An introduction to AI with Intel

Dr. Séverine Habert

Thursday, May 4th 2023

intel. lrz

Agenda

- A history of AI
- Concepts of Classical Machine Learning and Deep Learning
- Where classical ML shines
- Overview of Intel Hardware for AI
- Overview of Intel Software for AI

A history of AI

A definition of AI

- Concept:

Human thinking has the ability to both be replicated and mechanized.

Before AI

The history of Artificial Intelligence follows the history of computer

- **create a mechanical object capable of thinking like human.**

Dates back from the antiquity / Ancient Greece where:

- was first developed automatons
- was first theorized formal reasoning
- was fantasized more sophisticated artificial beings in myths and legends

Antikythera mechanism, oldest known example of an analogue computer used to predict astronomical positions and eclipses decades in advance

Talos, bronze automate in Greek mythology（coin dating from 300/280-270 BCE

Forward in time

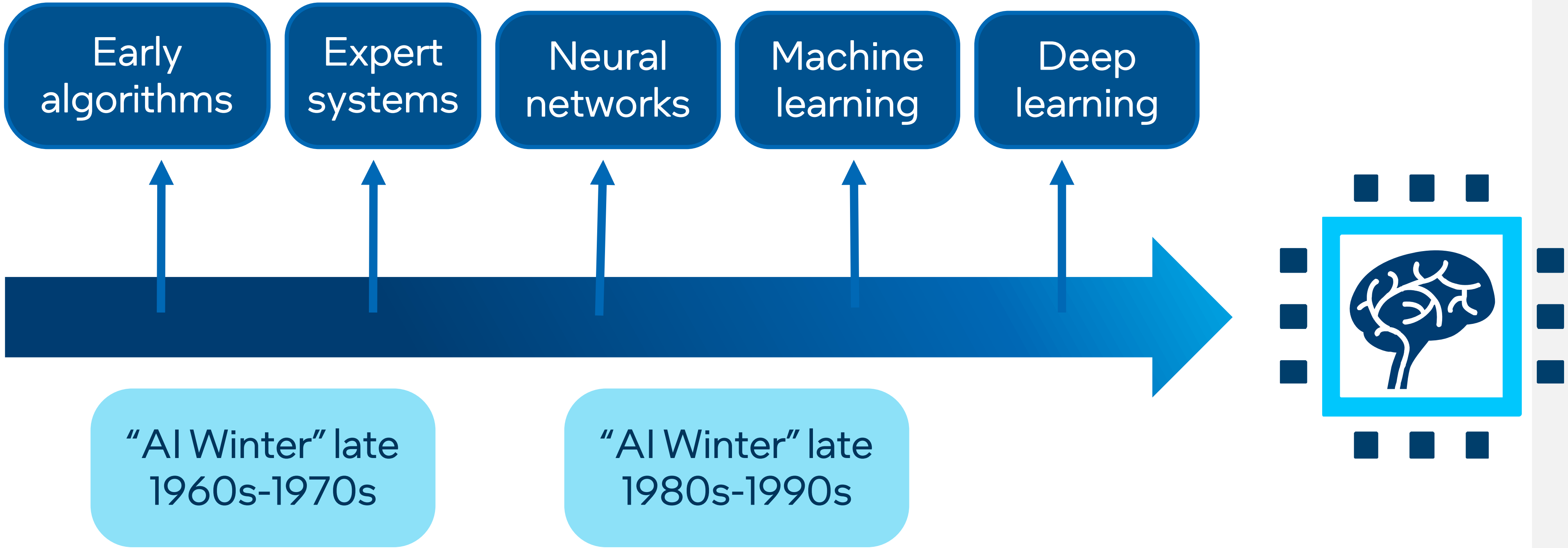
- Automata can be found through the history, Egypt, China, Renaissance

- Formal reasoning was also theorized by philosophers in various cultures : Al-Khwārizmī (algebra and algorithm), Descartes, Boole (mathematical logic)

- While “robots” start to appear in early 1900 in literature.

History in the XXth century

AI has experienced several hype cycles, where it has oscillated between periods of excitement and disappointment.



Early algorithms

Expert systems

Neural networks

Machine learning

Deep learning

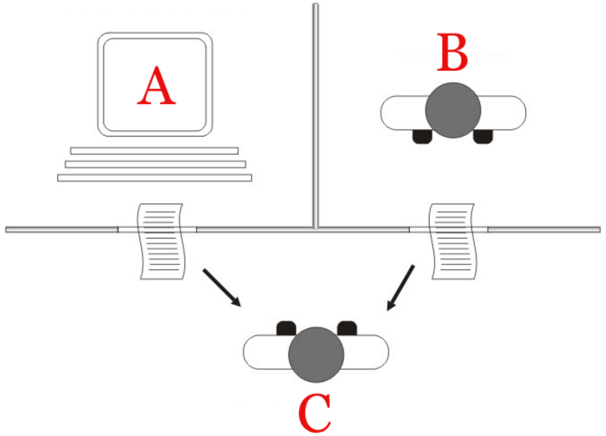
“AI Winter” late 1960s-1970s

“AI Winter” late 1980s-1990s

Cybernetics and Turing test

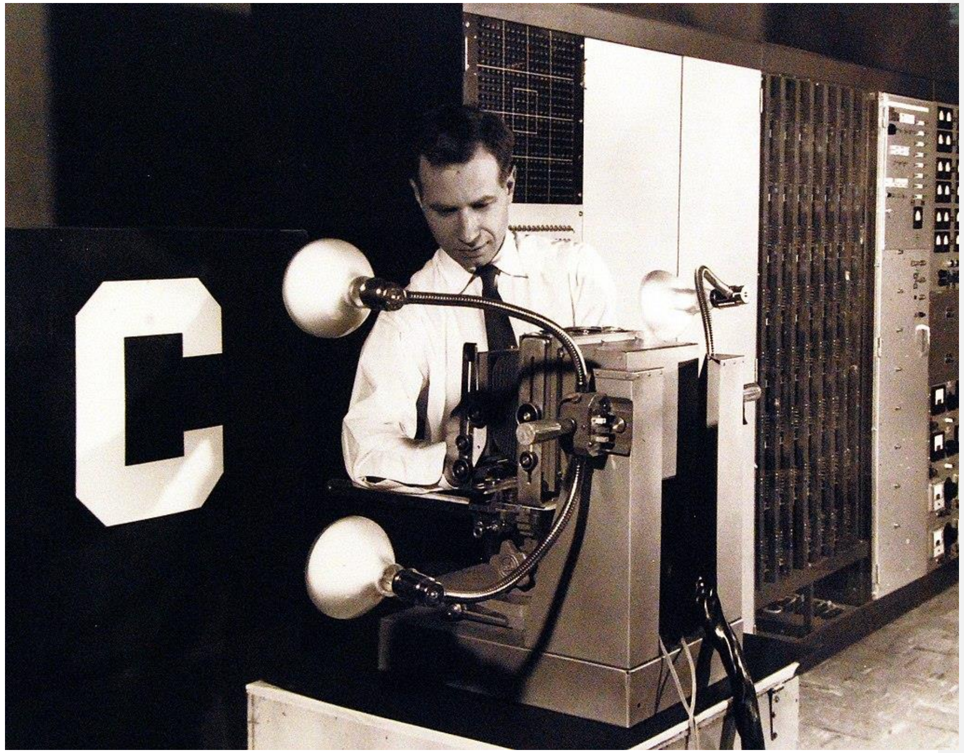
- **Very early prototype of neural networks**
- Early 1940, better understanding of the human brain as electrical network of neurons
- In 1943, network of artificial neurons - perceptron.

- **Turing test:**
- In 1950, Turing defines the Turing test to determine if a machine is intelligent
  - can it fool a human with its answers?



1950s: birth of AI

- 1956: Artificial Intelligence termed was coined at the Dartmouth Conference.
- 1957: Frank Rosenblatt implemented the perceptron algorithm.



Camera system of the Mark 1 Perceptron

The First “AI Winter”

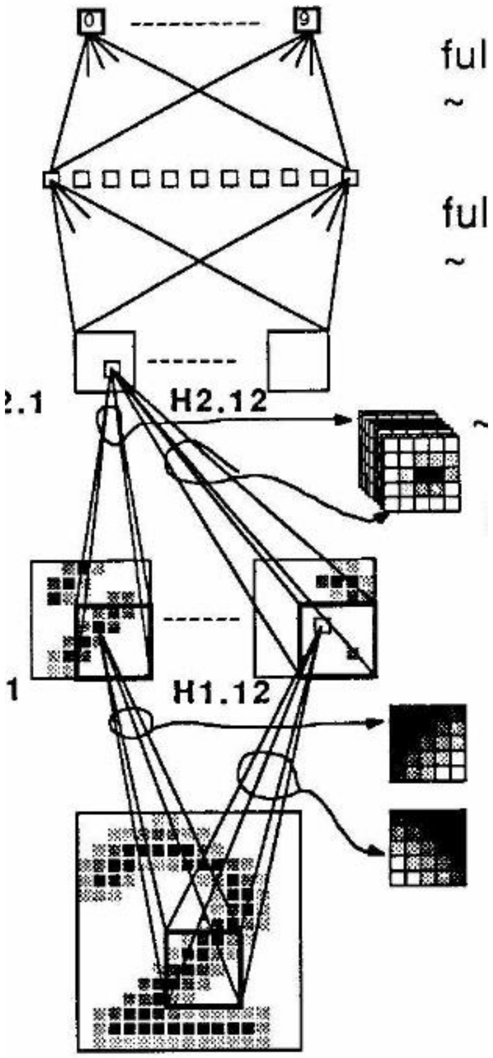
- 1969: limitations of the Perceptron algorithm
  - Can not represent non-linear functions

- 1966 & 1973: Two US government backed committee highlights AI’s failure to live up to promises.

- No more funding -> the first “AI Winter.”

1980's AI Boom

- Expert Systems become the first AI algorithm to be widely used in enterprise
- 1986: "Backpropogation" algorithm is able to train multi-layer perceptrons
- 1989: first theorical demonstration of backpropagation on Convolutional Neural Networks by Yann LeCun

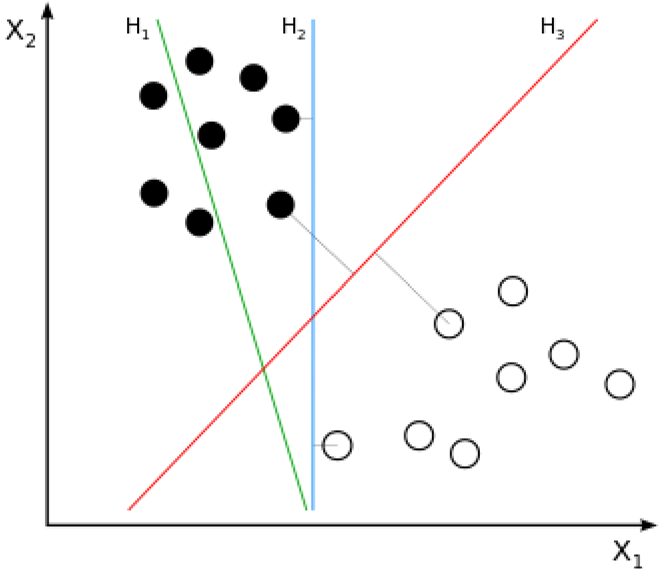


Another AI Winter (late 1980’s – early 1990s)

- Expert systems’ progress on solving business problems slowed.
- Neural networks didn’t scale to large problems.
- Interest in AI in business declined.

Late 1990's to early 2000's: Classical Machine Learning

- Advancements in the Support Vector Machine algorithm led to it becoming the machine learning method of choice.



- The Deep Blue chess expert system beat world chess champion Garry Kasparov.

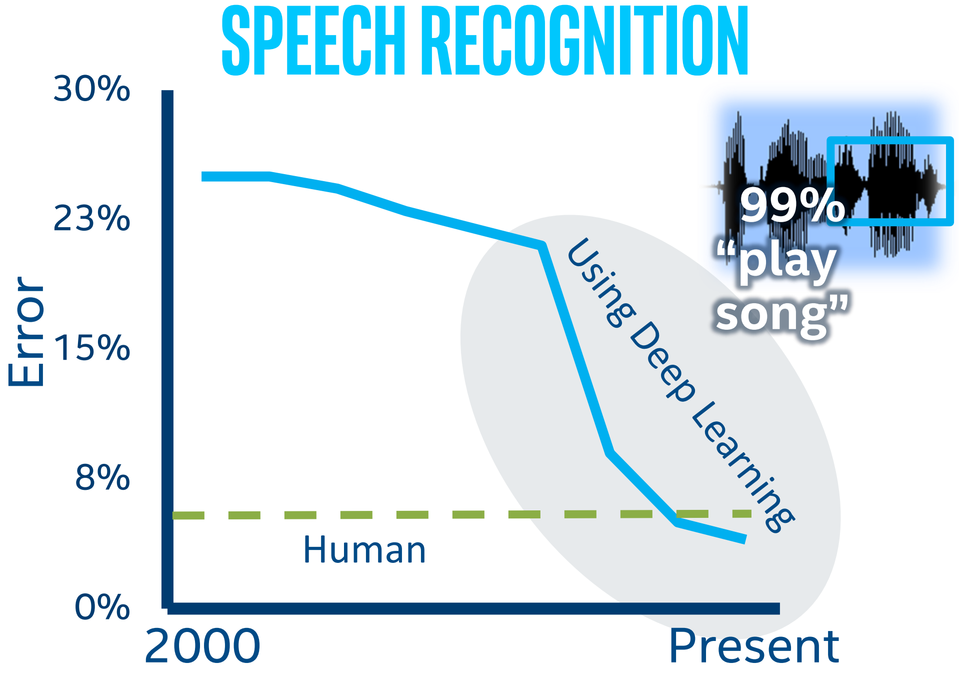
2000’s: Rise of Deep Learning

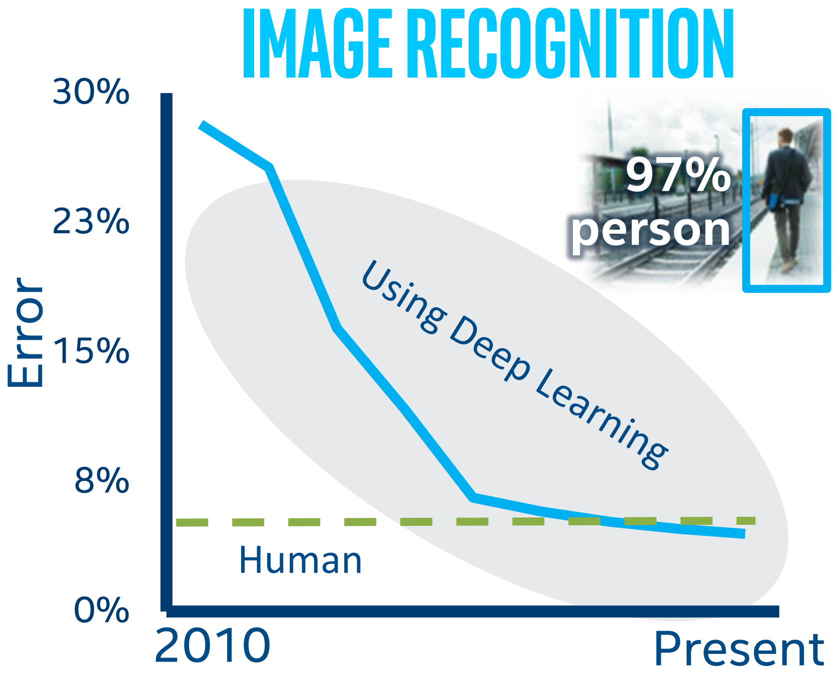
- Model architectures already exist since the 80-90’s such as CNN and RNN
- New architecture appears such as LSTM
- Bigger dataset and more powerful machines provide the essential tools for Deep Learning

- 2004: GPU implementation of Neural Networks
- 2009: The ImageNet database of human-tagged images is presented at the CVPR conference.

