Purdue University

Introduction to MLOps and LLM Pipelines

November 3rd, 2023

intel.

About our Speakers

Eduardo Alvarez – Senior AI Solutions Engineer, Intel

Eduardo Alvarez is a Senior AI Solutions Engineer at Intel, specializing in architecting AI/ML solutions, MLOps, and deep learning. With a background in the energy tech startup space, he managed a team focused on delivering SaaS applications for subsurface AI in hydrocarbon and renewable energy exploration and production. Now at Intel, Eduardo collaborates across technical teams, designing impactful solutions highlighting the Intel software and hardware stack's influence on Deep Learning and GenAI workloads. He is the author of Intel's MLOPs Professional Developer course, where he brings his expertise in the production deployments of AI tools to a broad audience of student and enterprise developers.

Agenda

- Introducing Intel® Certified Developer Program – MLOps Professional Certification
- Role of MLOps in Production AI/ML Solutions
- Considerations for Performance and Optimization
- Future of Operational AI, Sustainability, and Ethical AI
- How to get certified

Introducing Intel® Certified Developer – MLOps Professional

✓ Develop a marketable skillset focused on incorporating compute awareness into the AI solution design process

✓ Increase your hireability through the creation of a project that showcases competency in designing and implementing performant AI solutions

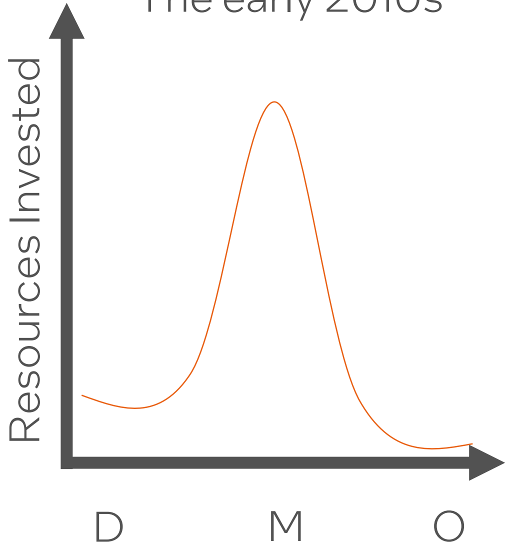
MLOps Professional Training Package

- 26 Video Lessons
- 8 Hands on Labs using the Intel® Developer Cloud
- 1 capstone project
- Office Hours
- Practice Certification Exam

Why Focus on Operational AI?

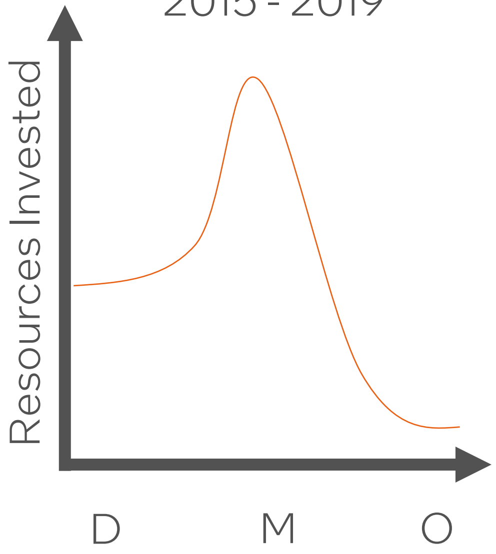
Operational AI Engineering Evolution Model = **D**ata **M**odel **O**perations

Traditional ML Era
The early 2010s



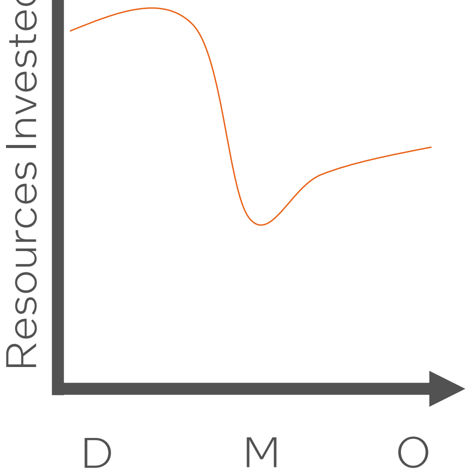
Heavy experimentation with forests, trees, linear regression, etc.

Neural Net Mania
2015 - 2019



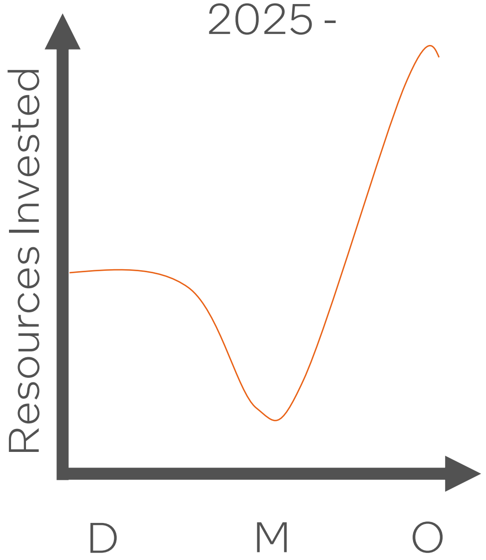
Stronger focus on data to train data-hungry neural networks (FFNN, LSTMs, RNNs, CNNs, etc.) with significant focus on model development and optimization.

GenAI Revolution
2020 - Current



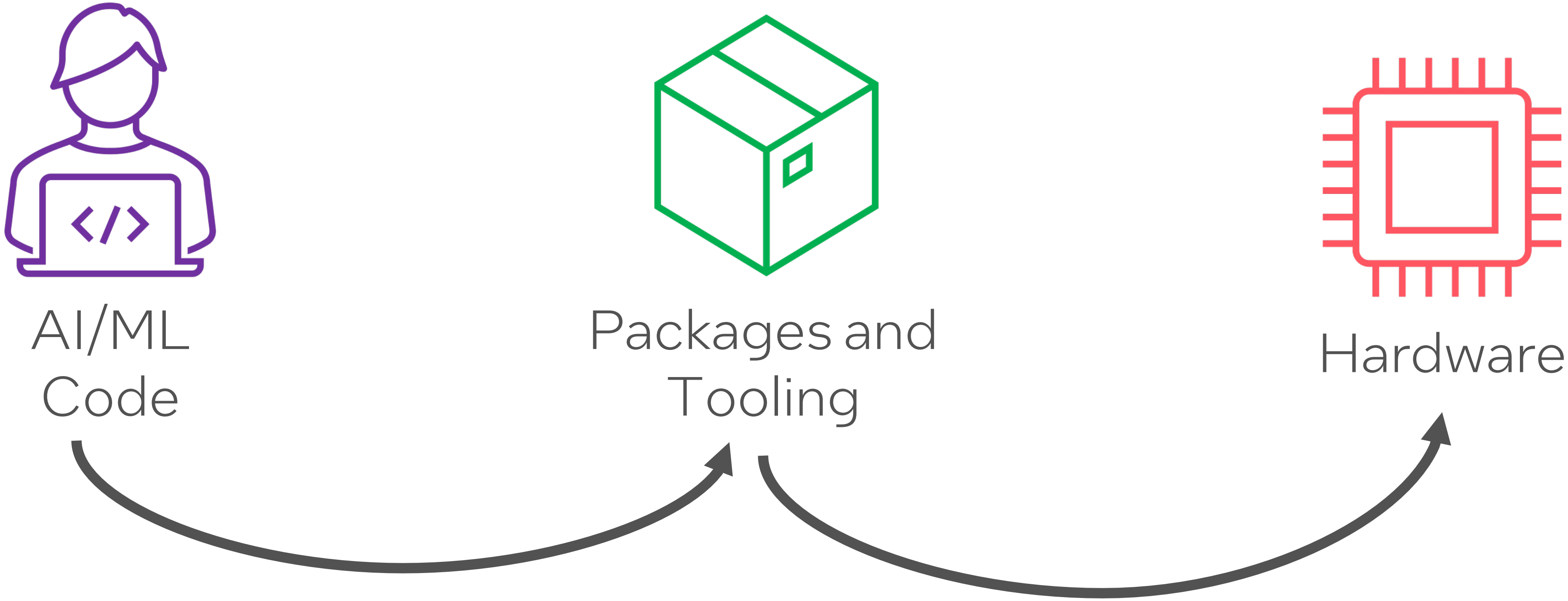
Open-source foundational models reduce focus on pre-training and architecture design. Data becomes a major focus in training LLMs and VT and for fine-tuning and RAG applications.

