tinyML® Summit

Miniature dreams can come true...

March 28-30, 2022 | San Francisco Bay Area

www.tinyML.org

Microsoft

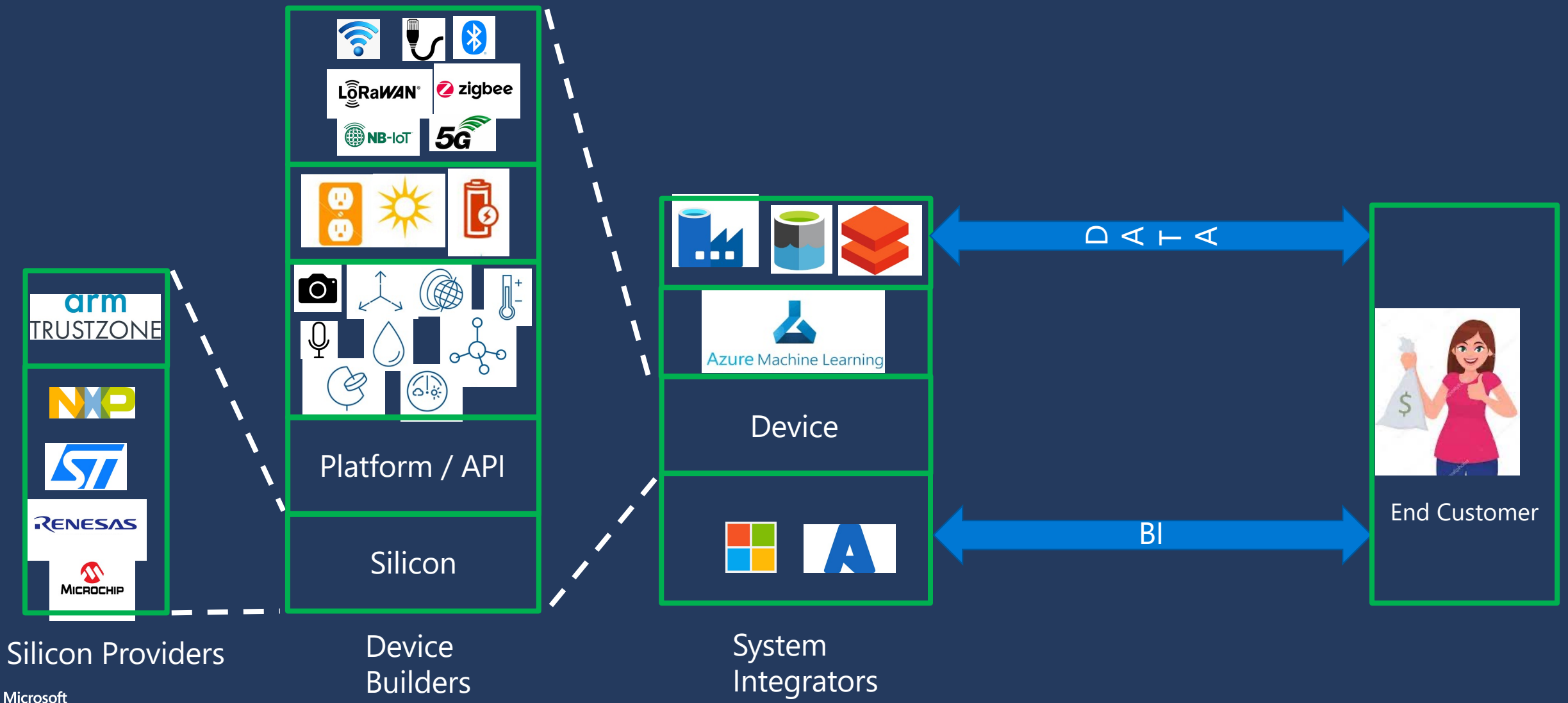
Challenges for Large Scale Deployment of Tiny ML Devices

