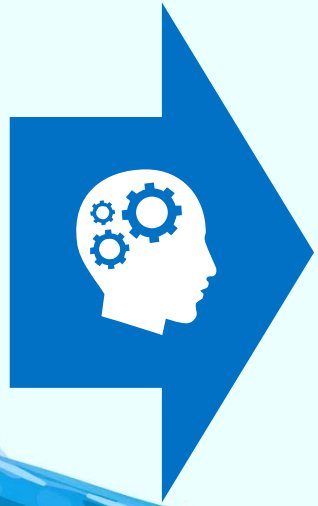
verage internet user 1.5 GB

Autonomous vehicle 4 TB

ONNECTED AIRPLANE 5 TB

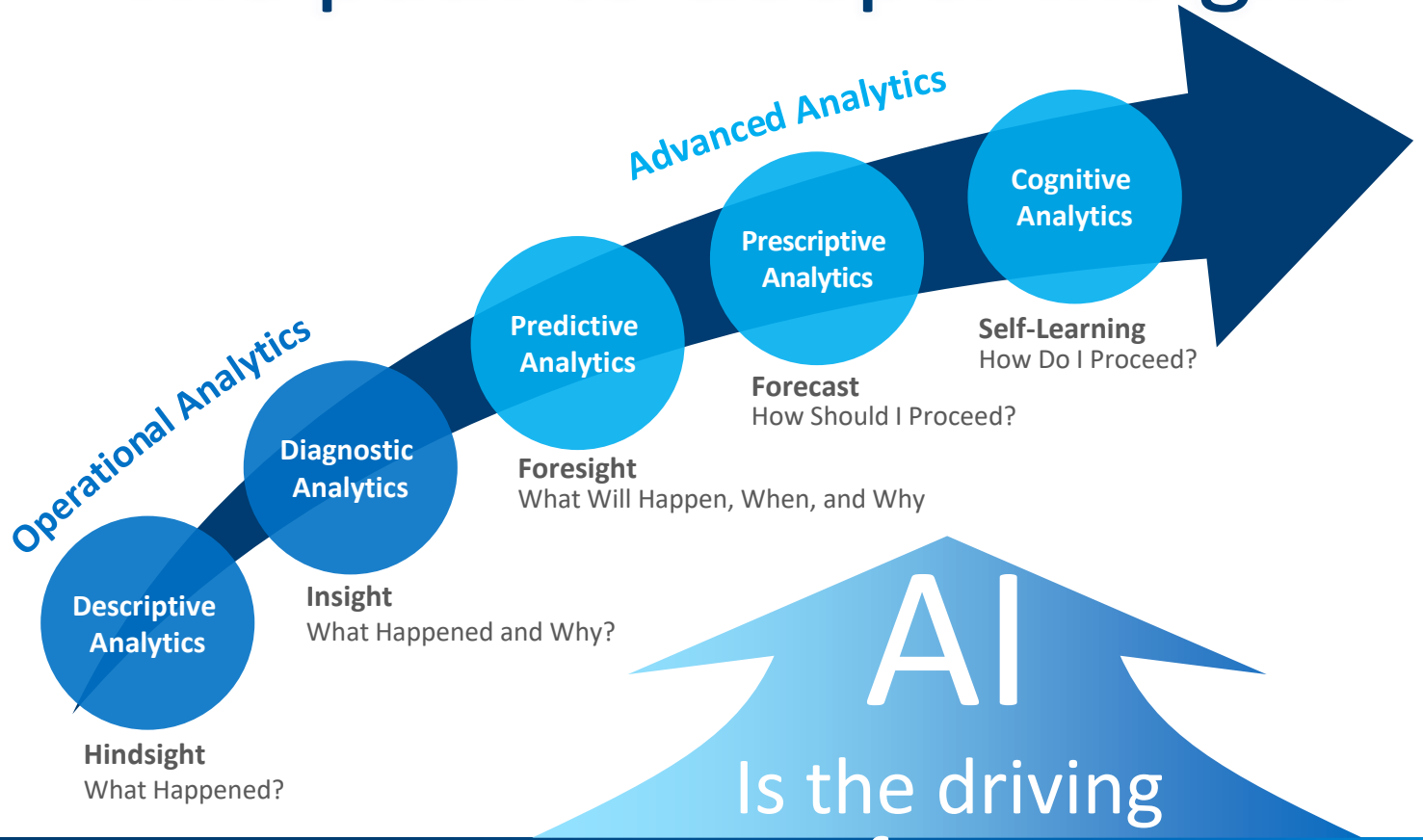
Smart Factory 1 PB

Cloud video Provider 750 pB

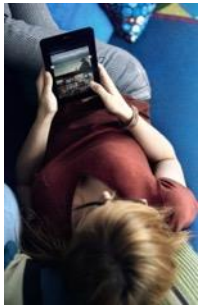


Source: Amalgamation of analyst data and Intel analysis.

The path to deeper insight



AI will transform



Consumer

- Smart Assistants
- Chatbots
- Search
- Personalization
- Augmented Reality
- Robots

Health

- Enhanced Diagnostics
- Drug Discovery
- Patient Care
- Research
- Sensory Aids

Finance

- Algorithmic Trading
- Fraud Detection
- Research
- Personal Finance
- Risk Mitigation

Retail

- Support Experience
- Marketing
- Merchandising
- Loyalty
- Supply Chain
- Security

Government

- Defense
- Data Insights
- Safety & Security
- Resident Engagement
- Smarter Cities

Energy

- Oil & Gas Exploration
- Smart Grid
- Operational Improvement
- Conservation

Transport

- In-Vehicle Experience
- Automated Driving
- Aerospace
- Shipping
- Search & Rescue

Industrial

- Factory Automation
- Predictive Maintenance
- Precision Agriculture
- Field Automation

Other

- Advertising
- Education
- Gaming
- Professional & IT Services
- Telco/Media
- Sports

Source: Intel forecast

Ai adoption is nascent

According to a recent Forrester Research survey…

58%

of business and technology professionals said they're researching AI, but **only**

12%

said they are currently using AI systems.

Source: Forrester Research – Artificial Intelligence: Fact, Fiction. How Enterprises Can Crush It; What's Possible for Enterprises in 2017

AI Opportunity assessment

Brainstorm and Prioritize Business Challenges

What business challenges am I facing today?

What business value is tied to each challenge?

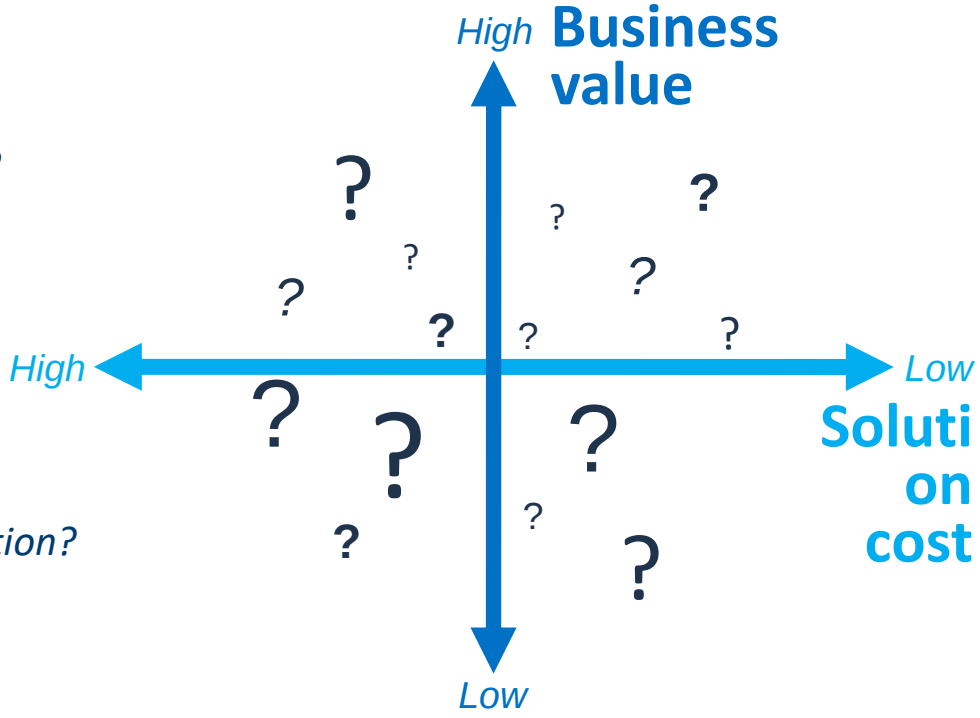
What are my solution requirements?

What data do I have at my disposal?

Do I know how to approach each challenge?

Do I have what I need to implement each solution?

How costly is it to implement each solution?



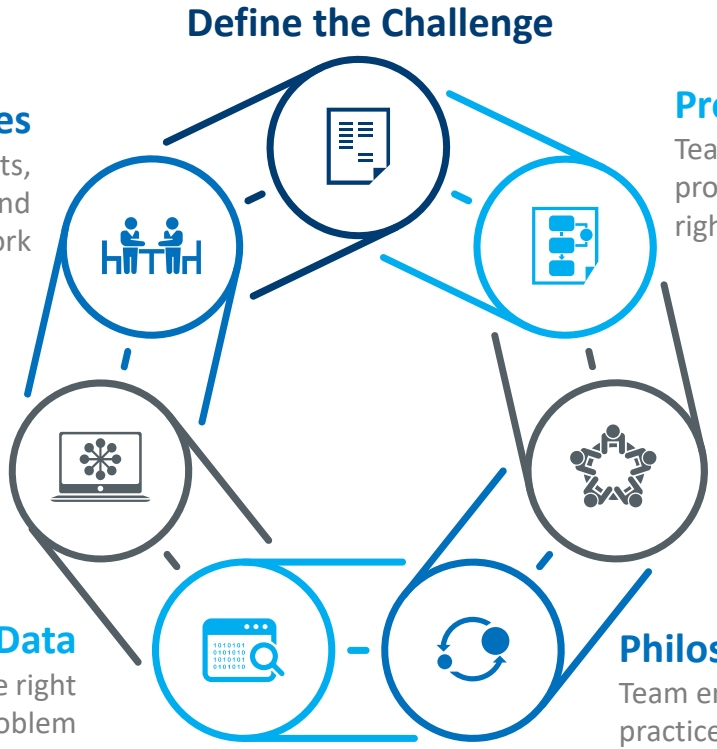
Which approach is right?

