
oneAPI
oneAPI
Developer
Summit at SC21
oneAPI AI Analytics – End to End
Antimo Musone - 14 November 21
intel
EY
Building a better working world

About Me

Info

- Senior Manager & Digital Innovator of EY Italy
- Intel Innovator Program
- Microsoft MVP AI
- Co-Founder of Fifth Ingenium

Contact

https://www.linkedin.com/in/antimo-musone/

https://twitter.com/AntimoMusone

antimo.musone@it.ey.com

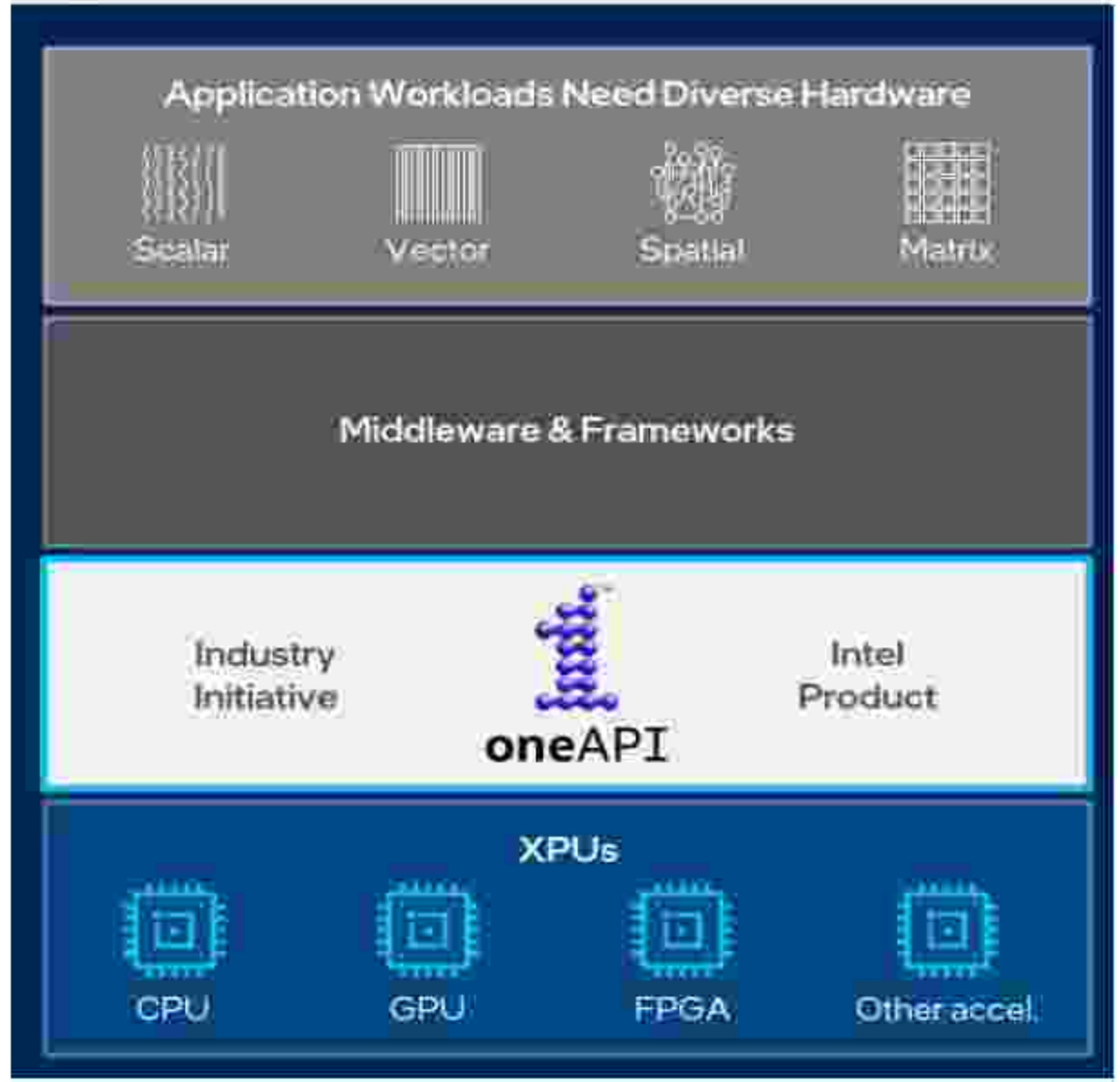
intel EY

Main Contents

1. Intel oneAPI AI Analytics Toolkit
2. MLOps
3. cnvrg.io

Intel oneAPI AI Analytics Toolkit
One Programming Model for Multiple Architectures & Vendors