Deep Learning – Breakthroughs (2012 – Present)

- 2011-2012: DanNet and AlexNet starts to win challenges in CV
- 2015: Deep learning platform TensorFlow is developed.



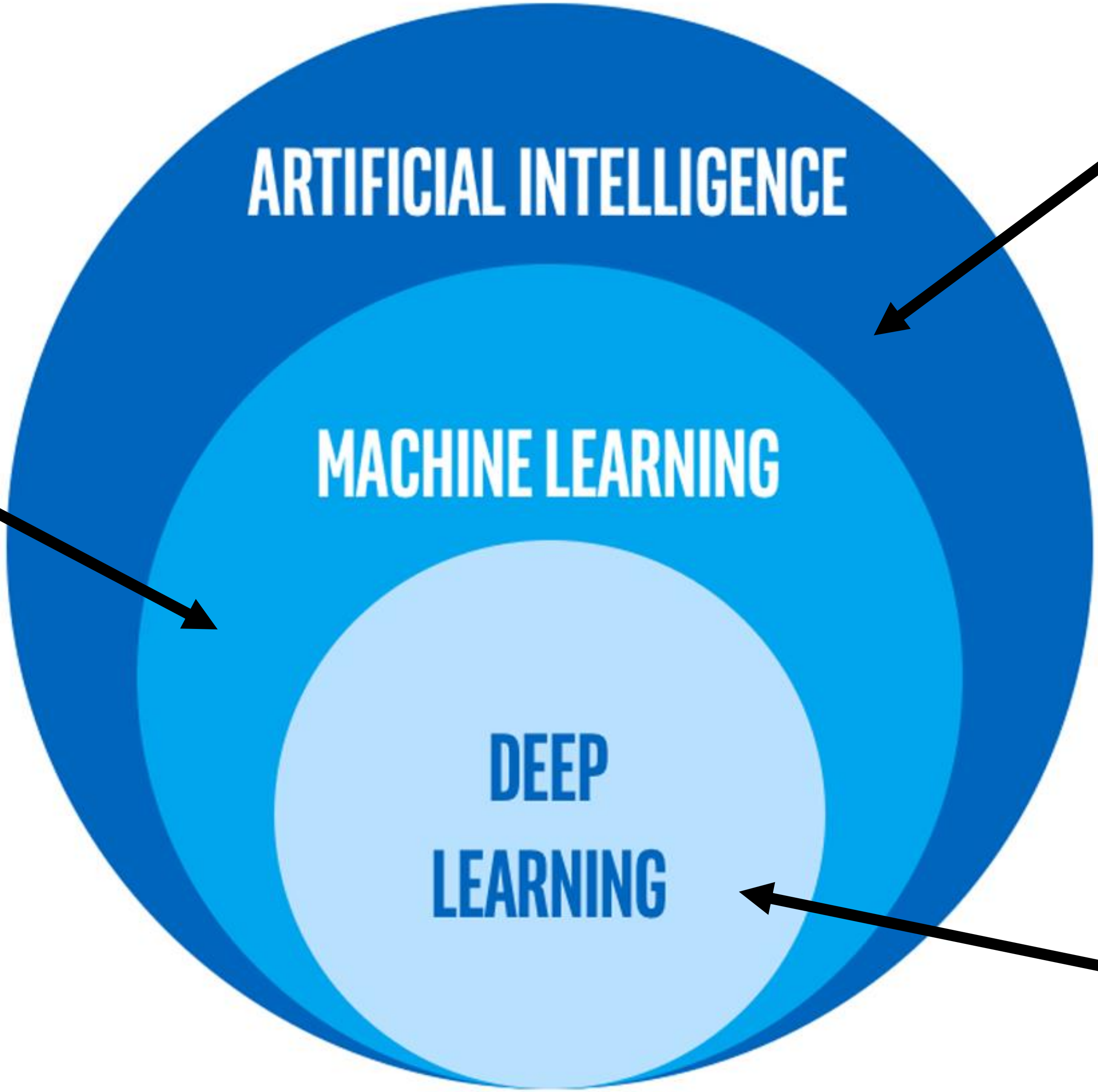


Latest developments

- Deep Learning underwent tremendous developments in the last 5 years, largely driven by transformers
- Used in every vertical: healthcare, manufacturing, agriculture, surveillance, …
- In 2022, ChatGPT and Stable Diffusion brought the attention of the public to AI

We are in the middle of an AI Summer !

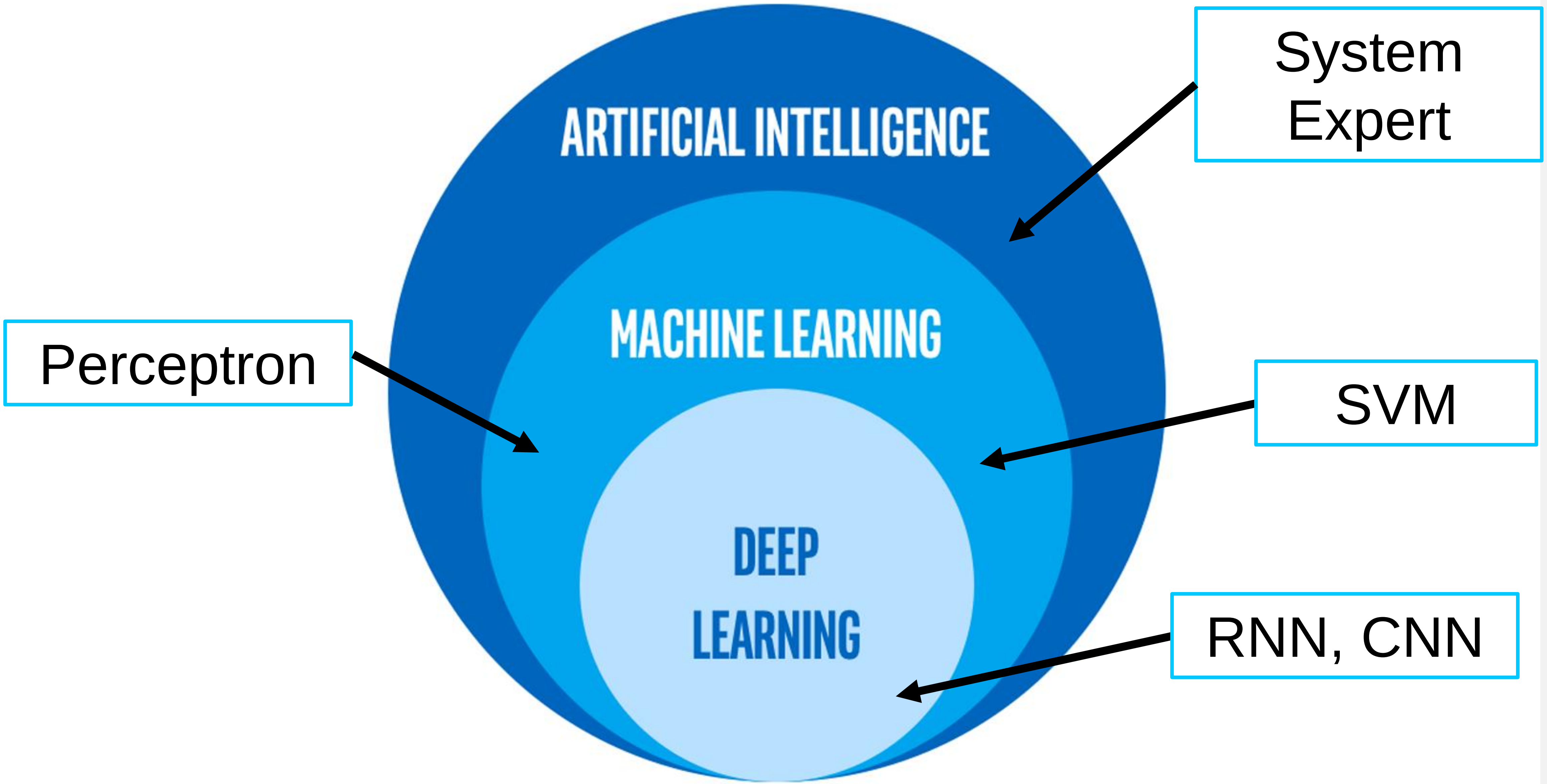
Definitions of AI concepts



System that leverage human knowledge for learning

System to perform a specific task without rule-based instruction

No more feature engineering, system learn hidden patterns



System Expert

Perceptron

SVM

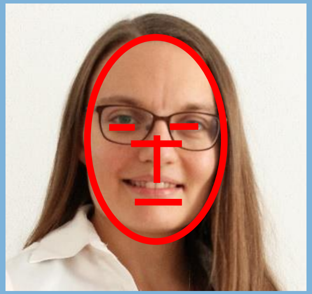
RNN, CNN

MACHINE LEARNING AND DEEP LEARNING

CLASSICAL MACHINE LEARNING

How do you engineer the best features?

$N \times N$



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

Roundness of face
Distance 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

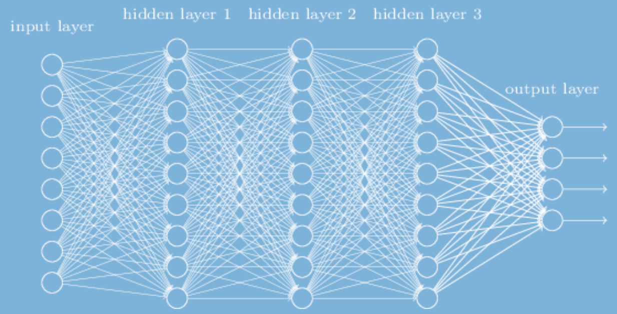
Séverine

DEEP LEARNING

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

$N \times N$

NEURAL NETWORK



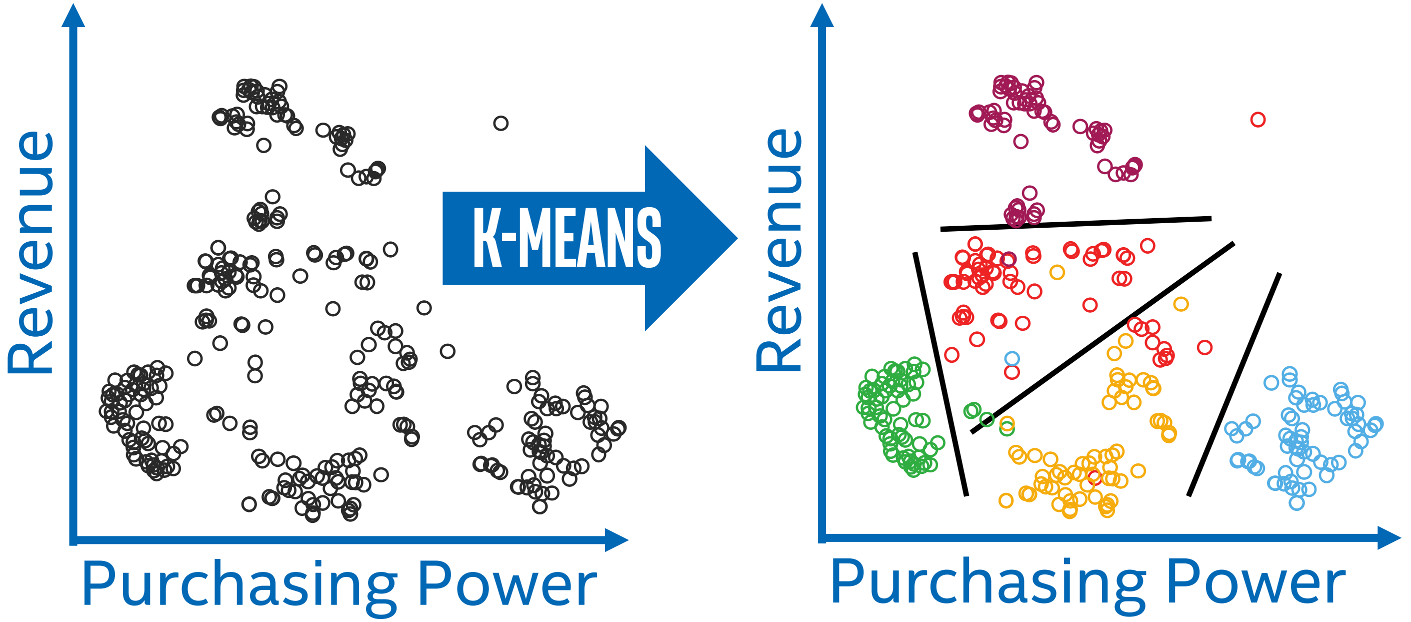
Séverine

Two Main Types of Machine Learning

	Goal	Common algorithms	Dataset
Supervised Learning	Make predictions	Linear regression, decision tree, NN	Labelled
Unsupervised Learning	Find structure in the data	Clustering, kmeans, NN in contrastive learning	Unlabelled