A large **manufacturer** uses data to improve their operations, with each challenge using a different approach to deliver maximum business value at the lowest possible cost

Challenge	Best approach	Approach	Answer
How many widgets should we manufacture?	Analyze historical supply/demand	Analytics/ Business Intelligence	10,000
What will our yield be?	Algorithm that correlates many variables to yield	Statistical/ Machine Learning	At current conditions, yield will be at 90% with 10% loss expected
Which widgets have visual defects?	Algorithm that learns to identify defects in images	Deep Learning	Widget 1003, Widget 1094 . . .
How do I resolve each defect?	Algorithms that learns associations between defect type/cause	Reasoning Systems	Widget 1003 – Tolerance issue, calibration . . . Widget 1094 – Etch issue, gas pressure warning . . .

Learn More in the Next Section

The AI lifecycle



Culture & Resources
Organization embraces data insights, sponsors properly resourced teams, and prioritizes analytic development work

Problem Solving Process
Team breaks down the defined business problem into workable steps to translate the right data to achieve results

Infrastructure
Organization secures hardware and software infrastructure that supports data processing in a timely manner

Expertise
A team of management sponsors, data scientists, data engineers, solution architects, and domain experts identifies the right data and works to translate the data to achieve results

Source Data
Team understands and obtains the right data that explains the business problem to achieve results

Philosophy
Team embraces fail-fast continuous improvement practices to evaluate their success in translating data to achieve results

Business
Imperative

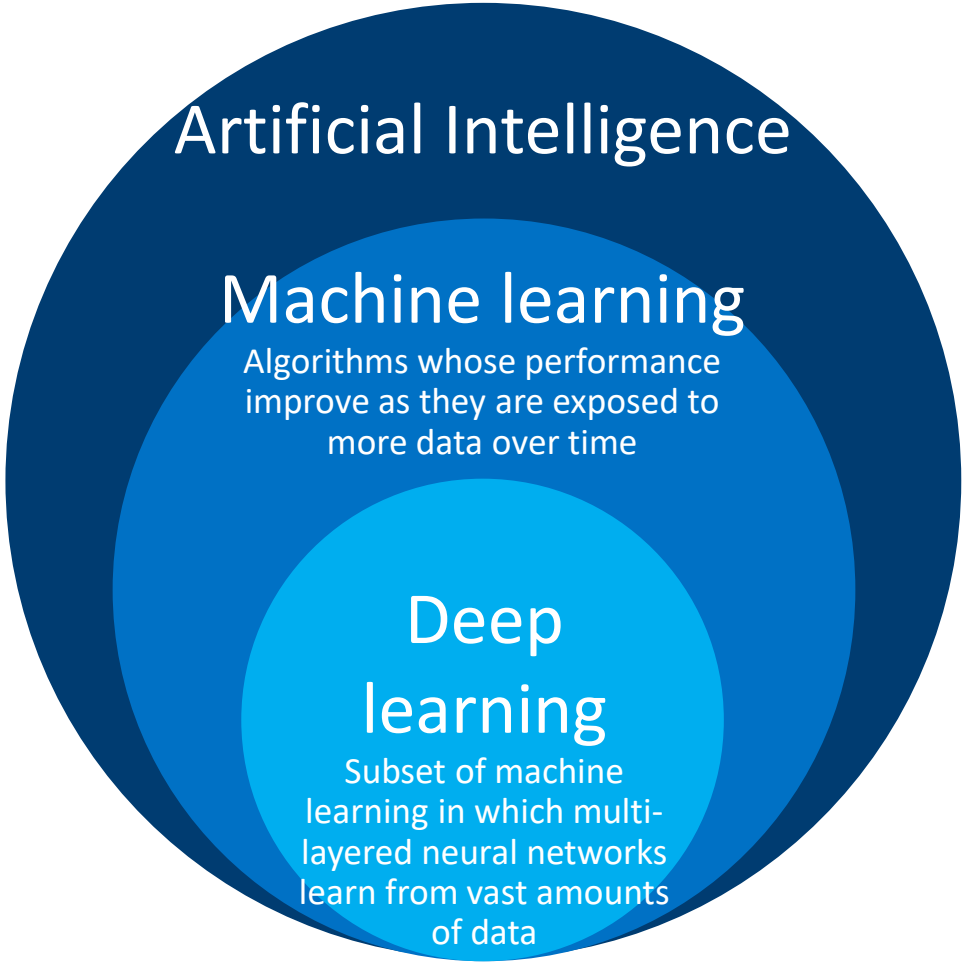

What is AI?

AI with Intel

Get started
today

Artificial Intelligence

is the ability of machines to learn from experience, without explicit programming, in order to perform cognitive functions associated with the human mind



Ai cl ser look

Machine learning

Algorithms designed to deliver better insight with more data

Regression (Linear/Logistic)

Classification (Support Vector Machines/SVM, Naïve Bayes)

Clustering (Hierarchical, Bayesian, K-Means, DBSCAN)

Decision Trees (RandomForest)

Extrapolation (Hidden Markov Models/HMM)

More…

Deep Learning

Neural networks used to infer meaning from large dense datasets

Image Recognition (Convolutional Neural Networks/CNN, Single-Shot Detector/SSD)

Speech Recognition (Recurrent Neural Network/RNN)

Natural Language Processing (Long-Short Term Memory/LSTM)

Data Generation (Generative Adversarial Networks/GAN)

Recommender System (Multi-Layer Perceptron/MLP)

Time-Series Analysis (LSTM, RNN)

Reinforcement Learning (CNN, RNN)

More…

reasoning

Hybrid of analytics & AI techniques designed to find meaning in diverse datasets

Associative Memory (Intel® Saffron AI memory base)

← **See also:** machine & deep learning techniques

More…

Machine vs. Deep learning

Machine Learning

How do you engineer the best features?

$N \times N$

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

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

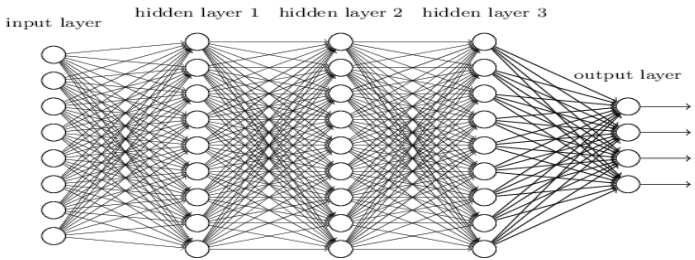
Arjun

Deep Learning

How do you guide the model to find the best features?

$N \times N$

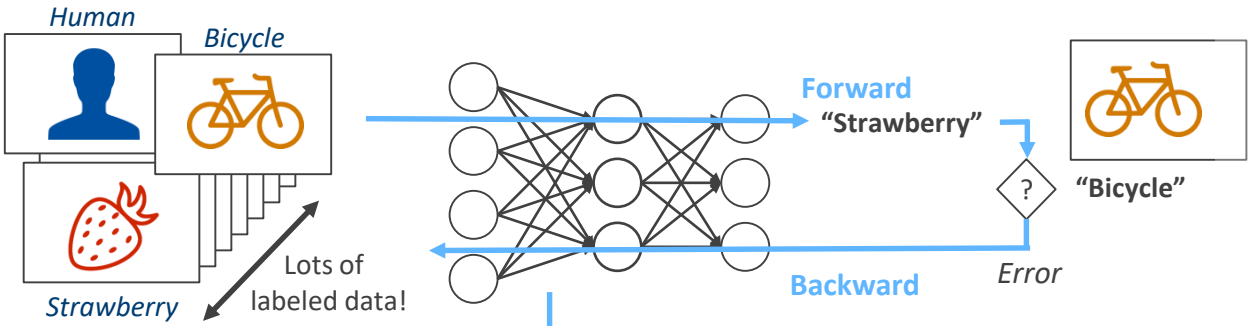
NEURAL NETWORK



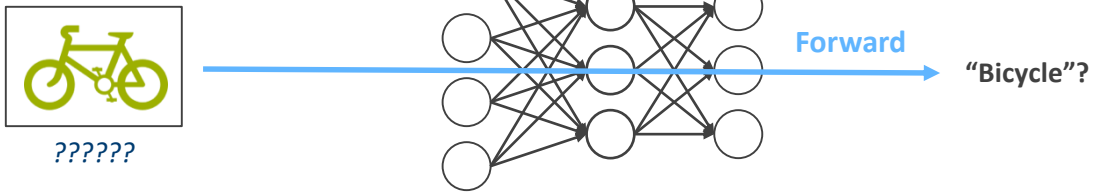
Arjun

Deep learning Basics

Training

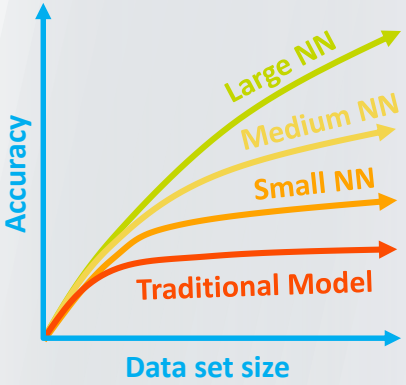


Inference



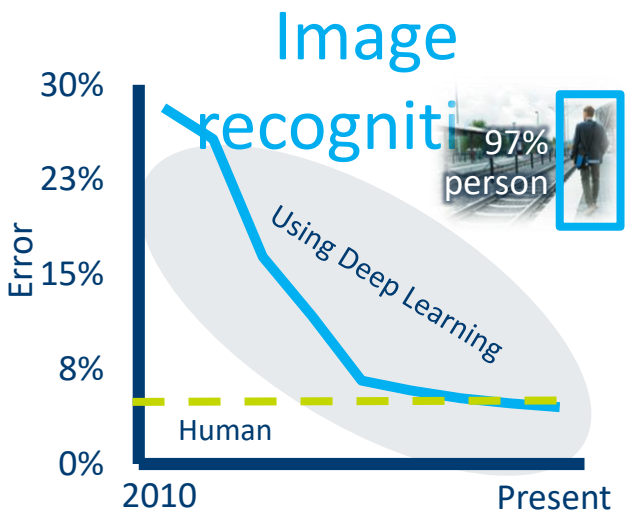
Did you know?

