AI DevSummit Workshop

Building AI Super Reference Kits

Presenters: Eduardo Alvarez, Rachel Oberman, and Paula Ramos

In this workshop:

- Learn about using Intel AI Reference Kits to prototype AI Solutions for real-world applications
- We will walk through the “Pharmaceutical Manufacturing Business” Super Kit codebase
- Participants will learn about the Intel optimizations in each component through a live hands-on pair programming session.
- Learn how each application is modularized to support an API endpoint
- The various application components will be deployed using DevOps tools like Docker and Docker Compose

What are AI Reference Kits?

In collaboration with Accenture*, Intel offers a series of downloadable AI reference kits to the open-source community to help enterprises accelerate their digital transformation journey. These kits are built upon the AI application tools that Intel provides to data scientists and developers.

https://www.intel.com/content/www/us/en/developer/topic-technology/artificial-intelligence/reference-kit-library.html

So, what is an AI Super Reference Kit?

An AI Super Reference Kit is born when we combine multiple kits for use in a particular vertical, domain, or use case. By innovating various components of a legacy process, AI Super Kits can have meaningful impacts across a business in a synergistic way.

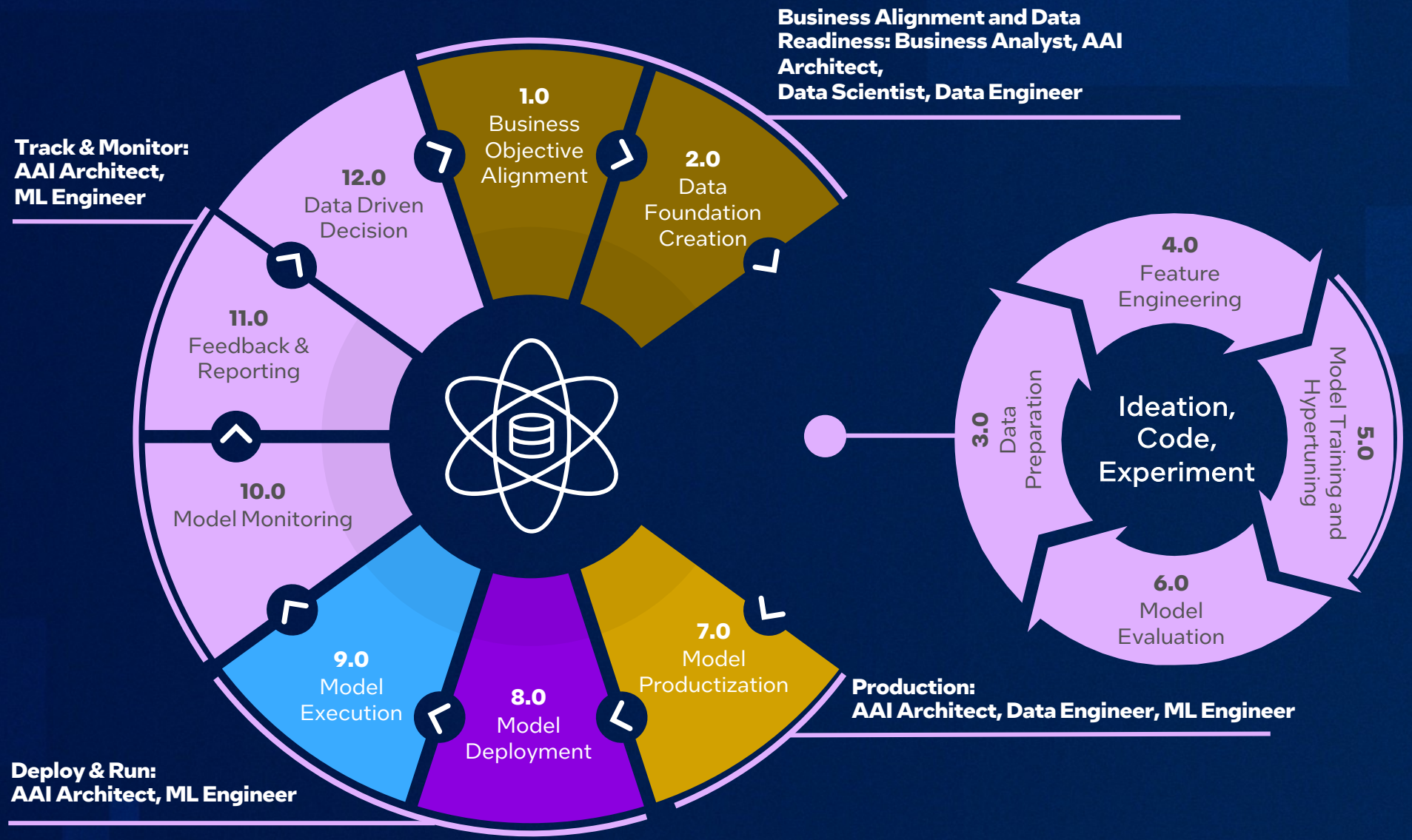
- Quickly prototype unified AI solutions with multiple reference kits
- Leverage sample architecture to plan for scale with essential DevOps tools
- Build with pre-baked Intel AI optimizations

