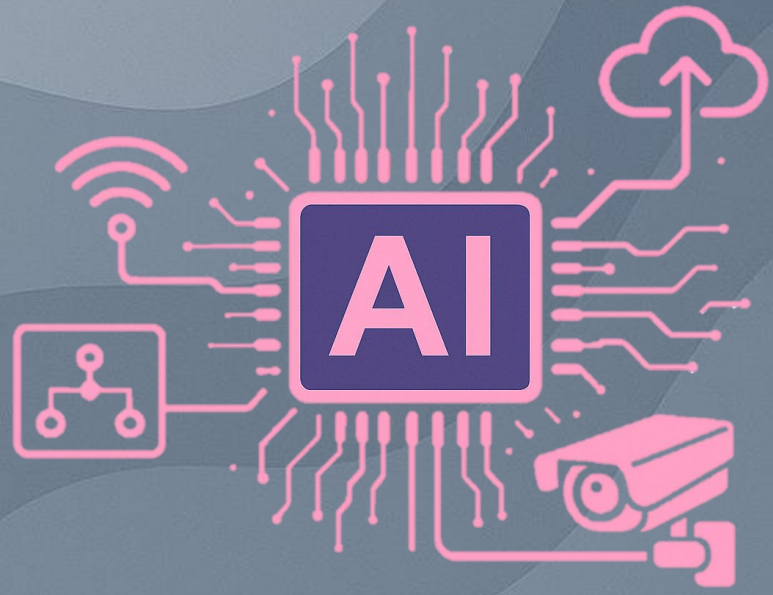
Edge AI

Smarter IoT for Instant Insights

Kim Berg
CTO Data & AI @Sogeti Sweden



Why Edge AI?

Real-Time Decision Making

Edge AI processes data instantly at the source, enabling faster, more effective decision-making with ultra-low latency.

Enhanced Privacy and Security

Keeping data local with Edge AI helps protect sensitive information and reduces security risks by minimizing exposure.

Reduced Bandwidth and Costs

Edge AI lowers bandwidth usage and cloud expenses by processing information locally instead of sending everything to the cloud.

Reliability in Remote Settings

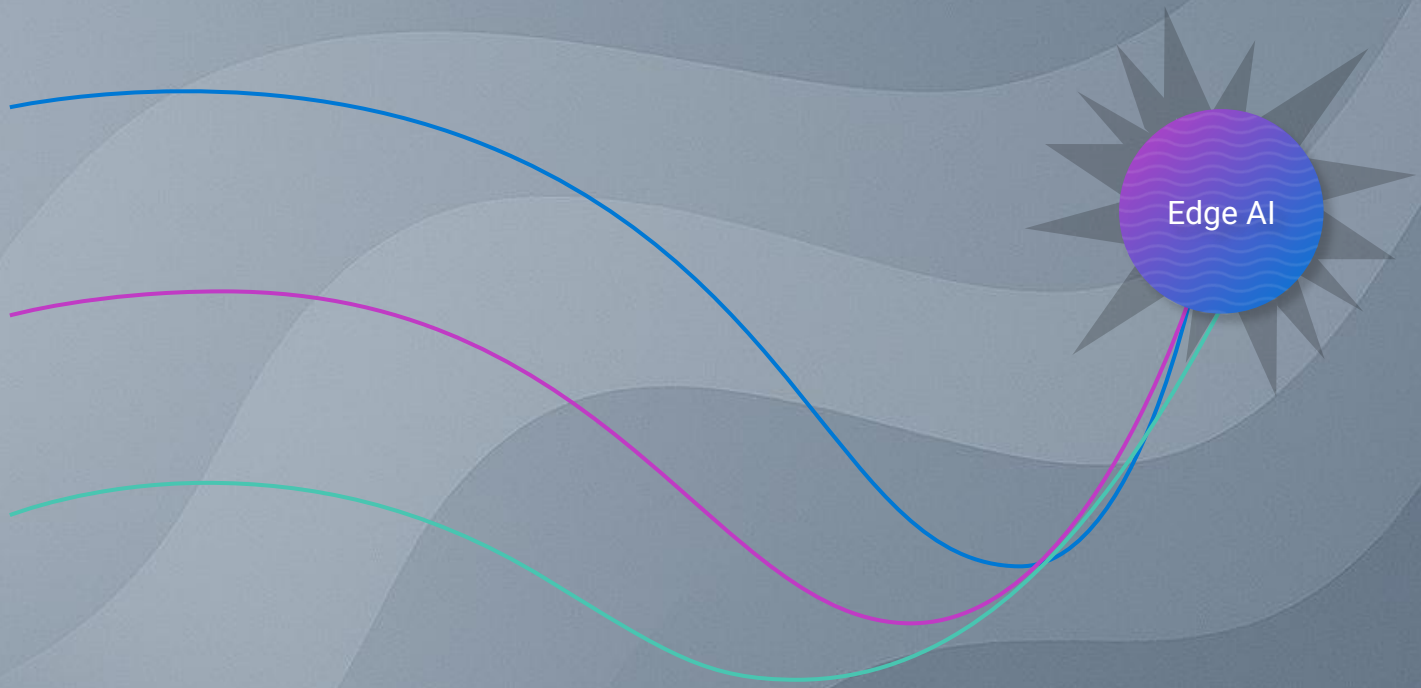
Local processing ensures dependable AI performance even in remote or offline environments without cloud access.