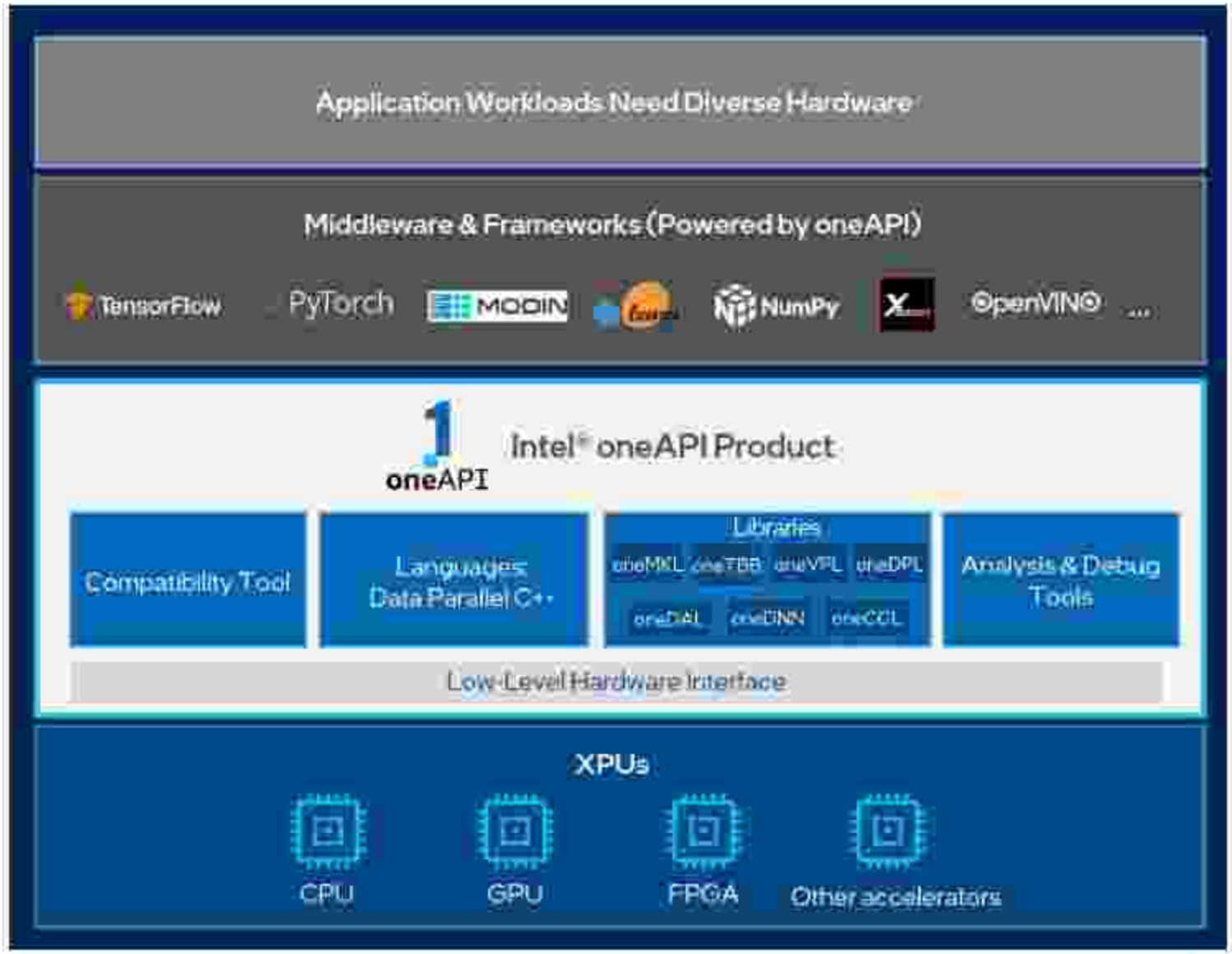
- **Freedom to Make Your Best Choice**
  - Choose the best accelerated technology the software doesn't decide for you
- **Realize all the Hardware Value**
  - Performance across CPU, GPUs, FPGAs, and other accelerators
- **Develop & Deploy Software with Peace of Mind**
  - Open industry standards provide a safe, clear path to the future
  - Compatible with existing languages and programming models including C++, Python, SYCL, OpenMP, Fortran, and MPI



Intel oneAPI AI Analytics Toolkit
Intel's oneAPI Implementation

- **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



intel EY

Intel oneAPI AI Analytics Toolkit
Intel oneAPI Software Tools for AI & Analytics

The Intel oneAPI AI Analytics Toolkit (AI Kit), a powerful set of familiar Python* tools to accelerate each step in the AI application pipeline.