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

Agenda

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

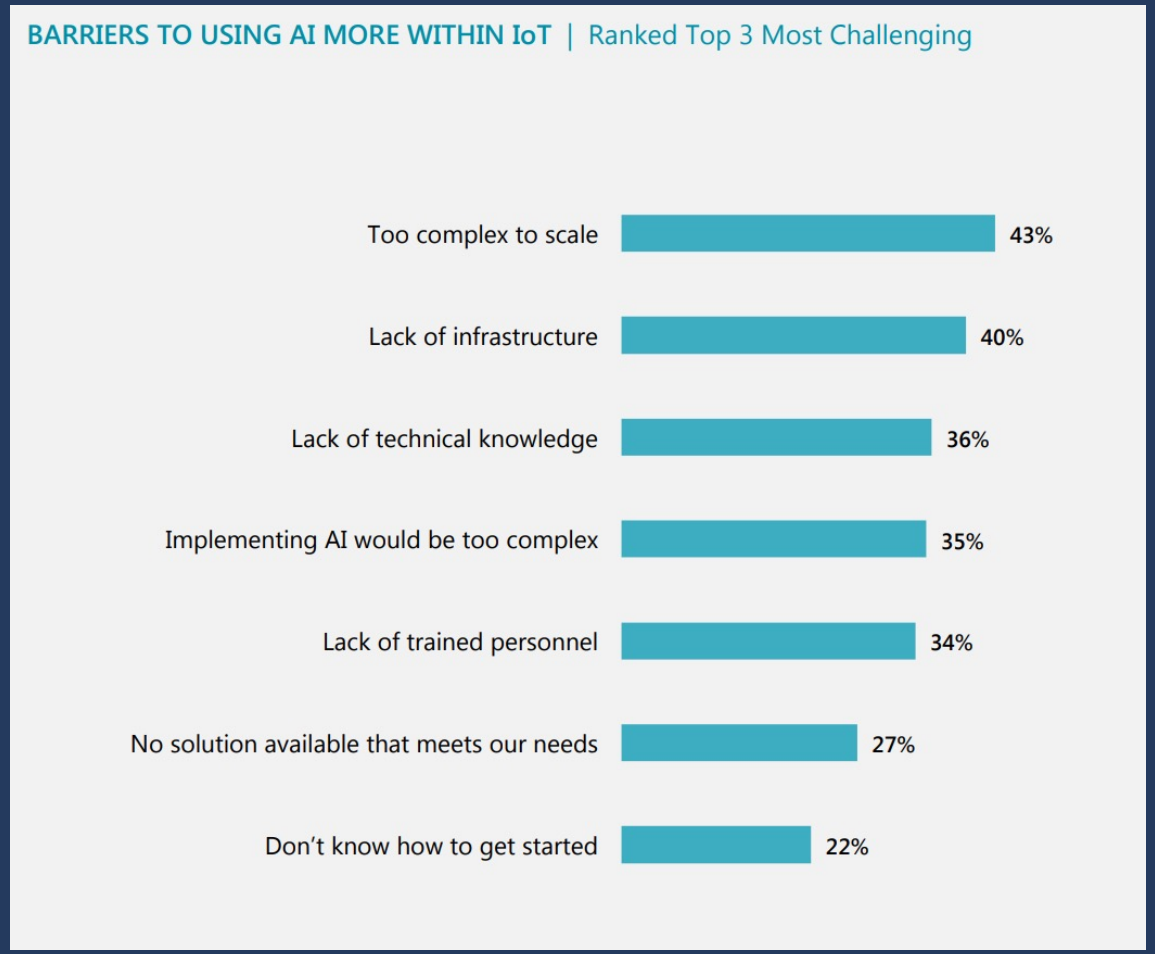
Tiny Device Ecosystem / Terminology