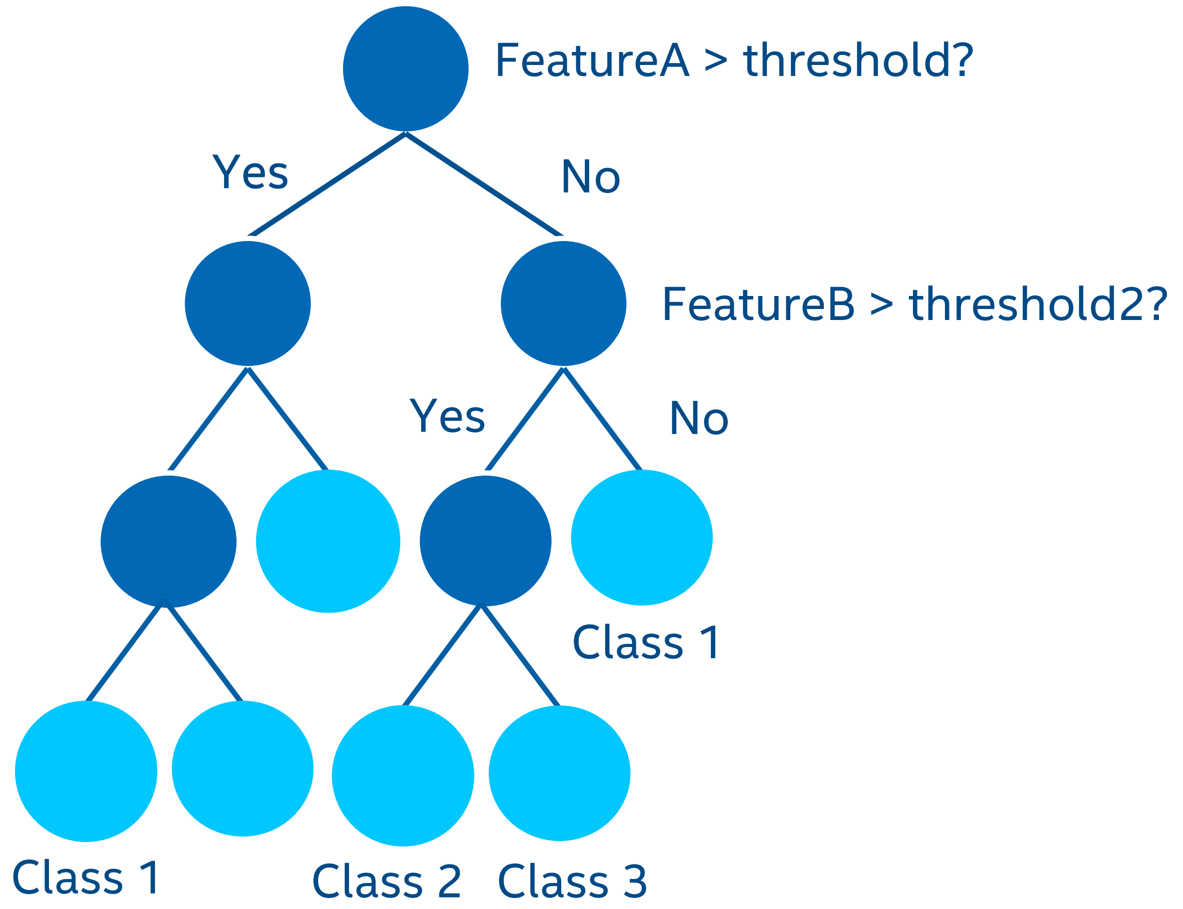
Classic ML example - Clustering

An 'Unsupervised Learning' Example



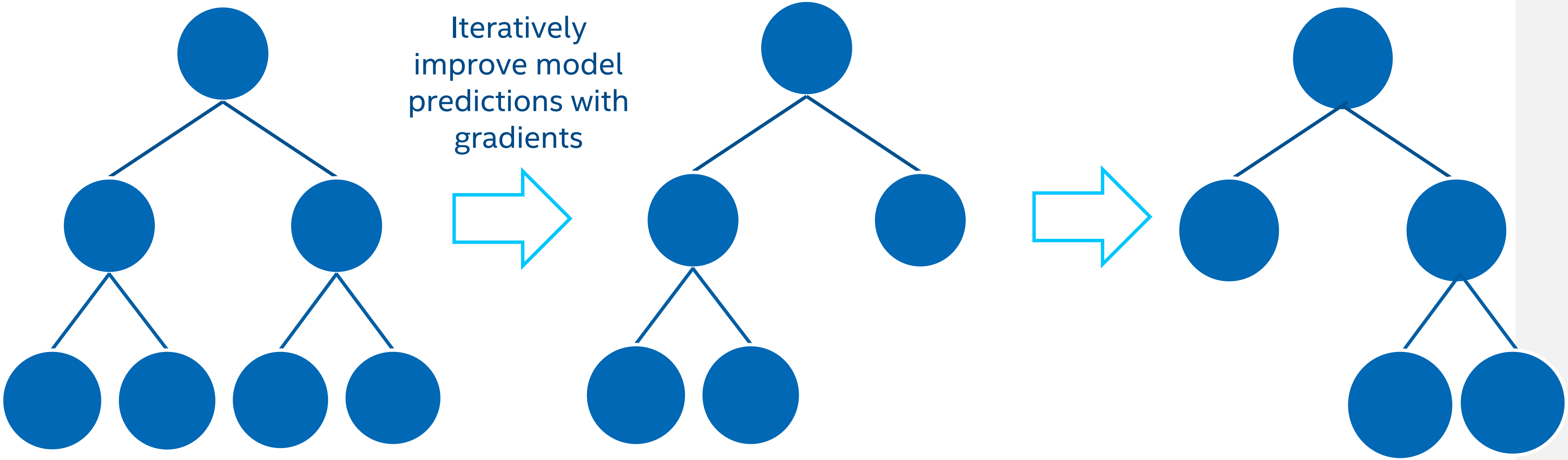
Classic ML example – Decision tree

A 'supervised Learning' Example

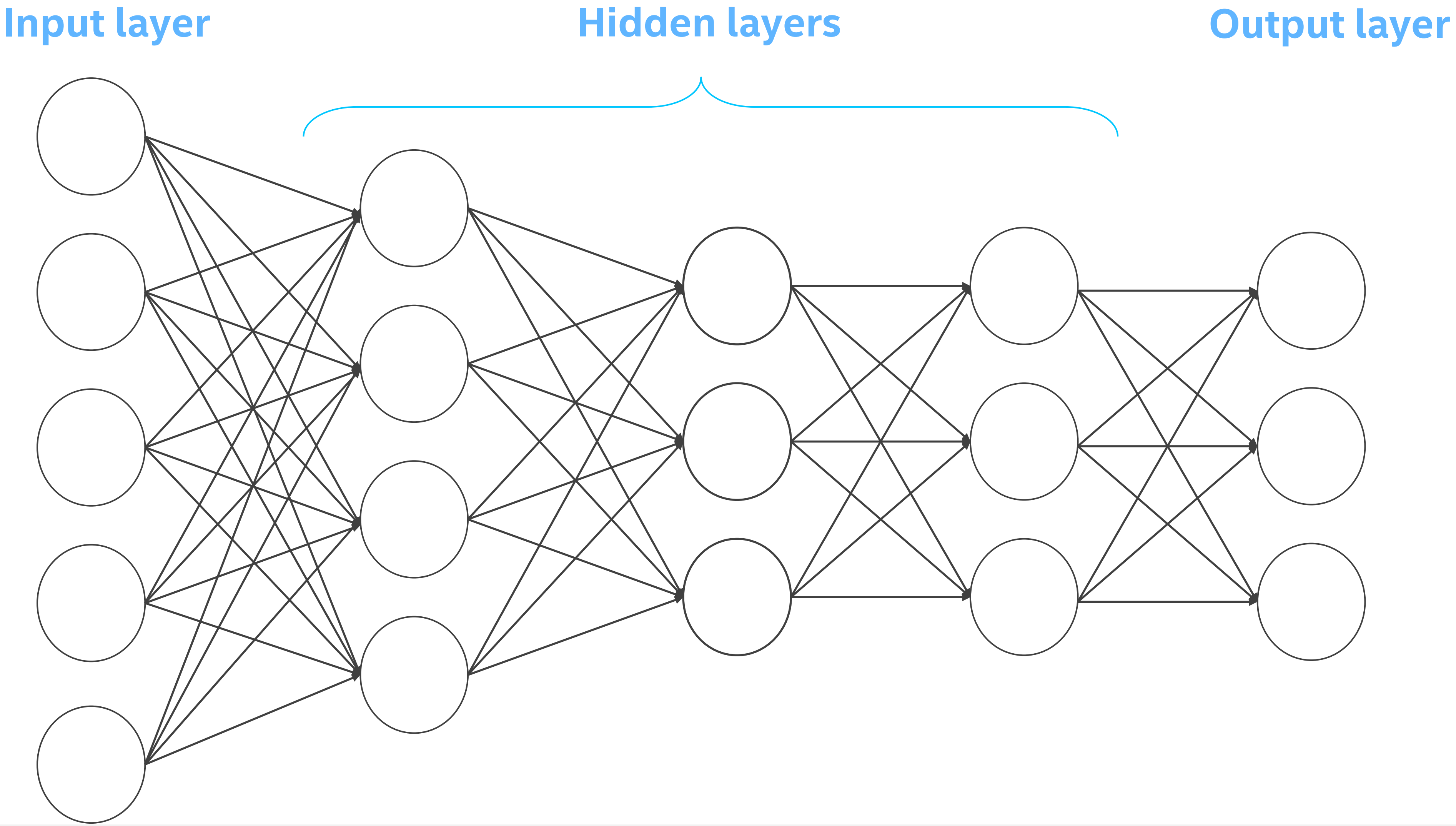


Classic ML example – Gradient Boosting

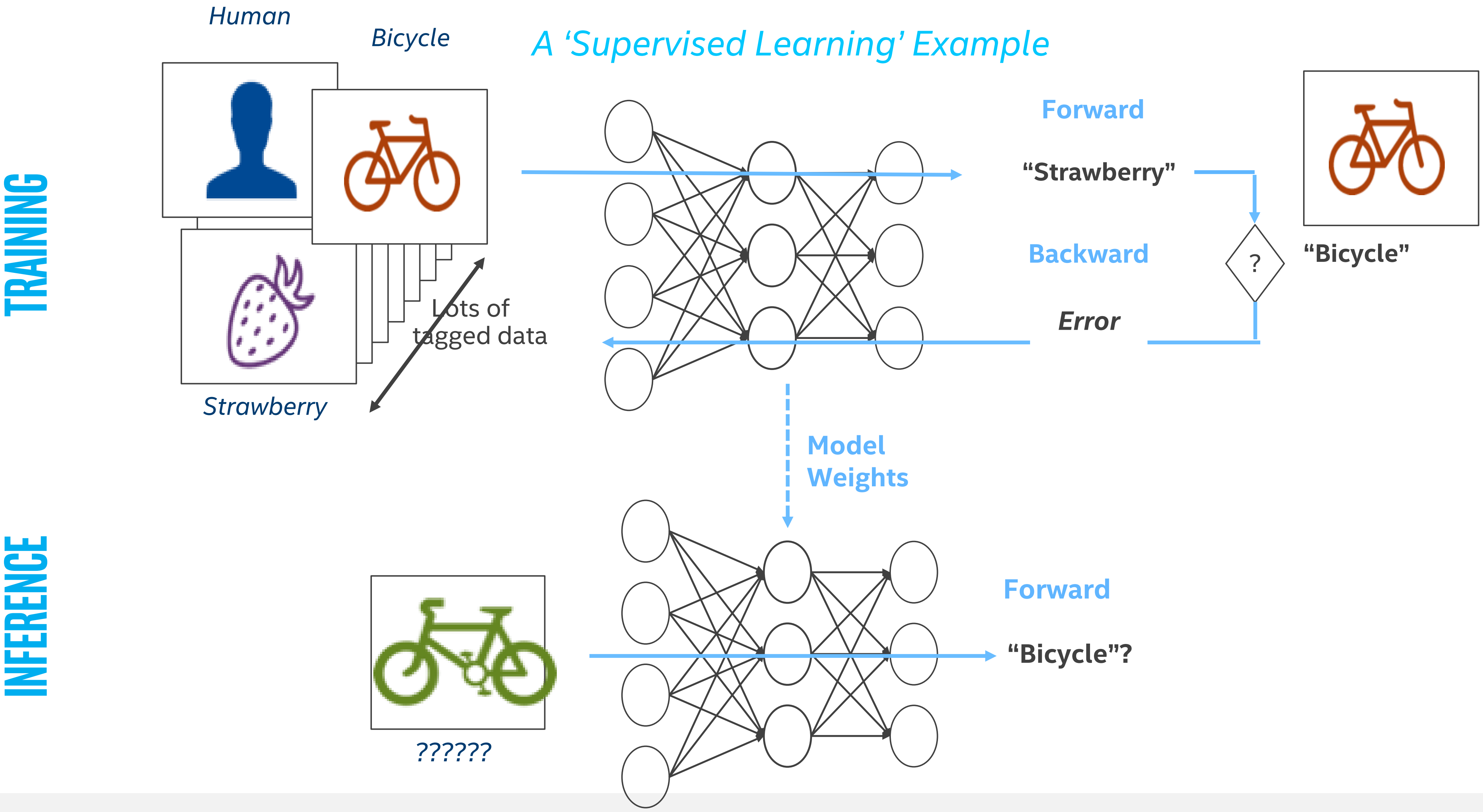
An ‘supervised Learning’ Example



Deep Learning – Neural Network



Deep Learning – Training + Inference



Deep Learning - typical domains

COMPUTER VISION

Classification

Object Detection

Instance or Semantic segmentation

NATURAL LANGUAGE PROCESSING

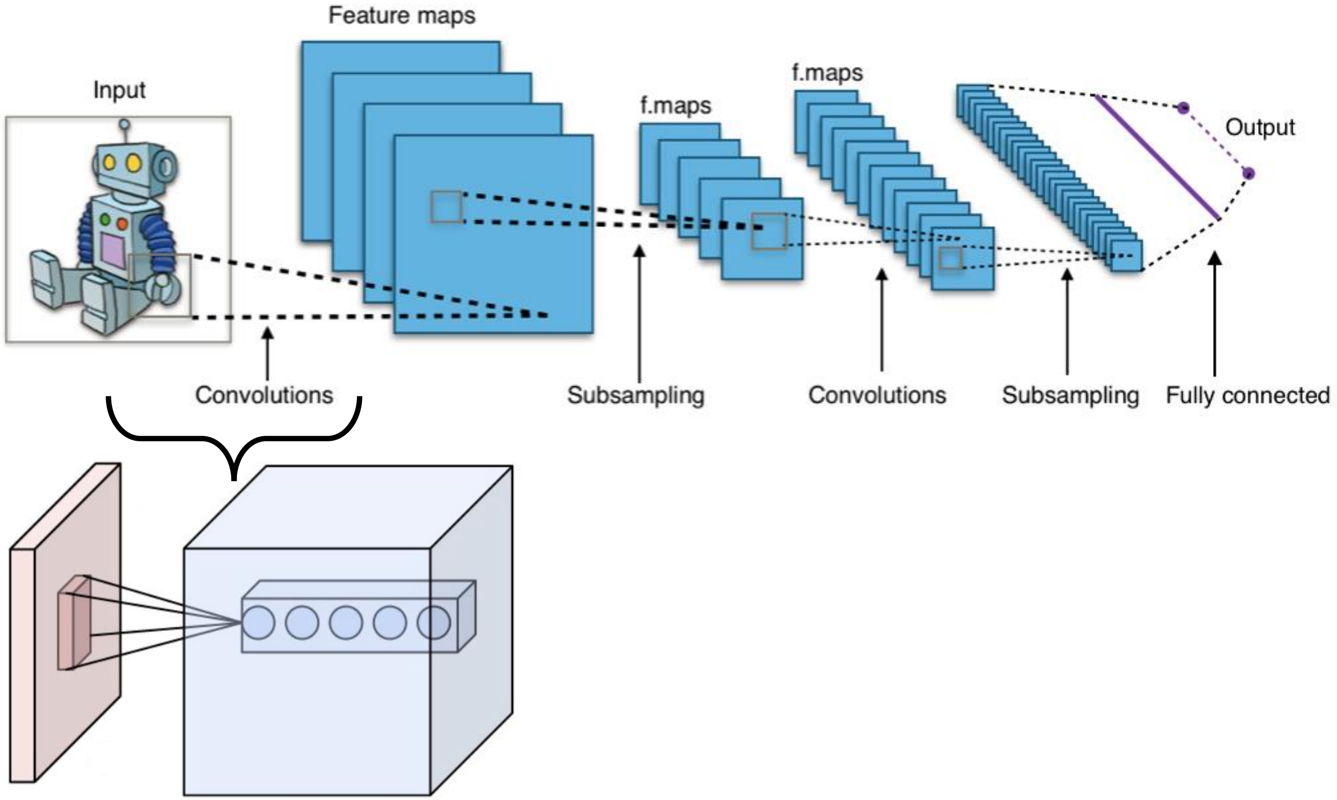
Speech and text tasks:

- Text Classification
- Named-entity recognition
- Machine Translation
- Chat-bots
- Automatic speech recognition

…

Deep Learning - CNN

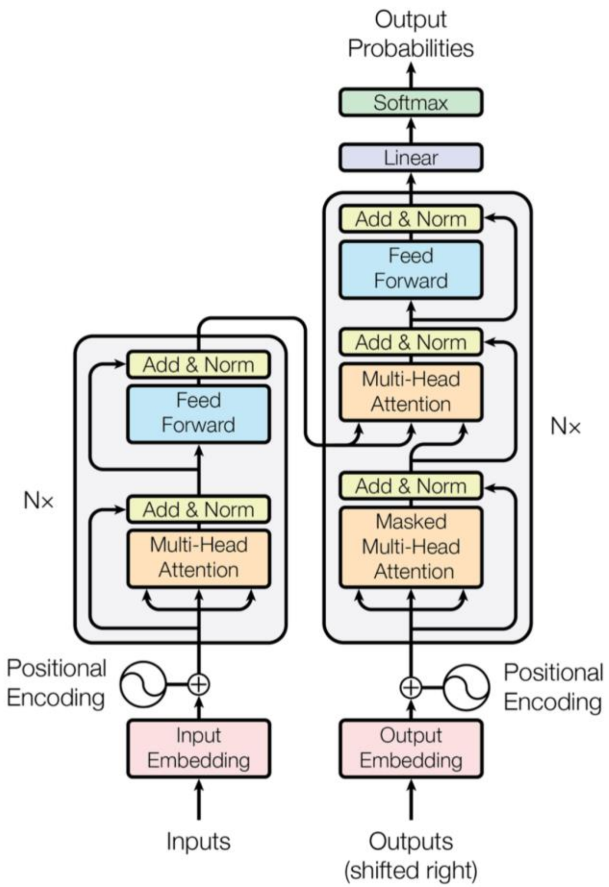
Convolutional Neural Networks remain the gold standard for Vision. Transformers for Vision are picking up.