Intel® oneAPI Toolkits

Intel® oneAPI AI Analytics Toolkit

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

Data Scientists, AI Researchers, ML/DL 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

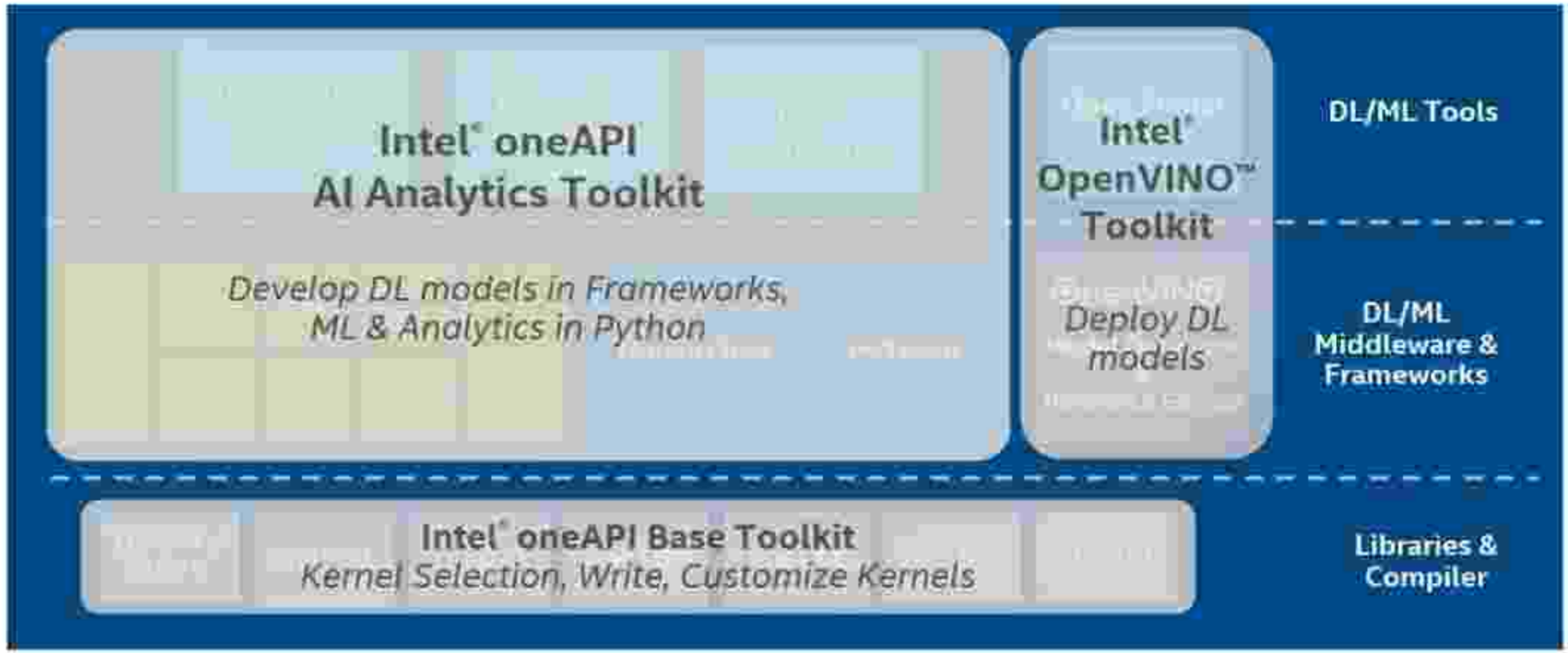
AI Application, Media & Vision Developers

OpenVINO

intel EY

Intel oneAPI AI Analytics Toolkit
AI Software Stack for Intel XPUs

Intel offers a robust software stack to maximize performance of diverse workloads



intel EY

Intel oneAPI AI Analytics Toolkit

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

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

intel EY

Intel oneAPI AI Analytics Toolkit
Key Features and Benefits – a little teaser

BENEFITS

Accelerate end-to-end AI and Data Science pipelines

Accelerate end-to-end AI and Data Science pipelines, achieve drop-in acceleration with optimized Python tools built using oneAPI libraries (i.e. oneMKL, oneDNN, oneCCL, oneDAL, and more)

High Performance

Achieve high-performance deep learning training and inference with Intel-optimized TensorFlow and PyTorch versions, and low-precision optimization with support for fp16, int8 and bfloat16

Expedite development

Expedite development using open source Intel-optimized pre-trained deep learning models for best performance via Model Zoo for Intel® Architecture (IA)

Support cross-architecture development and compute

Supports cross-architecture development (Intel® CPUs/GPUs) and compute

intel EY

Intel oneAPI AI Analytics Toolkit

High-Performance Deep Learning Using Intel Distribution of OpenVINO Toolkit

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

intel EY

Which tool should I use?
Use Both!

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

Which tool should I use?