Edge AI is now viable

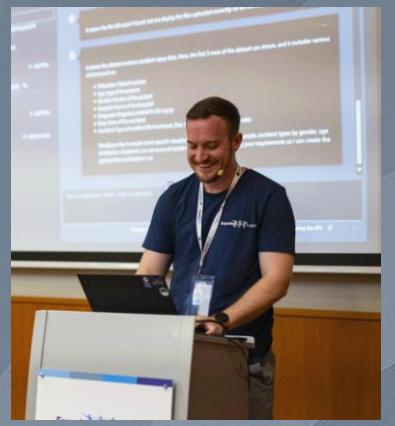
Powerful Hardware

Smaller, faster models

Runtime Optimizations








Love to dance and all sorts of outdoor activities

Over a decade in the tech industry with 9 years focusing on cloud

Part of Microsoft Responsible AI Innovation Center

Public speaker

Contributor to Open-source

MVP Microsoft® Most Valuable Professional

Bringing assets together to democratize AI

Foundry with cloud hosted models

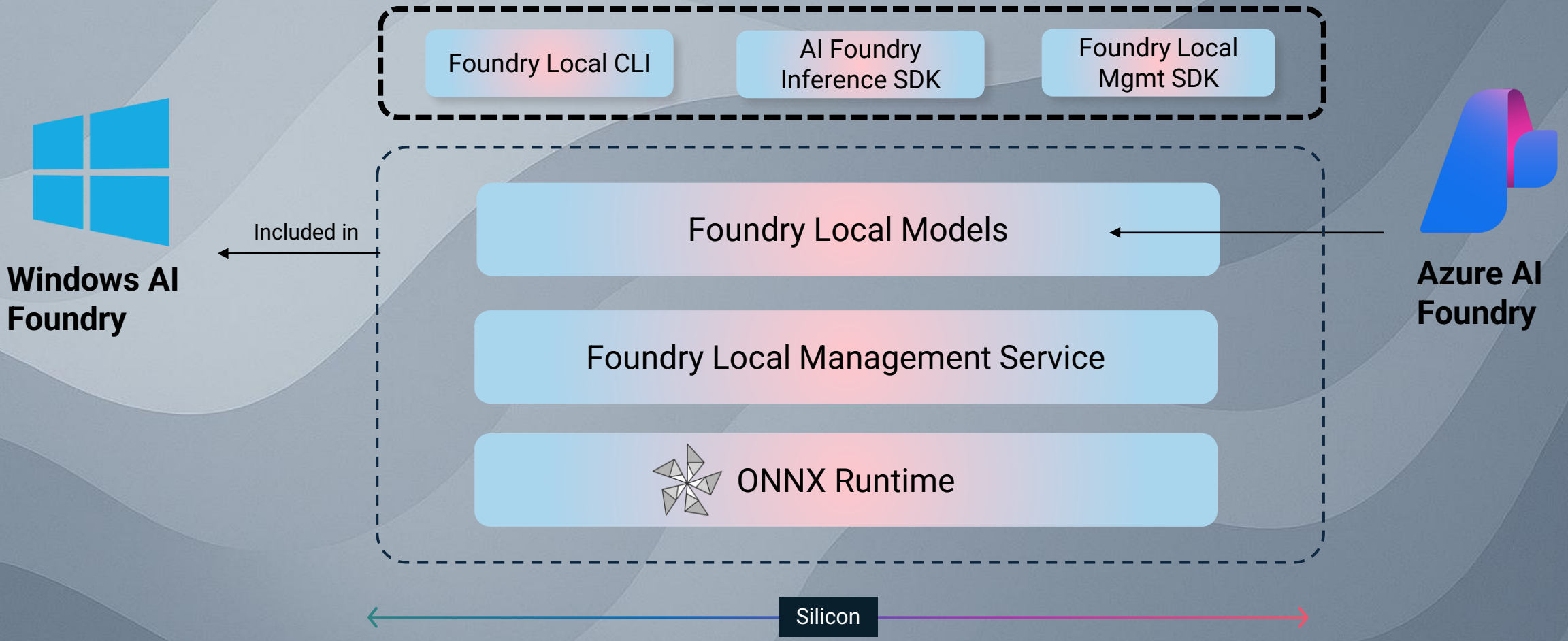
ONNX Runtime (ORT): cross-platform on-device inferencing