From https://segment-anything.com/

Deep Learning - Transformers

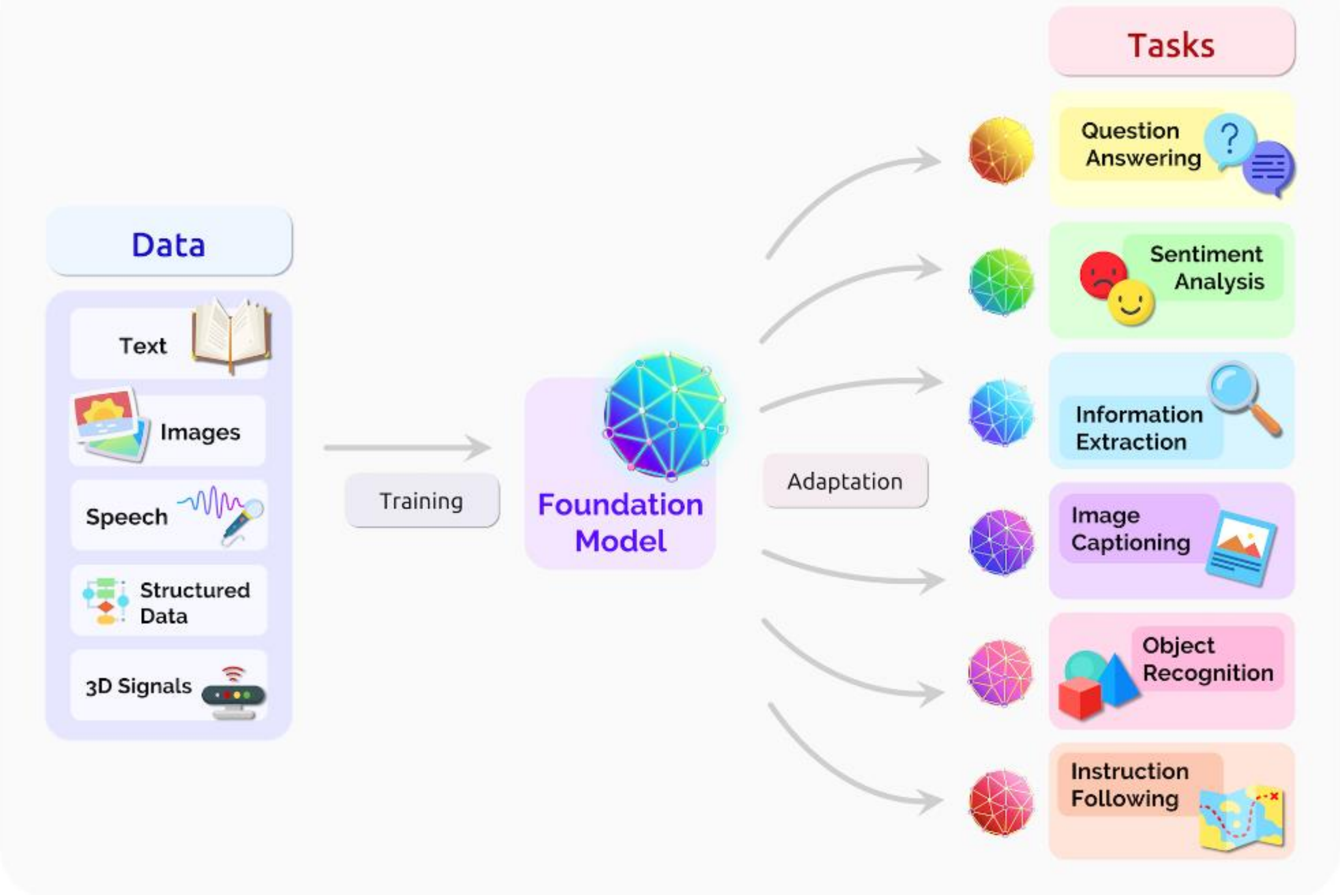
- Transformer architectures dominating (e.g., BERT, LLM, ...)
- Hugging Face = *de facto* standard for model and dataset repo



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

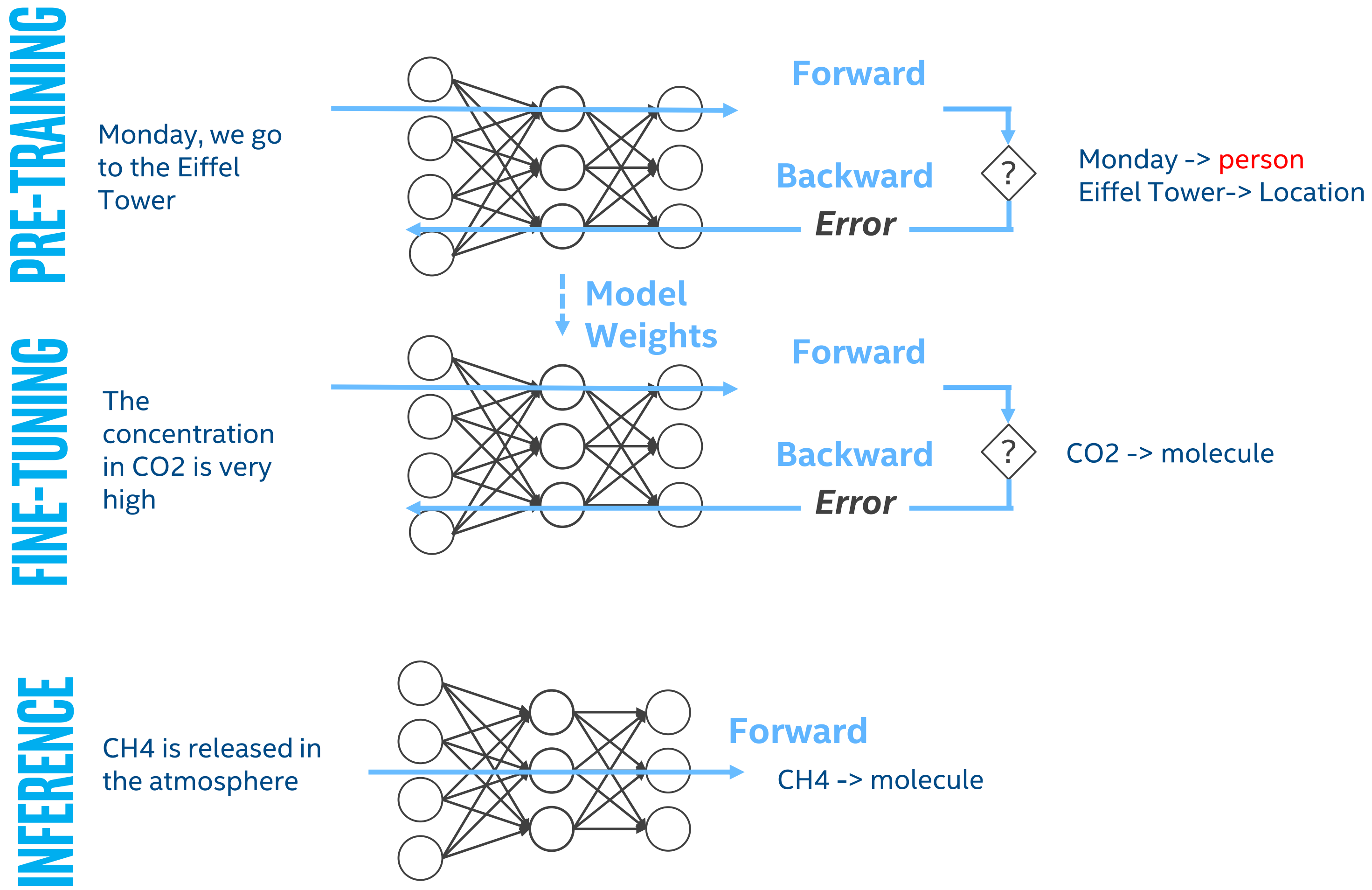
Deep Learning - Foundation models

- A Foundation Models is a large AI model trained on enormous quantities of unlabeled data
  - usually through self-supervised learning.
- This one model can then be adapted to a wide range of downstream tasks

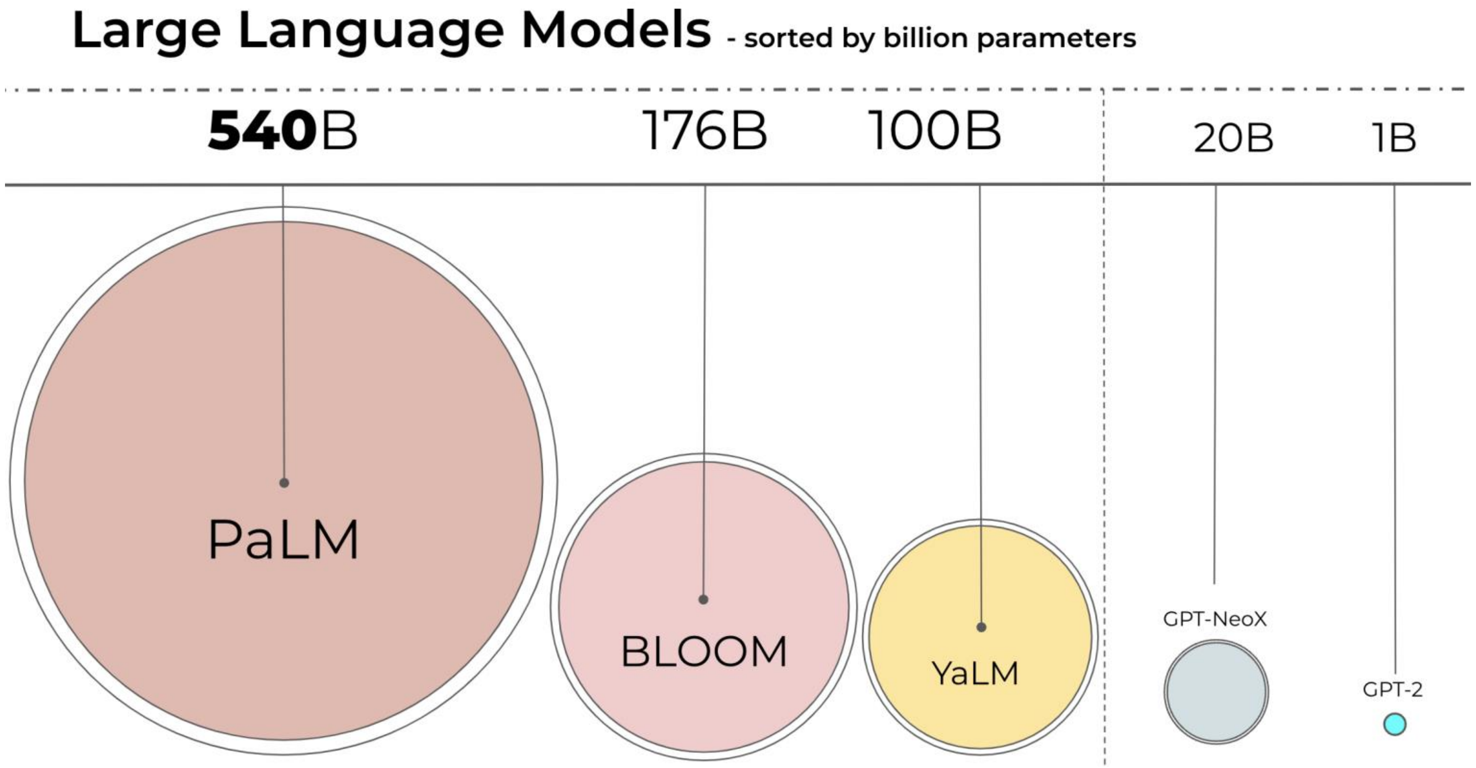


Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

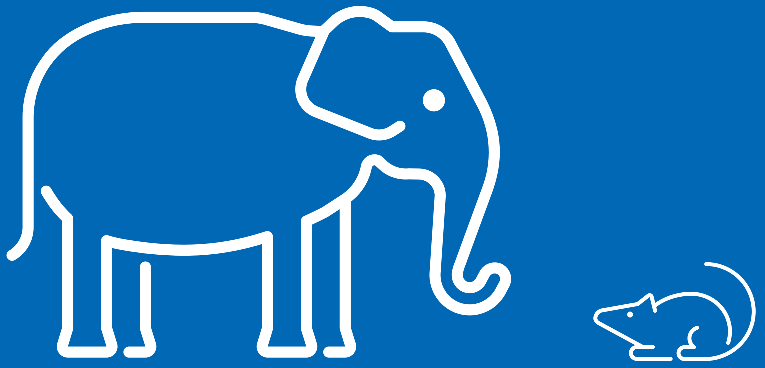
Deep Learning – Adaptation - Fine-tuning



Deep Learning – and now?



Where ML shines

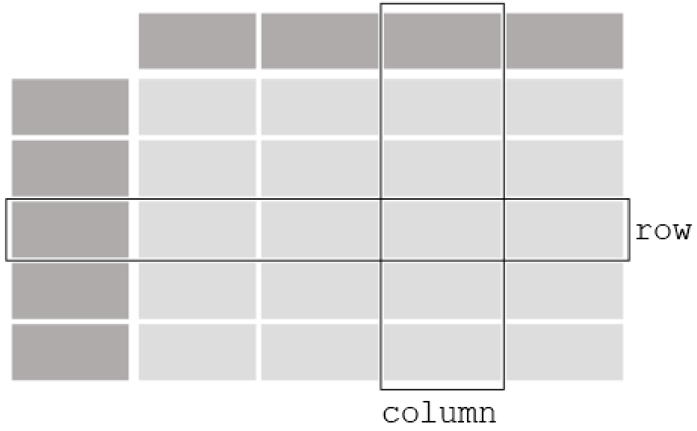


Benefits of using classical ML

- Classical ML complements Deep Learning methods
- Shine where DL is weak:
  - Tabular data
  - Small dataset
  - Explainability
  - Computationally efficient

Tabular data

- Most common data type, embedded in all relational database.

