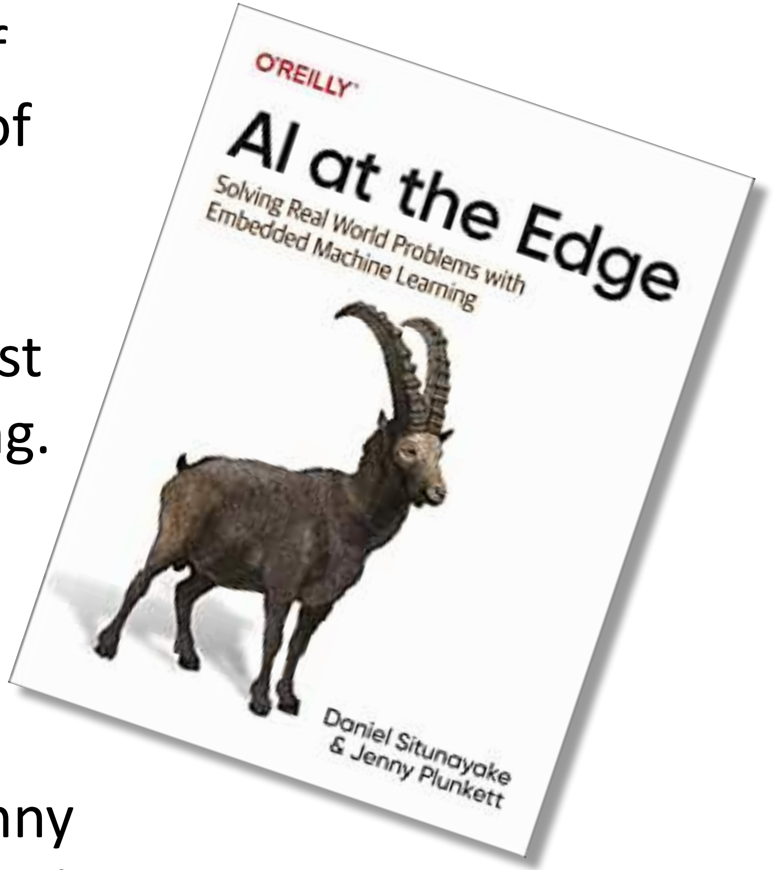
EdgeAI / TinyML

Bringing intelligence to sensors

Prof. Marcelo José Rovai
UNIFEI - Federal University of Itajubá, Brazil
TinyML4D Academic Network Co-Chair

UNIFEI

"**Edge AI** is a truly complete technology. As a topic, it makes use of knowledge from everything from the physical properties of semiconductor electronics all the way up to the engineering of high-level architectures that span devices and the cloud. It demands expertise in the most cutting-edge approaches to artificial intelligence and machine learning along with the most venerable skills of bare-metal embedded software engineering. It makes use of the entire history of computer science and electrical engineering, laid out end to end."