Windows on-device model APIs & ML runtime

70K+
Organizations

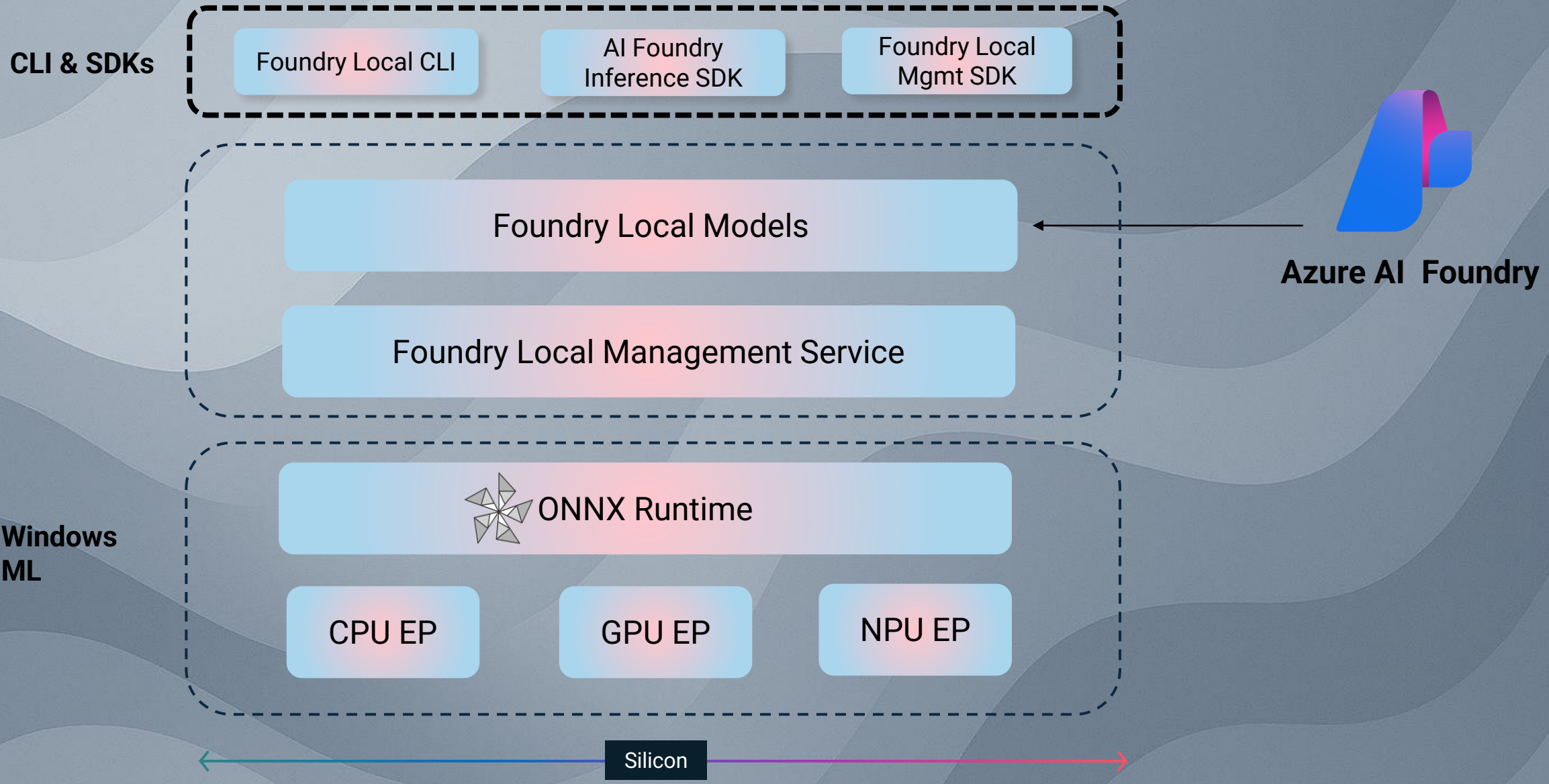
4x
Acceleration

100s of millions
Devices