Democratized AI
2025 -



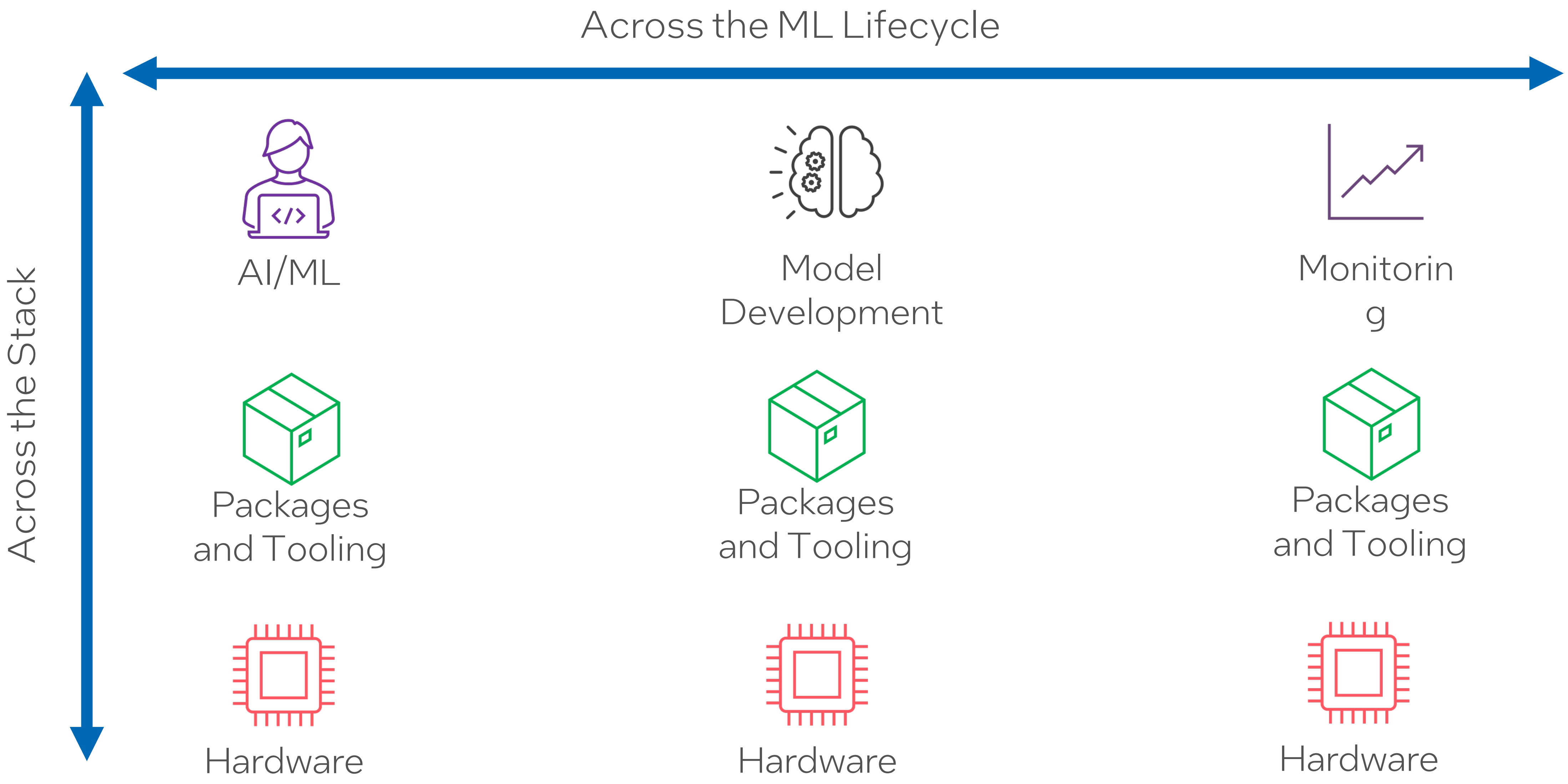
New high-quality data sources are scarce. Transformers technology has peaked, and the upper bound becomes operational optimizations driven by HW/SW.

This is how we teach Operational AI



With careful consideration for compute and software optimizations across the ML Lifecycle and the foundational knowledge required to architect and implement solutions in production.