	Intel oneAPI ai analytics toolkit	Intel distribution of Openvino toolkit
KEY VALUE	• Provides performance and easy integration across end-to-end data science pipeline • Maximum compatibility with open source FWKs and Libs	• Provides high performance and efficiency for DL inference solutions to deploy across Intel XPU architectures (cloud to edge) • Optimized package size for deployment based on memory requirements
USERS	Data Scientists, AI Researchers, DL/ML Developers	AI Application Developers, Media and Vision Developers
USE CASES	• Data Ingestion, data pre-processing, ETL operations • Model training and inference • Scaling to multi-core / multi-nodes / clusters	• Inference applications for vision, speech, text, NLP • Media streaming / encode, decode • Scale across hardware architectures – edge, cloud, datacenter, device
HARDWARE SUPPORT	• CPUs – Data center, server, workstation segments – Intel® Xeon® and Core™ processors • Future Intel Xe GPUs – Artic Sound/Ponte Vecchio	• CPU – Intel Xeon, Core and Atom processors • GPU – Intel® Processor Graphics (integrated), Intel® Iris® Xe Max • Graphics, Future Intel Xe architecture Artic Sound/Ponte Vecchio • VPU - NCS & Intel® Vision Accelerator Design Products • FPGA - Intel® Arria® 10 FPGA • GNA - Intel® Gaussian & Neural Accelerator

Getting Started with Intel oneAPI AI Analytics Toolkit

1 OVERVIEW

- Visit Intel® Distribution of oneAPI AI Analytics Toolkit for more details and up-to-date product information
- Release Notes

2 LEARNING

- Intel AI Blog site
- Webinars & articles
- Choose hardware option with Performance Benchmarks