Situnayake, Daniel; Plunkett, Jenny
AI at the Edge (pp. 215-216)
O'Reilly Media

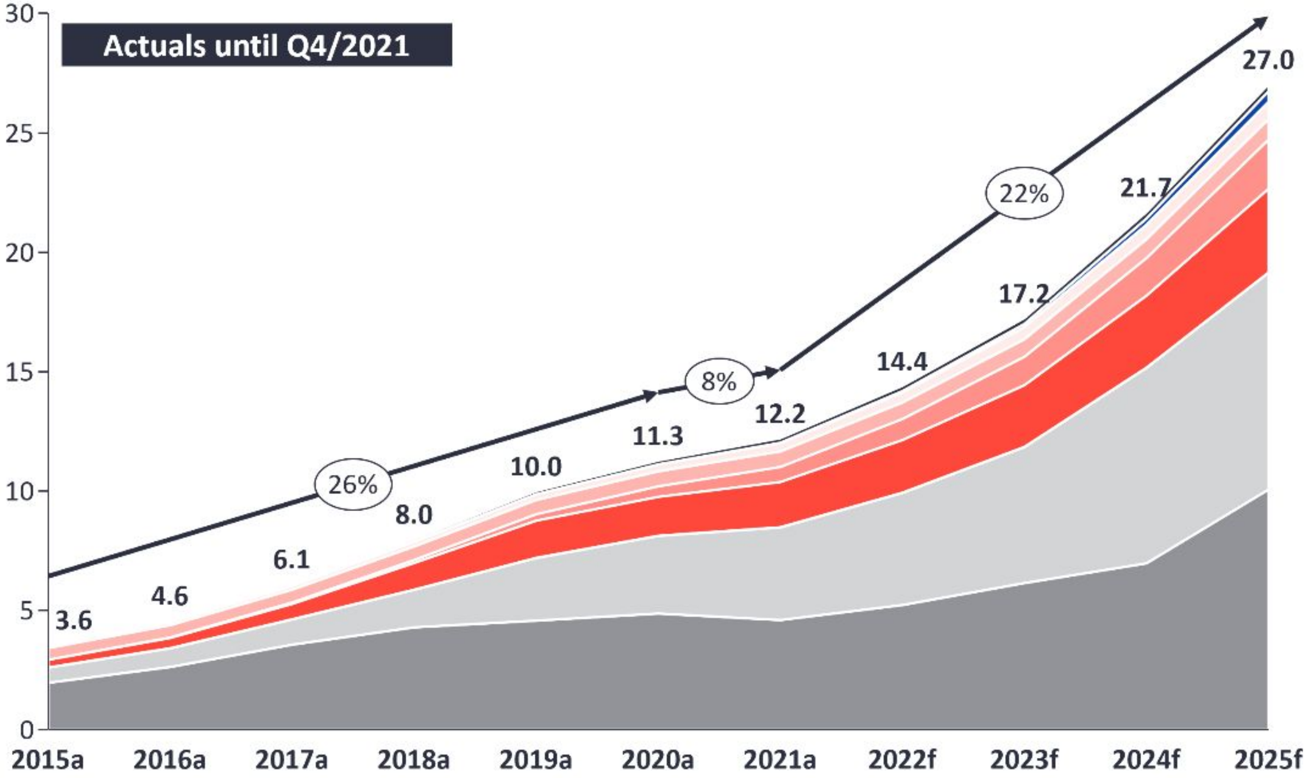
Summary

1. Internet of Things (IoT)
2. Embedded ML (TinyML) - Introduction
3. TinyML Applications - Examples
    - Sound
    - Vision
    - Vibration (Movement)
4. Other Sensors / MCUs / Models
5. The future of Edge AI
6. Q & A

Internet of Things (IoT)

Global IoT Market Forecast [in billion connected IoT devices]

Number of global active IoT Connections (installed base) in Bn



CONNECTIVITY TYPE	CAGR 20-21	CAGR 21-25
Wireless Neighborhood Area Networks (WNAN)	17%	11%
5G IoT	-	159%
Other	22%	20%
Wired IoT	4%	7%
LPWA	42%	34%
Legacy Cellular (2G/3G/4G)	16%	17%
Wireless Local Area Networks (WLAN)	19%	24%
Wireless Personal Area Networks (WPAN)	-6%	22%

