

tinyML[®] Summit

Miniature dreams can come true...

March 28-30, 2022 | San Francisco Bay Area



www.tinyML.org



Challenges for Large Scale Deployment of Tiny ML Devices

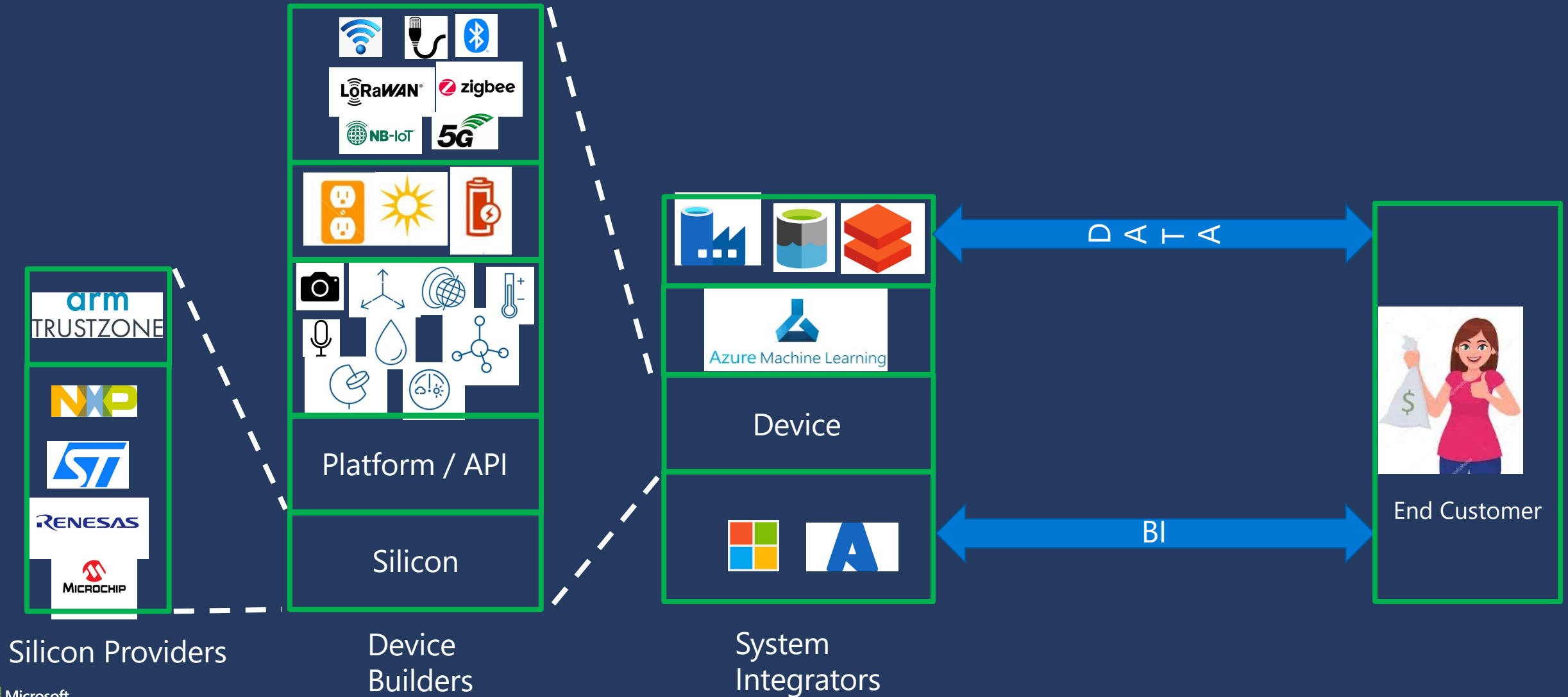
Gopal Raghavan, Jerome Schang, Jennifer Skinner-Gray, Pete Bernard
Azure Edge Devices, Platforms and Services