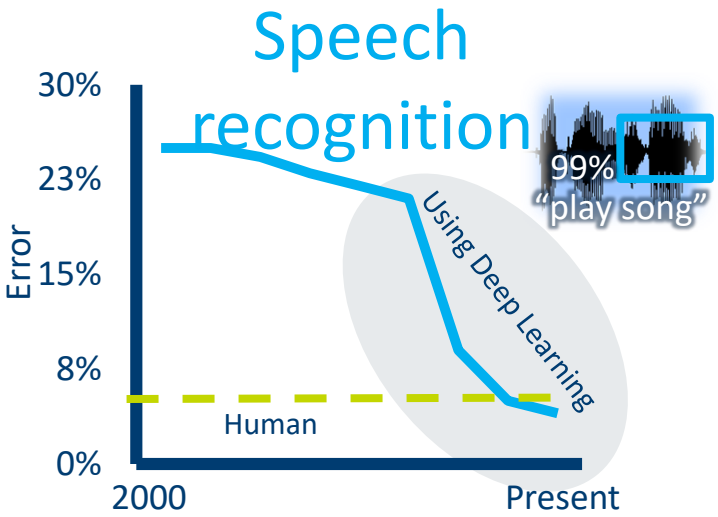
Training with a large data set AND deep (many layered) neural network often leads to the highest accuracy inference



Deep learning breakthroughs

Machines able to meet or exceed human image & speech recognition

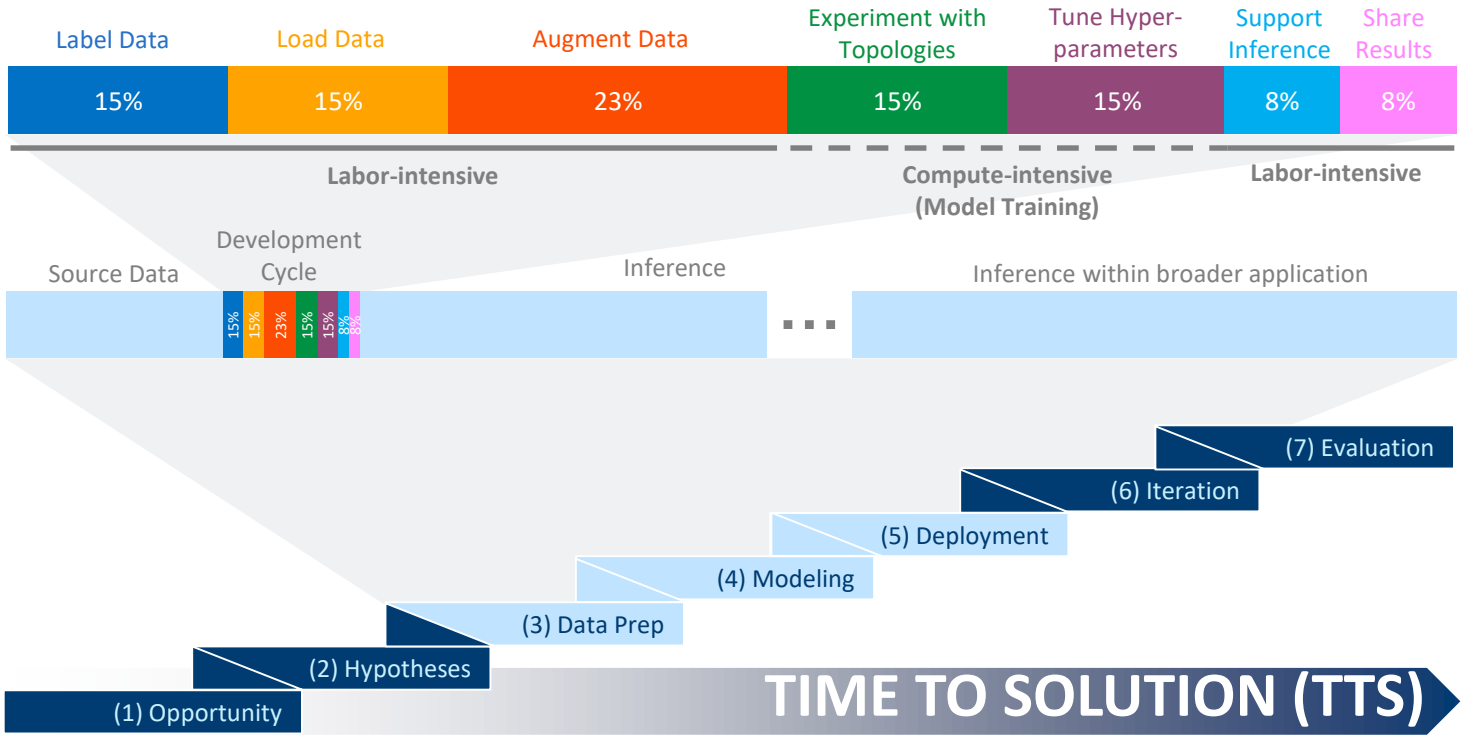




e.g.

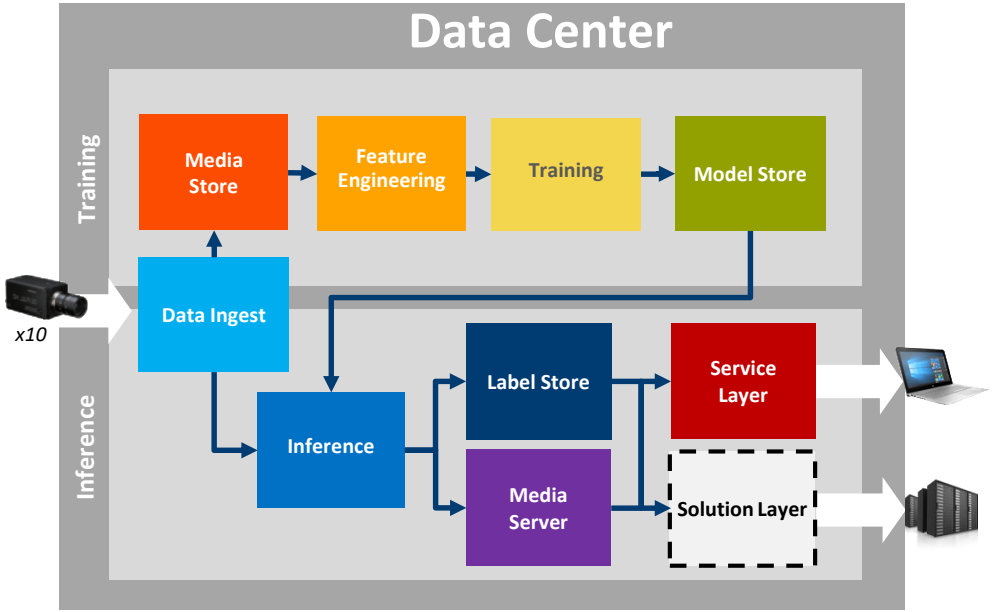
Source: ILSVRC ImageNet winning entry classification error rate each year 2010-2016 (Left), https://www.microsoft.com/en-us/research/blog/microsoft-researchers-achieve-new-conversational-speech-recognition-milestone/ (Right)

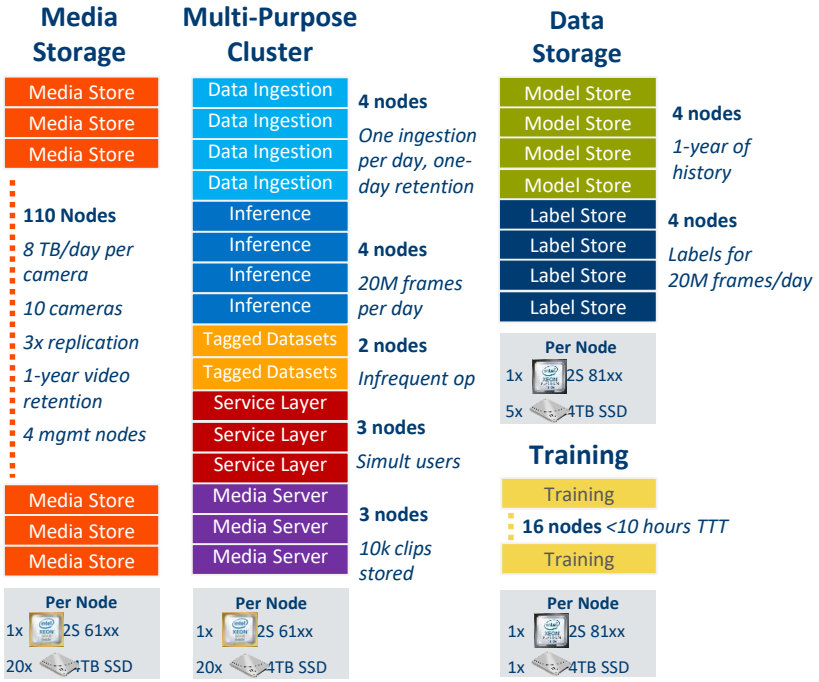
The Journey to Production AI



Deep learning in practice

AI deployments have many interconnected parts





Source: Intel customer engagement

Business
Imperative

What is AI?


AI with Intel

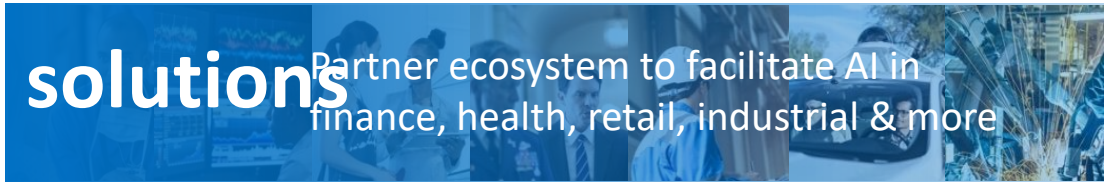
Get started
today

intel AI

Smarter AI Through the Industry's Most Comprehensive Platform

Data

Intel analytics ecosystem to get your data ready from integration to analysis



Future

Driving AI forward through R&D, investment and policy leadership

da ta

Intel analytics ecosystem to get your data ready from integration to analysis

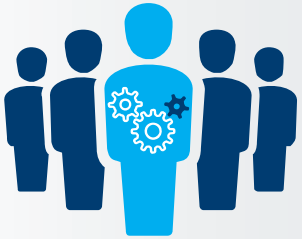
	Integrate	Store	Process	Manage	Analyze
Source(s)? Structured? Volume? URABILITY? Streaming? LOCALITY? VERNANCE? Other?	Tool for **live streaming data ingestion** from Internet of Things (IOT) sensors in endpoint devices *e.g. Kafka*, Sqoop*, MQTT*, WS*, REST*, Flume**	**File**, **block** or **object**-based storage solution given cost, access, volume and performance requirements *e.g. Lustre*, IBM* Spectrum* Scale* (GPFS), Dell/EMC* Isilon*, MySQL* (OLTP), Tera*data* (EDW), AWS* S3* (ODS), HDFS* (No-SQL), Hbase* (In-Mem DB)*	**Integration, cleaning, normalization and other** transformations on batch and/or streaming data *e.g. Hadoop* MapReduce*, Apache* Storm*, Beam**	**Job scheduling** and **storage management** framework for distributed computation in various domains *e.g. SLURM*, PBS*, YARN*, Mesos*, Kubernetes**	Applications in **HPC, Big Data, HPDA, AI** & more that have access to a common compute and data pool *e.g. MPI*, SHMEM*, Hadoop*, Spark*, Apache*, Flink*, TensorFlow*, MXNet**