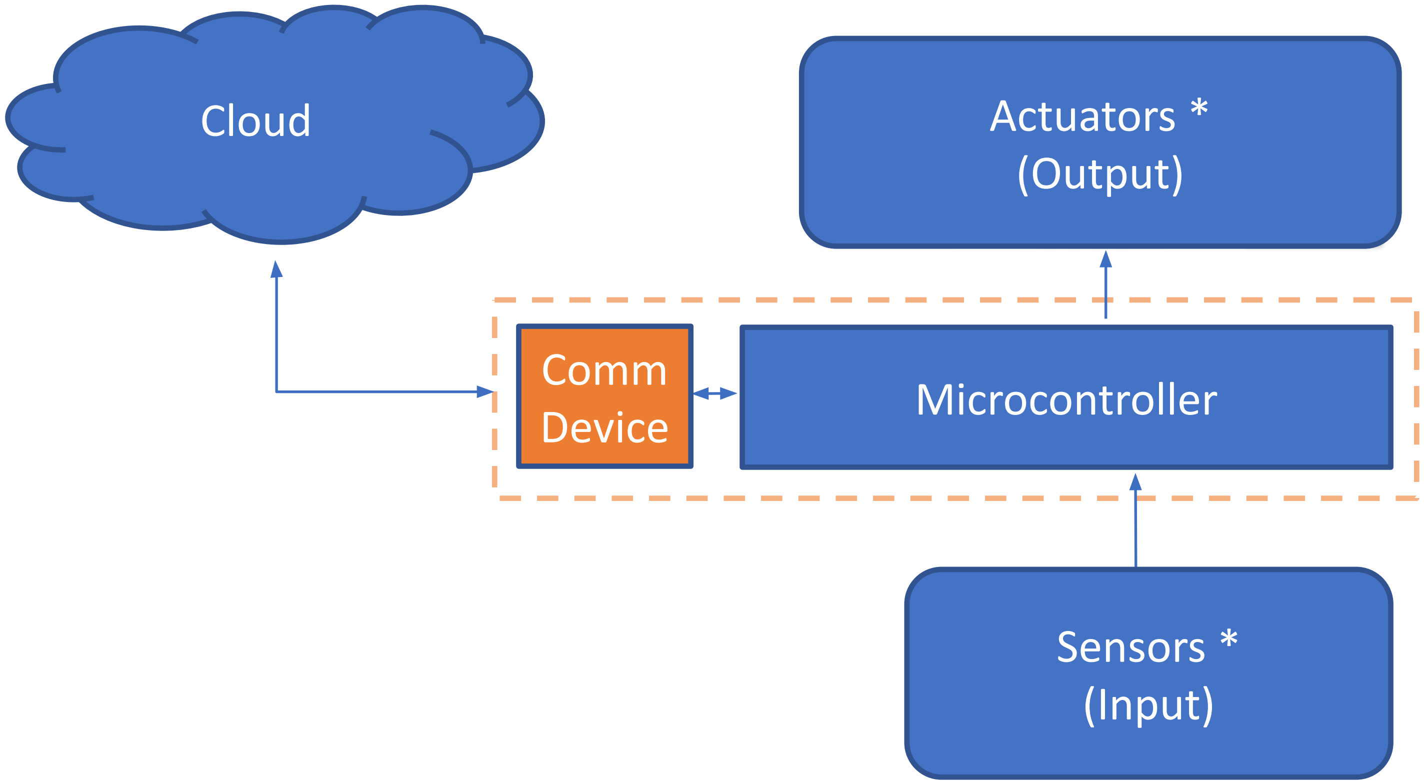
XX% = CAGR

Note: IoT Connections do not include any computers, laptops, fixed phones, cellphones or tablets. Counted are active nodes/devices or gateways that concentrate the end-sensors, not every sensor/actuator. Simple one-directional communications technology not considered (e.g., RFID, NFC). Wired includes Ethernet and Fieldbuses (e.g., connected industrial PLCs or I/O modules); Cellular includes 2G, 3G, 4G; LPWAN includes unlicensed and licensed low-power networks; WPAN includes Bluetooth, Zigbee, Z-Wave or similar; WLAN includes Wi-fi and related protocols; WNAN includes non-short range mesh, such as Wi-SUN; Other includes satellite and unclassified proprietary networks with any range.

Source: IoT Analytics Research 2022. We welcome republishing of images but ask for source citation with a link to the original post and company website.