3 DEVELOPMENT

4 DEPLOYMENT

5 CONSUME

MLOps + Intel oneAPI AI Analytics Toolkit

+ cnvrg.io

in the next slide we will focus on MLOps methodology and cnvrg.io tool

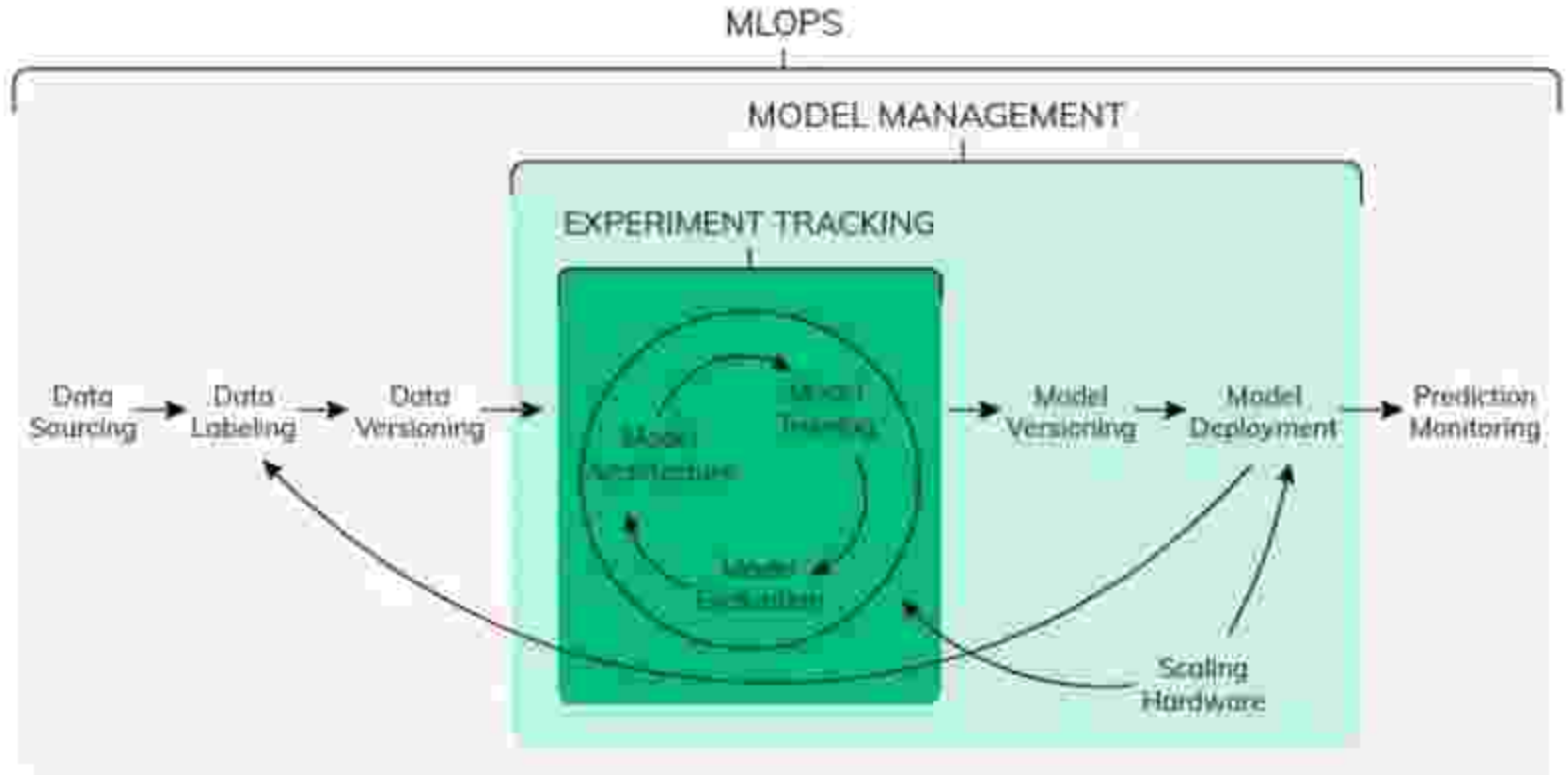
intel EY

MLOps
What it is, Why it matters and how to implement it

MLOps is a set of practices for collaboration and communication between data scientists and operations professionals to help manage production Machine Learning (or Deep Learning) lifecycle

The key phases of MLOps are:

1.
2. Data gathering
3. Data analysis
4. Data Transformation/Preparation
5. Model training & development
6. Model validation
7. Model serving
8. Model monitoring
9. Model re-training



intel EY

MLOps
Key Features

EXPERIMENTAL IN NATURE

- Implementing various features (hyperparameters, parameters, and models)
- Track and manage the data and the code base for reproducible results

HYBRID TEAM COMPOSITION

- Team usually includes data scientists or ML researchers, who focus on exploratory data analysis, model development, and experimentation

TESTING

- Testing an ML system involves model validation, model training, unit testing and integration testing

AUTOMATED DEPLOYMENT

- Can't just deploy an offline-trained ML model as a prediction service
- Need a multi-step pipeline to automatically retrain and deploy a model

PLANNING

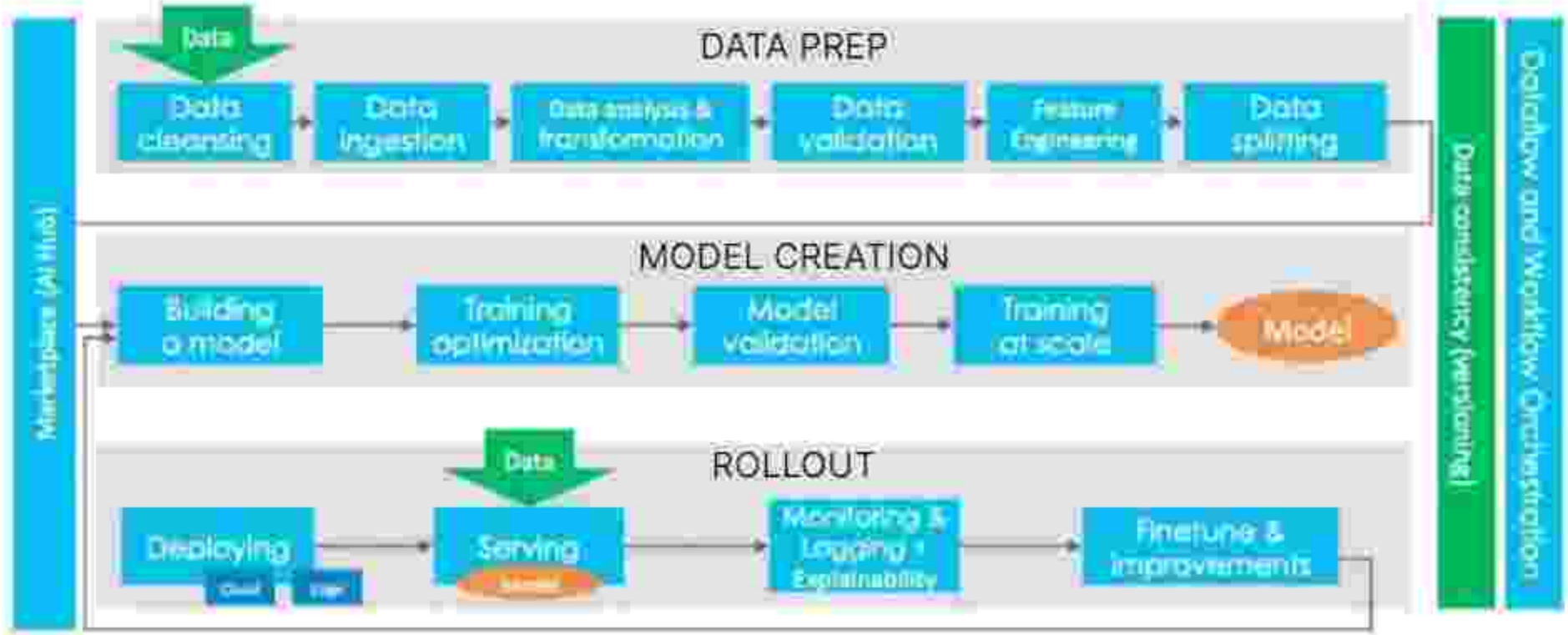
- Planning needs due to degraded system performance in production due to suboptimal coding, but also due to constantly changing data profiles