*Other names and brands may be claimed as the property of others. Non-exhaustive list of offerings in each category.

solutions

Partner ecosystem to facilitate AI in finance, health, retail, industrial & more

Intel AI Builders



Your one-stop-shop to find systems, software and solutions using Intel® AI technologies

builders.intel.com/ai/membership

Reference solutions

Get a head start using the many case studies, solution briefs and more reference collaterals spanning multiple applications

builders.intel.com/ai/solutionslibrary

SEE ALSO: AI Solution Deck (internal)

TOOLS

Portfolio of software tools to accelerate time-to-solution

TOOLKITS

App Developers

DEEP LEARNING DEPLOYMENT

OpenVINO™ †

Open Visual Inference & Neural Network Optimization toolkit for inference deployment on CPU/GPU/FPGA for TF, Caffe & MXNet**

Intel® Movidius™ SDK

Optimized inference deployment on Intel VPUs for TensorFlow & Caffe**

REASONING

Intel® Saffron™ AI

Cognitive solutions on CPU for anti-money laundering, predictive maintenance, more

DEEP LEARNING

Coming soon

Intel® Deep Learning Studio

Open-source tool to compress deep learning development cycle

libraries

Data Scientists

MACHINE LEARNING LIBS

Python
- Scikit-learn
- Pandas
- NumPy

R
- Cart
- Random Forest
- e1071

Distributed
- MLLib (on Spark)
- Mahout

DEEP LEARNING FRAMEWORKS

Coming soon

Now optimized for CPU

TensorFlow* MXNet* Caffe* BigDL/Spark*

Optimizations in progress

Caffe2* PyTorch* CNTK* PaddlePaddle*

foundation

Library Developers

ANALYTICS, MACHINE & DEEP LEARNING PRIMITIVES

Python

Intel distribution optimized for machine learning

DAAL

Intel® Data Analytics Acceleration Library (incl machine learning)

MKL-DNN **clDNN**

Open-source deep neural network functions for CPU / integrated graphics

DEEP LEARNING GRAPH COMPILER

Intel® nGraph™ Compiler (Alpha)

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

† *Formerly the Intel® Computer Vision SDK*

*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.

HARDWARE

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

intel AI

Mainstream

intensive

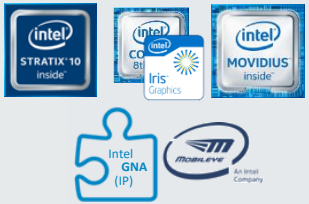
Deep Learning

Training

+

Inference

+



Most other AI

Ai compute continuum

Cloud / Data Center

Large scale data centers such as public cloud or comms service providers, gov’t and academia, large enterprise IT

Edge

Small scale data centers, small business IT infrastructure, to few on-premise server racks and workstations

device

User-touch endpoint devices with lower power requirements such as laptops, tablets, smart home devices, drones

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

Deep learning inference accelerators

Intel® FPGA

Custom deep learning inference

Intel® Movidius™ VPU

Low power computer vision & inference

Intel® Mobileye EyeQ

Autonomous driving inference platform

Intel® GNA IP[1]

Ultra low power speech & audio inference

Integrated graphics

Built-in deep learning inference

Data Center

Edge — *Small scale clusters to a few on-premise server & workstations*

Device — *User-touch end-devices typically with lower power requirements*

[1]GNA=Gaussian Mixture Model and Neural Network Accelerator

intel AI Solutions

Solutions
Solution Architects

TOOLKITS
App Developers

libraries
Data Scientists

foundation
Library Developers

Hardware
IT System Architects

ARTIFICIAL INTELLIGENCE

AI Solutions Catalog (Public & Internal)

Platforms | *Finance* | *Healthcare* | *Energy* | *Industrial* | *Transport* | *Retail* | *Home* | *More…*

DEEP LEARNING DEPLOYMENT

OpenVINO™
Open Visual Inference & Neural Network Optimization toolkit for inference deployment on CPU/GPU/FPGA for TF, Caffe & MXNet**

Intel® Movidius™ SDK
Optimized inference deployment on Intel VPUs for TensorFlow & Caffe**

REASONING

Intel® Saffron™ AI
Cognitive solutions on CPU for anti-money laundering, predictive maintenance, more

DEEP LEARNING

Coming soon

Intel® Deep Learning Studio
Open-source tool to compress deep learning development cycle

MACHINE LEARNING LIBRARIES

Python
- Scikit-learn
- Pandas
- NumPy

R
- Cart
- Random Forest
- e1071

Distributed
- MLlib (on Spark)
- Mahout

DEEP LEARNING FRAMEWORKS

Coming soon

Now optimized for CPU

TensorFlow* | MXNet* | Caffe* | BigDL/Spark*

Optimizations in progress

Caffe2* | PyTorch* | CNTK* | PaddlePaddle*

ANALYTICS, MACHINE & DEEP LEARNING PRIMITIVES

Python
Intel distribution optimized for machine learning

DAAL
Intel® Data Analytics Acceleration Library (incl machine learning)

MKL-DNN **clDNN**
Open-source deep neural network functions for CPU / integrated graphics

DEEP LEARNING GRAPH COMPILER

Intel® nGraph™ Compiler (Alpha)
Open-sourced compiler for deep learning model computations optimized for multiple devices (CPU, GPU, NNP) from multiple frameworks (TF, MXNet, ONNX)

AI FOUNDATION

DEEP LEARNING ACCELERATORS

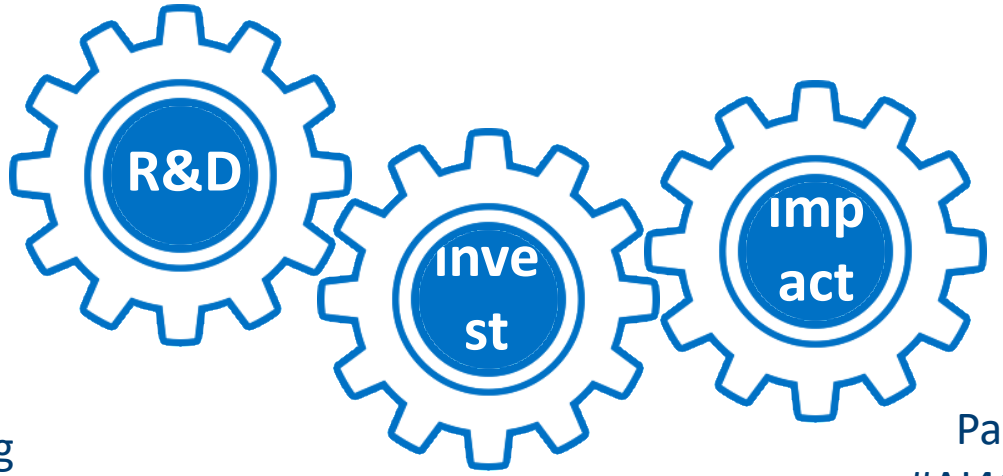
Coming soon

Intel Atom x7 | Intel Core i7 8th Gen | Intel Xeon Platinum | Data Center / Edge / Device | Intel Stratix 10 | Intel Arria 10 | Intel Movidius | Mobileye | Intel Iris Graphics | Intel GNA (IP) | Intel Nervana

Inference *Training*

Future

Driving AI forward through R&D, investment and policy leadership



- ✓ Image/Video/Audio
- ✓ Natural Language
- ✓ Autonomous Driving
- ✓ Reinforcement Learning
- ✓ Adversarial Learning
- ✓ Limited Precision/Sparsity
- ✓ More…

Over $1B invested in AI innovators[1]

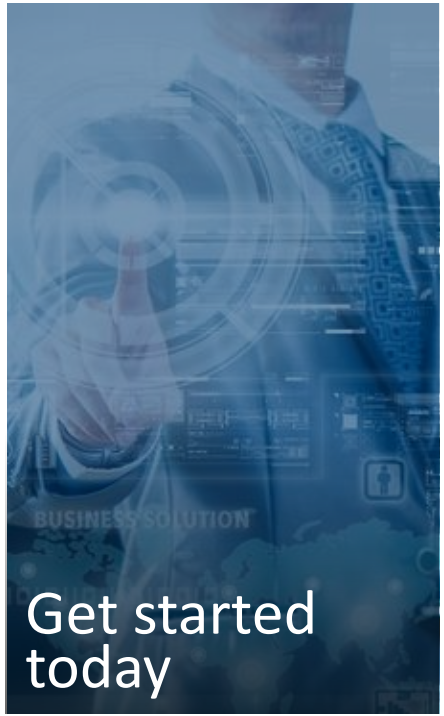
Partnering with #AI4GOOD to enrich the lives of every person on Earth through AI

[1]Source: https://newsroom.intel.com/editorials/intel-invests-1-billion-ai-ecosystem-fuel-adoption-product-innovation/

Business
Imperative

What is AI?