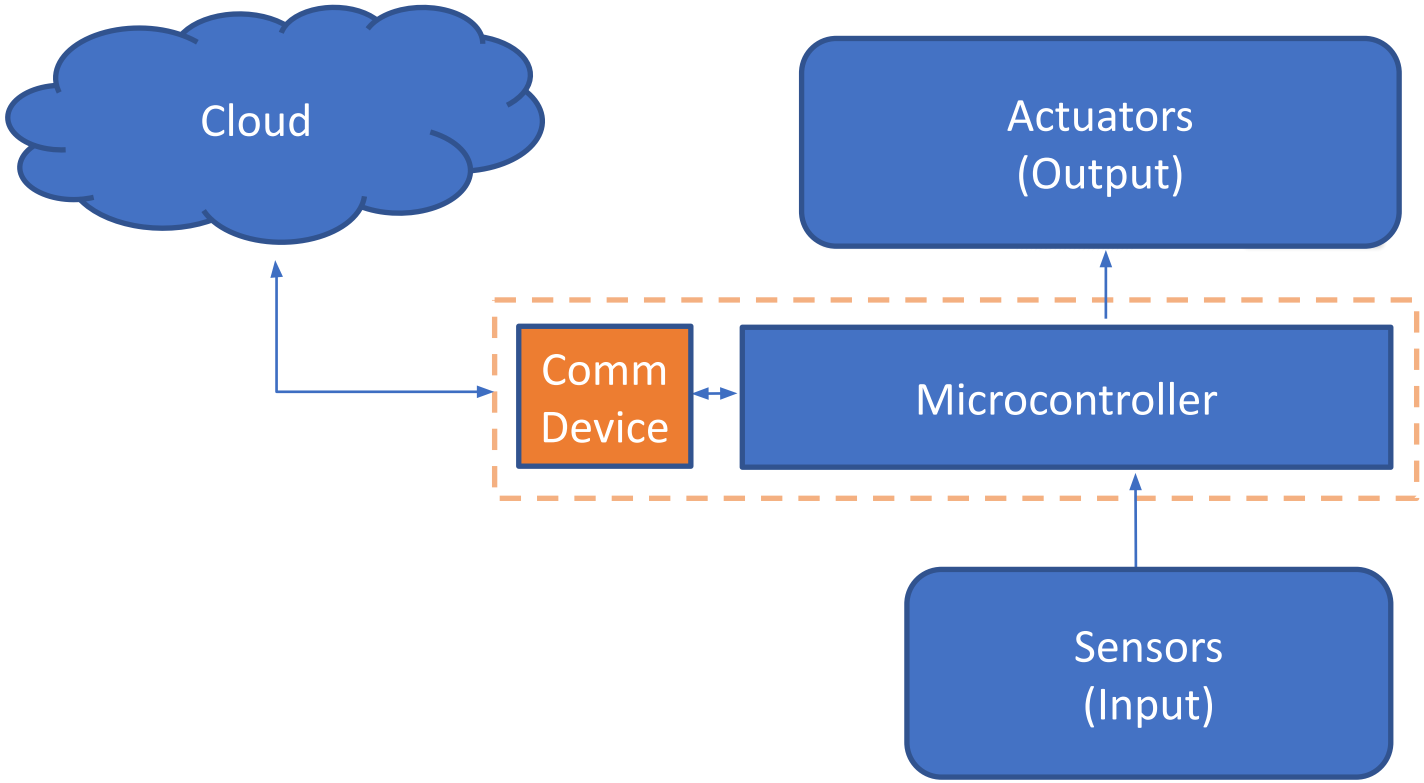
https://iot-analytics.com/number-connected-iot-devices

Typical IoT Project



* "Things"