Agenda

- Ecosystem / Blockers
- TinyML / Cloud integration
- Call to action



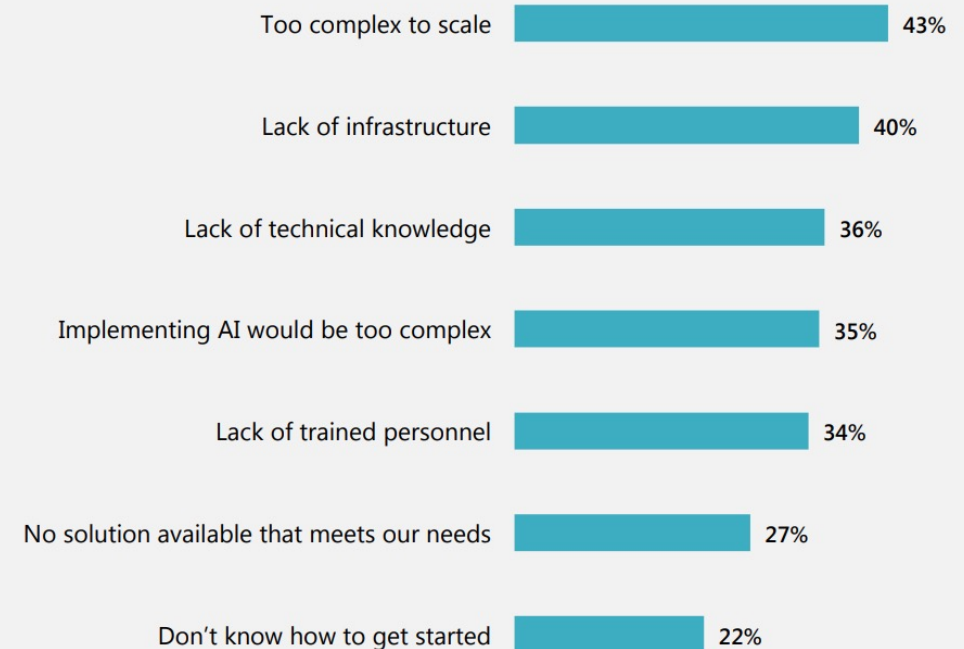
Tiny Device Ecosystem / Terminology



ML Edge Market Has Been Slow to Evolve

- IoT market and Intelligent Edge market predictions have been overoptimistic
- What can we do ?
 - Understand entire ecosystem
 - Identify and architect solutions to mitigate blockers

BARRIERS TO USING AI MORE WITHIN IoT | Ranked Top 3 Most Challenging



<https://aka.ms/IoTsignals>

Tiny Device Challenges

- Tiny devices are a piece of the Edge AI device spectrum
- Device Lifecycle Management
 - Secure deployments, Device updates...
- ML model lifecycle management
 - Diverse deployment conditions
 - Concept/Data drift require model updates
- Integration with Business Processes