Construct a Super Kit that uses customer churn prediction, demand forecasting, and delivery forecasting to automate/optimize customer analysis for a business.

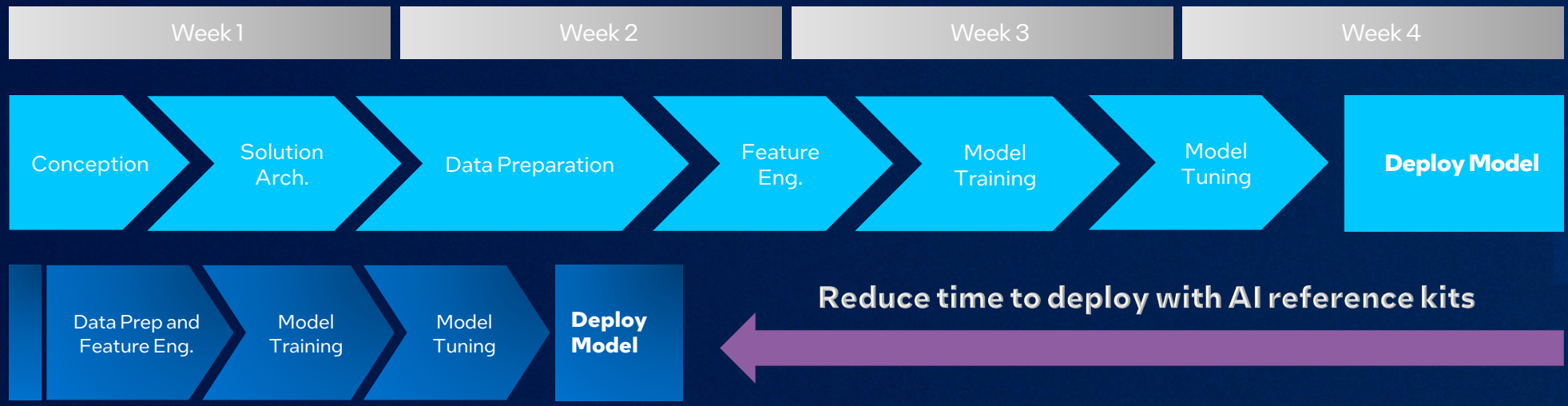
How do AI Super Reference Kits help developers?

Understanding Value to Developers

Development Cycles to Deploy AI Systems



Model Life Cycle Acceleration for Developers



- Fully functional Git repository
- Assuming around **2 days** to build code from scratch
- With the ref kit clone, fully functional code will be available in about **2 hours**

- Data processing: up to **2X** faster*
- Initial model training: up to **25X** faster*
- Faster model development or more training cycles

- Hyperparameter tuning up to **24%** faster*
- Reduction of model footprint by **75%***

* numbers are subjective and extracted from various Accenture benchmarks. These numbers depend on the accuracy target, framework, algorithm dataset, compute architecture, etc.

Hardware and Software Ecosystem Acceleration Opportunities

Hardware Acceleration

Achieve higher performance & throughput, lower latency and enable AI Inference at edge.

CPU CPU

HPU GPU

VPU FPGA

...

Software Optimization

Optimize data processing and ML/DL workloads across multiple hardware architectures.