Optimizing the AI/ML Lifecycle Full Stack

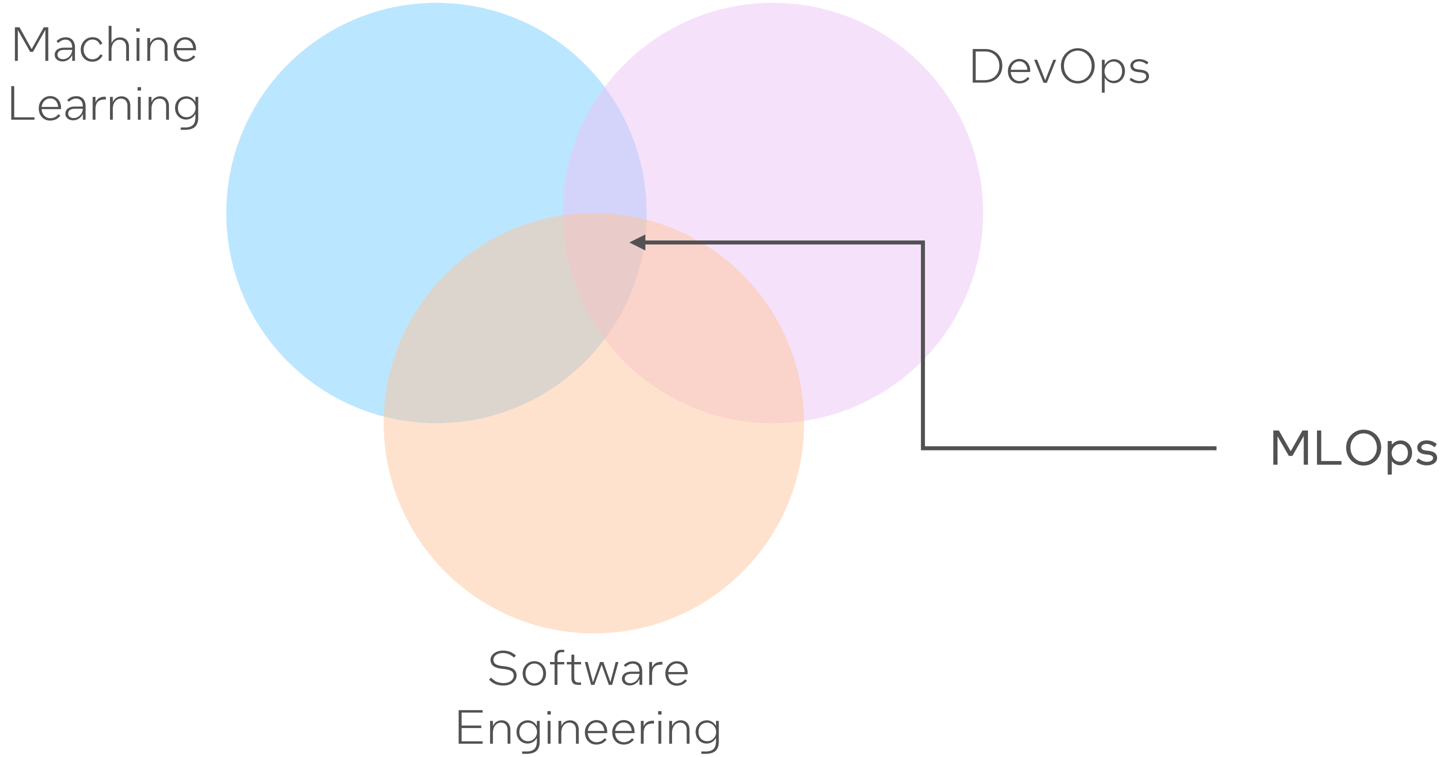


Understanding MLOps

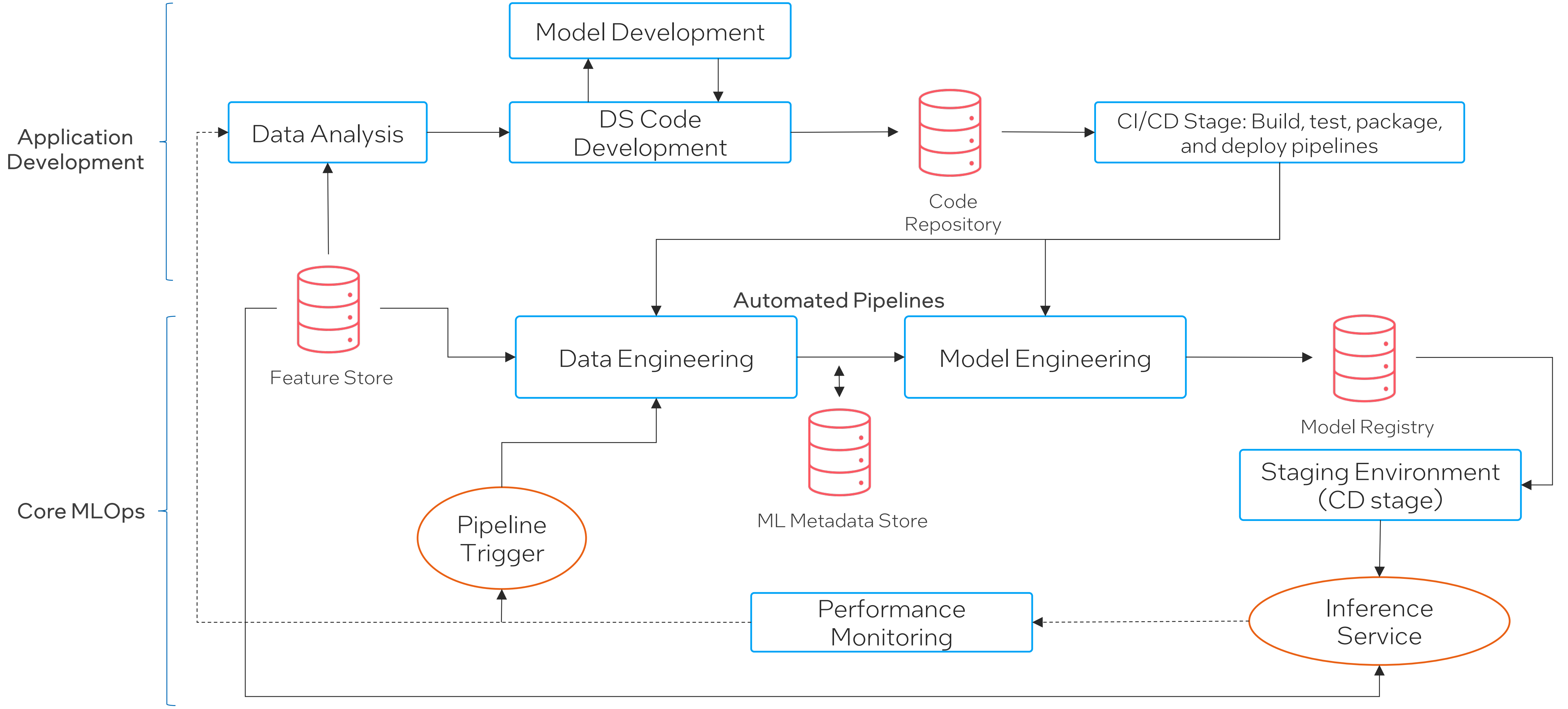
How It Enhances Application Development and Adds Value

What is MLOps?

MLOps, short for Machine Learning Operations, is a practice that focuses on the integration of machine learning models into operational processes to ensure the reliability, scalability, and maintainability of AI systems. It combines DevOps, software engineering, and machine learning principles to streamline and automate the end-to-end machine learning lifecycle.



MLOps Architecture



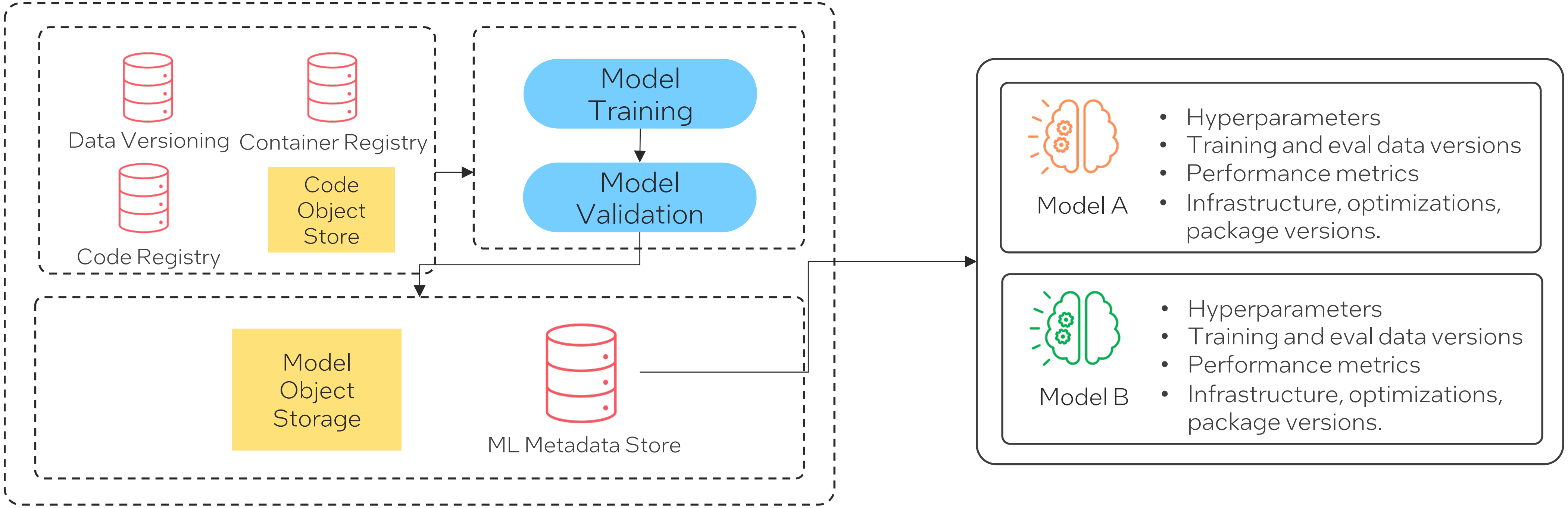
Initial Approach to Model Management

- Novices often save models as simple files without tracking their evolution.
- Basic tracking might include using spreadsheets for model names, parameters, and metrics.
- As AI initiatives expand, model management becomes more complex with frequent iterations and larger datasets.
- MLOps presents advanced practices to transition from basic to scalable and efficient AI system management.