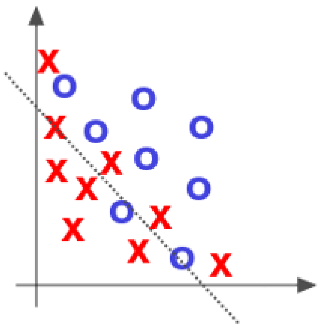


- Research has shown that gradient boosted tree ensemble models (XgBoost) are performing better than Deep learning on tabular data [1] and in particular on unseen datasets

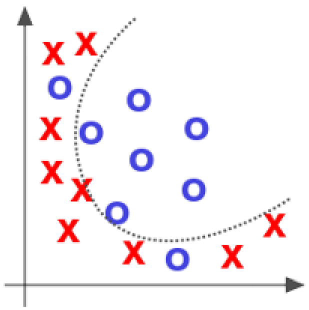
[1] Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84-90.

Small dataset

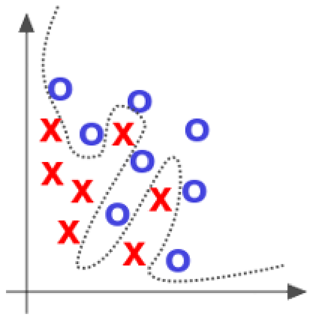
- Model with less parameters as found in classical ML will perform better on small dataset
- Less risk of overfitting



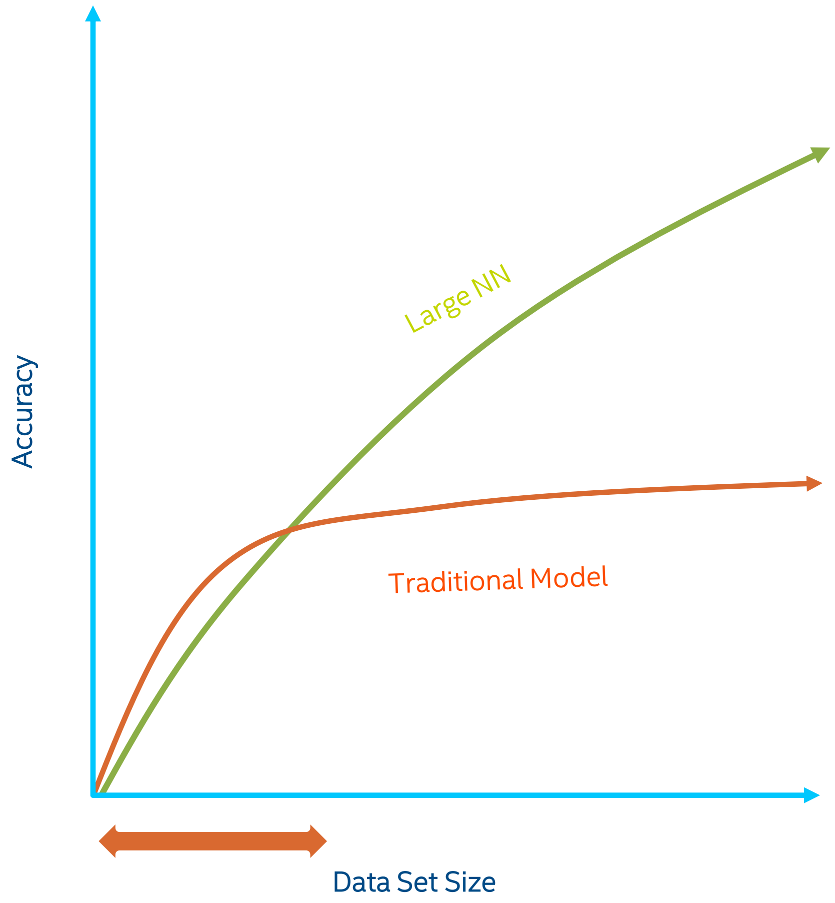
Under Fit



Appropriate

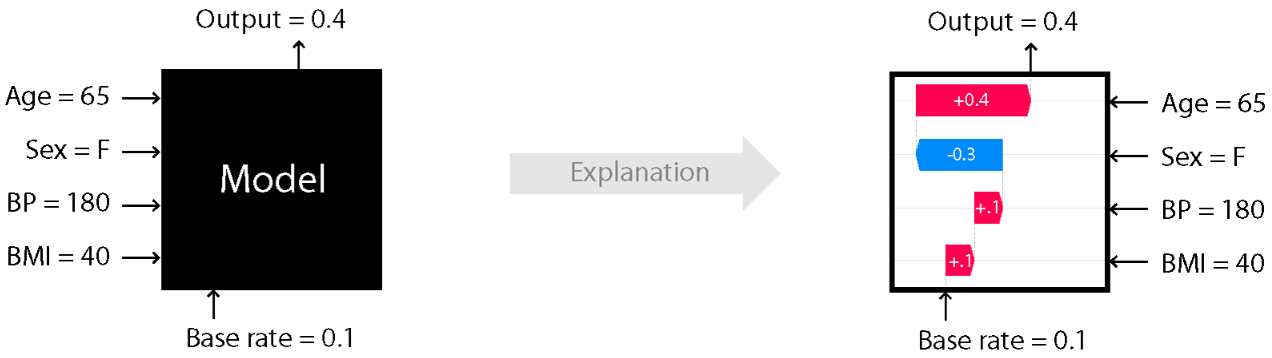


Over Fit

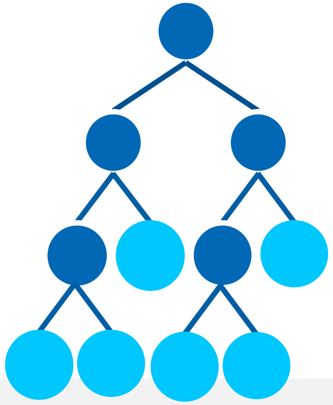


Explainibility

- Features used by the algorithm are available and their influence can be analyzed using different techniques such as SHAP



- Classical model can also be easier to apprehend



Computationally efficient

- Can be trained on a CPU and run inference fast
- In order of minutes and not days as with DL
- Faster prototyping

Intel Hardware for AI

Flexible AI Acceleration

CPU *only*

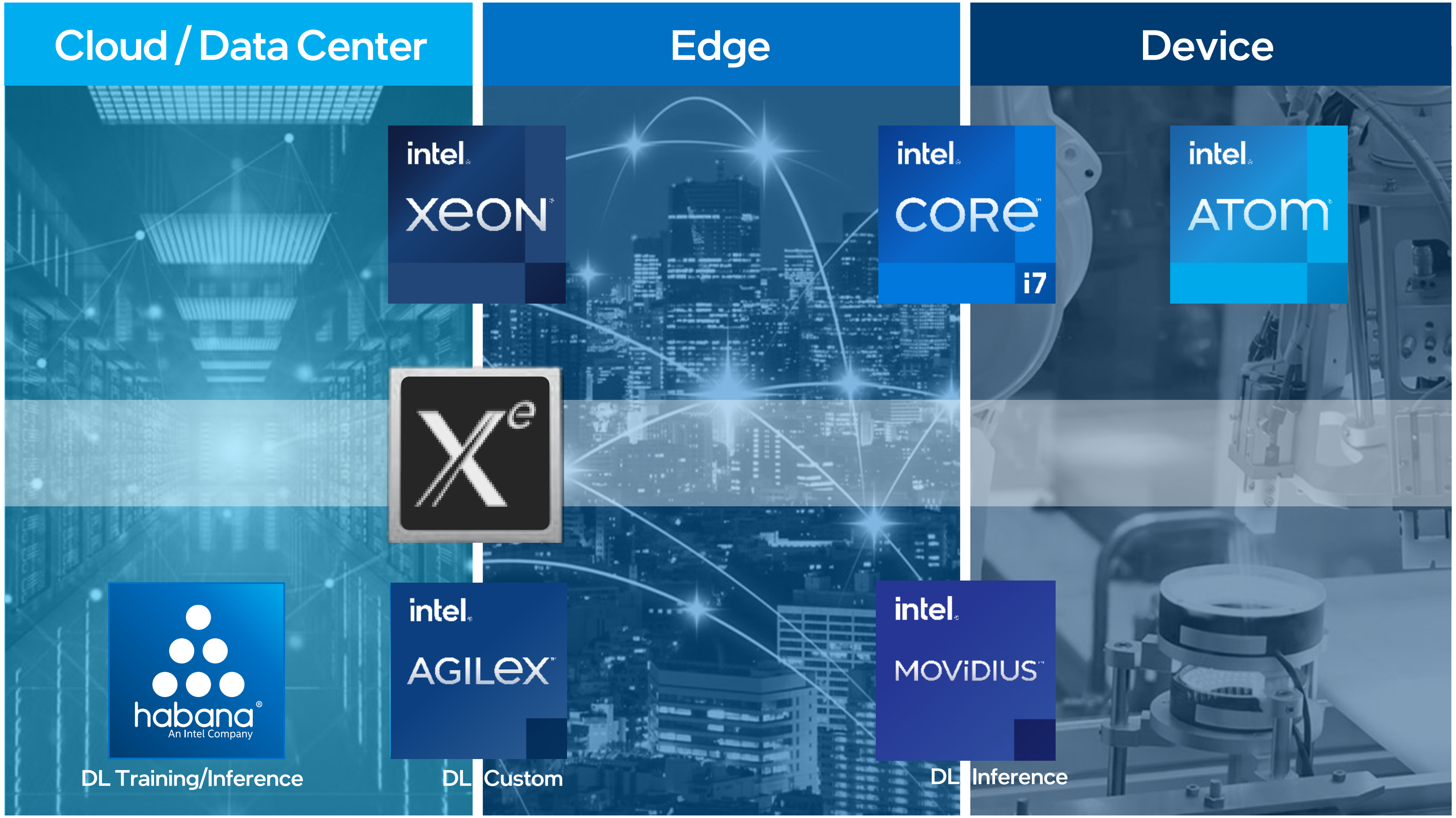
Built-in AI acceleration for mainstream AI use cases

CPU + GPU

When compute is dominated by AI, HPC, graphics, and/or real-time media

CPU + custom

When compute is dominated by deep learning (DL)



Intel® Xeon® Scalable Processors

The Only Data Center CPU with Built-in AI Acceleration

Intel Advanced Vector Extensions 512

Intel Deep Learning Boost (Intel DL Boost)

Intel Advanced Matrix Extensions

Shipping

Cascade Lake

New Intel DL Boost (VNNI)
New memory storage hierarchy

Cooper Lake

Intel DL Boost (BFLOAT16)

April 2021

Ice Lake

Intel DL Boost (VNNI) and new Intel Software Guard Extensions (Intel® SGX) that enable new AI use cases like federated learning

January 2023

Sapphire Rapids

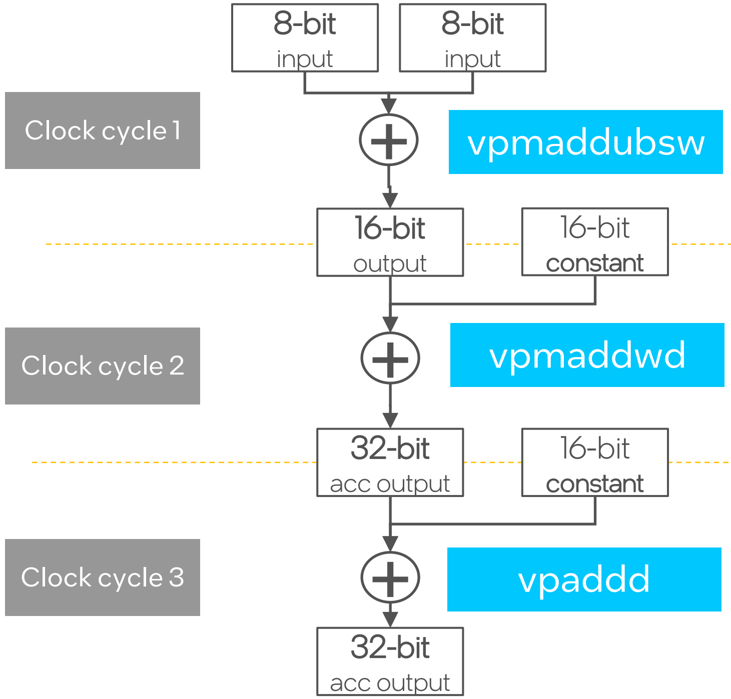
Intel Advanced Matrix Extensions (AMX) extends built-in AI acceleration capabilities on Xeon Scalable

Leadership performance

One Processor for Scalar, Vector, and Matrix

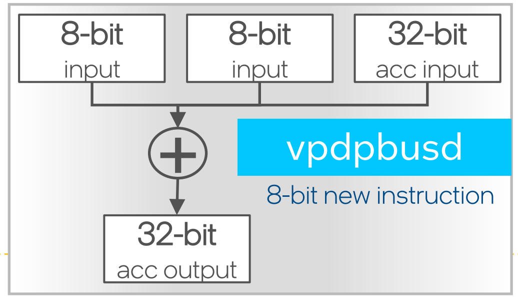
Intel® AVX-512

85 int8 ops/cycle/core
with 2 FMA



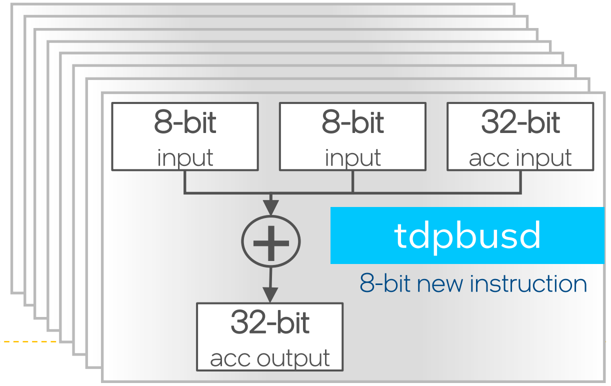
Intel® AVX-512 (VNNI)

256 int8 ops/cycle/core
with 2 FMAs



Intel® AMX

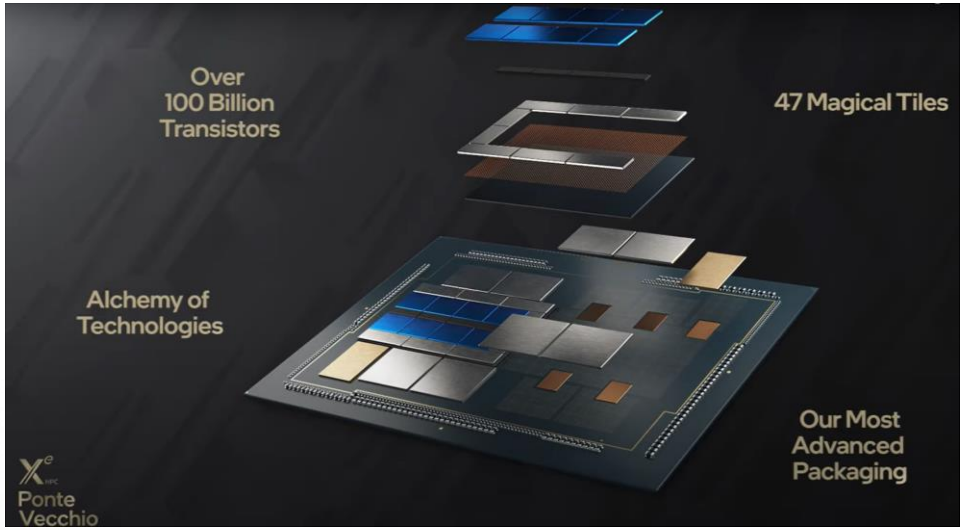
2048 INT8 ops/cycle/core
Multi-fold MACs in one instruction