Consistent Security, Identity, Management, Data and AI

Tiny Edge: MCU

Light Edge: MPU

Heavy Edge: Multi-core

Edge Silicon

IoT Endpoints

Edge Gateways

Edge Compute

Distributed Cloud

Public Cloud



Azure Sphere
Azure RTOS



Azure IoT Edge
Edge AI Devices



Windows IoT
Edge AI Gateways



Azure Stack HCI
Azure Stack Edge
Azure Stack Hub

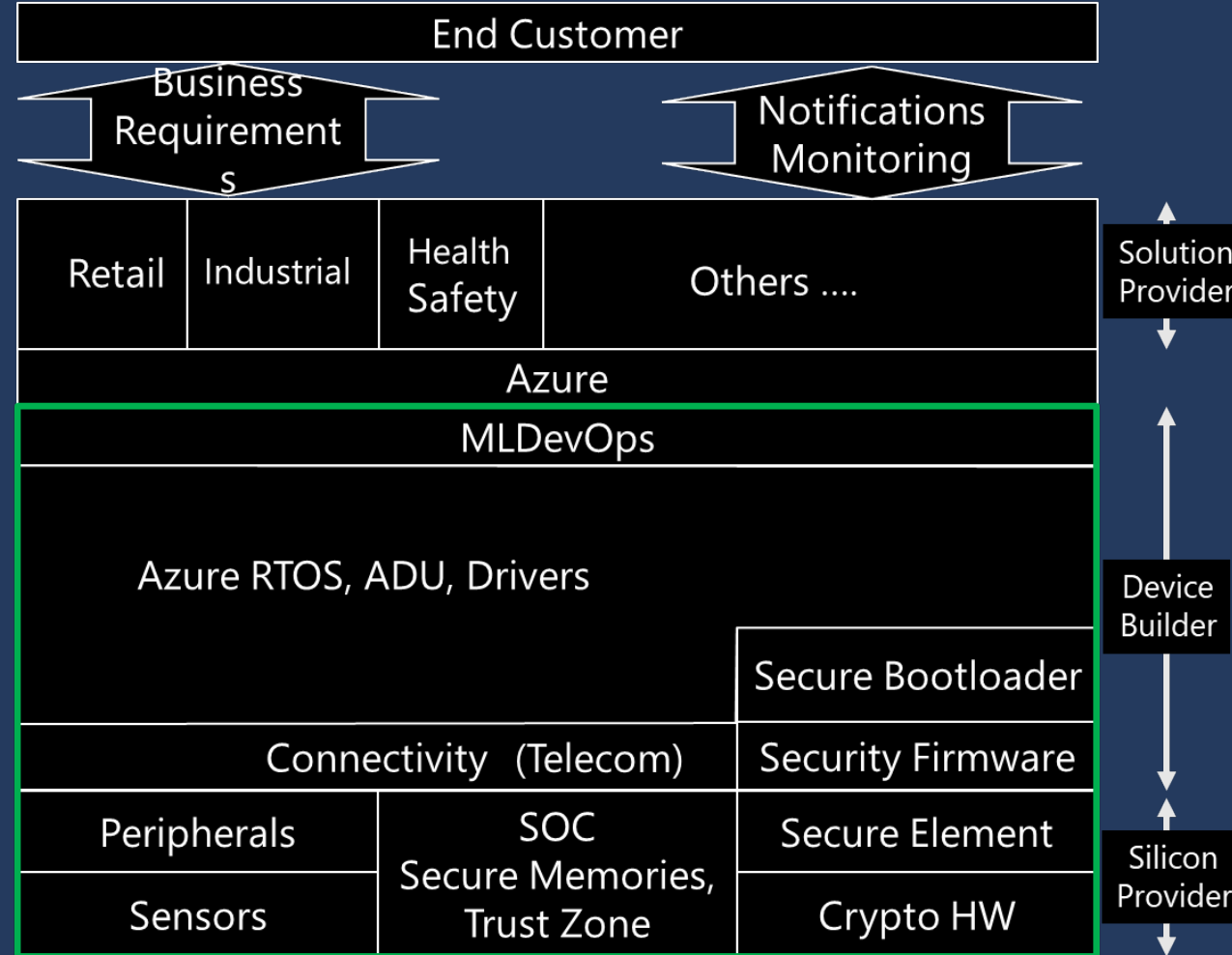


Azure Edge Zones

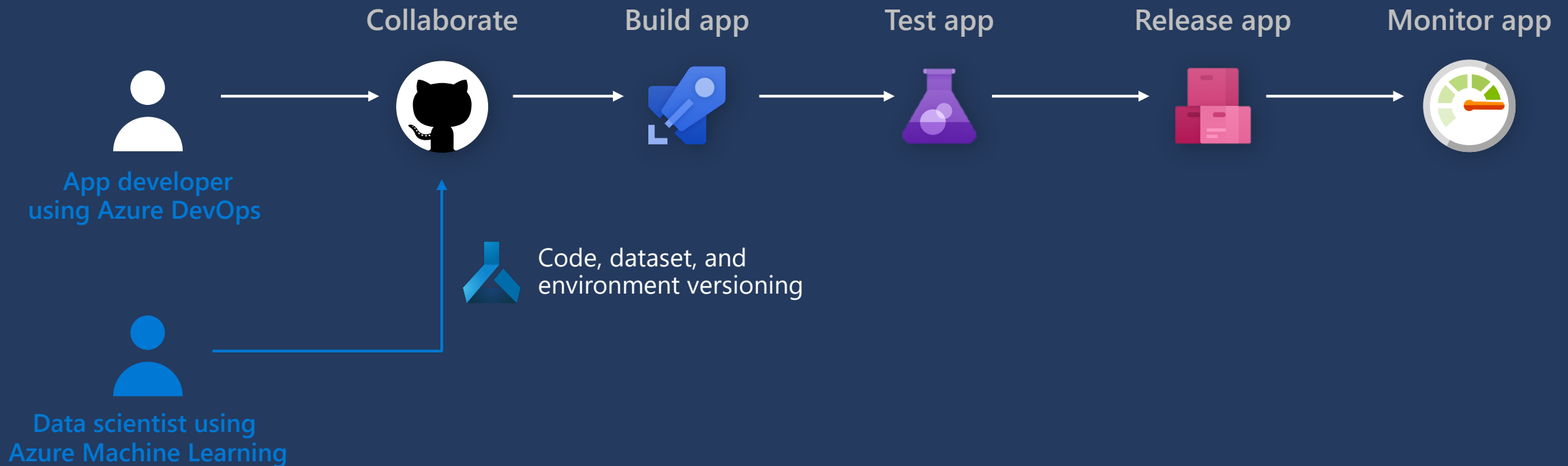


Azure

Certified Devices (Platform) Allow Rapid PoCs and Subsequent Large-Scale Deployment