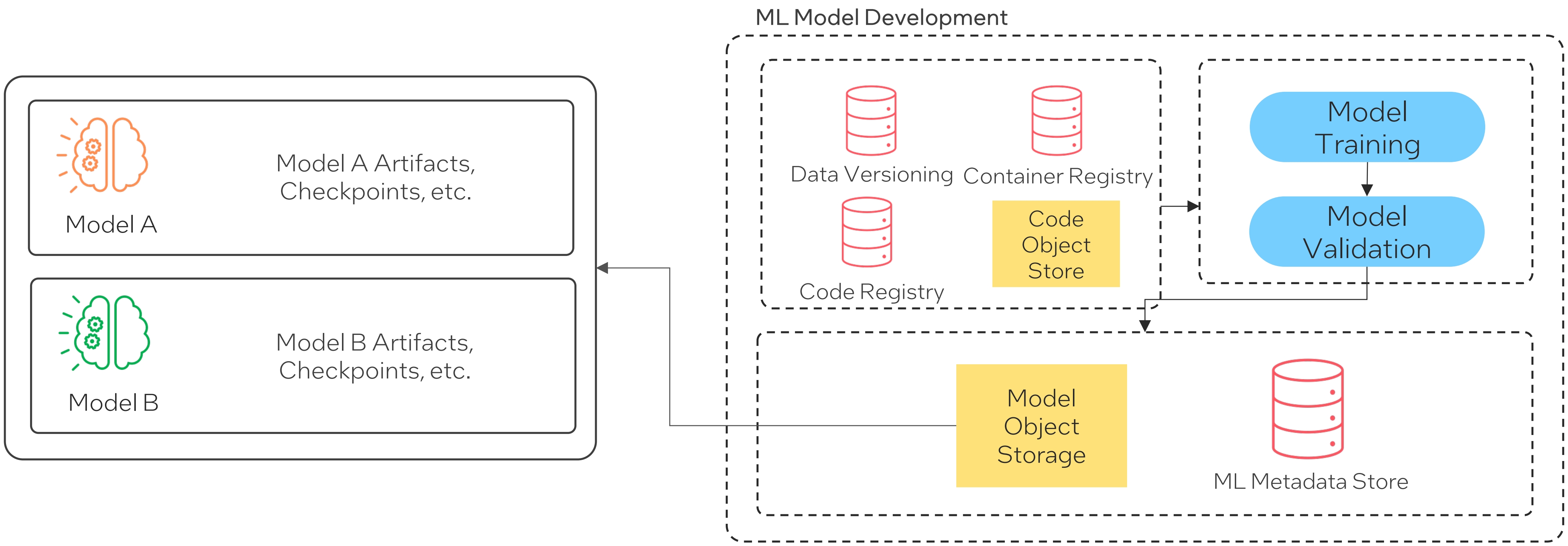
model_development_tracker.xlxs

Name	Accuracy	Data Version	Path
Model A	86%	1.0	C://home/my_models/modelA.h5
Model B	67%	1.4	C://home/my_models/modelB.h5
Model C	78%	3.0	C://home/my_models/modelC.h5

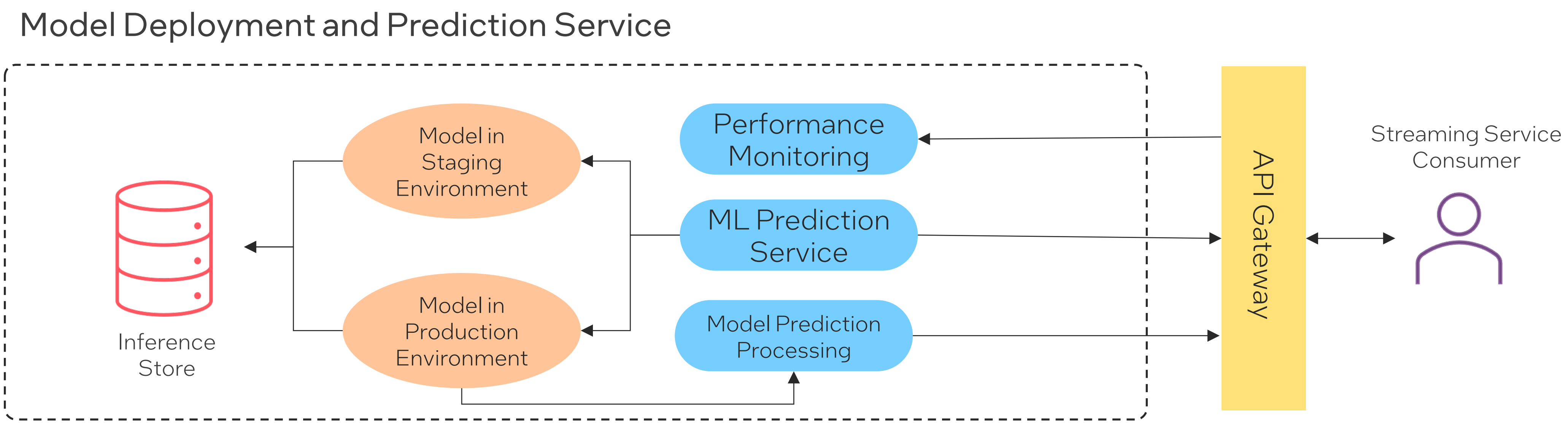
Model Metadata



Storing Models



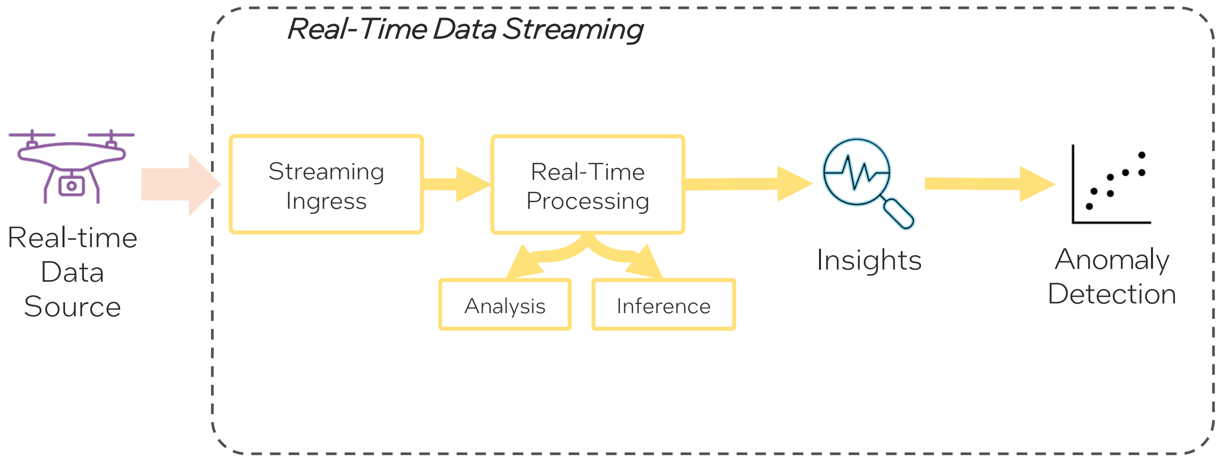
ML Prediction Service



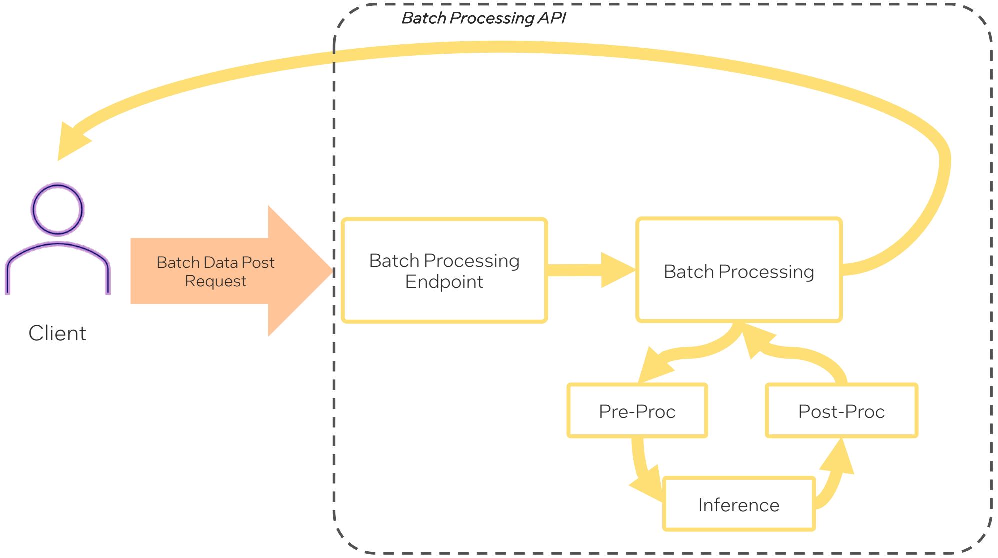
Inference Modes

The two primary inference modes are batch prediction and real-time/online prediction. Batch prediction processes multiple inputs together, which is beneficial for large datasets or cost-saving. Real-time prediction focuses on low-latency responses to individual data points.