Typical IoT Project



5 Quintillion

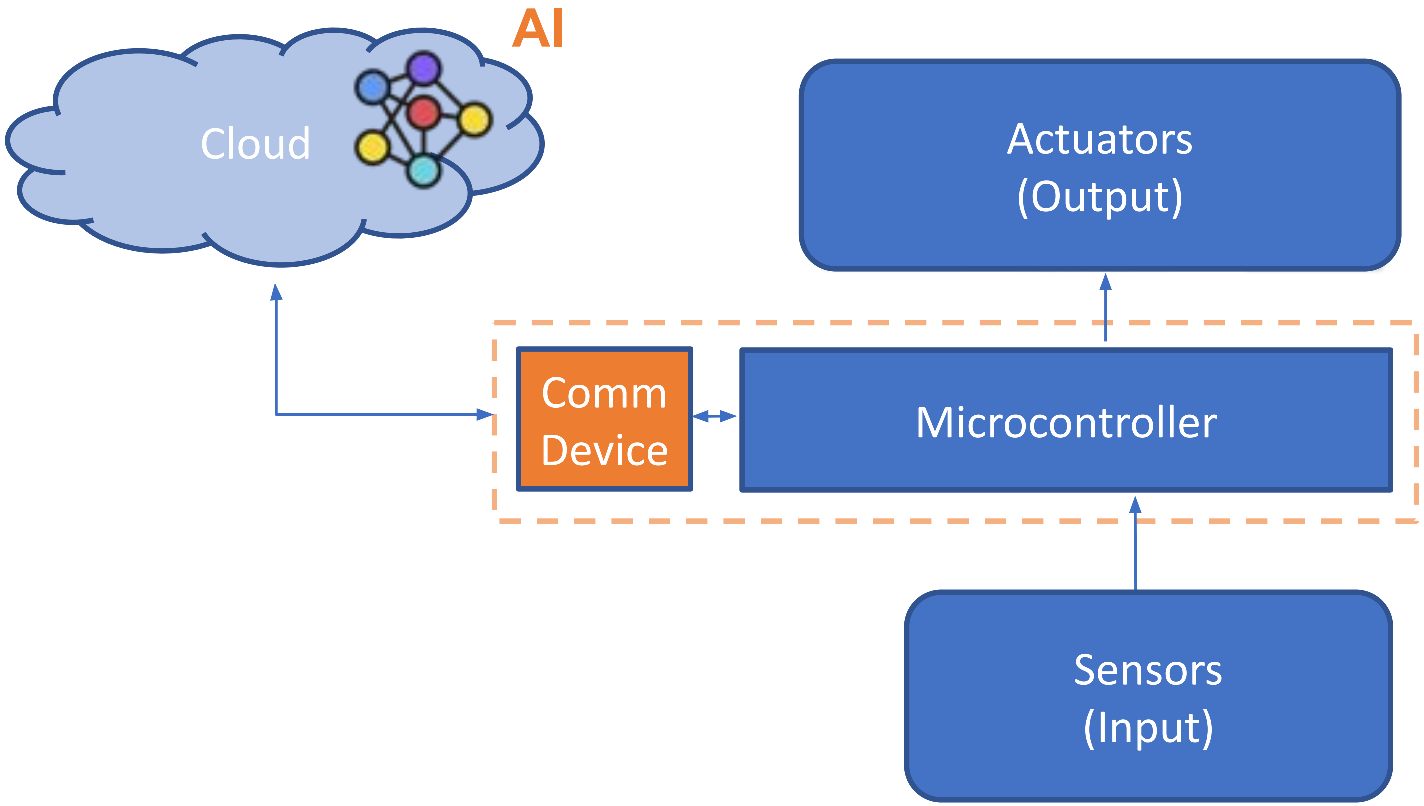
bytes of data produced every day by IoT

<1%

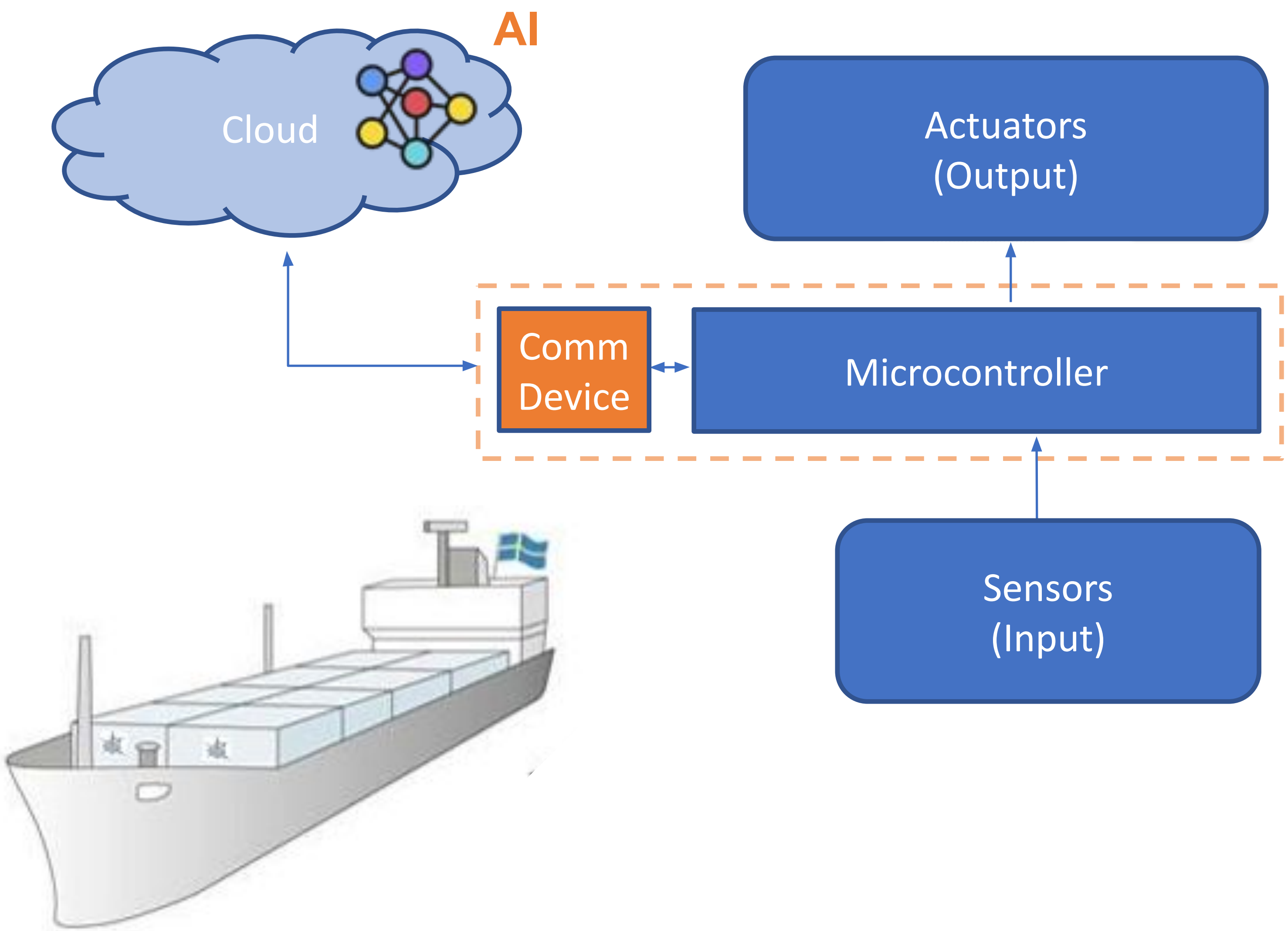
of unstructured data is analyzed or used at all

Source: Harvard Business Review, [What's Your Data Strategy?](), April 18, 2017
Cisco, [Internet of Things (IoT) Data Continues to Explode Exponentially. Who Is Using That Data and How?](), Feb 5, 2018