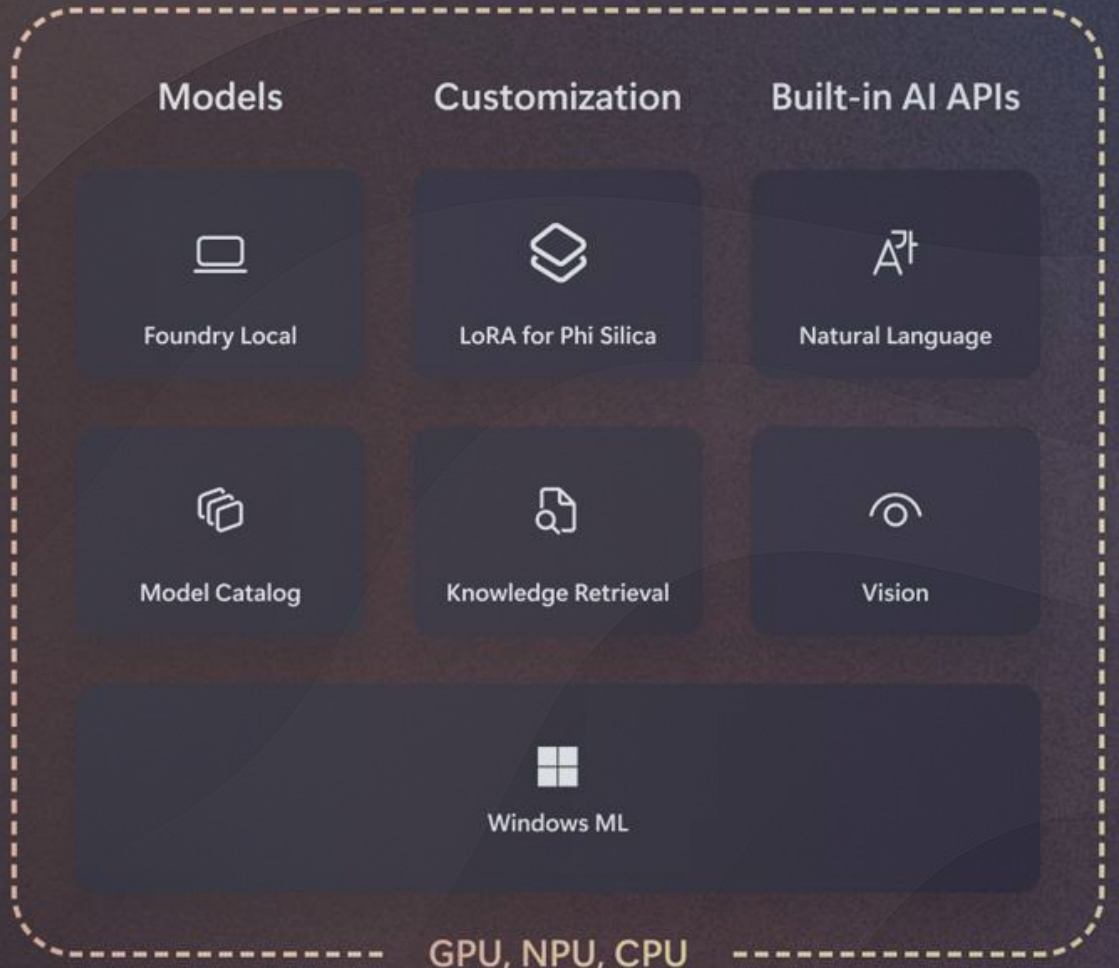


Available today on Windows and macOS

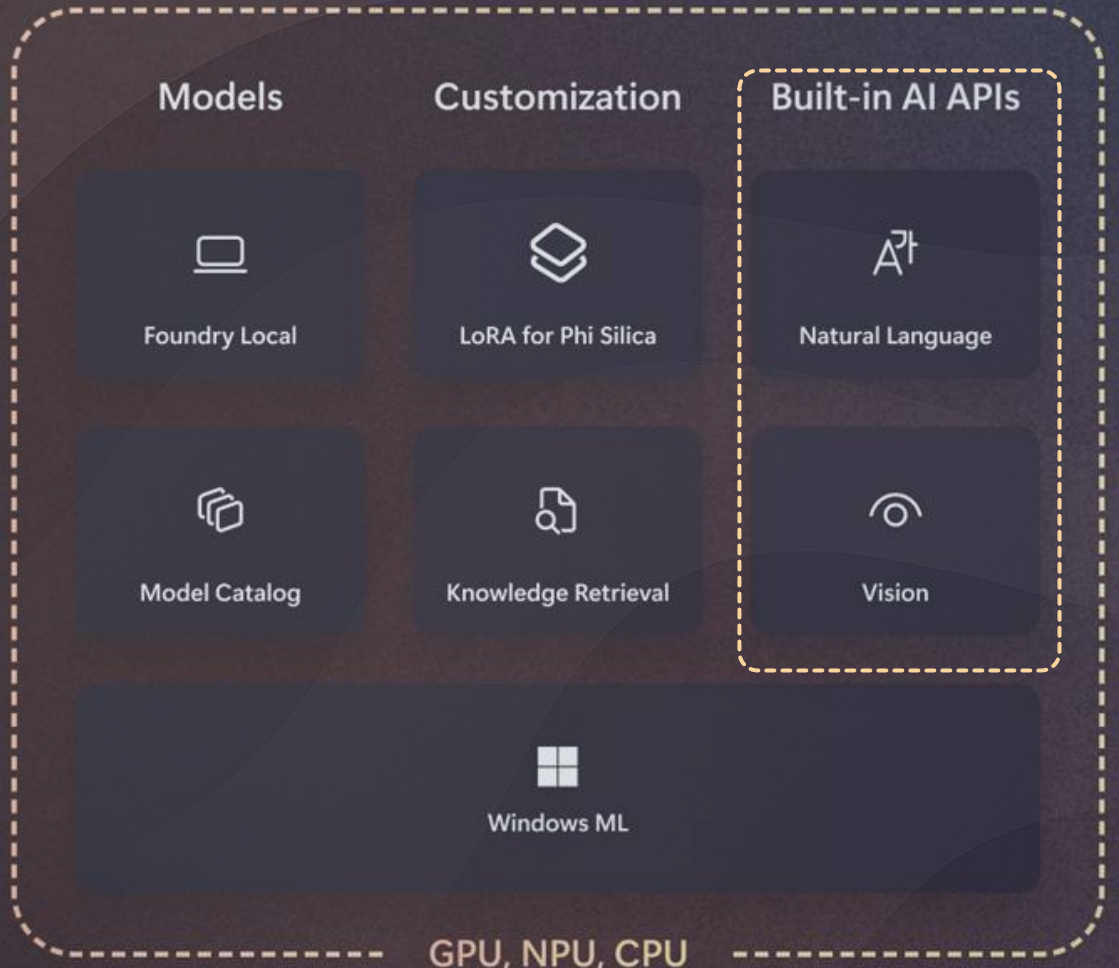
Foundry Local is integrated in Windows AI Foundry