Real-time or "Online" Inference

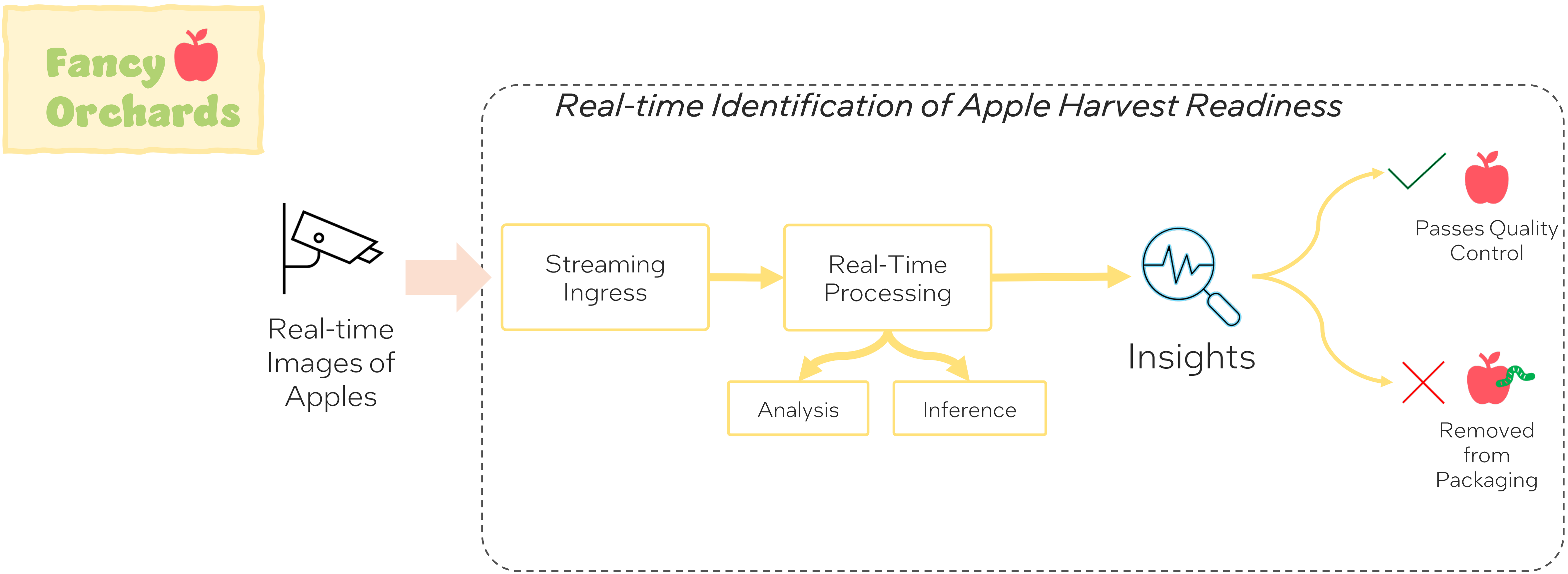


Batch Inference



Fancy Orchards: Real-time Inference

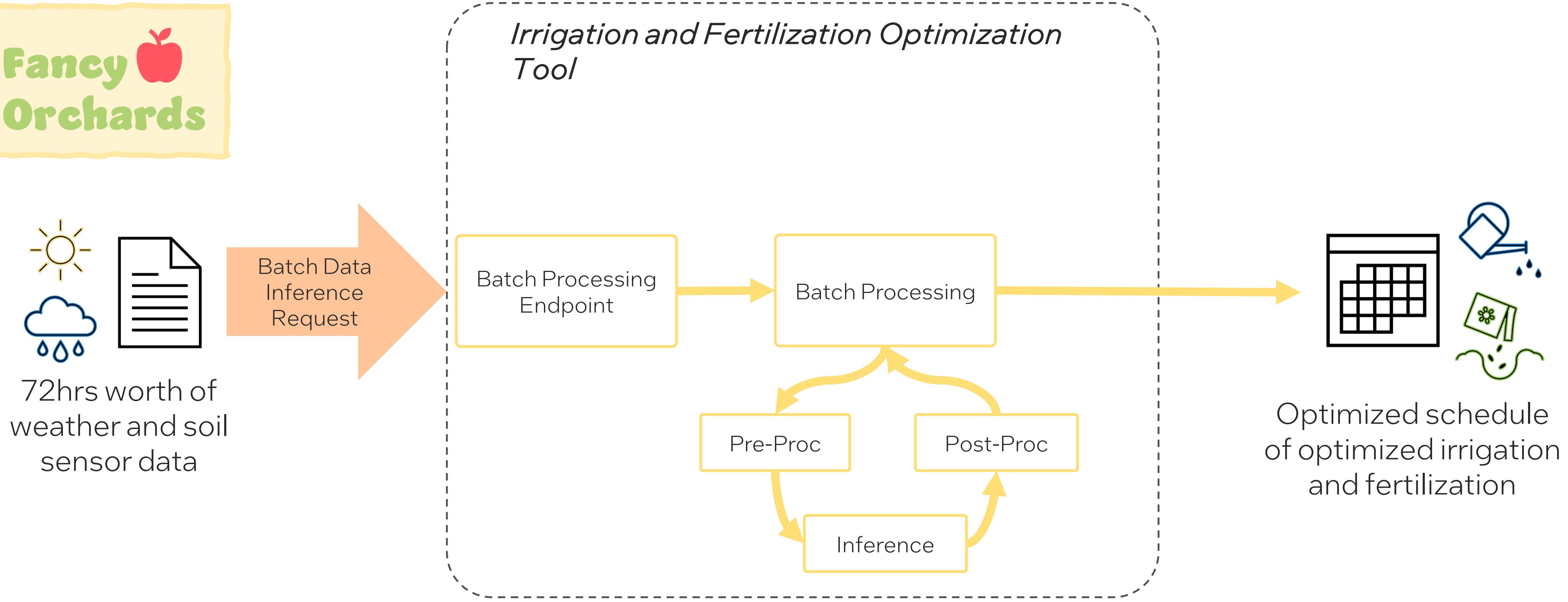
- Fancy Orchards" uses real-time inference for production line QA/QC processes.
- Camera sensors continuously capture apple images, which the AI system instantly analyzes for defects.
- Swift real-time decisions enable rapid detection and removal of defective apples.



Fancy Orchards: Batch Inference

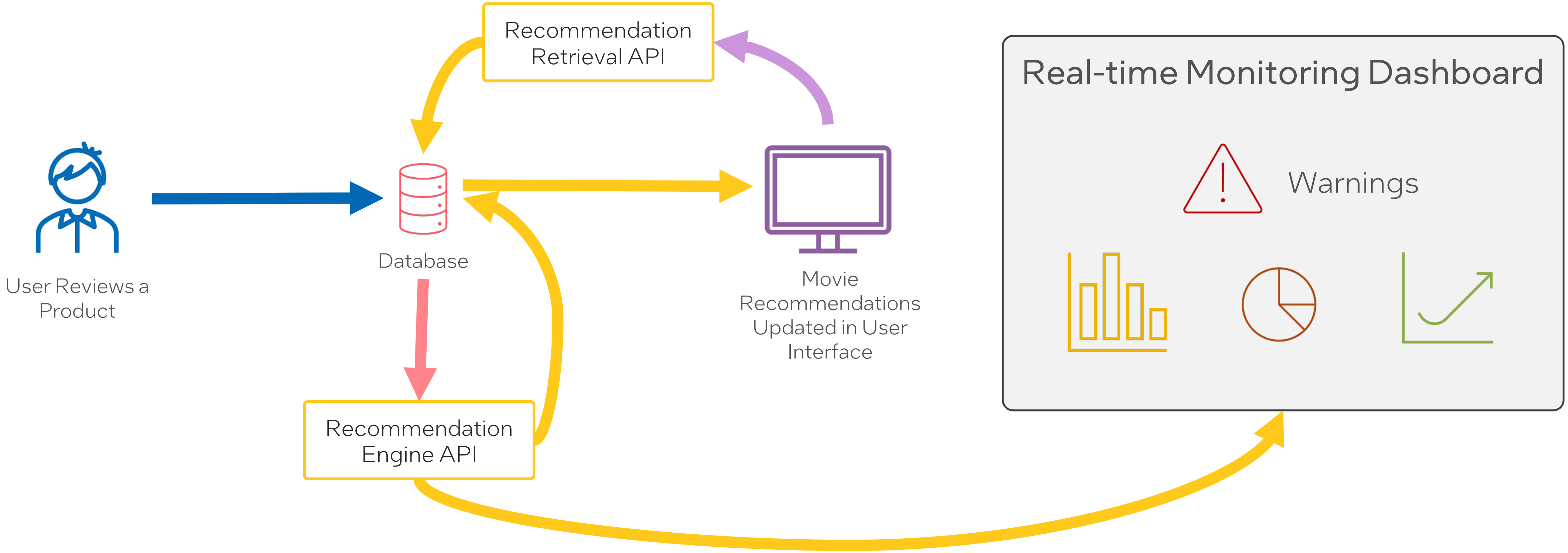
- "Fancy Orchards" uses batch inference to enhance apple cultivation based on historical weather patterns.
- The model processes chunks of new weather data for water and fertilizer recommendations.
- This batch processing optimizes resources by processing data in larger, more efficient sets.