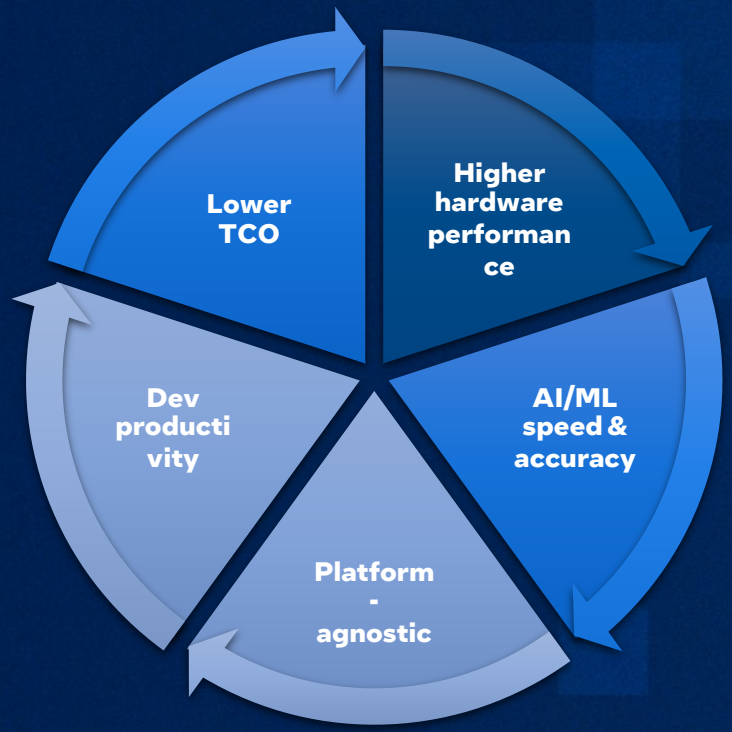
oneAPI

SIGOPT cnvrg.io OpenVINO

Scikit learn TensorFlow PyTorch

pandas NumPy python

...



Pharmaceuticals Manufacturing Business

The First AI Super Reference Kit

Pharmaceuticals Manufacturing Business "Super Kit"

Demand Forecasting
The demand forecasting component uses a time series prediction CNN-LSTM model to predict the demand for multiple products across multiple locations. It leverages the Intel® Extensions for TensorFlow on Intel® 4th Generation Xeon® Scalable processors.

Predictive Asset Maintenance
The demand predictive asset maintenance component uses an XGBoost classifier to flag assets that need maintenance. It leverages the Intel® Extension for Scikit-Learn, XGBoost, and daal4py on Intel® 4th Generation Xeon® Scalable processors.

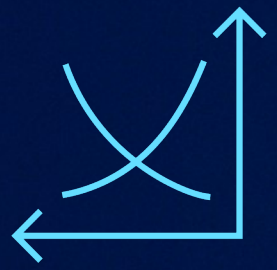
Visual Anomaly Detection
The generative AI chatbot component uses a GPT4all-J LLM and Retrieval Augmented Generation (RAG) to respond to queries associated with fictitious robotic maintenance scenarios. It leverages the PyTorch 2.0, LangChain, and Hugging Face Transformers on Intel® 4th Generation Xeon® Scalable processors.

Generative AI Chatbot
The visual anomaly detection component uses a binary classification computer vision model based on VGG-16 or Padim to produce a flag if the product passes or fails visual inspection. It leverages the OpenVINO, Anomalib, Intel® Extension for PyTorch, and Hugging Face Transformers on Intel® 4th Generation Xeon® Scalable processors OR Intel ARC GPUs.

Frontend	Streamlit
DevOps	docker, FastAPI
AI Tools and Libraries	LangChain, Transformers, Intel oneAPI AI Analytics Toolkit, OpenVINO, ANOMALiB, PyTorch, scikit-learn, TensorFlow
Platform	Intel Developer Cloud
Hardware	Intel Xeon, Intel ARC Graphics