X^e HPC (Ponte Vecchio)

Leadership Performance for Data-level Parallel AI Workloads

>40 active tiles, over 100 billion transistors integrated into a single package



Powering New Phase of SuperMUC-NG at Leibniz Supercomputing Centre (LRZ)

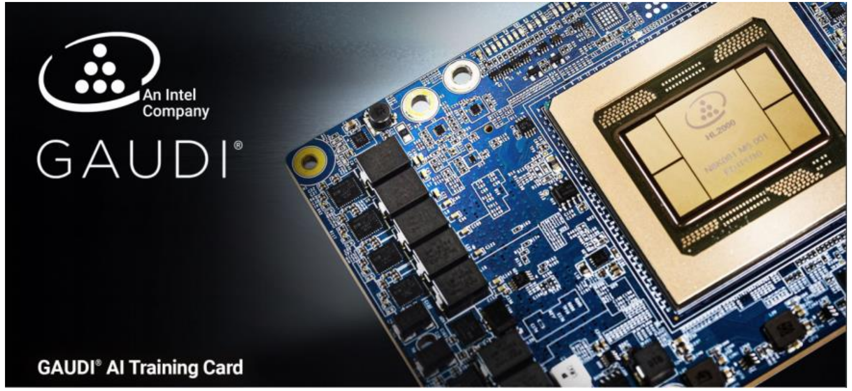
https://www.youtube.com/watch?v=JzbN1IOAcwY

Habana – an Intel Company

intel.

habana®

Deep Learning ASIC for Training and Inference



Gaudi accelerators in AWS EC2 instances, leverages up to 8 Gaudi accelerators and deliver up to 40% better price performance than current GPU-based EC2 instances for training

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

AI Software Stack for Intel XPUs

oneAPI

One Programming Model for Multiple Architectures & Vendors

Freedom to Make Your Best Choice

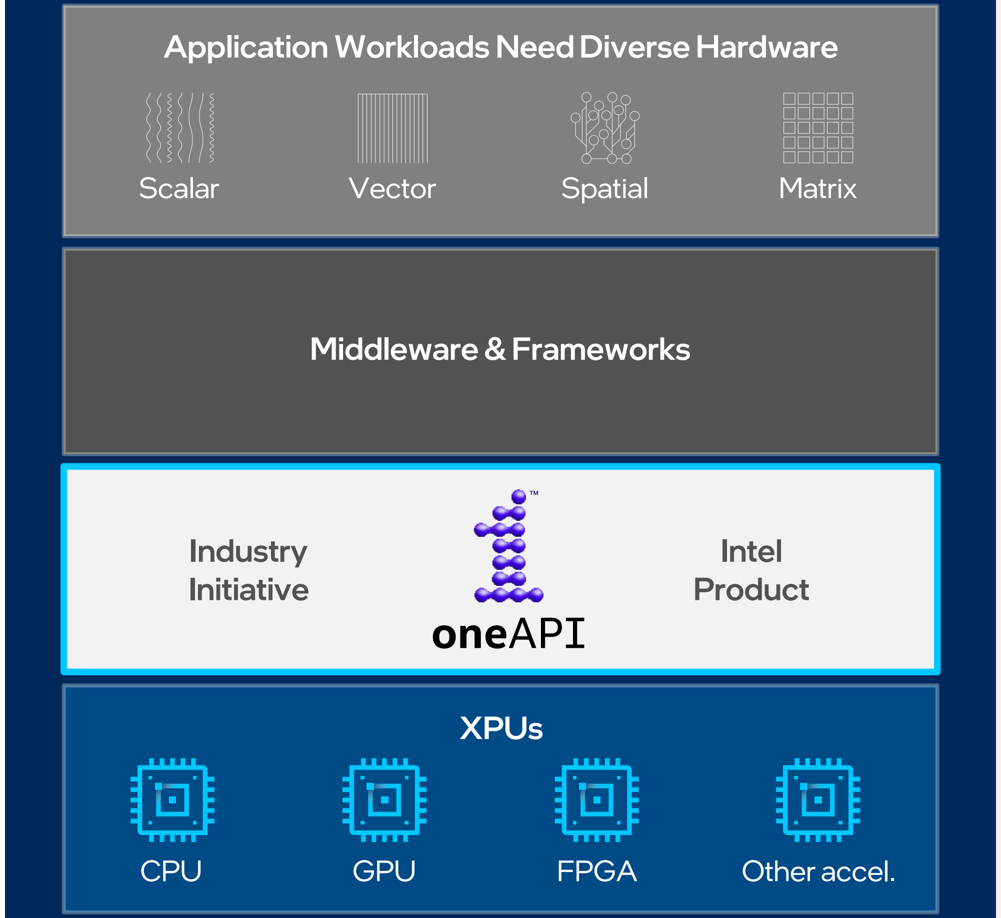
- Choose the best accelerated technology the software doesn't decide for you

Realize all the Hardware Value

- Performance across CPU, GPUs, FPGAs, and other accelerators

Develop & Deploy Software with Peace of Mind

- Open industry standards provide a safe, clear path to the future
- Compatible with existing languages and programming models including C++, Python, SYCL, OpenMP, Fortran, and MPI



Intel's oneAPI Ecosystem

Built on Intel's Rich Heritage of CPU Tools Expanded to XPUs

oneAPI

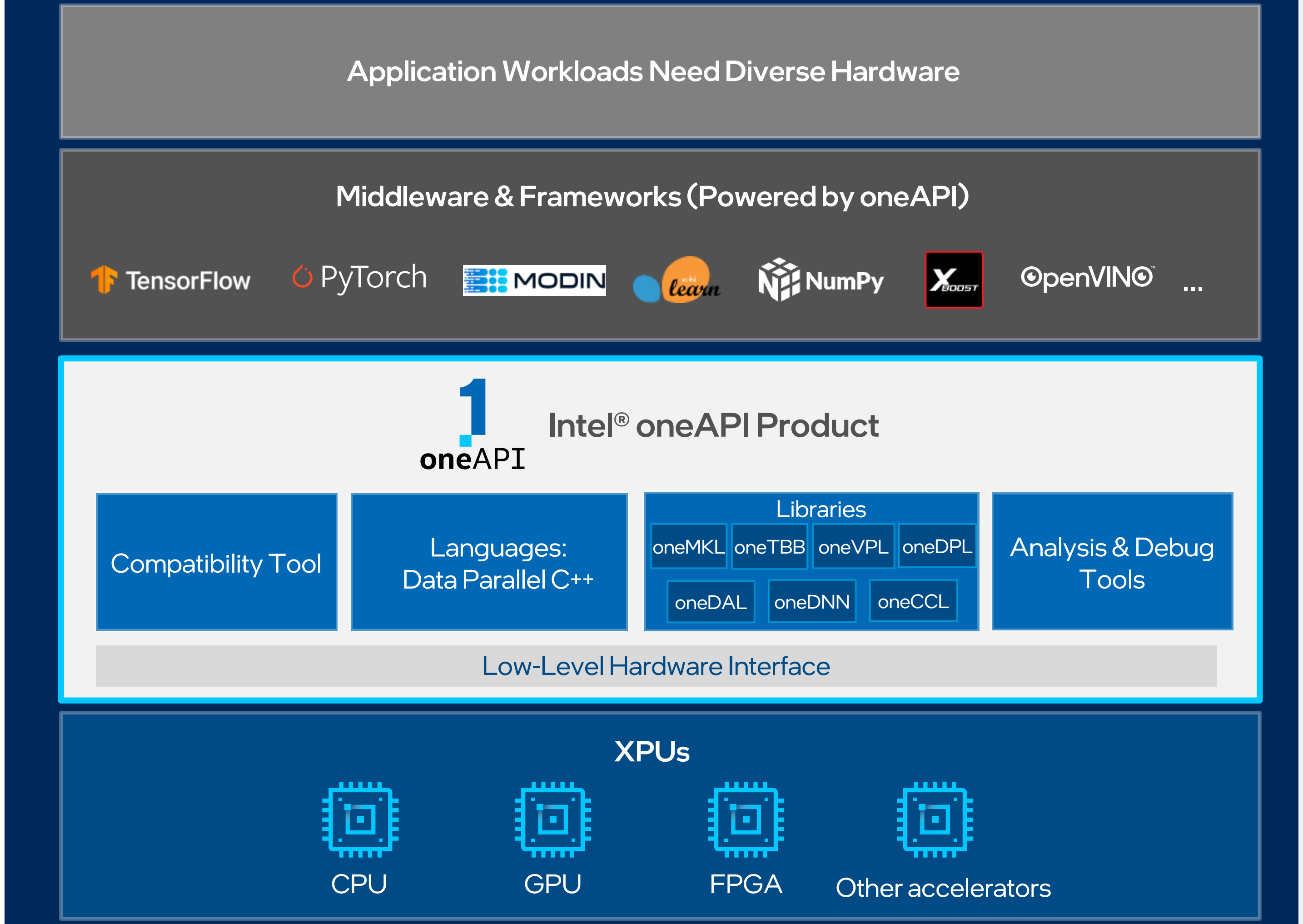
A cross-architecture language based on C++ and SYCL standards

Powerful libraries designed for acceleration of domain-specific functions

A complete set of advanced compilers, libraries, and porting, analysis and debugger tools

Powered by oneAPI

Frameworks and middleware that are built using one or more of the oneAPI industry specification elements, the DPC++ language, and libraries listed on oneapi.com.



Available Now

Visit software.intel.com/oneapi for more details
Some capabilities may differ per architecture and custom-tuning will still be required. Other accelerators to be supported in the future.

Intel® AI Analytics Toolkit

Powered by oneAPI

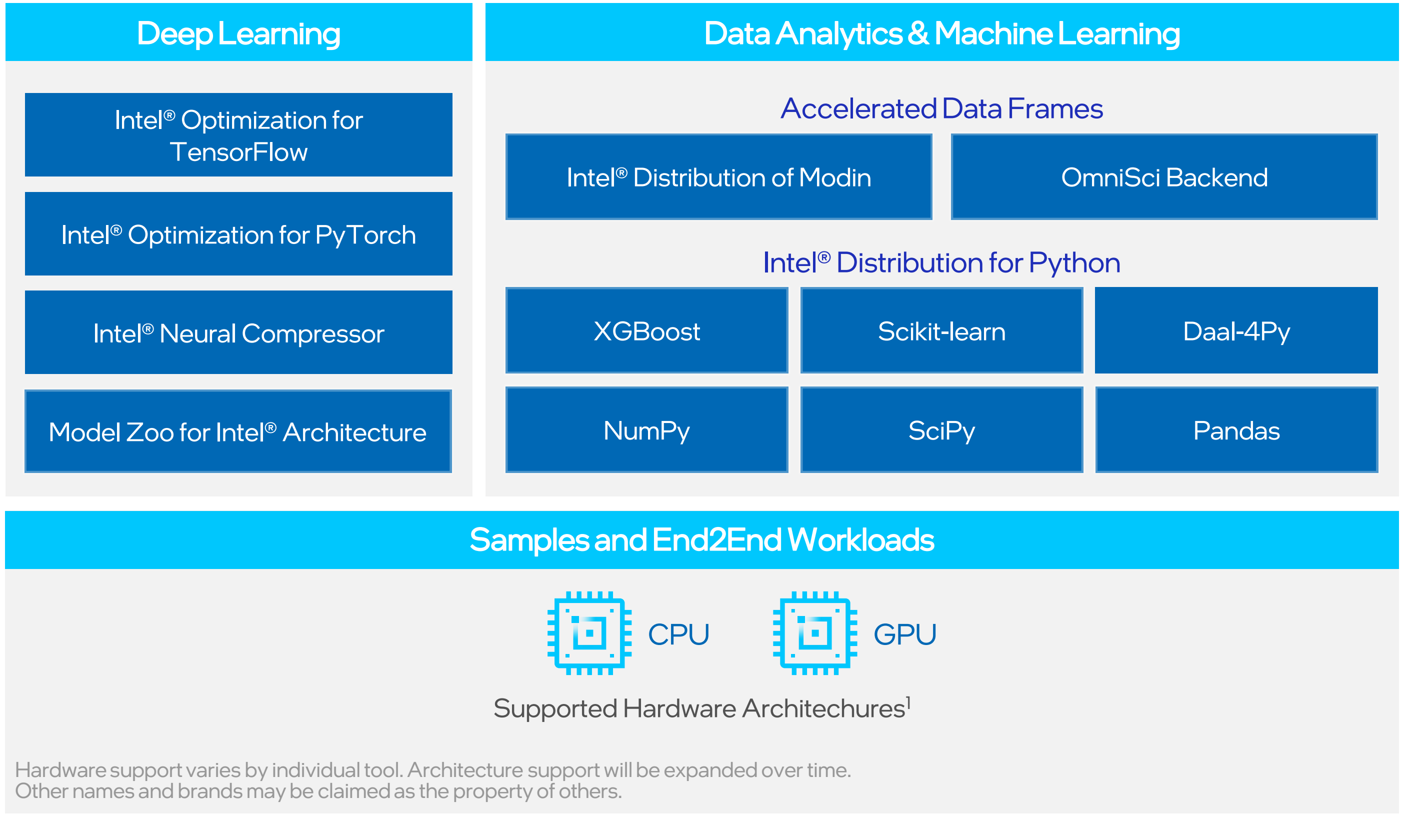
Accelerate end-to-end AI and data analytics pipelines with libraries optimized for Intel® architectures

Who Uses It?

Data scientists, AI researchers, ML and DL developers, AI application developers

Top Features/Benefits

- Deep learning performance for training and inference with Intel optimized DL frameworks and tools
- Drop-in acceleration for data analytics and machine learning workflows with compute-intensive Python packages



Deep Learning

Intel® Optimization for TensorFlow

Intel® Optimization for PyTorch

Intel® Neural Compressor

Model Zoo for Intel® Architecture

Data Analytics & Machine Learning

Accelerated Data Frames

Intel® Distribution of Modin

OmniSci Backend

Intel® Distribution for Python

XGBoost

Scikit-learn

Daal-4Py

NumPy

SciPy

Pandas

Samples and End2End Workloads

CPU

GPU

Supported Hardware Architectures[1]

Hardware support varies by individual tool. Architecture support will be expanded over time.
Other names and brands may be claimed as the property of others.