Fancy Orchards



Real-time Monitoring for AI in Production

Monitoring AI workloads provides continuous insights into performance, resource utilization, and system health. Utilizing monitoring tools, metrics such as GPU utilization, memory consumption, and latency can be tracked, with alerts set up for a proactive response.



Runtime data is logged, processed, and displayed on the telemetry dashboard.

Hands-on #1 - Setup

1. Navigate to cloud.intel.com
2. In the Training and Workshops Section, select "Launch Jupyter Lab"
3. Clone Repository: *git clone https://github.com/intel/certified-developer.git*
4. Navigate 01_model_development_basics in the workshops folder and open model_development_basics.ipynb
5. Start workshop!

Common Bottlenecks in AI Workloads and Opportunities for Optimization

Exploring Optimizations for AI Workloads

Performance Bottlenecks

92% Thread Utilization

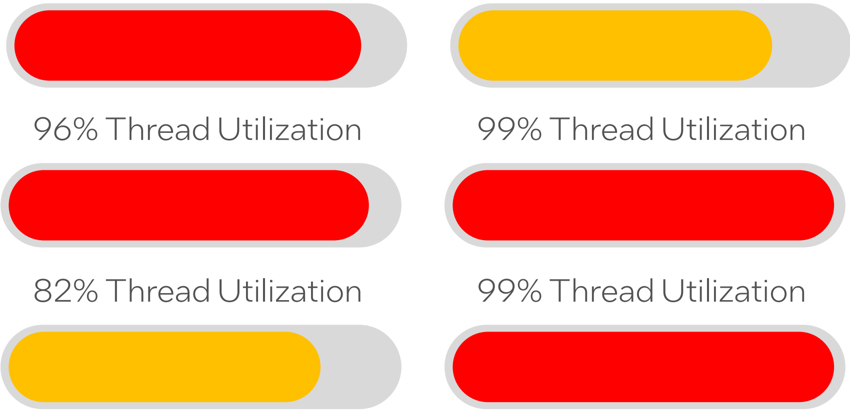
85% Thread Utilization

96% Thread Utilization

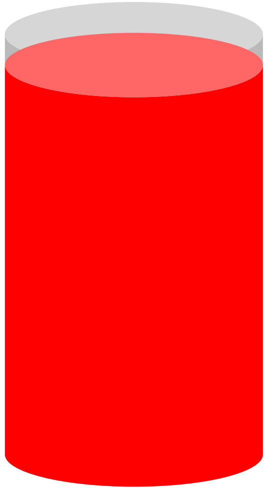
99% Thread Utilization

82% Thread Utilization

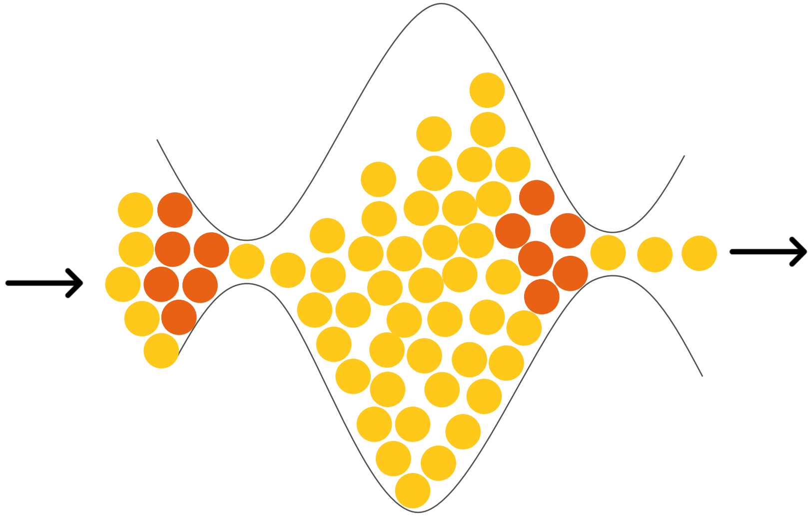
99% Thread Utilization



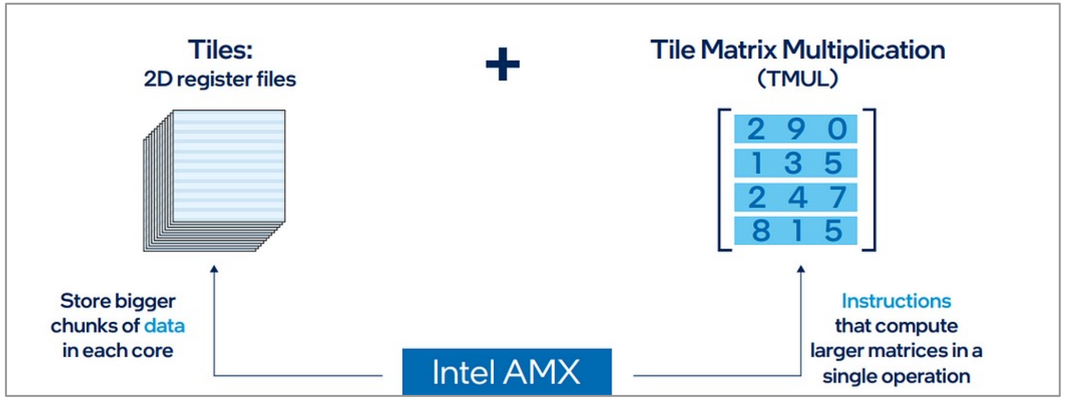
97%% Memory Utilization



Inputs/Outputs are Bottlenecked



Software-Hardware Co-Design



Advanced Matrix Extensions, available in Intel 4th Generation Xeon processors, have dedicated instructions to improve memory management and matrix operations for deep learning workloads.

SOFTWARE- HARDWARE
CO-DESIGN EXAMPLE

```
with torch.backends.mkldnn.verbose(torch.backends.mkldnn.VERBOSE_ON):
    with torch.cpu.amp.autocast():
        model(input)
```

Upstreamed optimizations from intel into PyTorch, enabling Advanced Matrix Extensions optimizations during model training with auto mixed precision.

Engineer Data | Create Machine Learning & Deep Learning Models | Deploy

Container Repository
oneContainer

oneAPI powered
AI Reference Kits

MLOps
Cnvrg.io

Developer Sandbox
Intel® Developer Cloud

Annotation/Training/Optimization
Intel® GETi

Connect AI to Big Data — Apache Spark — BigDL (previously "Analytics Zoo")

Accelerate End-to-End Data Science and AI — AI Analytics Toolkit

Data Analytics Scale

MODIN
SciPy
pandas
NumPy

Optimized Frameworks and Middleware

TensorFlow
PyTorch
mxnet
PaddlePaddle
scikit learn
ONNX
LightGBM
XGBoost
CatBoost

Optimize Models

Automate Model Tuning AutoML

Automate Low-Precision Optimization

SigOpt

Intel Neural Compressor

OpenVINO™ Toolkit

Write Once

Deploy Auto-Optimized

Anywhere

w/ Intel Optimizations

SYCLomatic | oneDAL | oneDNN | oneCCL | oneMKL | SynapseAI™

intel ATOM | intel CORE | intel XEON | intel ARC GRAPHICS | intel DATA CENTER GPU | intel habana

Note: not all components are necessarily compatible with all other components in other layers

Example: Intel® Optimization for PyTorch

ECOSYSTEM: torchvision | TorchServe | Hugging Face | PyTorch Lightning | ...