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



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

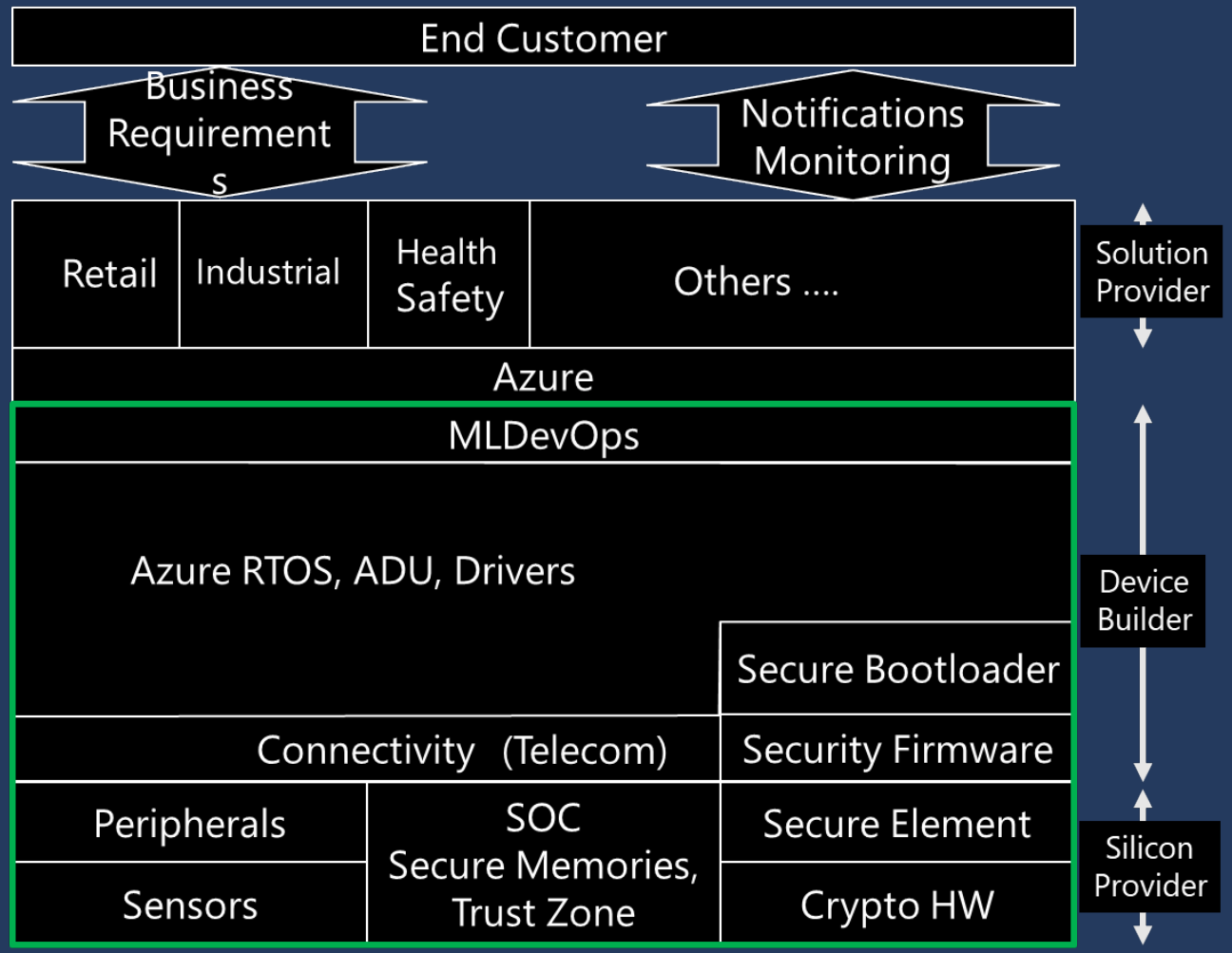
Tiny Edge: MCU

Light Edge: MPU

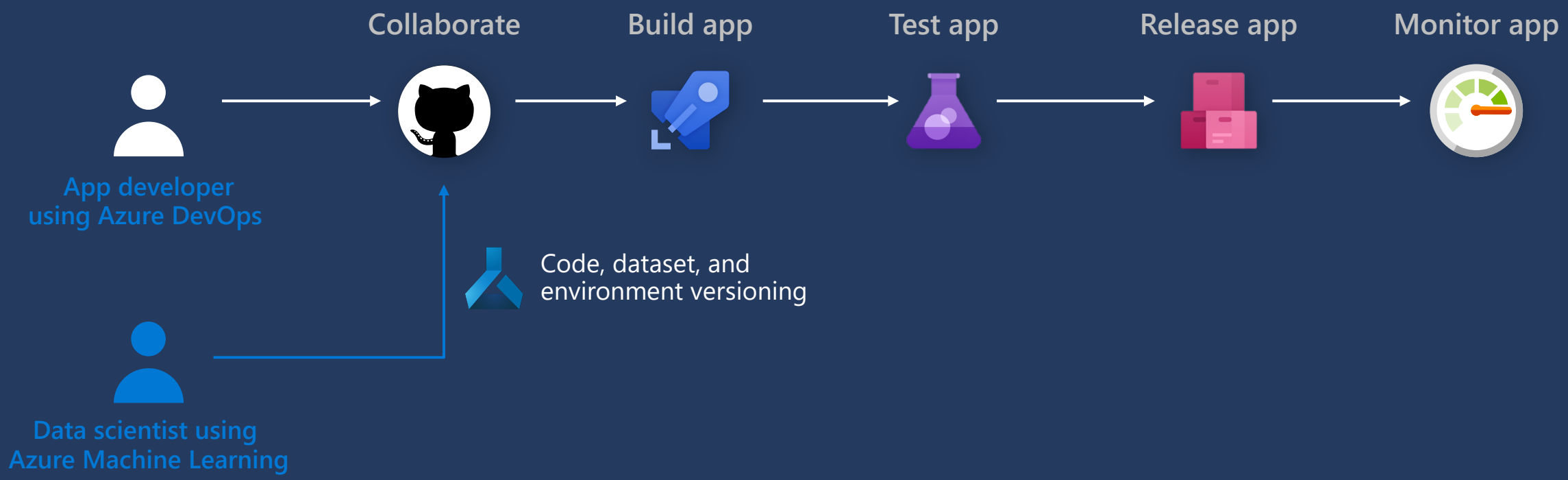
Heavy Edge: Multi-core