MONITORING

- Models in production and summary statistics of data that build model need to be monitored
- Need notifications or a roll-back process when values deviate from your expectations

MLOps
Focus on Continuous Integration, Continuous Deployment & ContinuousTesting

- **Continuous Integration** (CI) concerns testing and validating both code and components, as well as data, data schemas and models
- **Continuous Deployment** (CD) concerns a system (an ML training pipeline) that should automatically distribute another service (model prediction service) or roll back changes to a model
- **Continuous Testing** (CT) concerns the automatically retraining and serving the models

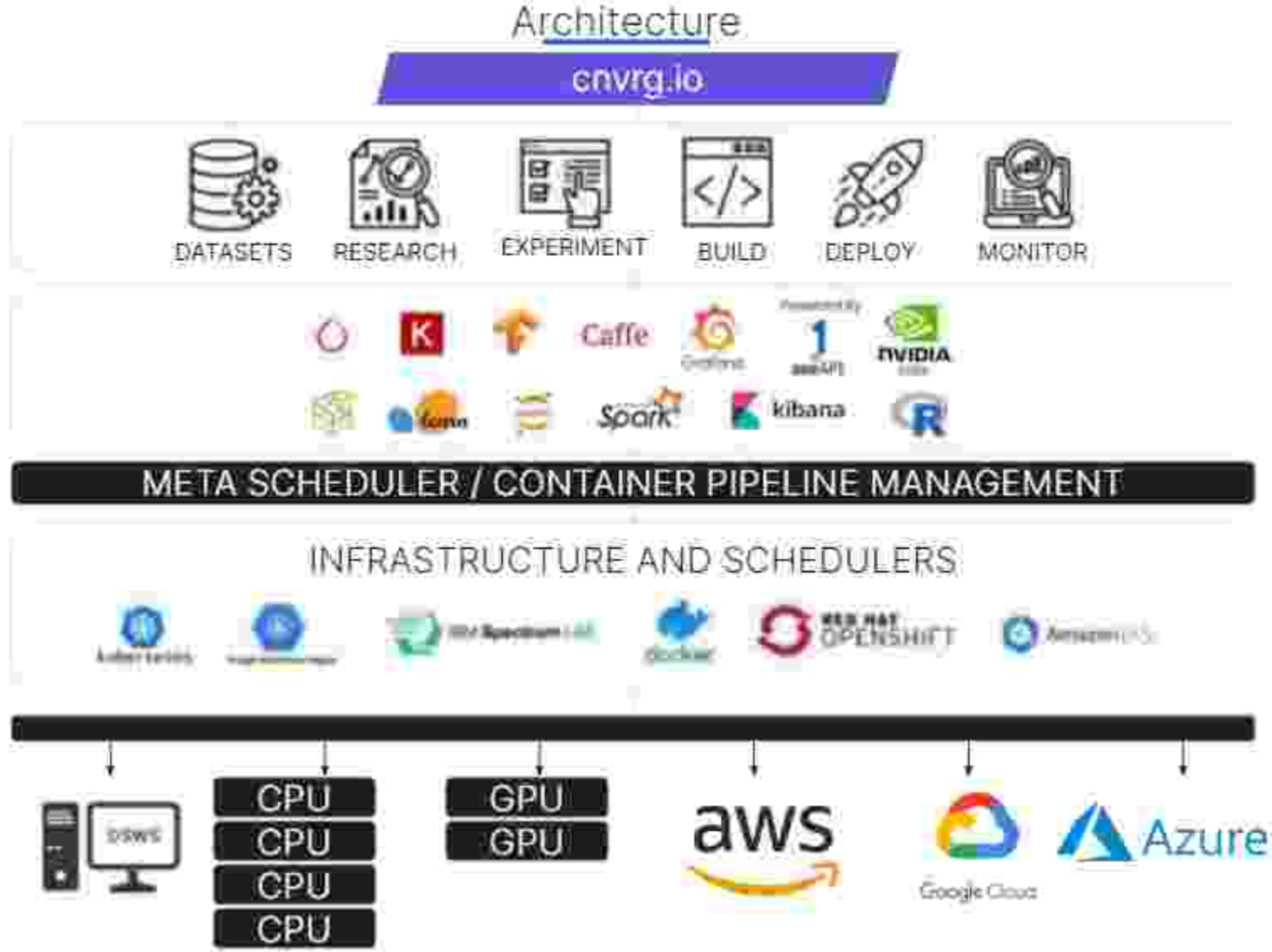


intel EY

cnvrg.io
A Full Stack Machine Learning Operating System

- cnvrg.io was built by data scientists, for data scientists to streamline the machine learning process in order to help teams to:
  - manage, build and automate machine learning
  - bridge science and engineering

- cnvrg.io is the world's most flexible end-to-end machine learning operating system built to empower AI developers to build high impact models, faster, on any AI infrastructure. With cnvrg.io, AI developers are given the freedom to run AI workloads where it is faster and most cost effective, in half the time.