FRAMEWORKS: PyTorch | Intel® Extension for PyTorch*

LIBRARIES: oneAPI oneDNN | oneAPI oneCCL

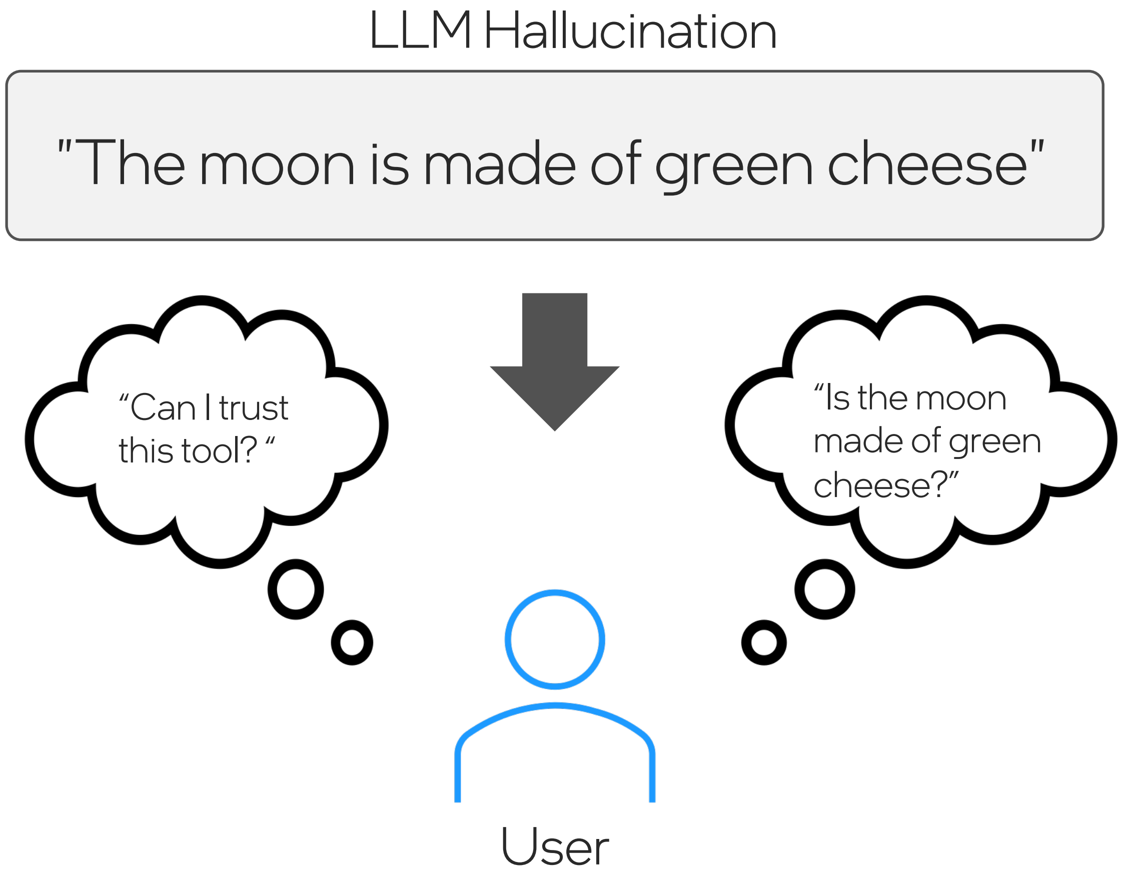
24

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

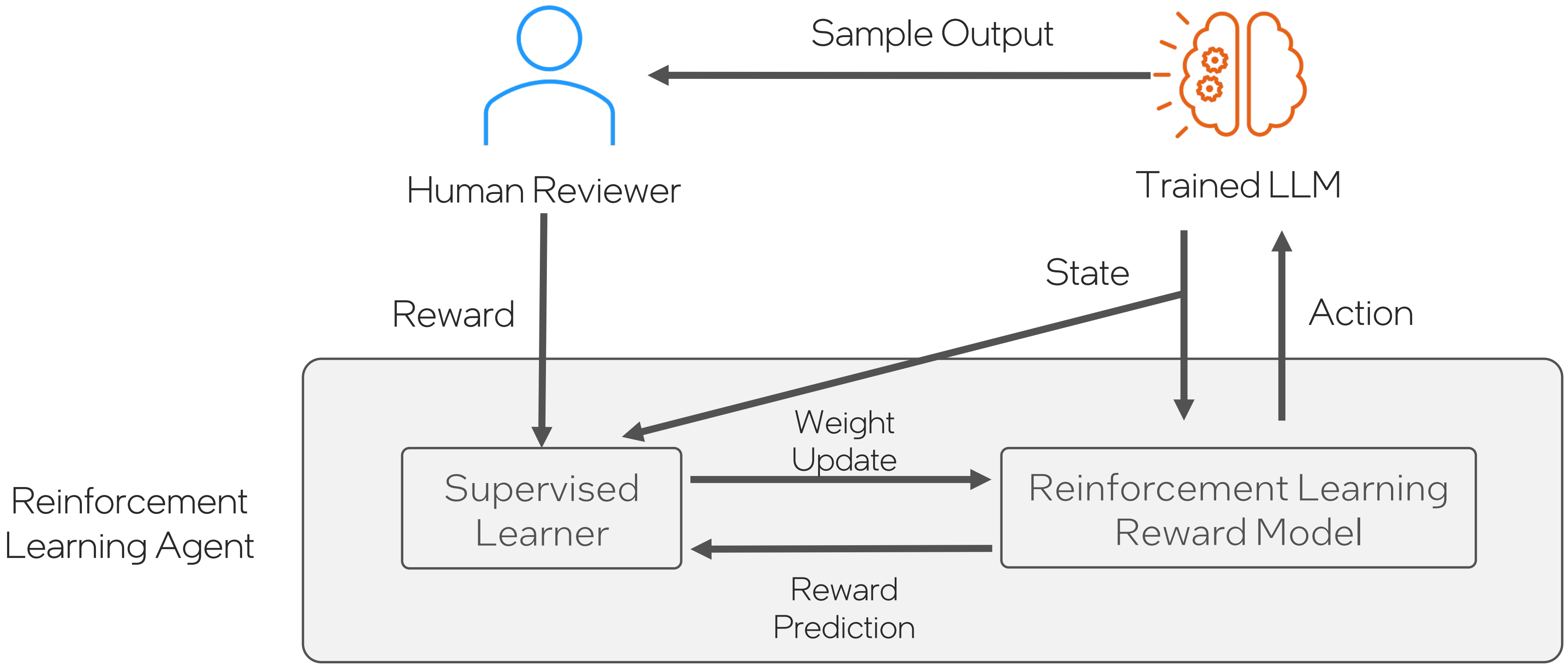
Future of Operational AI

Considerations for LLMs

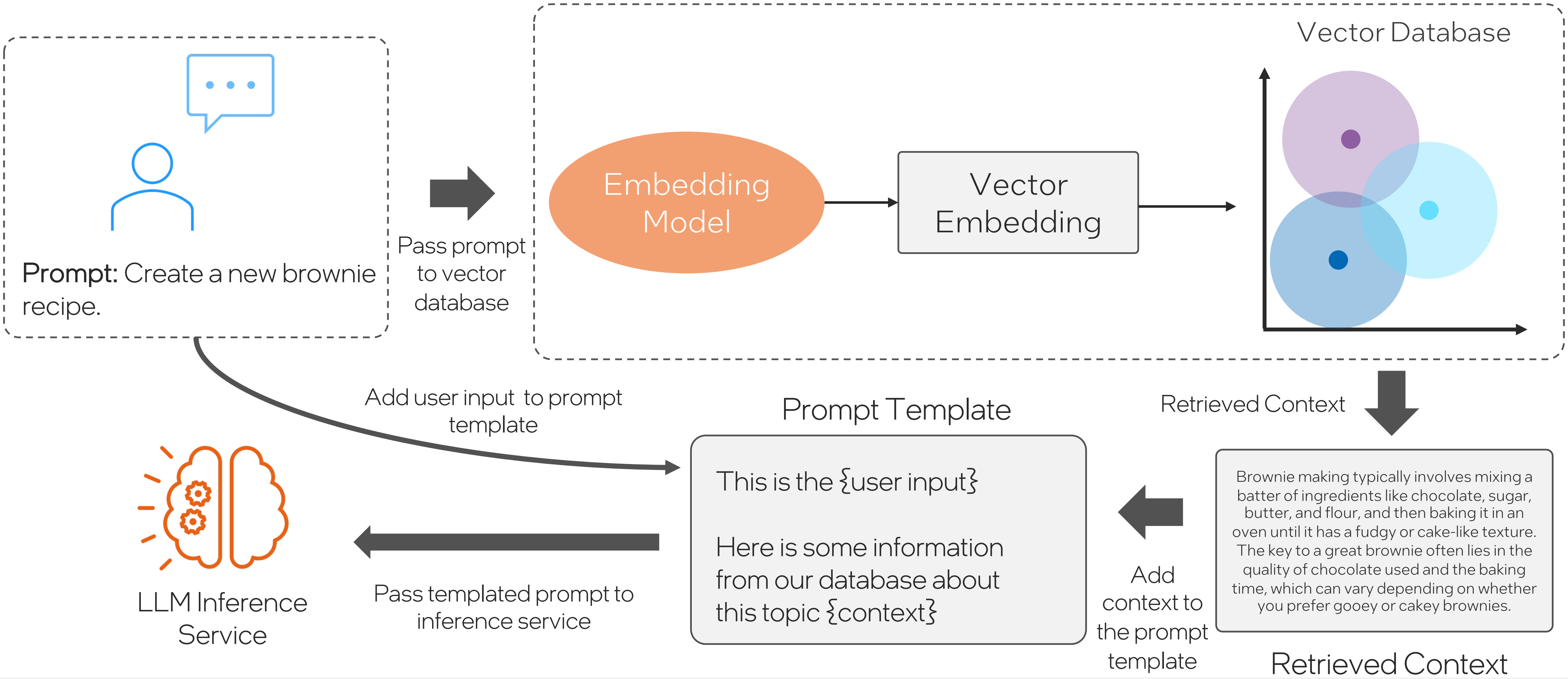
Understanding Hallucinations in Generative AI



Role of RLHF in Addressing Hallucinations



Advanced LLM Chains & Retrieval Augmented Generation and In-Context Learning



Hands-on #2 - Setup

1. Navigate to cloud.intel.com
2. In the Training and Workshops Section, select “Launch Jupyter Lab”
3. Clone Repository: *git clone https://github.com/intel/certified-developer.git*
4. Navigate 02_llm_pipelines in the workshops folder and open llm_pipelines.ipynb
5. Start workshop!

Impact of MLOps on Sustainability & Ethical AI

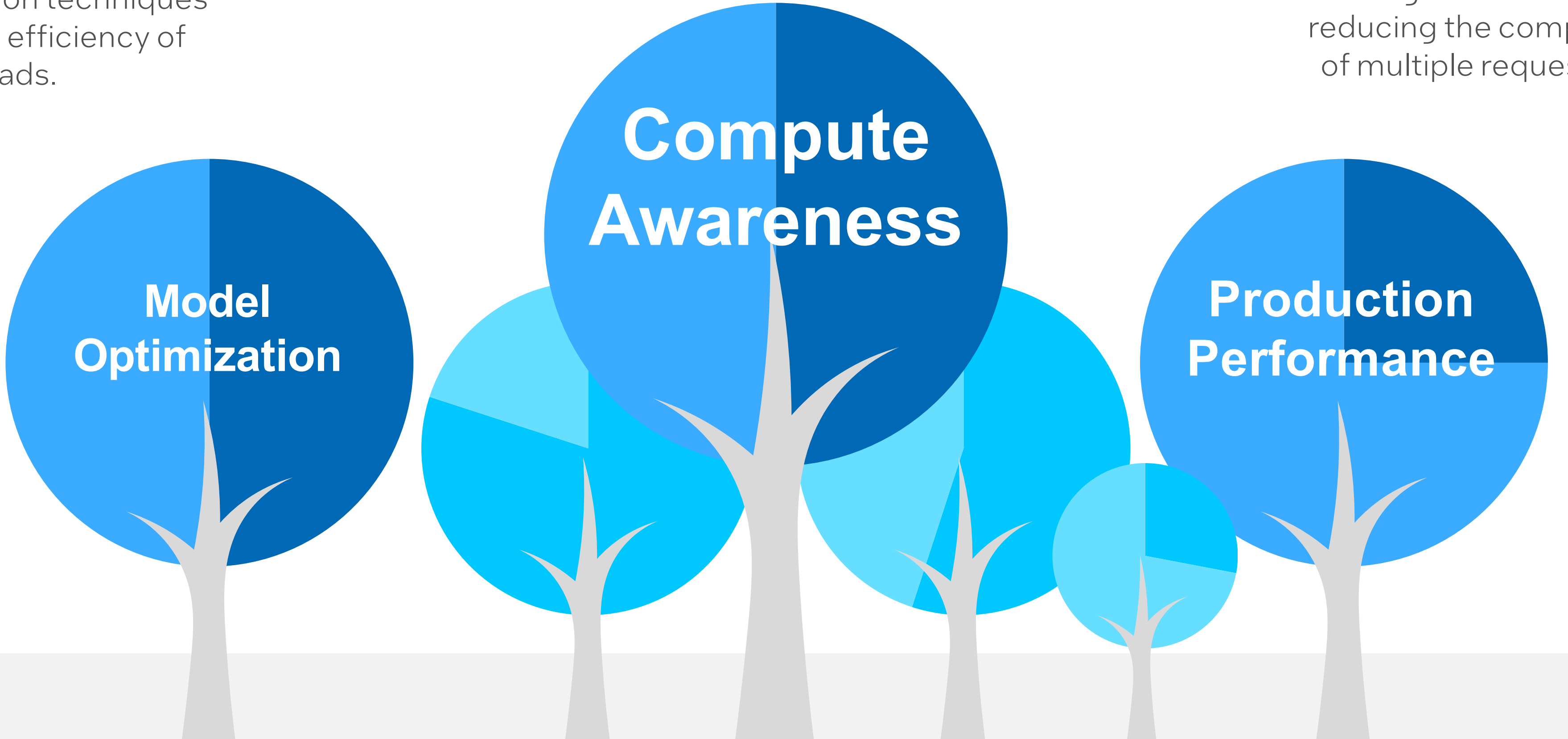
Role of MLOps in Responsible AI

How does MLOps Impact Sustainability of a Solution?

Using compute-aware principles, AI/ML developers can help in the right-sizing of computational resources to reduce the idle time of servers. Balance hardware and software from edge to cloud—utilizing a heterogenous infrastructure with a combination of AI computing chipsets that meet specific application needs can ensure that compute is optimized, and energy used efficiently

Optimized models with built-in Intel® AI Engines like Intel® Advanced Matrix Extensions (Intel® AMX) or model compression techniques improve energy efficiency of workloads.

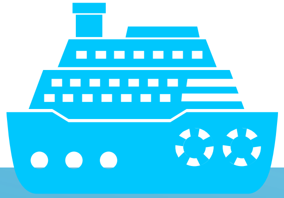
Ensuring models are performant in production so that users get the right answer the first time, reducing the computational load of multiple requests to servers.



How does MLOps Impact Ethical AI?

Clear benefits

- Standardizes model development for ethical compliance.
- Enables traceability through version control.
- Automated checks for bias and ethical behavior.



Profound Impacts

- Enables deep, scientific inquiry into AI ethics.
- Facilitates ethical 'peer review' through collaboration.
- Allows real-time ethical evaluation metrics.
- Promotes a culture of continuous ethical reflection.
- Makes ethics an integral part of AI, not just a checklist.

intel
certified
developer

MLOps Professional

Scan to Register
for the Course