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

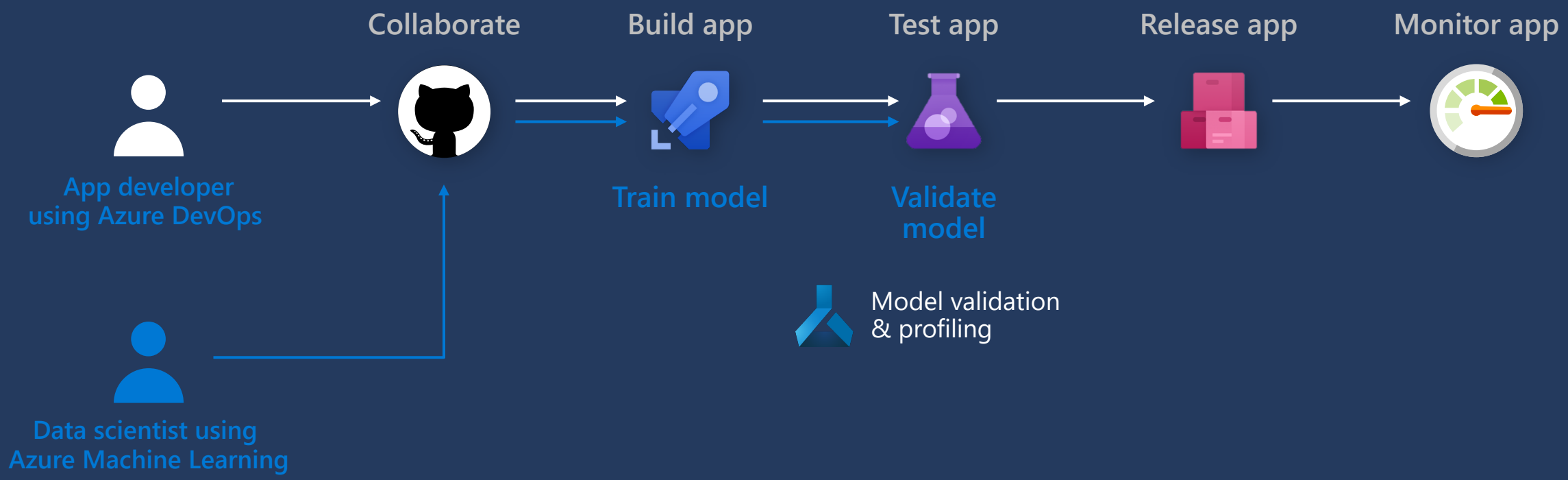


MLDevOps with Azure Machine Learning



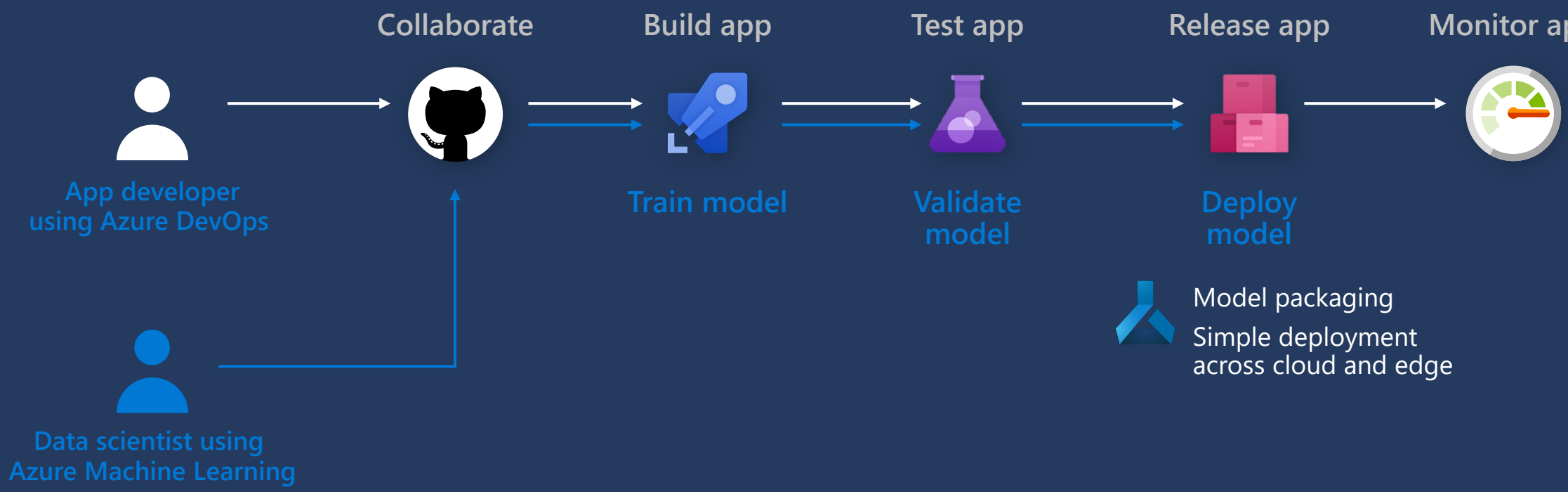
- [x] Model reproducibility
- [ ] Model validation