intel EY

cnvrg.io
Key Features

Mlops utilization dashboard

- Improve visibility
- Increase infrastructure utilization by up to 80% with advanced resource management and visibility across all ML runs
- Monitor utilization, properly size the compute components and visualize who is using what, with extensive visualization of cluster utilization

Open and flexible platform

- Users can develop their own models, run experiments of their choice and modify behavior and code with Jupyter notebooks (or RStudio, VSCode)
- New tools and utilities can be easily integrated into the platform

Heterogeneous compute pipelines

- Launch and manage end-to-end heterogeneous ML pipelines where each component or stage (in a single pipeline) can run on a different compute architecture that is optimized for the specific use-case

High-performance infrastructure

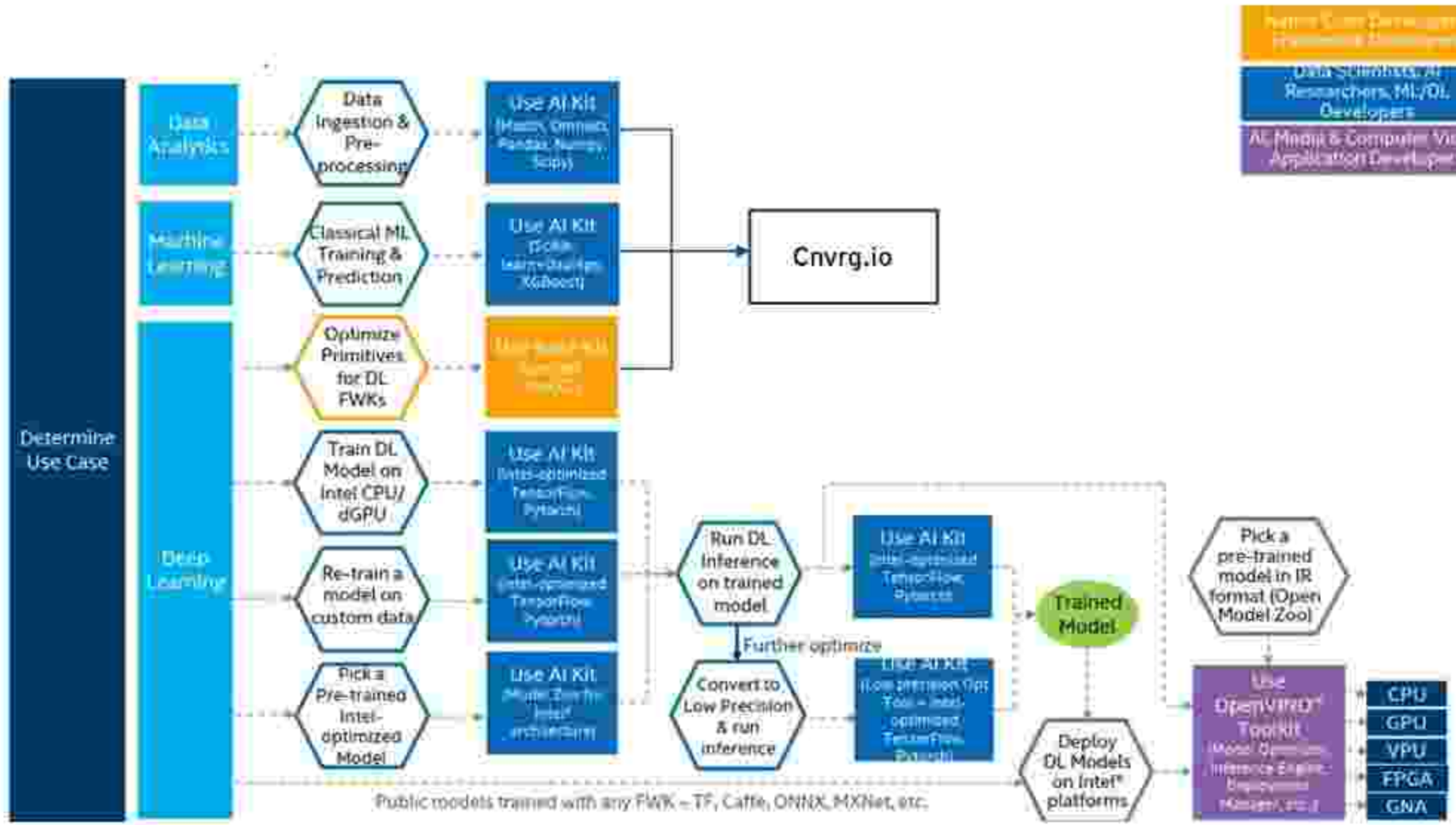
- Better utilization of your infrastructure
- Hardware vendor bring the highest performance for both your AI training and Inferencing workloads.