AI with Intel


BUSINESS SOLUTION
Get started
today

Intel® Xeon® processor Platform

Now Ready For Deep Learning Performance

INFERENCE THROUGHPUT

Up to **277x**[1]

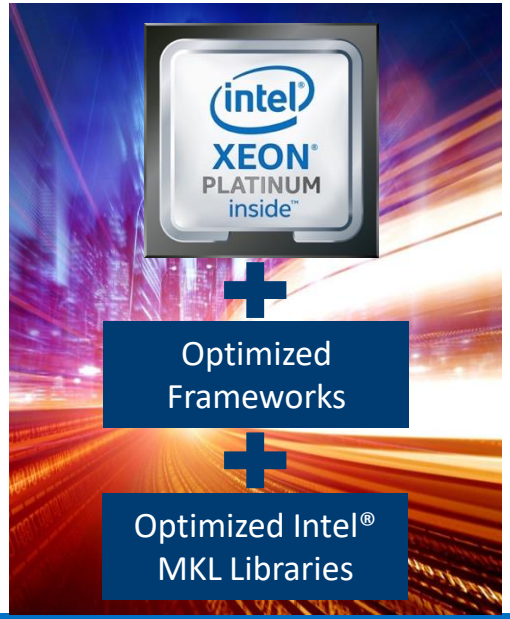
Intel® Xeon® Platinum 8180 Processor
higher Intel optimized Caffe GoogleNet v1 with Intel® MKL inference throughput compared to
Intel® Xeon® Processor E5-2699 v3 with BVLC-Caffe

Inference and training throughput uses FP32 instructions

TRAINING THROUGHPUT

Up to **241x**[1]

Intel® Xeon® Platinum 8180 Processor
higher Intel Optimized Caffe AlexNet with Intel® MKL training throughput compared to
Intel® Xeon® Processor E5-2699 v3 with BVLC-Caffe

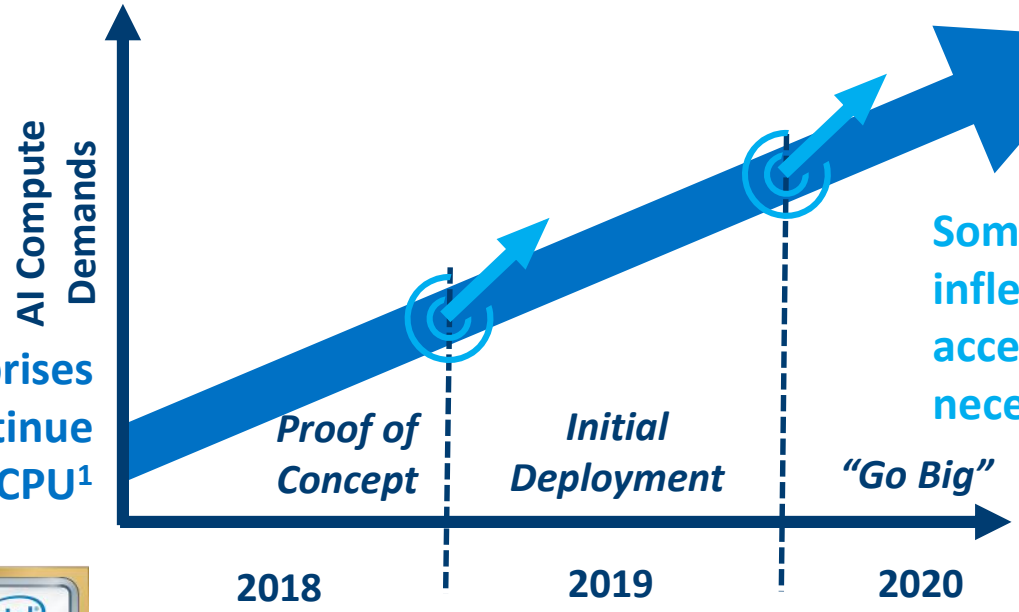


Deliver significant AI performance with hardware and software optimizations on Intel® Xeon® Scalable Family

[1] The benchmark results may need to be revised as additional testing is conducted. The results depend on the specific platform configurations and workloads utilized in the testing, and may not be applicable to any particular user's components, computer system or workloads. The results are not necessarily representative of other benchmarks and other benchmark results may show greater or lesser impact from mitigations. 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 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 Source: Intel measured as of June 2018. Configurations: See slide 4.

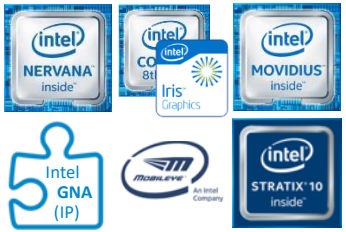
Build on your AI Foundation

Evaluate AI Acceleration Needs Moving Forward



Most enterprises are and will continue using the CPU[1]

Some enterprises reach an inflection point where acceleration becomes necessary



[1]Based on Intel survey of enterprise customers

Intel® Xeon® Processor Scalable Family

Now build the AI you want on the CPU you know

your
FOUNDATION
for AI

Get maximum utilization

running data center and AI workloads side-by-side

Break memory barriers

in order to apply AI to large data sets and models

Train complex models

through efficient scaling to many nodes

Access optimized tools

including continuous performance gains for TensorFlow, MXNet, more

Run in the cloud

including AWS, Microsoft, Alibaba, TenCent, Google, Baidu, more

Build on The premier AI portfolio

from multi-purpose to purpose-built and cloud to device

Intel® Nervana™ neural network processor (NNP)‡

Fastest time-to-train for intensive deep learning environments

purpose-built for deep learning

Blazingly-fast Data Access

Using high bandwidth memory and separate compute and data pipelines

High Degree of Parallelism

New numerical format (Bfloat16) for enhanced performance and conversion

New Levels of Scalability

Massive bi-directional data transfer through

‡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.

Deep learning frameworks

Popular DL Frameworks are now optimized for CPU

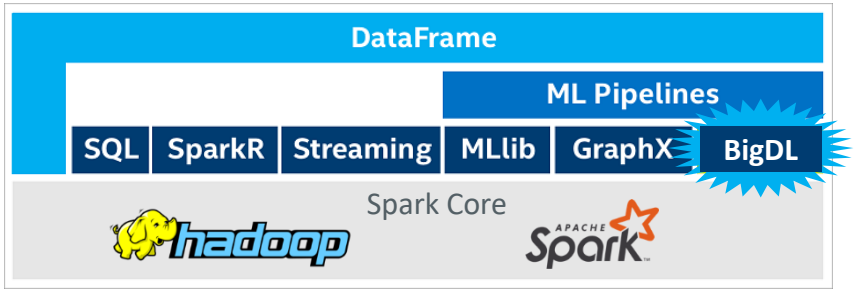
See installation guides at ai.intel.com/framework-optimizations/

More under optimization: Caffe2* PYTORCH* Microsoft CNTK* PaddlePaddle* *and more…*

SEE ALSO: Machine Learning Libraries for Python (Scikit-learn, Pandas, NumPy), R (Cart, randomForest, e1071), Distributed (MlLib on Spark, Mahout)
**Limited availability today*

BigDL

High Performance Deep Learning for FREE on CPU Infrastructure1



BigDL is a distributed deep learning library for Apache Spark* that can run directly on top of existing Spark or Apache Hadoop* clusters with direct access to stored data and tool/workflow consistency!

No need to deploy costly accelerators, duplicate data, or suffer through scaling headaches!

Feature Parity with Caffe* and Torch*

Lower TCO and improved ease of use with existing infrastructure

Deep Learning on Big Data Platform, Enabling **Efficient Scale-Out**

software.intel.com/bigdl

[1]*Open-source software is available for download at no cost; 'free' is also contingent upon running on existing idle CPU infrastructure where the operating cost is treated as a 'sunk' cost*

Intel® AI academy

For developers, students, instructors and startups

Get smarter using online tutorials, webinars, student kits and support forums

Educate others using available course materials, hands-on labs, and more

Get 4-weeks FREE access to the Intel® AI DevCloud or use your existing Intel® Xeon® Processor-based cluster

Showcase your innovation at industry & academic events and online via the Intel AI community forum

software.intel.com/ai

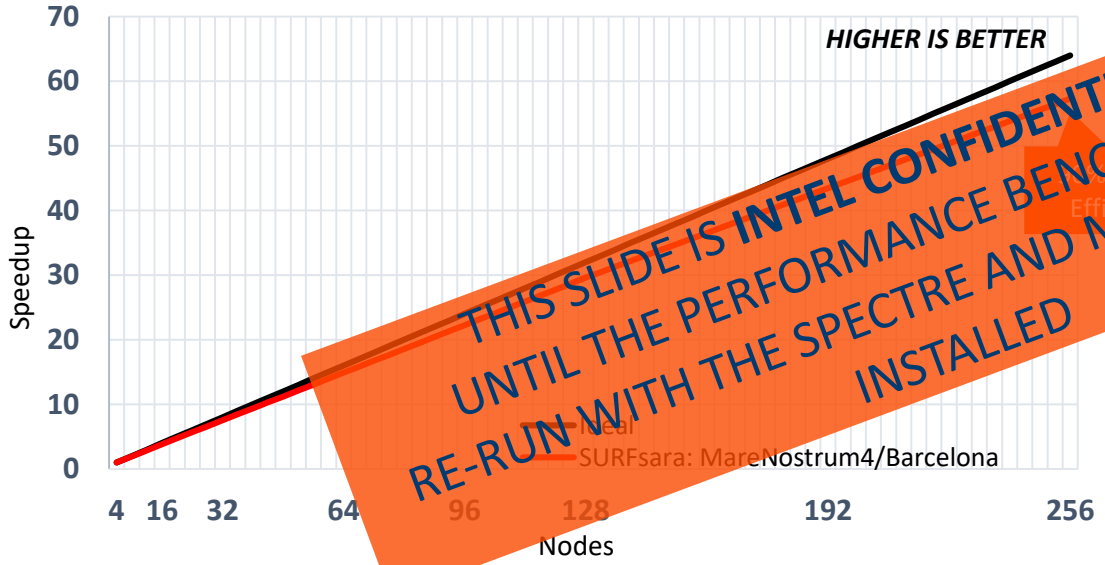
Learn more at

ai.intel.com

intel AI

Fast & Efficient DL scaling on CPU

Intel® - SURFsara* Research Collaboration - Multi-Node Intel® Caffe ResNet-50
Scaling Efficiency on 2S Intel® Xeon® Platinum 8160 Processor Cluster



- MareNostrum4 Barcelona Supercomputing Center
- ImageNet-1K
- 256 nodes
- 90% scaling efficiency
- Top-1/Top-5 > 74%/92%
- Batch size of 32 per node
- Global BS=8192
- Throughput: 15170 Images/sec

Time-To-Train: 70 minutes (50 Epochs)

90% scaling efficiency with up to 74% Top-1 accuracy on 256 nodes

Configuration Details: see end of presentation

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system. 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 Source: Intel measured as of June 2017

Intel® Stratix® 10 FPGA

Custom DL inference

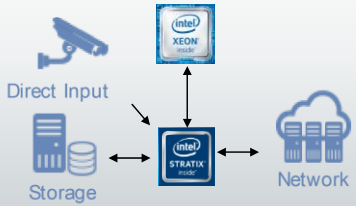
Scalable acceleration for deep learning inference in real-time with higher efficiency, and wide range of workloads & configurations