- [ ] Model deployment
- [ ] Model retraining

MLDevOps with Azure Machine Learning



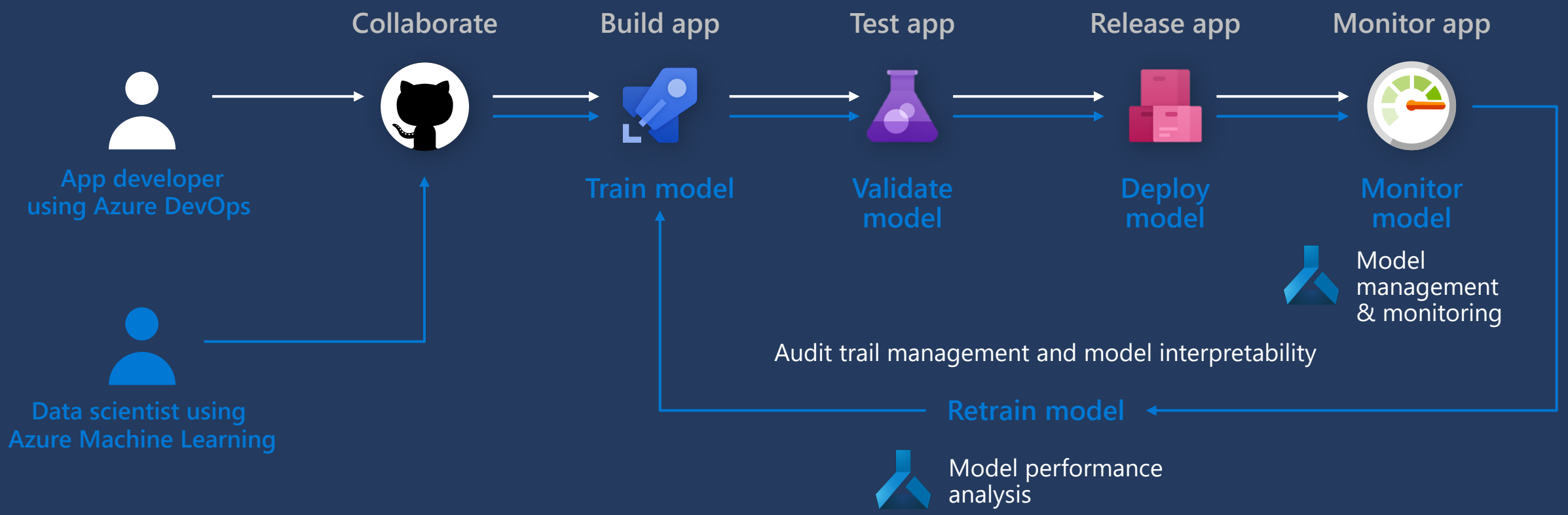
- [x] Model reproducibility
- [x] Model validation
- [ ] Model deployment
- [ ] Model retraining