intel EY

cnvrg.io
Benefits

intel EY

AI Development & Deployment Workflow



intel EY

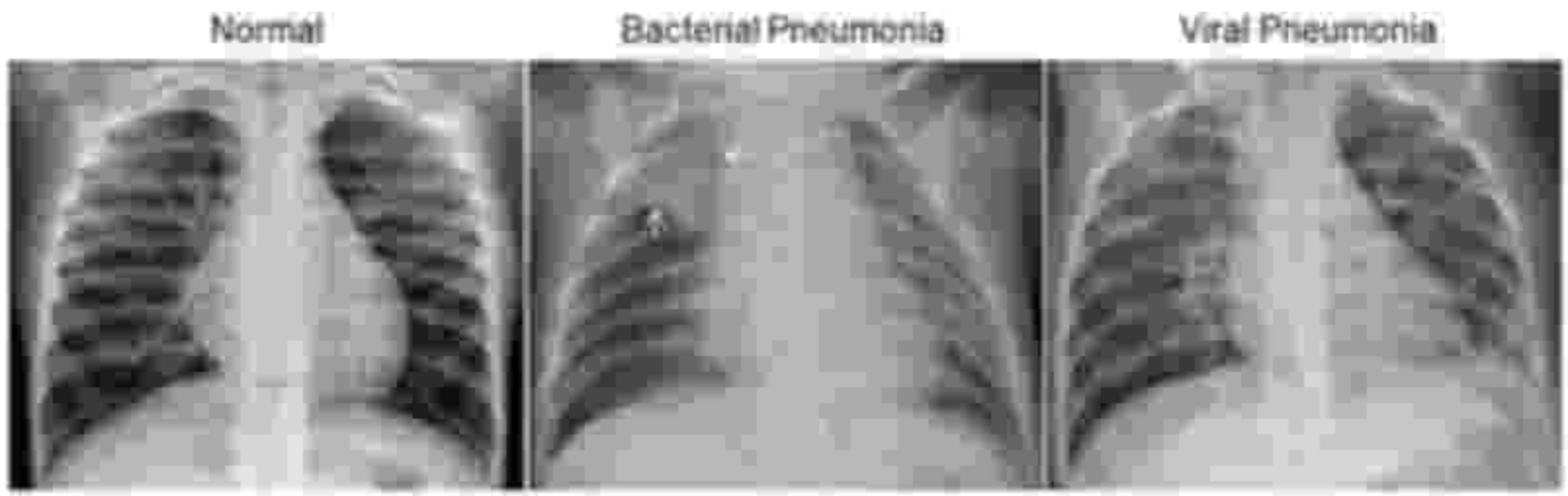
Demo
cnvrg.io & MLOps & oneAPI AI Analytics Toolkit in action

Computer Vision Model Training with ResNet50 and the Chest X-ray Dataset

Main steps

1. Set up the Compute Resources
2. Load Data
3. Catch Data
4. Build an ML Pipeline with Cached Data
5. Building the Chest X-ray Flow: Setting Data
6. Building the Chest X-ray Flow: Setting Training Model: ResNet50
7. Define the Compute Resource for ResNet50
8. Tracking and Monitoring Results

X ray classification



intel EY

Call to Actions

Download tools, setup the environment and try the Demo

Intel oneAPI AI Analytics Toolkit

Download the Intel® oneAPI AI Analytics Toolkit

Setup cnvrg.io

Downlad an configure cnvrg

Try a Demo

https://github.com/oneapi-src/oneAPI-samples/tree/master/AI-and-Analytics

intel EY

THANKS AND SEE YOU SOON!

Contact

s https://www.linkedin.com/in/antimo-musone/

antimo.musone@it.ey.com