Modernizing Manufacturing Business Processes

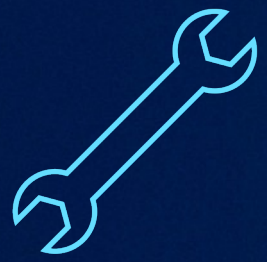
Predict demand



Machine learning
- Tensorflow

Check equipment

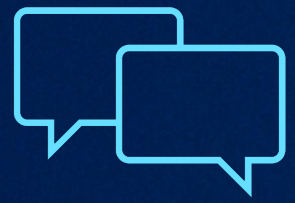
Predictive maintenance



Machine learning – XGBoost, Scikit-Learn*, daal4py

Service the Production Line

Customer Support



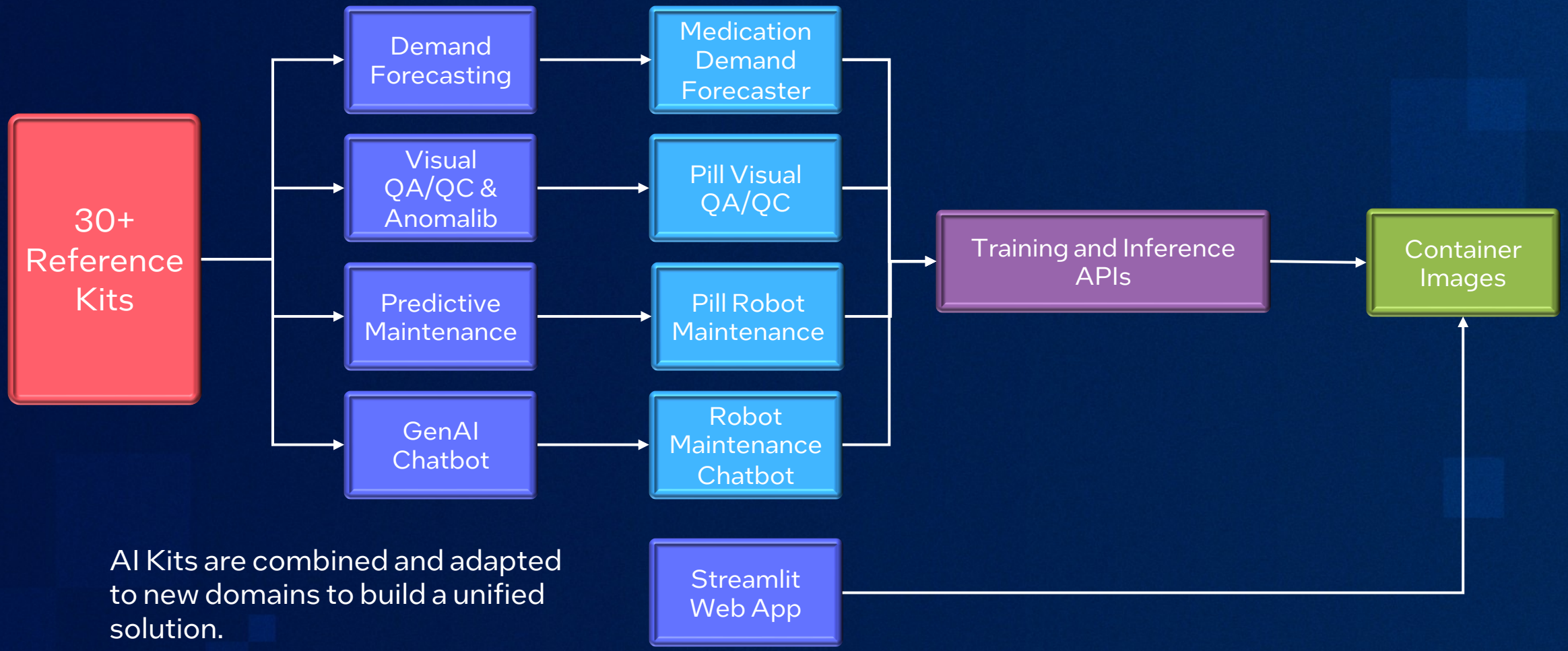
Generative AI – Q&A Chatbot w/ Retrieval Augmented Generation