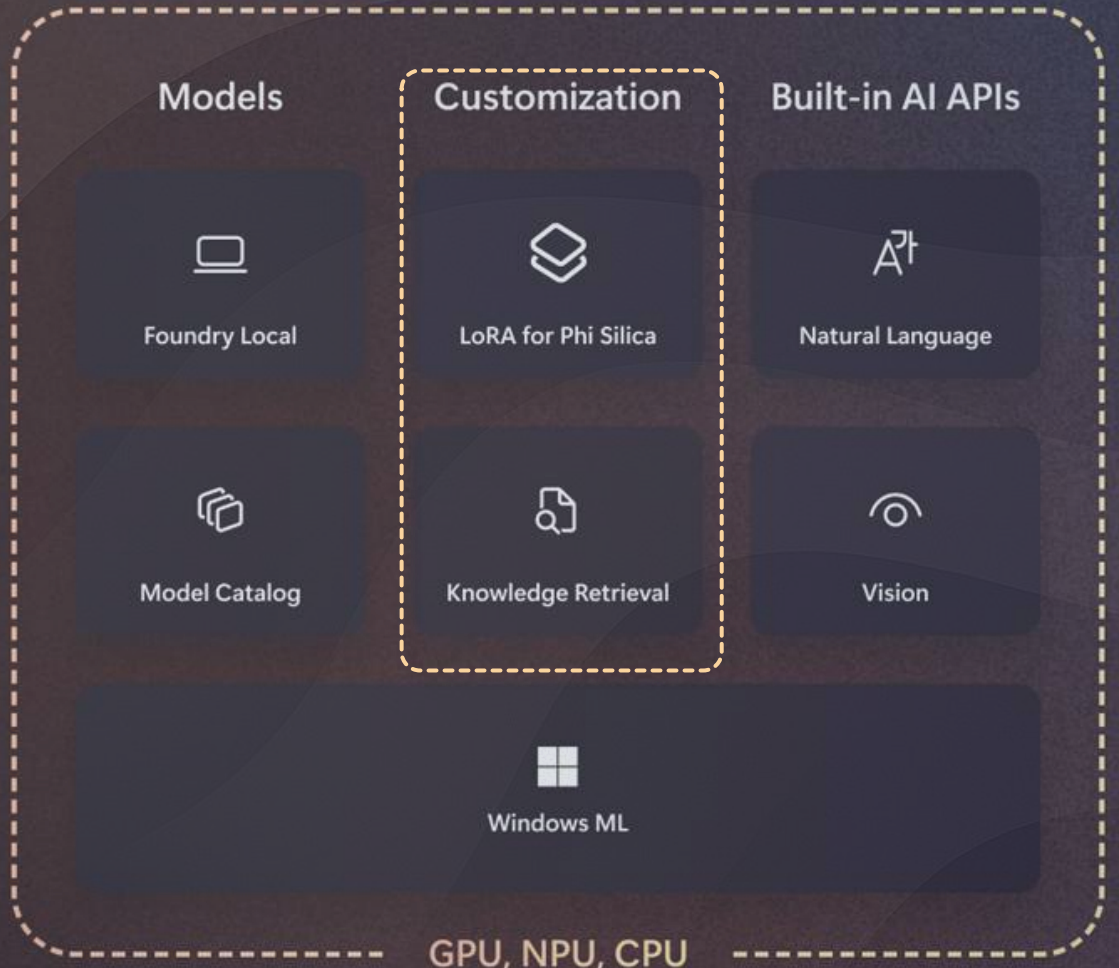
Windows AI Foundry



Windows AI Foundry

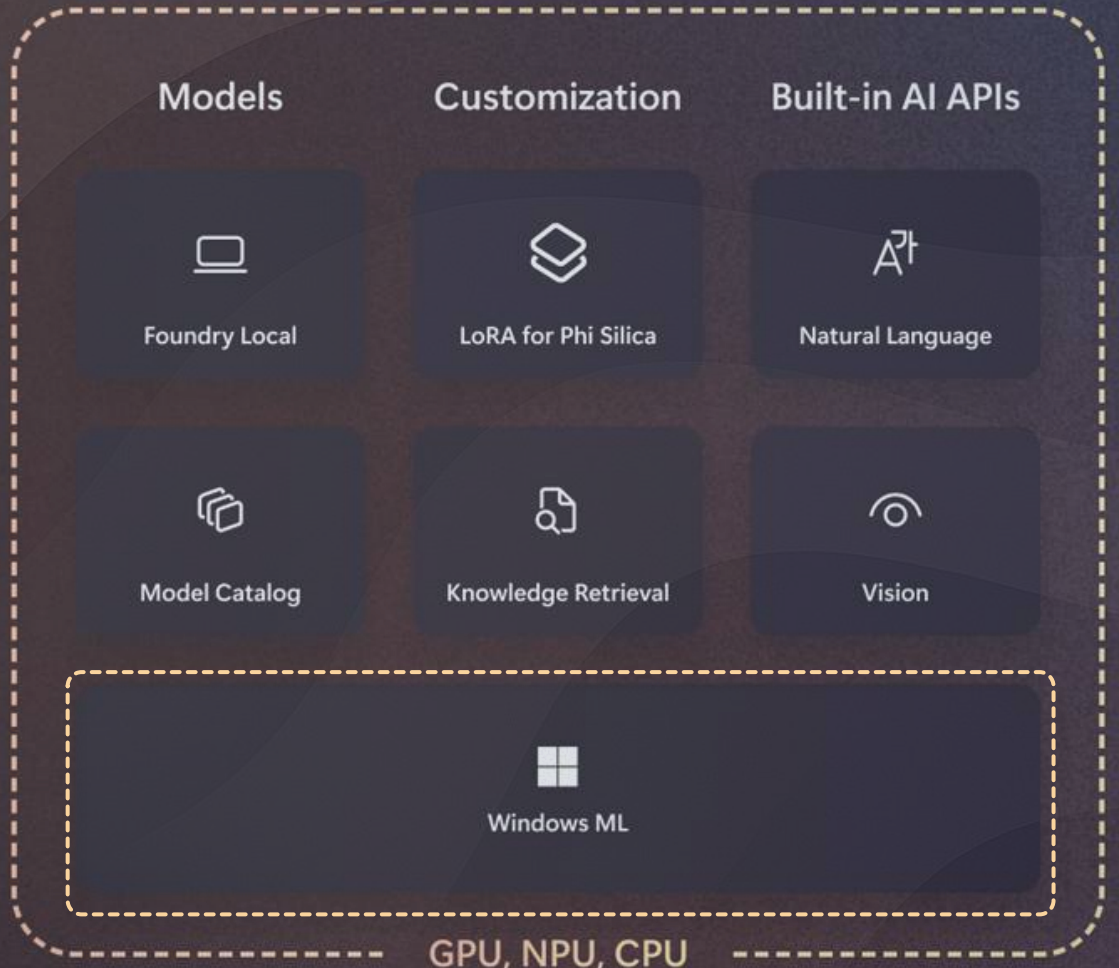


Windows AI Foundry



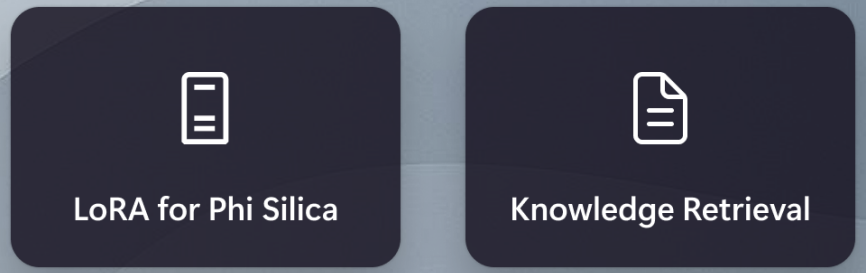
Windows AI Foundry

Windows AI Foundry



Windows AI APIs Customization, and Why?