MLDevOps with Azure Machine Learning



- [x] Model reproducibility
- [x] Model validation
- [x] Model deployment
- [ ] Model retraining

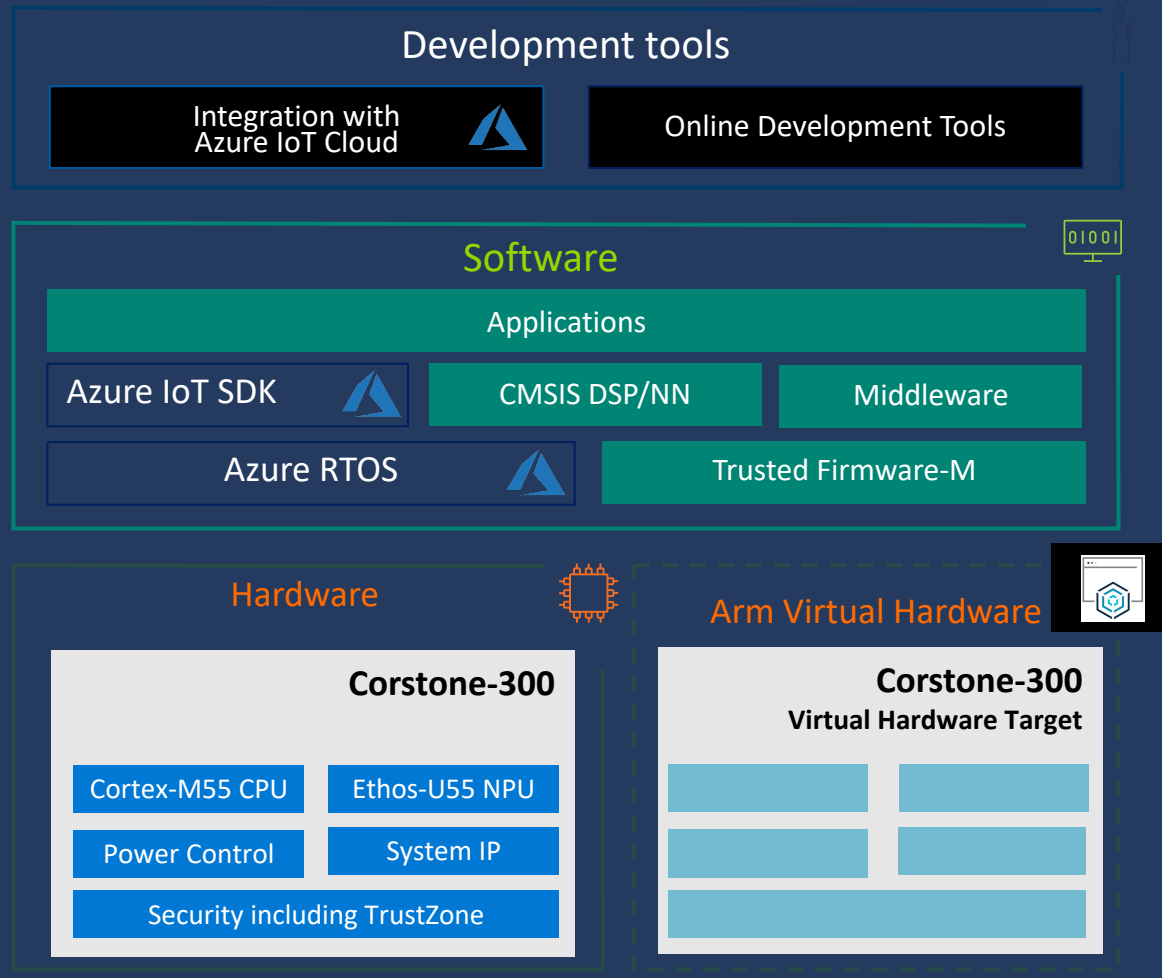
MLDevOps with Azure Machine Learning



- 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

tinyML Summit 2022 Sponsors

Copyright Notice

www.tinyml.org