Efficient and low latency

- Up to **80%** reduction in power consumption (vs. Intel® Xeon® processor)[1]
- Deterministic low latency, for real-time inline processing of streaming data without buffering (as low as **<1ms**[2] latency)

flexibility

- Reconfigurable for a variety of configurations & fast workload switching



Future-ready

- Future proof for new neural network topologies, arbitrary precision data types (FloatP32 => FixedP2, sparsity, weight sharing), inline & offload processing

Note: available as discrete or Xeon with Integrated FPGA (Broadwell Proof of Concept)
[1] Configuration details on final slides
[2]Projected latency (GoogleNet, FP11, batch size=2, memory banks=4, 0.3 ms)
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 Source: Intel measured as of November 2016
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.
Notice Revision #20110804

Intel® Movidius™ Vision processing unit (VPU)

Power-Efficient Image Processing, Computer Vision & Deep Learning for Devices

SERVICE ROBOTS

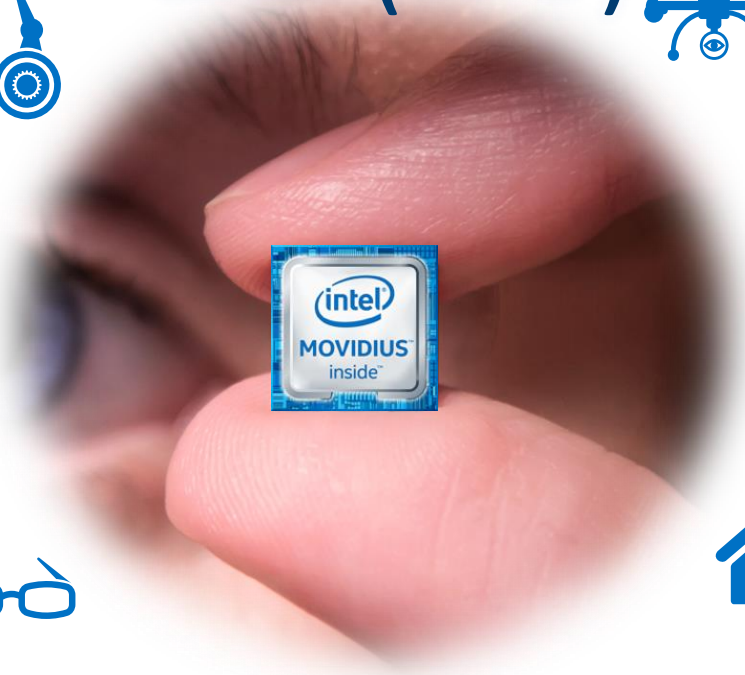
- Navigation
- 3D Vol. Mapping
- Multi-Modal Sensing

SURVEILLANCE

- Detection and Classification
- Identification
- Multi-Nodal Systems
- Multi-Modal Sensing
- Video, Image Capture

WEARABLES

- Detection, Tracking
- Recognition
- Video, Image, Session Capture



DRONES

- Sense & Avoid
- GPS Denied Hovering
- Pixel Labeling
- Video, Image Capture

AR-VR HMD

- 6DOF Pose, Position, mapping
- Gaze, Eye Tracking
- Gesture Tracking, Recognition
- See-Through Camera

SMART HOME

- Detection, Tracking
- Perimeter, Presence Monitoring
- Recognition, Classification
- Multi-Nodal Systems
- Multi-Modal Sensing
- Video, Image Capture

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

Intel® Gaussian neural accelerator (GNA)

Streaming Co-Processor for Low-Power Audio Inference & More

Ample throughput

For speech, language & other sensing inference

Low power

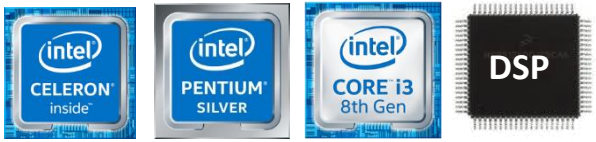
<100 mW power consumption for always-on applications

Flexibility

Gaussian Mixture Model (GMM) & Neural Network Inference support

Intel GNA (IP)

+



Try it TODAY!

Intel® Speech Enabling Developer Kit

https://software.intel.com/en-us/iot/speech-enabling-dev-kit

Learn more: https://sigport.org/sites/default/files/docs/PosterFinal.pdf

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

Intel integrated processor graphics

Built-in Deep Learning Inference Acceleration

Ubiquity/Scalability

- Shipped in > 1billion Intel SOCs
- Broad choice of performance/power offering across Intel® Atom™ , Intel® Core™ and Intel® Xeon™ processors

Powerful & Flexible Architecture

Rich data type support for 32bitFP, 16bitFP, 32bitInteger, 16bitInteger with SIMD multiply-accumulate instructions

Media Leadership

- Intel® Quick Synch Video – fixed function media blocks to improve power and performance
- Intel® Media SDK - API that provides access to hardware-accelerated codecs

Memory Architecture

Shared memory architecture on die between CPU and GPU to enable lower latency and power

Hardware integration

Software support

MacOS (CoreML and MPS[1])
Windows O/S (WinML)
OpenVINO™ Toolkit (Win, Linux)
clDNN

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

Visit www.mobileye.com

Intel distribution for python

Advancing Python* Performance Closer to Native Speeds

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

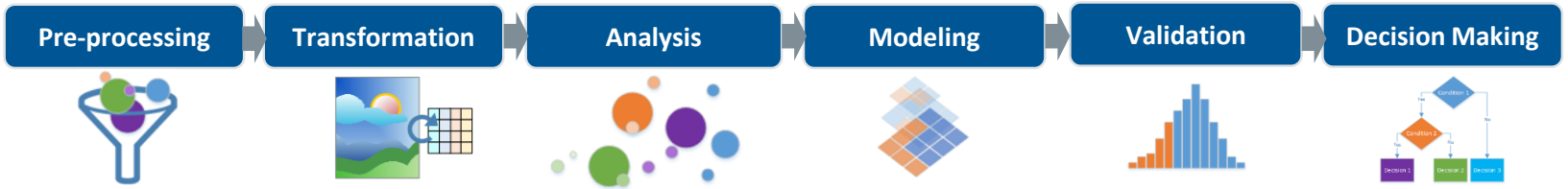
Faster Access to Latest Optimizations for Intel Architecture

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

Intel® Data Analytics Acceleration Library (Intel® DAAL)

High Performance Machine Learning and Data Analytics Library

Building blocks for all data analytics stages, including data preparation, data mining & machine learning



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.)

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

Intel® MKL-dnN

Intel's Open-Source Math Kernel Library for Deep Neural Networks

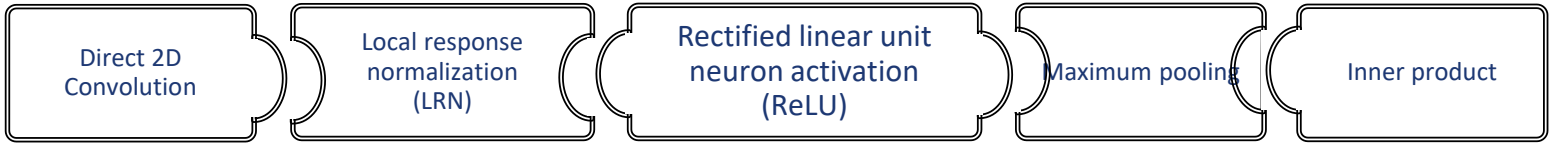
For developers of deep learning frameworks featuring optimized performance on Intel hardware

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:



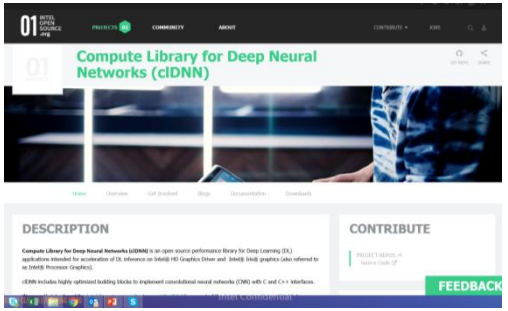
All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

Intel® clDnn

Compute Library for Deep Neural Networks on Intel Integrated Graphics

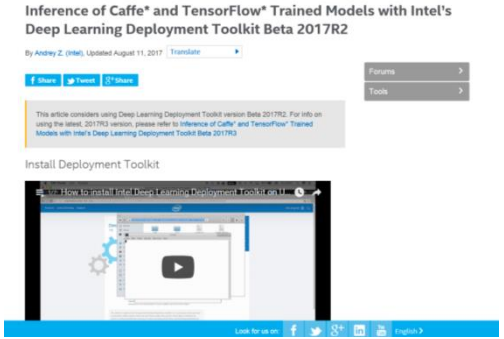
clDNN – Intel GPU DL acceleration middleware

- Open-sourced as of beginning May @ GitHub
- Official public page: https://01.org/cldnn
- Intel clDNN and Deep Learning Toolkit whitepaper
- DL Toolkit how to guide