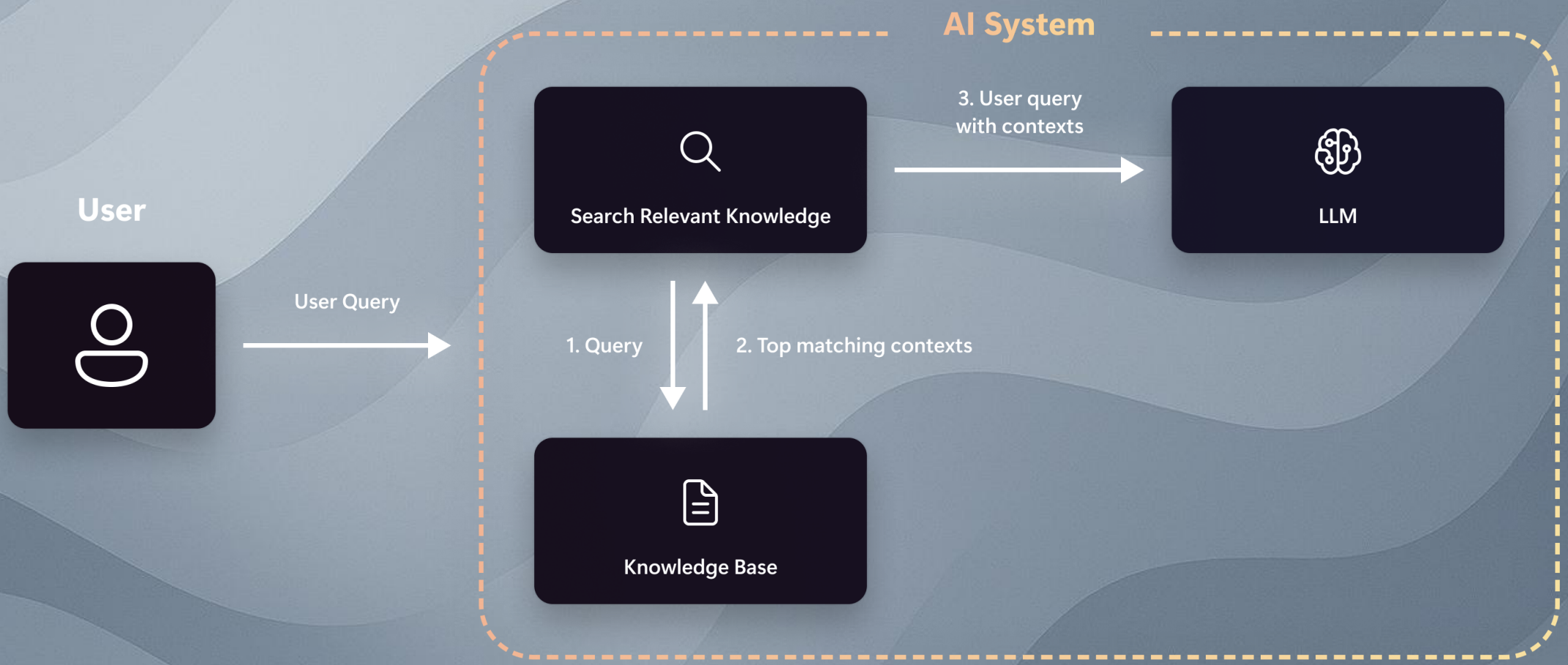
Making inbox models work better for your domain



Intent	Example	Customization Path to Consider
Brand Voice	“Make it sound like *my* company”	LoRA
Task Adaptation	“Optimize/format for *my* workflows”	LoRA
Domain-Specific Language	“Understand legal/industrial/technical terms”	LoRA and/or Knowledge Retrieval
Private Knowledge Grounding	“Answer based on *my* docs and content”	Knowledge Retrieval