Service the product

Visual QA/QC

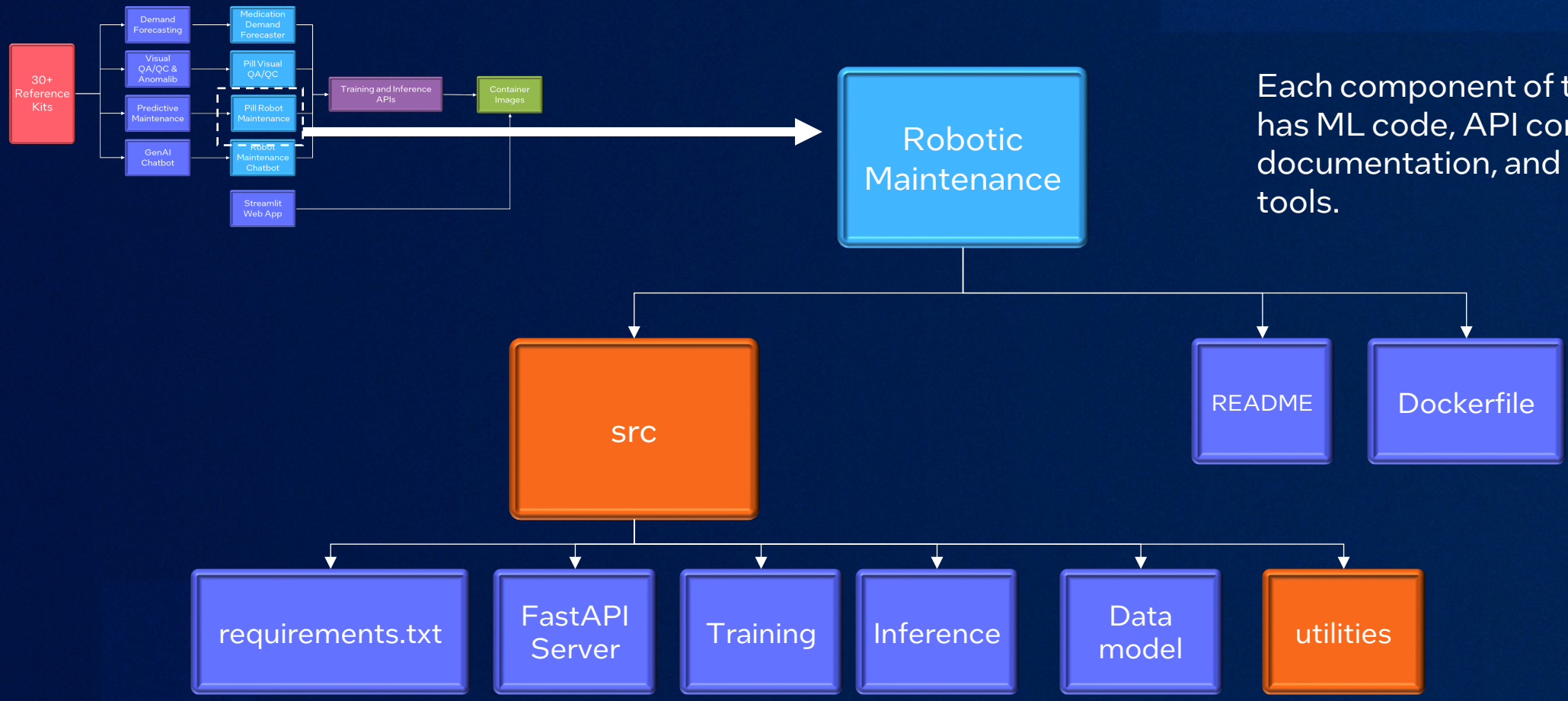
Anomalib and Intel® Extension for PyTorch* Computer Vision

Super AI Reference Kit Adaption Flow – Pharmaceutical Manufacturing Business Example



AI Kits are combined and adapted to new domains to build a unified solution.

Example Package Architecture Diagram



Each component of the super kit has ML code, API configs, utilities, documentation, and deployment tools.

Begin Workshop

Workshop Steps

1. User access for Intel Developer Cloud
2. Provision a Tiny VM
3. Connect to Tiny VM Using VSCode
4. Clone the Super Kit Repository (15 min)
5. Start a high-level code overview
6. Demand Forecast Deep Dive and Package Layout Reinforcement
7. Deployment Mechanics
8. Optimizations and Overview of Robotics Maintenance, Visual QA/QC, and Chatbot Components
9. Once Launched, go into UI and train/inference with Demand Forecast + Robotics Maintenance
10. Recap of the Learnings

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details.
No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

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

In Closing

- We have learned how AI Super Reference Kits provide:
  - Quickly prototype unified AI solutions with multiple reference kits
  - Leverage sample architecture to plan for scale with essential DevOps tools
  - Build with pre-baked intel AI optimizations
- We reviewed the “Pharmaceutical Manufacturing Business” Super Kit codebase and developed an understanding of how the architectural elements of the application

In Closing (continued)

- Implementation overview of Intel AI Optimizations:
  - Intel® Optimizations for Tensorflow*
  - Intel® Extension for Scikit-Learn*
  - Intel® Optimizations for XGBoost
  - Intel® oneAPI Data Analytics Library (oneDAL) for Python* API - aka daal4py
  - OpenVINO: Anomalib
  - Intel® Extension for PyTorch*
  - PyTorch 2.0 – Upstreamed Intel Optimizations
- Learn how each application is modularized to support an API endpoint
- The various application components will be deployed using DevOps tools like Docker and Docker Compose

Call to Action

AI Reference Kit
Main Page

Download the
AI Super Reference Kit

Download Intel AI
Reference Toolkits for free

Q&A


intel