Frameworks integration

- Part of Deep-Learning Deployment Toolkit
- Part of Intel® OpenVINO™ Toolkit

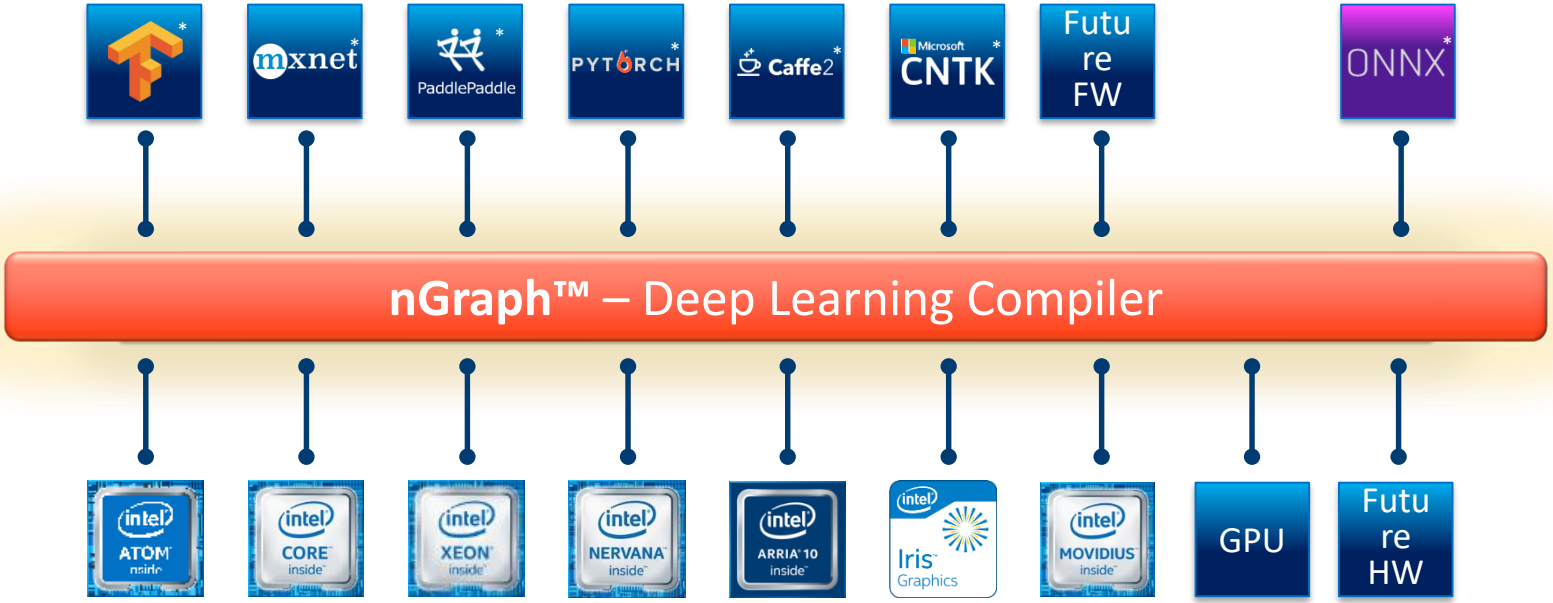


All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

Intel® ngraph™ compiler

AVAILABLE IN ALPHA

Open-source compiler enabling flexibility to run models across a variety of frameworks and hardware



*Other names and brands may be claimed as the property of others.
All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

Openvino™ toolkit

Cross-Platform Tool to Accelerate Computer Vision & Deep Learning Inference Performance

Intel® Deep Learning Deployment Toolkit

Model Optimizer
Convert & Optimize

IR

Inference Engine
Optimized Inference

Code Samples & 10 Pre-trained Models

IR = Intermediate Representation file

Traditional Computer Vision Tools & Libraries

Optimized Libraries

OpenCV* | **OpenVX*** | **Photography Vision**

Code Samples

For Intel® CPU & CPU with integrated graphics

Increase Media/Video/Graphics Performance

Intel® Media SDK
Open Source version

OpenCL™ Drivers & Runtimes

For CPU with integrated graphics

Optimize Intel® FPGA

FPGA RunTime Environment
(from Intel® FPGA SDK for OpenCL™)

Bitstreams

FPGA – Linux* only

Intel® Architecture-Based Platforms Support

CELERON inside | ATOM inside | CORE inside | XEON inside | ARRIA 10 inside | MOVIDIUS inside | IRIS Pro GRAPHICS

software.intel.com/openvino-toolkit

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

OS Support CentOS* 7.4 (64 bit) Ubuntu* 16.04.3 LTS (64 bit) Microsoft Windows* 10 (64 bit) Yocto Project* version Poky Jethro v2.0.3 (64 bit)

Intel® Deep Learning Deployment Toolkit (DLDT)

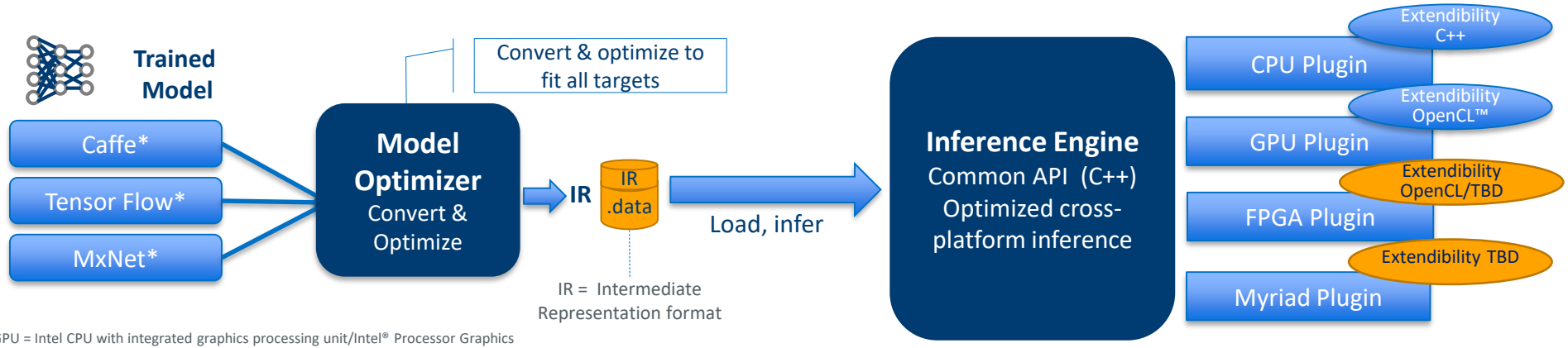
Take Full Advantage of the Power of Intel® Architecture for Deep Learning

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.



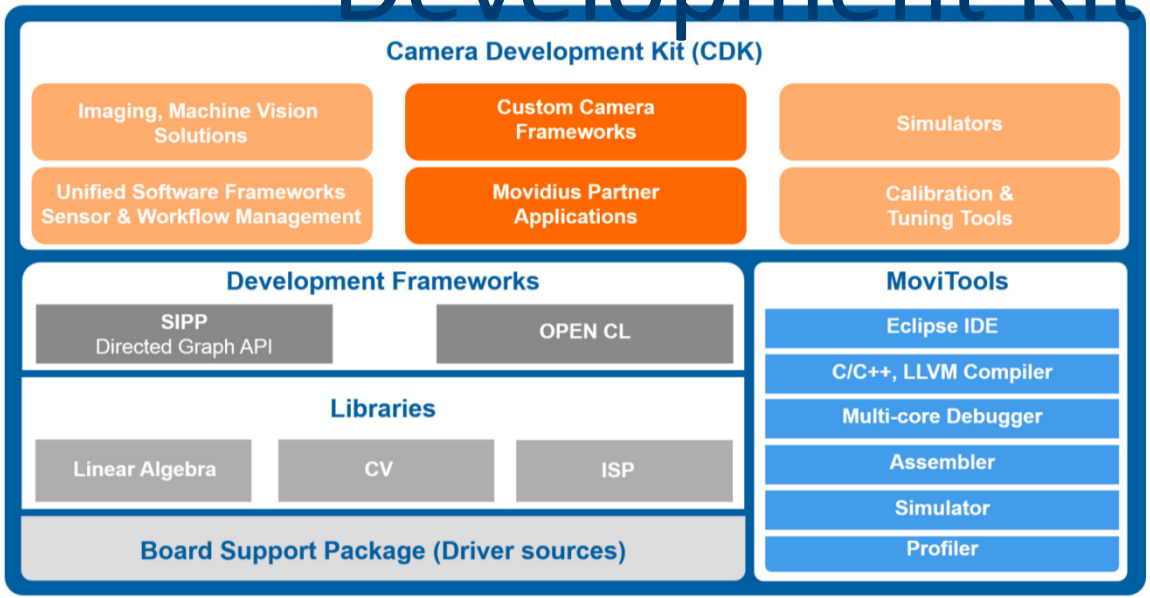
GPU = Intel CPU with integrated graphics processing unit/Intel® Processor Graphics

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

Intel® Movidius™ Software

Comprehensive Software Development Suite for Video Processing Units (VPUs)

Development Kit (SDK)



rapid prototyping

Remove time and complexity using built-in directed graph framework

rapid prototyping

Remove time and complexity using built-in directed graph framework

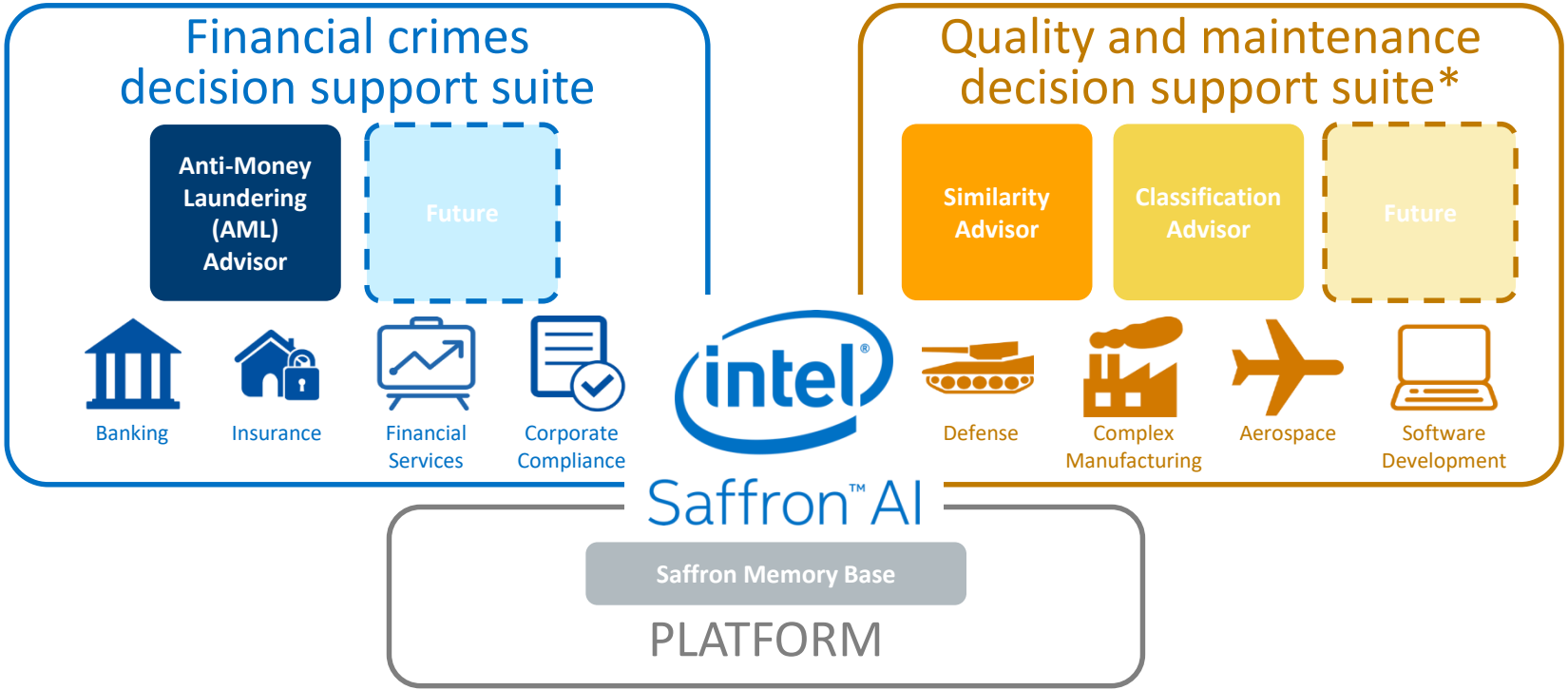
Flexible development

From C/C++ to graphical development with rich tool suite

Learn more: https://uploads.movidius.com/1463156704-2016-04-29_MDK_ProductBrief.pdf