Get the Toolkit HERE or via these locations

Intel Installer | Docker | Apt, Yum | Conda | Intel® DevCloud

Learn More: software.intel.com/oneapi/ai-kit

Back to Domain-specific Toolkits for Specialized Workloads

Intel oneAPI Software Tools for AI & Analytics

Intel® oneAPI Toolkits

Intel® AI Analytics Toolkit

Accelerate machine learning & data science pipelines with optimized deep learning frameworks & high-performing Python libraries

Data Scientists, AI Researchers, DL/ML Developers

Intel® oneAPI Base Toolkit

Incl. Intel® oneAPI Deep Neural Network Library (oneDNN), Intel® oneAPI Collective Communications Library (oneCCL), & Intel® oneAPI Data Analytics Library (oneDAL)

Optimize primitives for algorithms and framework development

DL Framework Developers - Optimize algorithms for Machine Learning & Analytics

Toolkit Powered by oneAPI

Intel® Distribution of OpenVINO™ Toolkit

Deploy high performance inference & applications from edge to cloud

AI Application, Media, & Vision Developers

oneAPI Ecosystem Endorsements for AI domain

The industry needs a programming model where developers can take advantage of an array of innovative hardware architectures. The goal of oneAPI is to provide increased choice of hardware vendors, processor architectures, and faster support of next-generation accelerators. Microsoft has been using oneAPI elements across Intel hardware offerings as part of its initiatives and supports the open standards-based specification. We are excited to support our customers with choice and accelerate the growth of AI and machine learning.
- Tim Harris, Principal Architect, Azure AI, Microsoft

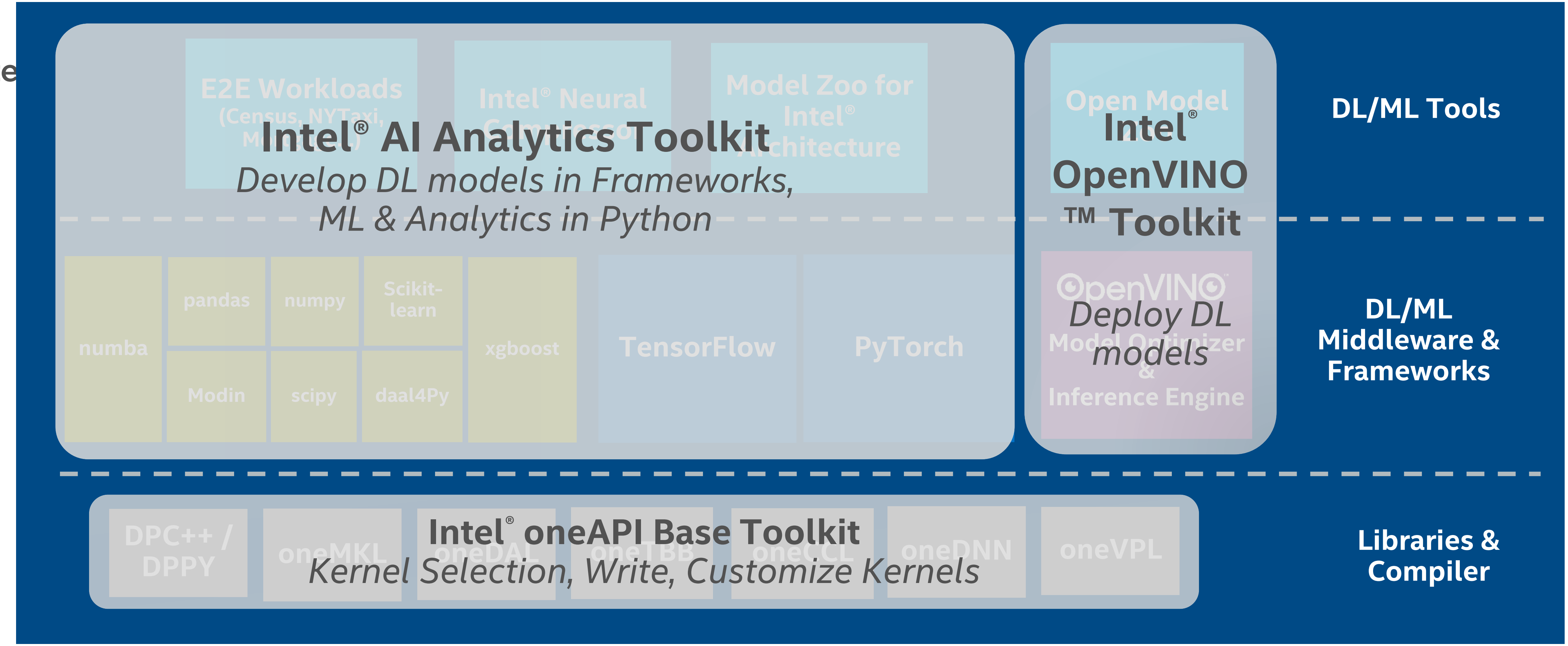
Microsoft Azure

With the growth of AI, machine learning, and data-centric applications, the industry needs a programming model that allows developers to take advantage of rapid innovation in processor architectures. TensorFlow supports the oneAPI industry initiative and its standards-based open specification. oneAPI complements TensorFlow's modular design and provides increased choice of hardware vendor and processor architecture, and faster support of next-generation accelerators. TensorFlow uses oneAPI today on Xeon processors and we look forward to using oneAPI to run on future Intel architectures.

TensorFlow

AI Software Stack for Intel XPUs

- Inte



E2E Workloads (Census, NYTaxi, Mortgage…)

Intel® Neural Compressor

Model Zoo for Intel® Architecture

Intel® AI Analytics Toolkit

Develop DL models in Frameworks, ML & Analytics in Python

numba

pandas

numpy

Scikit-learn

Modin

scipy

daal4Py

xgboost

TensorFlow

PyTorch

Open Model Zoo

Intel® OpenVINO™ Toolkit

OpenVINO™

Deploy DL models

Model Optimizer & Inference Engine

DPC++ / DPPY

oneMKL

oneDAL

oneTBB

oneCCL

oneDNN

oneVPL

Intel® oneAPI Base Toolkit

Kernel Selection, Write, Customize Kernels

DL/ML Tools

DL/ML Middleware & Frameworks

Libraries & Compiler

Full Set of Intel oneAPI cross-architecture AI ML & DL Software Solutions oneAPI Available on

Intel® DevCloud for oneAPI

A development sandbox to develop, test and run workloads across a range of Intel CPUs, GPUs, and FPGAs using Intel's oneAPI software.

Get Up & Running In Seconds!

Sign up at:
software.intel.com/devcloud/oneapi

intel
DevCloud

- 1 Minute to Code
- No Hardware Acquisition
- No Download, Install or Configuration
- Easy Access to Samples & Tutorials
- Support for Jupyter Notebooks, Visual Studio Code

High-Performance Deep Learning Using Intel® Distribution of OpenVINO™ toolkit - Powered by oneAPI

A toolkit for fast, more accurate real-world results using high-performance AI and computer vision inference deployed into production on Intel XPU architectures (CPU, GPU, FPGA, VPU) from edge to cloud

Who needs this product?

AI application developers, OEMs, ISVs, System Integrators, Vision and Media developers

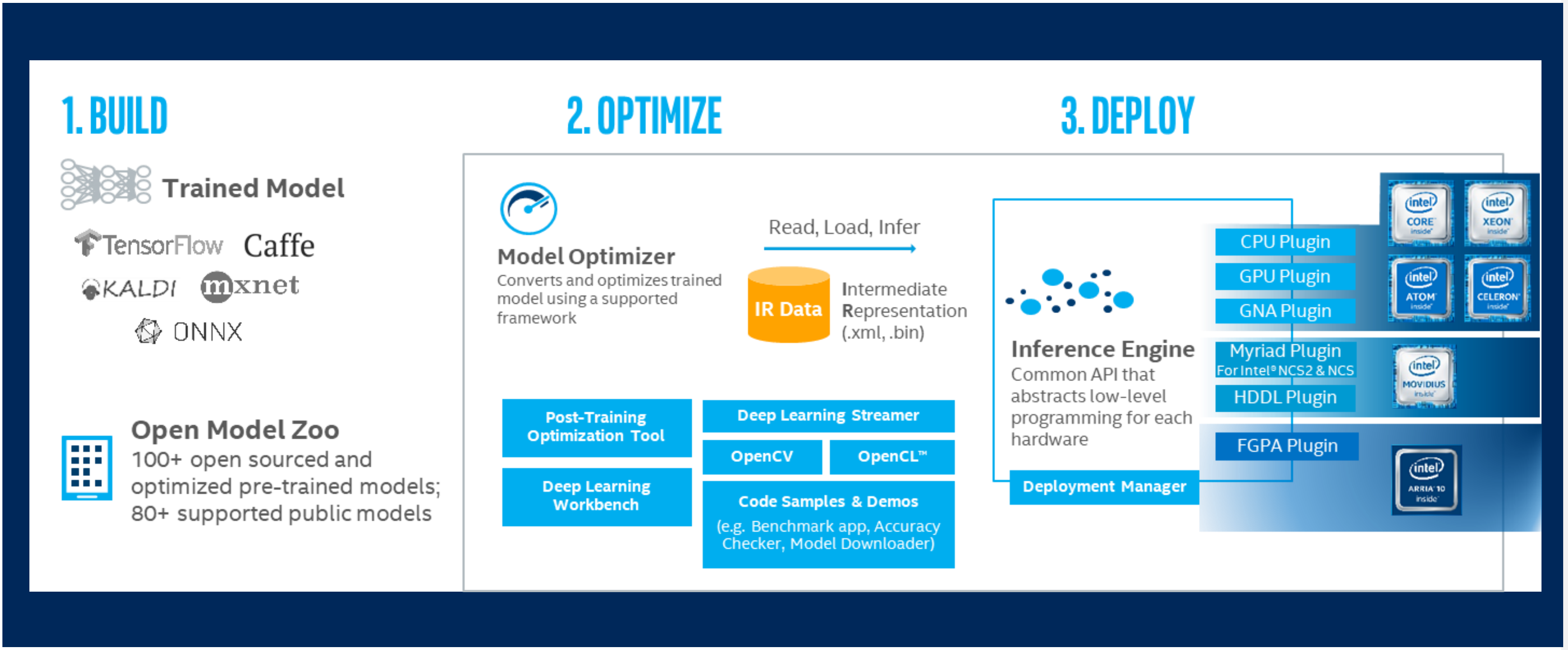
Top Features/Benefits

High-performance, deep learning inference deployment

Streamlined development; ease of use

Write once, deploy anywhere

Proven, industry-leading accelerated technology



1. BUILD

Trained Model

TensorFlow Caffe

KALDI mxnet

ONNX

Open Model Zoo

100+ open sourced and optimized pre-trained models; 80+ supported public models

2. OPTIMIZE

Model Optimizer

Converts and optimizes trained model using a supported framework

Read, Load, Infer

IR Data

Intermediate Representation (.xml, .bin)

Post-Training Optimization Tool

Deep Learning Workbench

Deep Learning Streamer

OpenCV

OpenCL™

Code Samples & Demos

(e.g. Benchmark app, Accuracy Checker, Model Downloader)

3. DEPLOY

Inference Engine

Common API that abstracts low-level programming for each hardware

Deployment Manager

CPU Plugin

GPU Plugin

GNA Plugin

Myriad Plugin For Intel® NCS2 & NCS

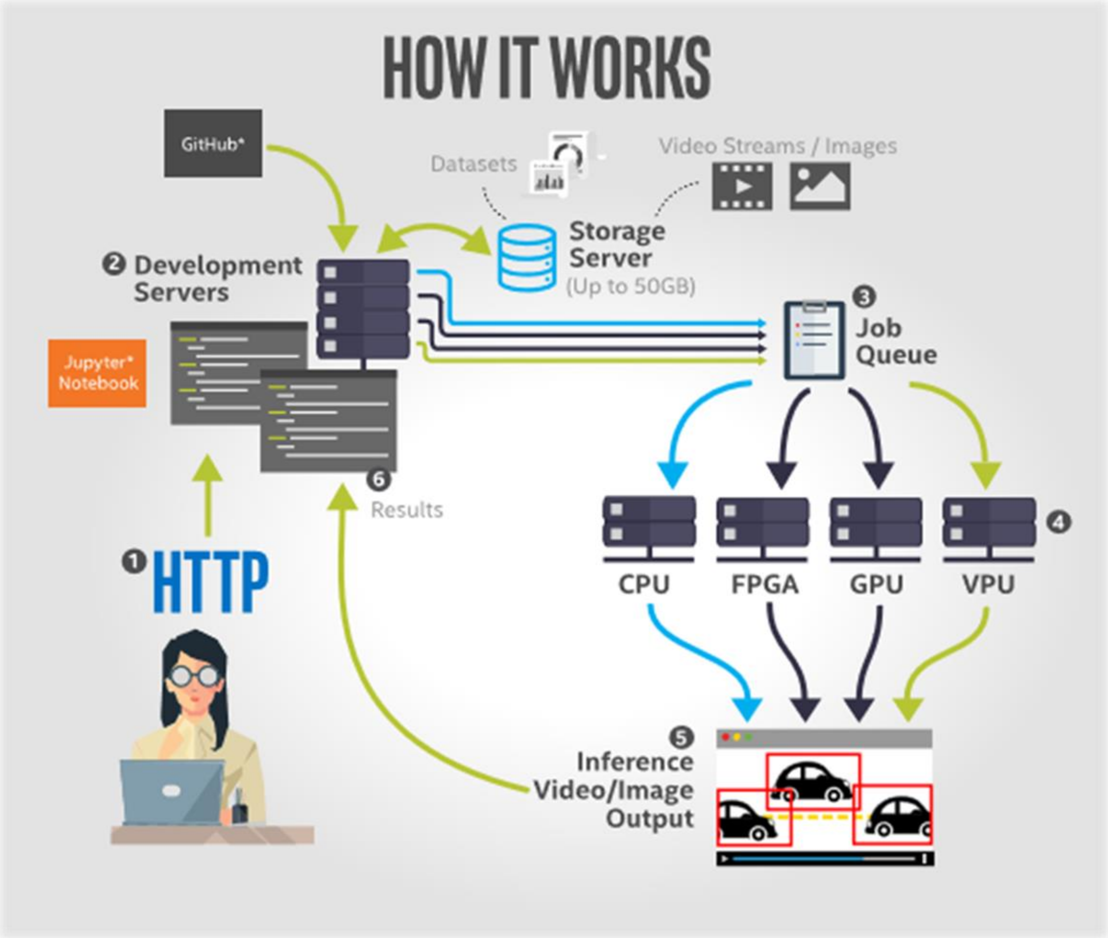
HDDL Plugin

FGPA Plugin

software.intel.com/openvino-toolkit

Accelerate Time to Production with Intel® DevCloud for the Edge

See immediate AI Model performance across Intel's vast array of Edge Solutions



- **Instant, Global Access**
  Run AI applications from anywhere in the world
- **Prototype on the Latest Hardware and Software**
  Develop knowing you're using the latest Intel technology
- **Benchmark your Customized AI Application**
  Immediate feedback - frames per second, performance
- **Reduce Development Time and Cost**
  Quickly find the right compute for your edge solution