MLDevOps with Azure Machine Learning



Model reproducibility



Model validation

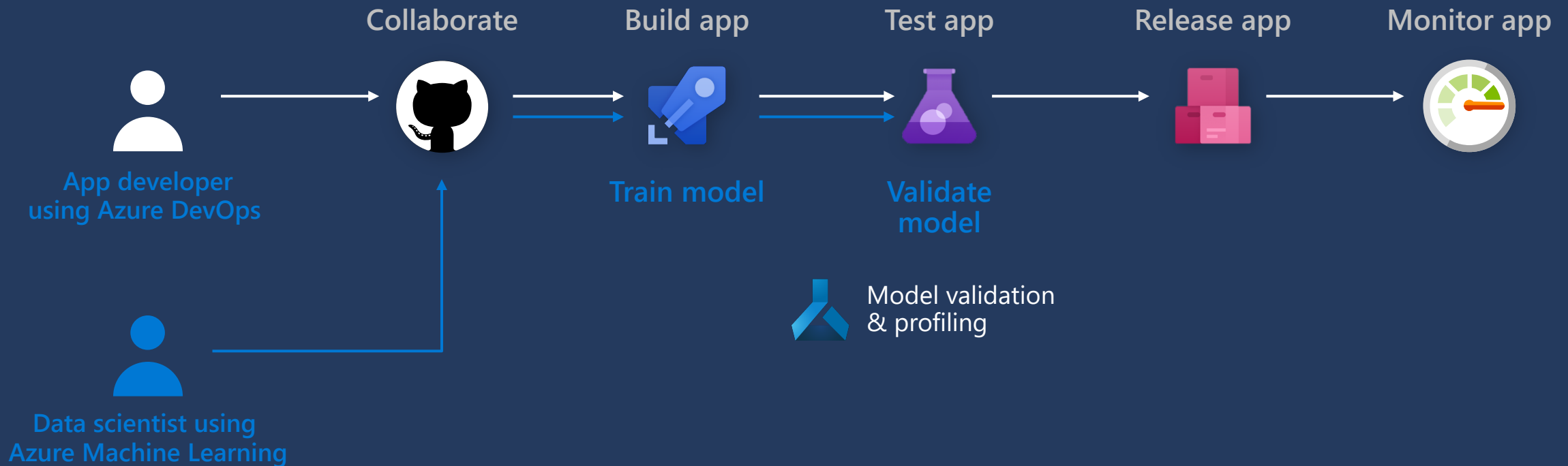


Model deployment



Model retraining

MLDevOps with Azure Machine Learning



Model reproducibility



Model validation

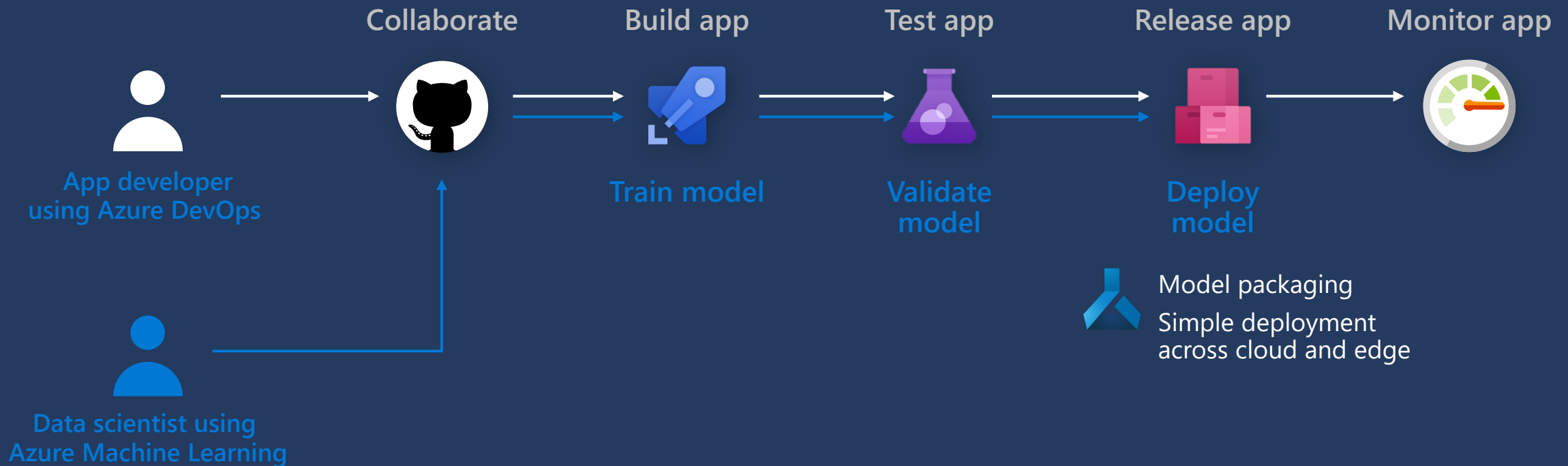


Model deployment



Model retraining

MLDevOps with Azure Machine Learning



Model reproducibility



Model validation

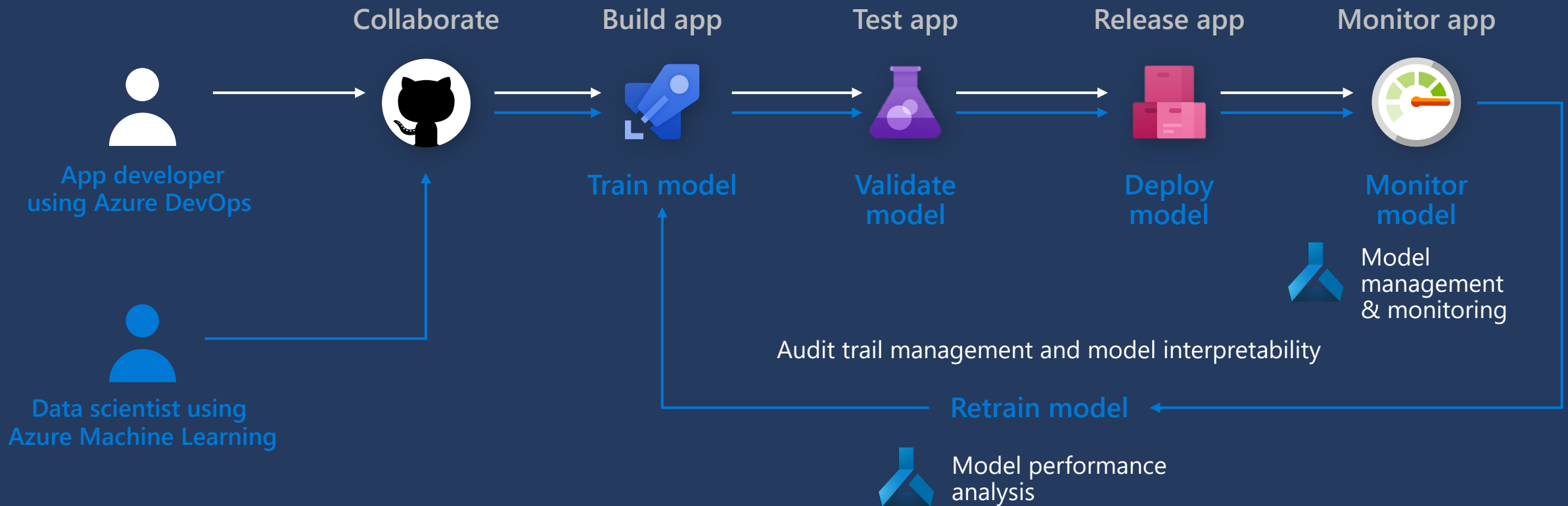


Model deployment



Model retraining

MLDevOps with Azure Machine Learning



✓ Model reproducibility

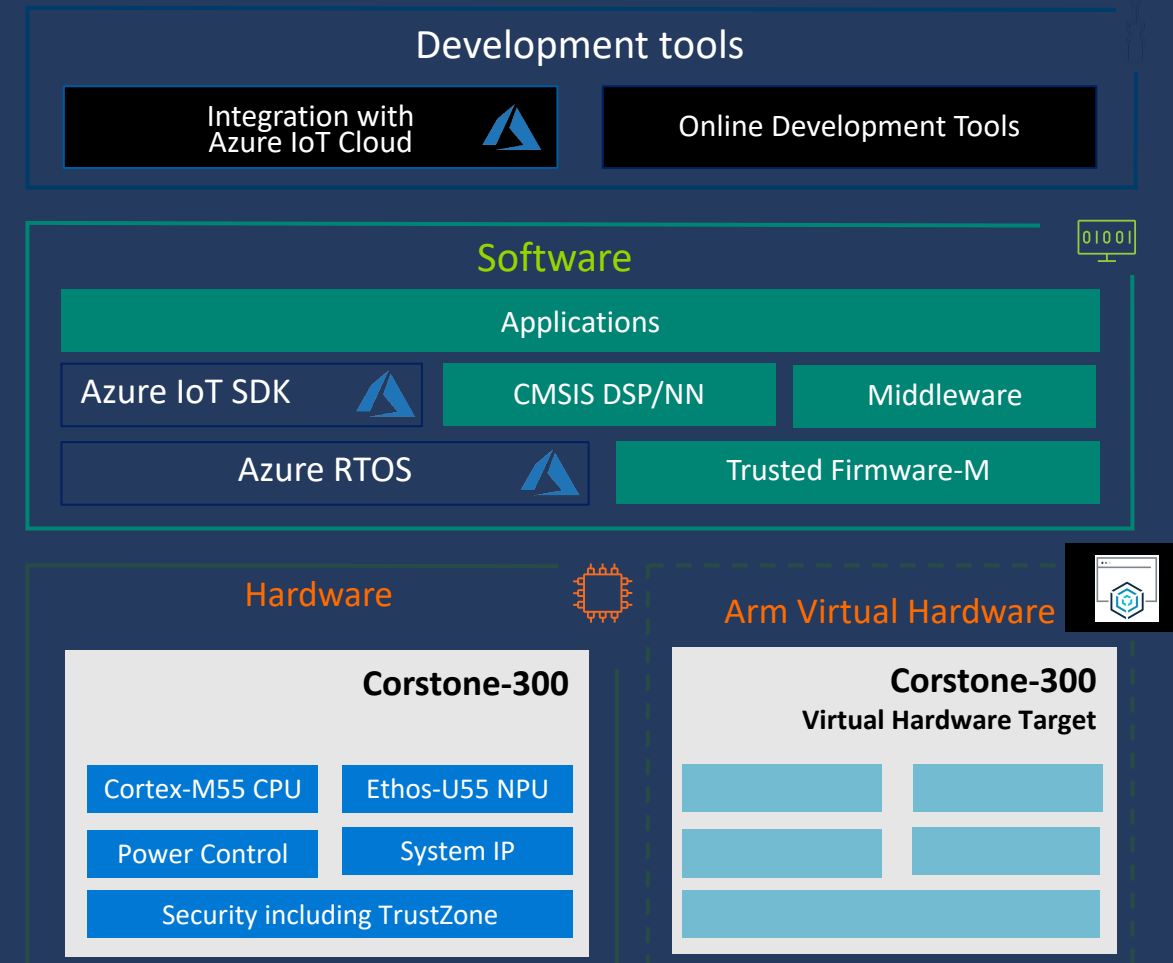
✓ Model validation

✓ Model deployment

✓ Model retraining

Azure + Arm Virtual Hardware for System Emulation

- Hardware accurate, high-speed emulation hosted on Azure
- Evaluate performance on a variety of partner silicon to determine hardware choice
- Emulate entire deployment process including digital twins