Intel® saffron™ ai

Cognitive Reasoning Solutions for Complex Real-World Challenges



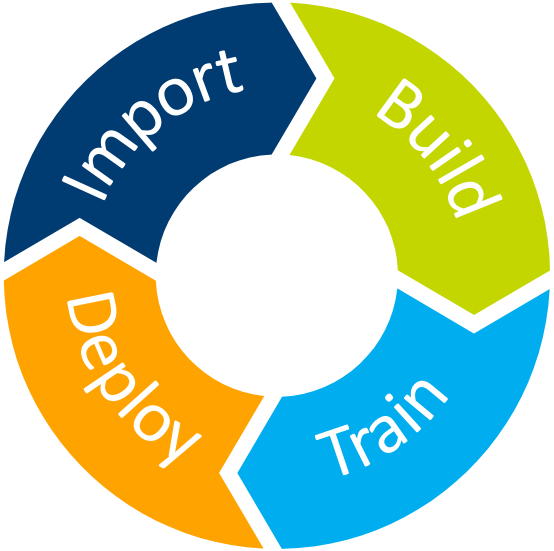
All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

Intel® Deep learning studio

Coming Soon

Compress the Development Cycle to Accelerate Time-to-Solution

Data curation/processing
Data partitioning
Data labeling



Multi-user collaboration
Interactive sessions
Model library

Fast training
Batch training
Experiment tracking
Multi-node distribution
Analytics & visualization
Hyperparameter optimization

Batch inference
Model compression
Inference deployment
Export to edge devices

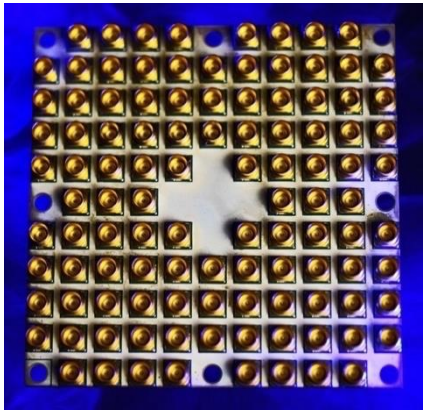
Coming to the Intel® Deep Learning System

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

Leading AI research

Choose a partner on the cutting-edge of AI breakthroughs

Neuromorphic Computing Test Chip Codenamed "Loihi"



Quantum Computing 49-Qubit Test Chip Codenamed "Tangle-Lake"

NEW AI Technologies @Intel Labs

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

Configuration details

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer. No computer system 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. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

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.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

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, the Intel logo, Intel inside, the Intel inside logo, Xeon, the Xeon logo, Xeon Phi, the Xeon Phi logo, Core, the Core logo, Atom, the Atom logo, Movidius, the Movidius logo, Stratix, the Stratix logo, Arria, the Arria logo, Myriad, Nervana and others are trademarks of Intel Corporation in the U.S. and/or other countries.
*Other names and brands may be claimed as the property of others.

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. Notice Revision #20110804.

Results have been estimated or simulated 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.

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

Configuration details (Cont'd)

Configuration: AI Performance – Software + Hardware

INFERENCE using FP32 Batch Size Caffe GoogleNet v1 128 AlexNet 256.

The benchmark results may need to be revised as additional testing is conducted. The results depend on the specific platform configurations and workloads utilized in the testing, and may not be applicable to any particular user's components, computer system or workloads. The results are not necessarily representative of other benchmarks and other benchmark results may show greater or lesser impact from mitigations. 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 Source: Intel measured as of June 2017 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.

Configurations for Inference throughput

Platform :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.28GB (12slots / 32 GB / 2666 MHz),4 instances of the framework, CentOS Linux-7.3.1611-Core , SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework caffe version: a3d5b022fe026e9092fc7abc7654b1162ab9940d Topology:GoogleNet v1 BIOS:SE5C620.86B.00.01.0004.071220170215 MKLDNN: version: 464c268e544bae26f9b85a2acb9122c766a4c396 NoDataLayer. Measured: 1449.9 imgs/sec vs Platform: 2S Intel® Xeon® CPU E5-2699 v3 @ 2.30GHz (18 cores), HT enabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 64GB DDR4-2133 ECC RAM. BIOS: SE5C610.86B.01.01.0024.021320181901, CentOS Linux-7.5.1804(Core) kernel 3.10.0-862.3.2.el7.x86_64, SSD sdb INTEL SSDSC2BW24 SSD 223.6GB. Framework BVLC-Caffe: https://github.com/BVLC/caffe, Inference & Training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. BVLC Caffe (http://github.com/BVLC/caffe), revision 2a1c552b66f026c7508d390b526f2495ed3be594

Configuration for training throughput:

Platform :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.28GB (12slots / 32 GB / 2666 MHz),4 instances of the framework, CentOS Linux-7.3.1611-Core , SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework caffe version: a3d5b022fe026e9092fc7abc765b1162ab9940d Topology:alexnet BIOS:SE5C620.86B.00.01.0004.071220170215 MKLDNN: version: 464c268e544bae26f9b85a2acb9122c766a4c396 NoDataLayer. Measured: 1257 imgs/sec vs Platform: 2S Intel® Xeon® CPU E5-2699 v3 @ 2.30GHz (18 cores), HT enabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 64GB DDR4-2133 ECC RAM. BIOS: SE5C610.86B.01.01.0024.021320181901, CentOS Linux-7.5.1804(Core) kernel 3.10.0-862.3.2.el7.x86_64, SSD sdb INTEL SSDSC2BW24 SSD 223.6GB. Framework BVLC-Caffe: https://github.com/BVLC/caffe, Inference & Training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. BVLC Caffe (http://github.com/BVLC/caffe), revision 2a1c552b66f026c7508d390b526f2495ed3be594

Configuration details (cont'd)

Intel® and SURFsara* Research Collaboration MareNostrum4/BSC* Configuration Details

***MareNostrum4/Barcelona Supercomputing Center**: https://www.bsc.es/

Compute Nodes: 2 sockets Intel® Xeon® Platinum 8160 CPU with 24 cores each @ 2.10GHz for a total of 48 cores per node, 2 Threads per core; 32K; L1i cache 32K; L2 cache 1024K; L3 cache 33792K, 96 GB of DDR4, Intel® Omni-Path Host Fabric Interface, dual-rail. Software: Intel® MPI Library 2017 Update 4Intel® MPI Library 2019 Technical Preview Gold 1.5.0PSM2 w/ Multi-EP, 10 Gbit Ethernet, 200 GB local SSD, Red Hat* Enterprise Linux 6.7.

Intel® Caffe: Intel® version of Caffe; http://github.com/intel/caffe/, revision 8012927bf2bf70231cbc7ff55de0b1b [illegible]
Intel® MKL version: mklml_lnx_2018.0.20170425; Intel® MLSL version: l_mlsl_2017.1.016

Model: Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models [illegible] (ResNet-50) modified for 1K classes. ResNet-50. Batch size as stated in the performance chart

Time-To-Train: measured using "train" command. Data copied to memory on all nodes in the cluster before training. No input image data transferred over the fabric while training;

Performance measured with:
export OMP_NUM_THREADS=44 (the remaining 4 cores are used for driving communication), export I_MPI_FABRICS=tmi, export I_MPI_TMI_PROVIDER=psm2

OMP_NUM_THREADS=44 KMP_AFFINITY="proclist=[0-87:granularity=thread,exclusive" KMP_HW_SUBSET=1t MLSL_NUM_SERVERS=4 mpiexec.hydra -PSM2 -l -n $SLURM_JOB_NUMBER_NODES -ppn 1 -f hosts2 -genv OMP_NUM_THREADS 44 -env KMP_AFFINITY "proclist=[0-87:granularity=thread,exclusive" -env KMP_HW_SUBSET 1t -genv I_MPI_FABRICS tmi -genv I_MPI_HYDRA_BRANCH_COUNT $SLURM_JOB_NUM_NODES -genv I_MPI_HYDRA_PMI_CONNECT alltoall sh -c 'cat /ilsvrc12_train_lmdb_striped_64/data.mdb > /dev/null ; cat /ilsvrc12_val_lmdb_striped_64/data.mdb > /dev/null ; ulimit -u 8192 ; ulimit -a ; numactl -H ; /caffe/build/tools/caffe train --solver=/caffe/models/intel_optimized_models/multinode/resnet_50_256_nodes_8k_batch/solver_poly_quick_large.prototxt -engine "MKL2017"

SURFsara blog: https://blog.surf.nl/en/imagenet-1k-training-on-intel-xeon-in-less-than-40-minutes/ ; Researchers: Valeriu Codreanu, Ph.D. (PI).; Damian Podareanu, MSc; SURFsara* & Vikram Saletore, Ph.D. (co-PI): Intel Corp.

*SURFsara B.V. is the Dutch national high-performance computing and e-Science support center. Amsterdam Science Park, Amsterdam, The Netherlands.

Configuration details (cont’d)

Intel® Arria 10 – 1150 FPGA energy efficiency on Caffe/AlexNet up to 25 img/s/w with FP16 at 297MHz

Vanilla AlexNet Classification Implementation as specified by http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf, Training Parameters taken from Caffe open-source Framework are 224x224x3 Input, 1000x1 Output, FP16 with Shared Block-Exponents, All compute layers (incl. Fully Connected) done on the FPGA except for Softmax, Arria 10-1150 FPGA, -1 Speed Grade on Altera PCIe DevKit with x72 DDR4 @ 1333 MHz, Power measured through on-board power monitor (FPGA POWER ONLY), ACDS 16.1 Internal Builds + OpenCL SDK 16.1 Internal Build, Compute machine is an HP Z620 Workstation, Xeon E5-1660 at 3.3 GHz with 32GB RAM. The Xeon is not used for compute.