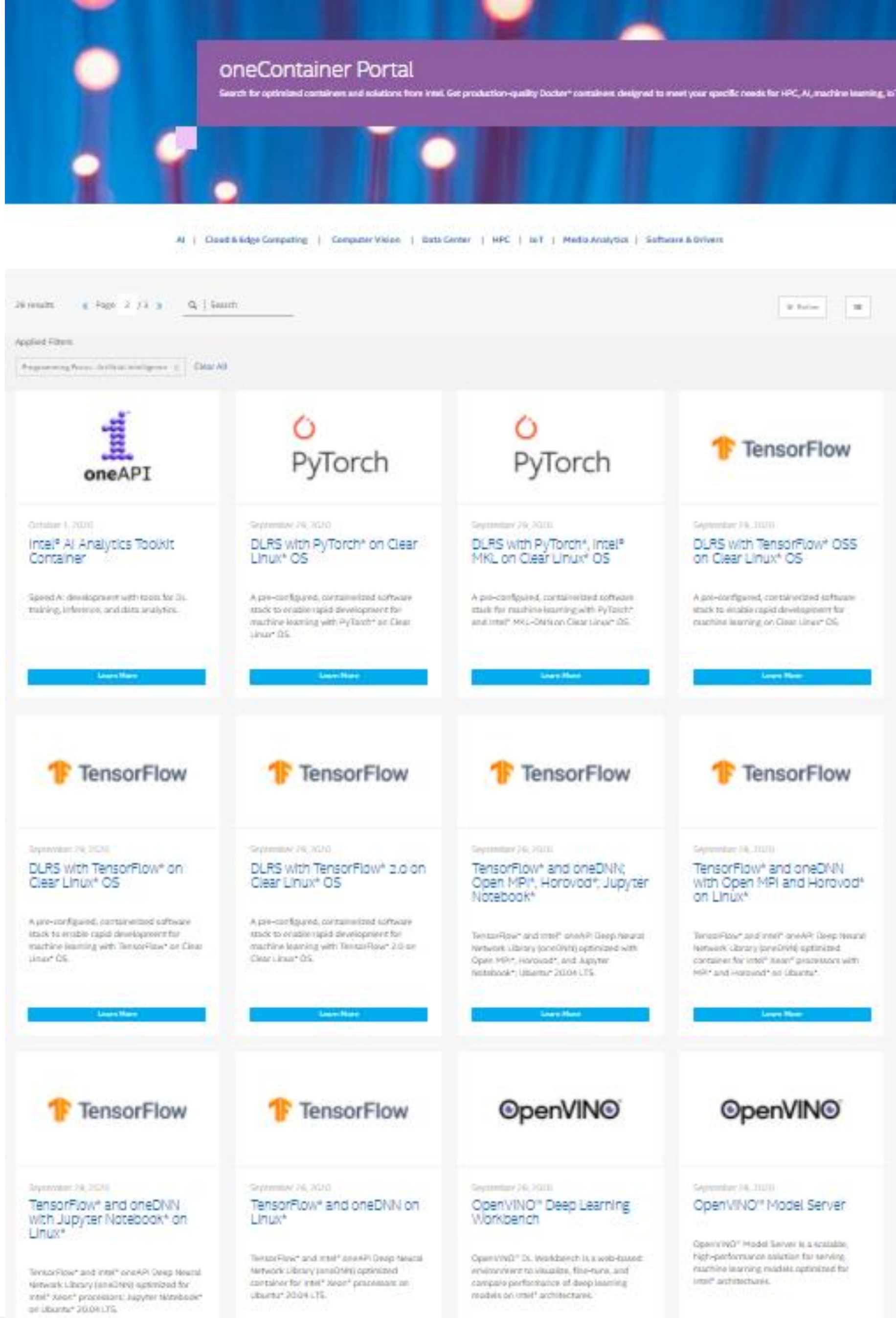
Sign up now for access

AI Containers for Flexibility

- Optimized, validated, deployable AI containers
- Available via Docker containers. Will expand to include Kubernetes orchestrations, Helm charts
- Access from oneContainer Portal
  - Include containers with ready-to-use AI software stacks
  - And containers with full AI workloads (including models)

Topology	Frameworks
DLRM	PYT
ResNet50	PYT, TF, OV
BERT-large	PYT, TF, OV
Transformer-LT	PYT, TF
MobileNet-v1	PYT, TF, OV
SSD-Mobilenet-v1	PYT, TF, OV
SSD-Resnet34	PYT, TF, OV
WaveNet*	TF

Topology	Framework
Mask R-CNN	PYT, TF, OV
RNN-T	PYT, TF, OV
3D-UNet	TF, OV
DIEN	TF
Wide & Deep	PYT, TF
RNX101	
Yolo-V3	PYT, TF, OV
NCF*	TF



oneContainer Portal

Search for optimized containers and solutions from Intel. Get production-quality Docker* containers designed to meet your specific needs for HPC, AI, machine learning, IoT

AI | Cloud & Edge Computing | Computer Vision | Data Center | HPC | IoT | Media Analytics | Software & Drivers
26 results
Page 2 / 3
Search
Applied Filters:
Programming Focus: Artificial Intelligence
Clear All
oneAPI
PyTorch
TensorFlow
OpenVINO
Learn More

October 1, 2020
Intel® AI Analytics Toolkit Container
Speed AI development with tools for DL training, inference, and data analytics.

September 29, 2020
DLRS with PyTorch* on Clear Linux* OS
A pre-configured, containerized software stack to enable rapid development for machine learning with PyTorch* on Clear Linux* OS.

September 29, 2020
DLRS with PyTorch*, Intel® MKL on Clear Linux* OS
A pre-configured, containerized software stack for machine learning with PyTorch* and Intel® MKL-DNN on Clear Linux* OS.

September 18, 2020
DLRS with TensorFlow* OSS on Clear Linux* OS
A pre-configured, containerized software stack to enable rapid development for machine learning on Clear Linux* OS.

September 29, 2020
DLRS with TensorFlow* on Clear Linux* OS
A pre-configured, containerized software stack to enable rapid development for machine learning with TensorFlow* on Clear Linux* OS.

September 29, 2020
DLRS with TensorFlow* 2.0 on Clear Linux* OS
A pre-configured, containerized software stack to enable rapid development for machine learning with TensorFlow* 2.0 on Clear Linux* OS.

September 26, 2020
TensorFlow* and oneDNN; Open MPI*, Horovod*; Jupyter Notebook*
TensorFlow* and Intel® oneAPI Deep Neural Network Library (oneDNN) optimized with Open MPI*, Horovod*, and Jupyter Notebook*; Ubuntu* 20.04 LTS.

September 18, 2020
TensorFlow* and oneDNN with Open MPI and Horovod* on Linux*
TensorFlow* and Intel® oneAPI Deep Neural Network Library (oneDNN) optimized container for Intel® Xeon® processors with MPI* and Horovod* on Ubuntu*.

September 18, 2020
TensorFlow* and oneDNN with Jupyter Notebook* on Linux*
TensorFlow* and Intel® oneAPI Deep Neural Network Library (oneDNN) optimized for Intel® Xeon® processors; Jupyter Notebook* on Ubuntu* 20.04 LTS.

September 26, 2020
TensorFlow* and oneDNN on Linux*
TensorFlow* and Intel® oneAPI Deep Neural Network Library (oneDNN) optimized container for Intel® Xeon® processors on Ubuntu* 20.04 LTS.

September 26, 2020
OpenVINO™ Deep Learning Workbench
OpenVINO™ DL Workbench is a web-based environment to visualize, fine-tune, and compare performance of deep learning models on Intel® architectures.

September 18, 2020
OpenVINO™ Model Server
OpenVINO™ Model Server is a scalable, high-performance solution for serving machine learning models optimized for Intel® architectures.

Which Toolkit Should I Use

Use Both!

Intel® oneAPI Analytics Toolkit & Intel® Distribution of OpenVINO™ toolkit

Toolkits are complementary to each other and recommendation is to use them both based on your current phase of AI Journey

- *I am **exploring and analyzing data;** I am **developing models***
- *I want **performance and compatibility** with frameworks and libraries I use*
- *I would like to have **drop-in acceleration** with little to no additional code changes*
- *I prefer **not to learn any new tools** or languages*

Data Scientist/ML Developer
Intel® oneAPI AI Analytics Toolkit

- *I am **deploying models***
- *I want **leading performance and efficiency** across multiple target HW*
- *I'm concerned about **having lower memory footprint**, which is critical for deployment*
- *I am **comfortable with learning and adopting a new tool or API** to do so*

App Developer
Intel® Distribution of OpenVINO™ toolkit

If you prefer working on primitives and to optimize kernels and algorithms directly using oneAPI libraries (oneDNN, oneCCL & oneDAL), then use Intel® oneAPI Base Toolkit

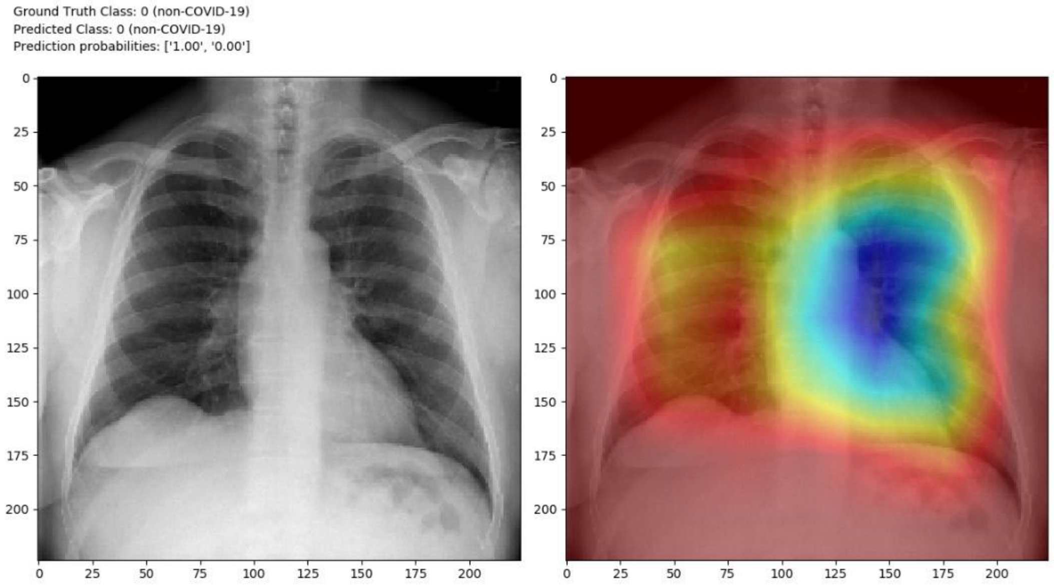
Customer Success Story

Accrad AI-based Solution Helps Accelerate COVID-19 Diagnosis

Optimized by Intel® oneAPI Analytics Toolkit & Intel® Distribution of OpenVINO™ toolkit

CheXRad helps radiologists and physicians identify COVID-19, viral pneumonia and other diseases on chest X-ray images, and predict the need for ventilators.

- *CheXRad* comes pre-configured with a COVID-19 and viral pneumonia classification neural network.
- To architect, train and validate the neural network, Accrad used **Intel Tensorflow from AI Analytics Toolkit** and the **Intel oneAPI DevCloud** to develop the model.
- To optimize its model for deployment, Accrad used **OpenVINO™ toolkit** and **Intel® DevCloud for Edge**.
- *CheXRad* could classify pathologies in 140 chest x-rays in just **90 seconds** —up to **160x faster** than radiologists, at comparable levels of accuracy, sensitivity and specificity.



```
Ground Truth Class: 0 (non-COVID-19)
Predicted Class: 0 (non-COVID-19)
Prediction probabilities: ['1.00', '0.00']
```

Learn more in this solution brief

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Key Takeaways & Call to Action

- Intel toolkits are **FREE**, complementary & work seamlessly together
- They help achieve performance & efficiency across different stages of AI Journey
- Recommend the toolkits based on current phase of customer pipeline

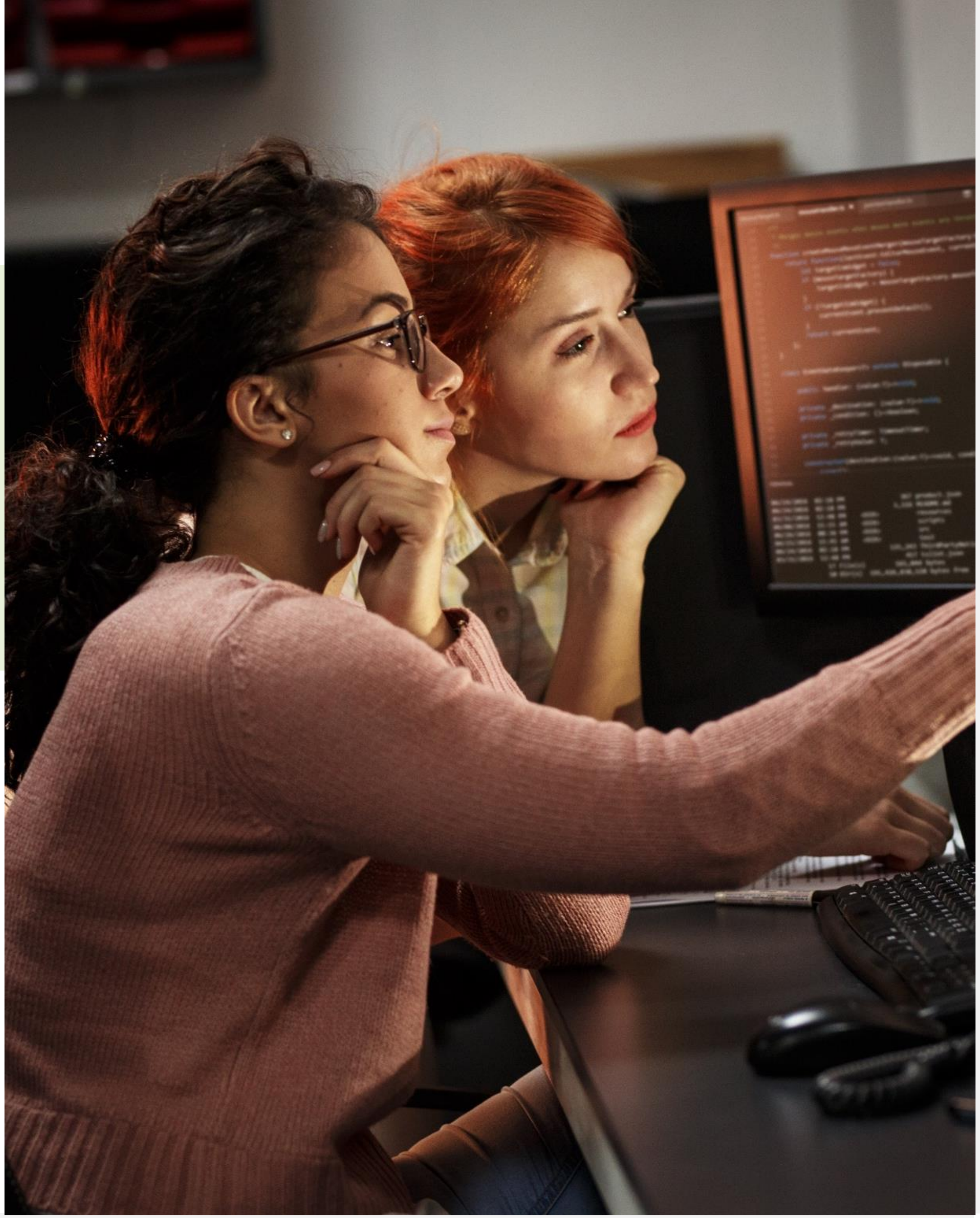
Download the toolkits

Intel® oneAPI AI Analytics Toolkit

Intel® Distribution of OpenVINO ™ toolkit

Intel® oneAPI Base Toolkit

Learn more about Intel® oneAPI Toolkits
intel.com/oneAPI-AllToolkits



May 16th
Intel Deep Learning workshop
@ LRZ

link

Survey

https://survey.lrz.de/index.php/582495?lang=en