- Emulate complete data pipeline from sensor to Azure business insights



Cloud Integration is Key to Scale TinyML Deployments

- Business intelligence and decision making motivate TinyML deployments
- Device lifecycle management of billions of devices
 - Security
- ML model development and performance monitoring
 - Features vs Models
- Improving model accuracy
 - Use large models for critical inference and edge device as trigger (Increase FP rate)
 - Decision making across fleets of devices
 - Federated learning

Call To Action

- Silicon providers : Support cloud integration in your eval kits
 - Integrate physical connectivity
 - Support cloud connectivity (Azure RTOS / Azure Device Update)
 - Get Azure Certified (<https://aka.ms/IoTCertificationsBasics>)
- Toolchain / Model developers: Automate model generation and optimization
 - Containerize for rapid evaluation
 - Model zoo with same model targeted to different devices
 - Low code / no code experience
- Invest in standardized comparison of performance metrics (e.g. MLCommons)
 - Across device and model conversion toolchain



AONdevices

arm

ASPINITY

brainchip
The Neuromorphic Computing Company

CEVA®

Deeplite

EDGE IMPULSE

emza
visual sense

FotaHub

GREENWAVES
TECHNOLOGIES

Grovetly Inc.

Himax

HOTC

imagimob

infineon

itemis

KLIKA·TECH
GLOBAL IOT SOLUTIONS

LatentAI

LATTICE
SEMICONDUCTOR

Micro.ai

OmniML

NXP

POI

Plumerai

PROPHESSEE

Qeexo

Qualcomm

Rackner

RealityAI®
Engineering Solutions for the Edge

REXEN
technology

RENESAS

SAP

seeed
The IoT Hardware Enabler

SensiML

Sony Semiconductor
Solutions
Corporation

ST
life.augmented

SA STREAM ANALYZE

synaptics®

SynSense

SYNTIANT

Tensil.ai

TensorFlow

XMOS



Copyright Notice

This presentation in this publication was presented as a tinyML® Summit 2022. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org