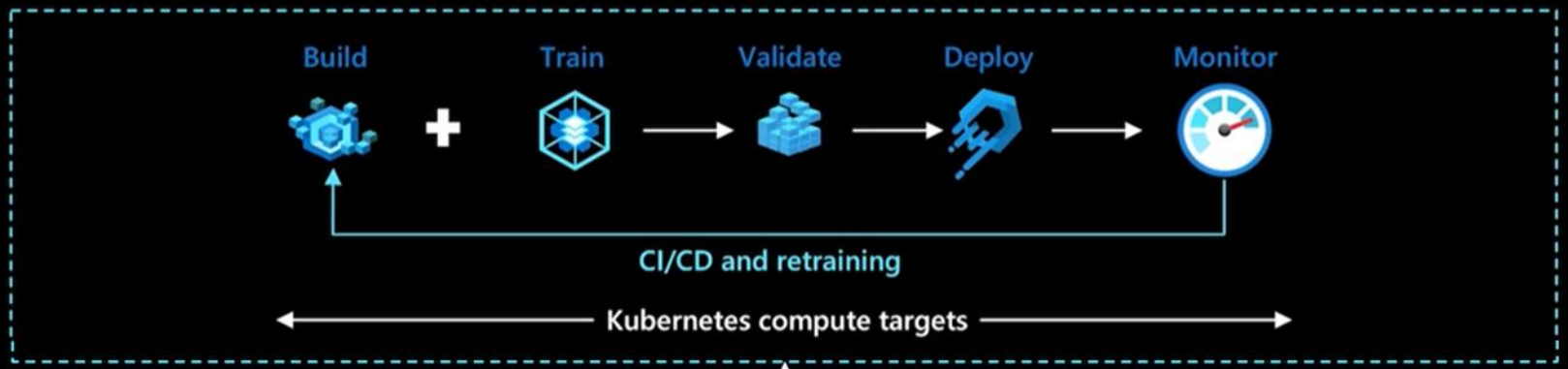
Knowledge Retrieval

Customizing what the model knows

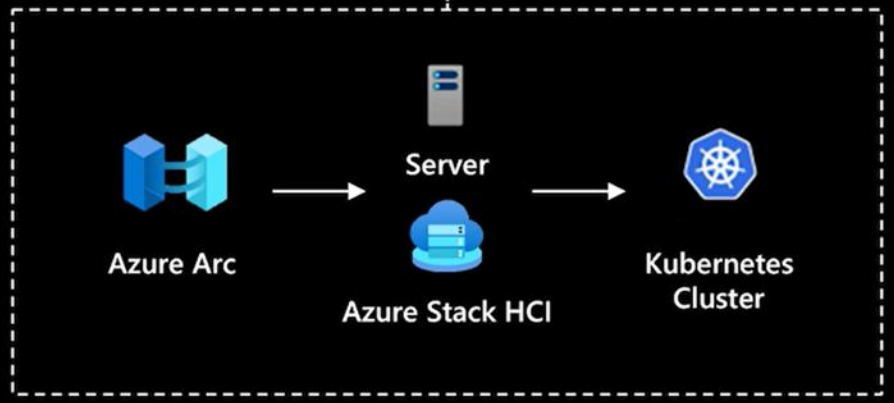


Machine learning anywhere

Azure Machine Learning



Machine Learning Kubernetes Extension



On-premises

Azure Kubernetes Service

Multi-cloud

Vision Companion

Problem:

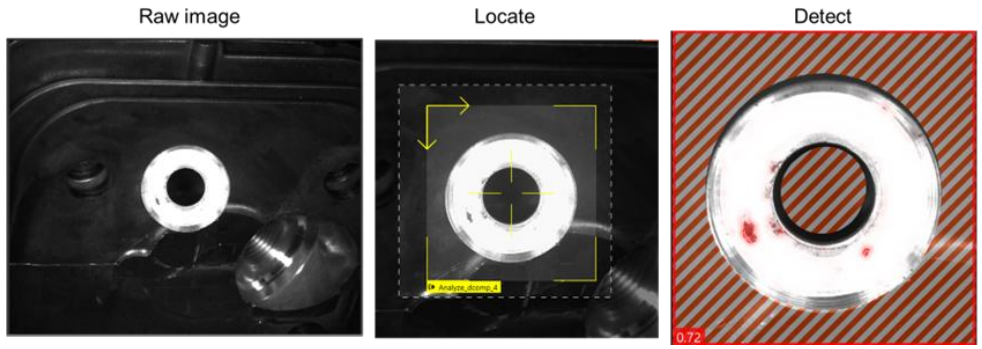
During the casting process we sometimes are left with airgaps inside the metal. Many times, these are not shown after the casting process, but during the machining process they appear.

Currently there is a manual quality check that happens later in the process flow. However, at that stage there is no possibility in being able to remelt the part, which means that there is a costly scrapping process.

Cylinder machining is done in 8 different fully automated machine cells.

In Husqvarna we are casting many different parts with similar issues..

Cylinder



Flywheel

Bar

Husqvarna Group

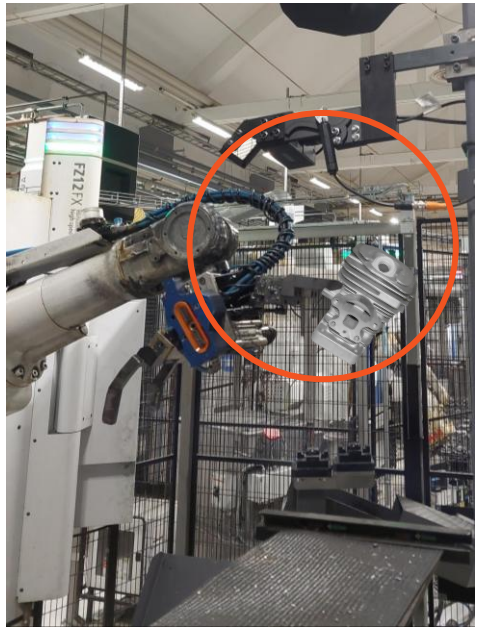
Requirements

Evaluation must be completed in 30s

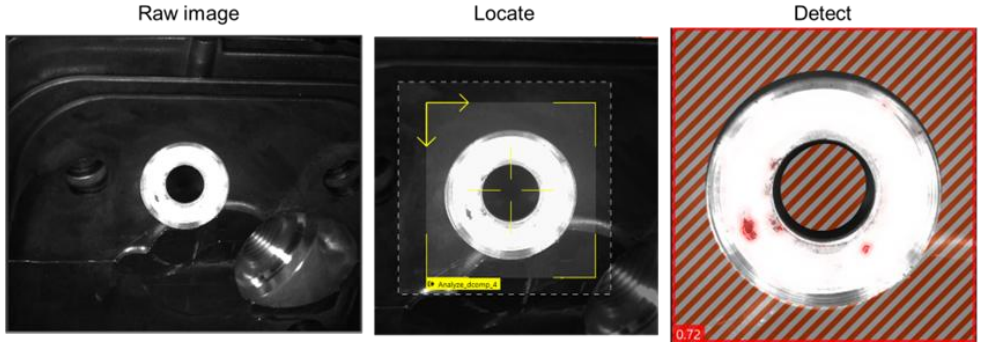
6-8 angles to evaluate depending on part

8x machines running in parallel

Min model performance:
95% Precision, 65% Recall



Cylinder



Husqvarna

Questions?

Thanks for listening!

https://github.com/Kim-Berg/AI-on-the-Edge