Typical AIoT Project

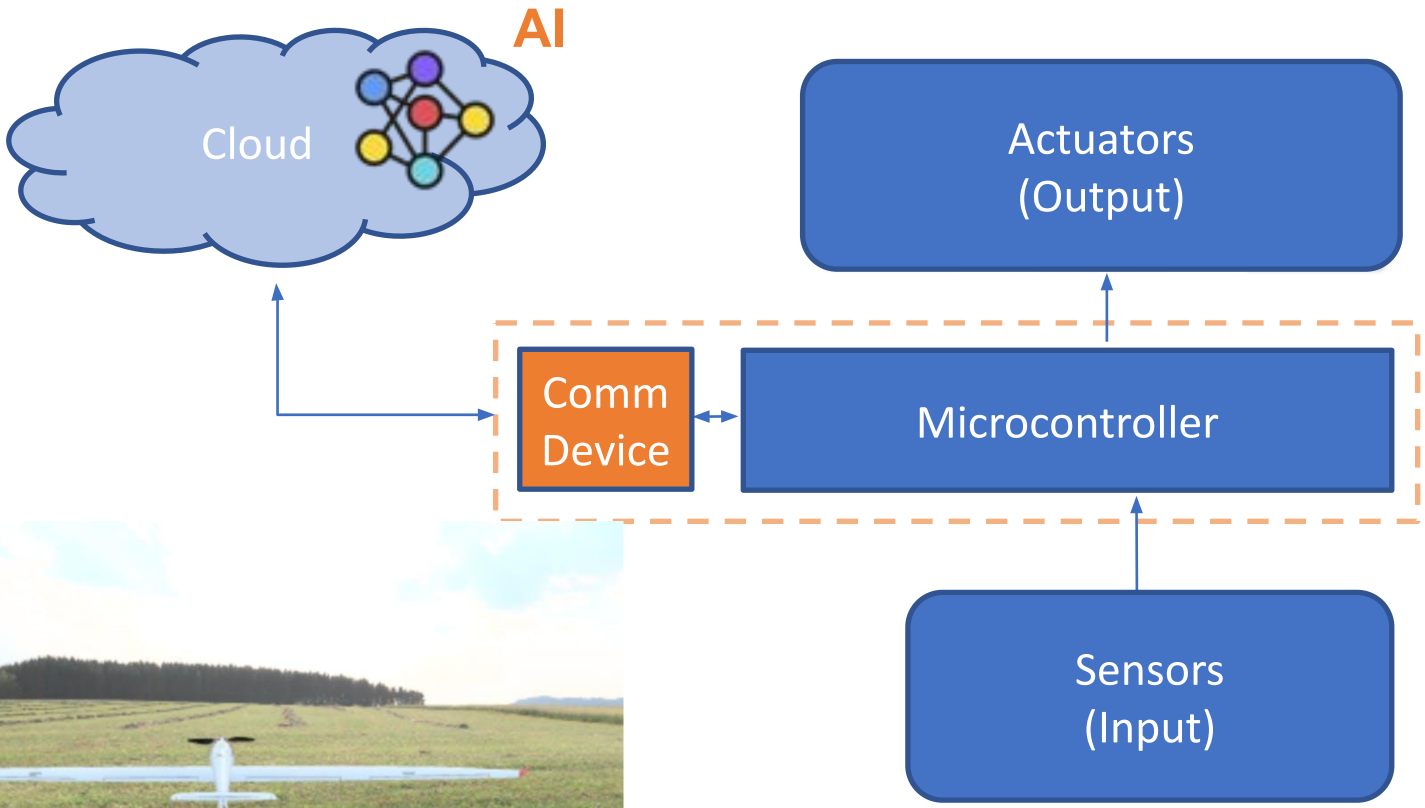


Typical AIoT Project … … Issues



Bandwidth

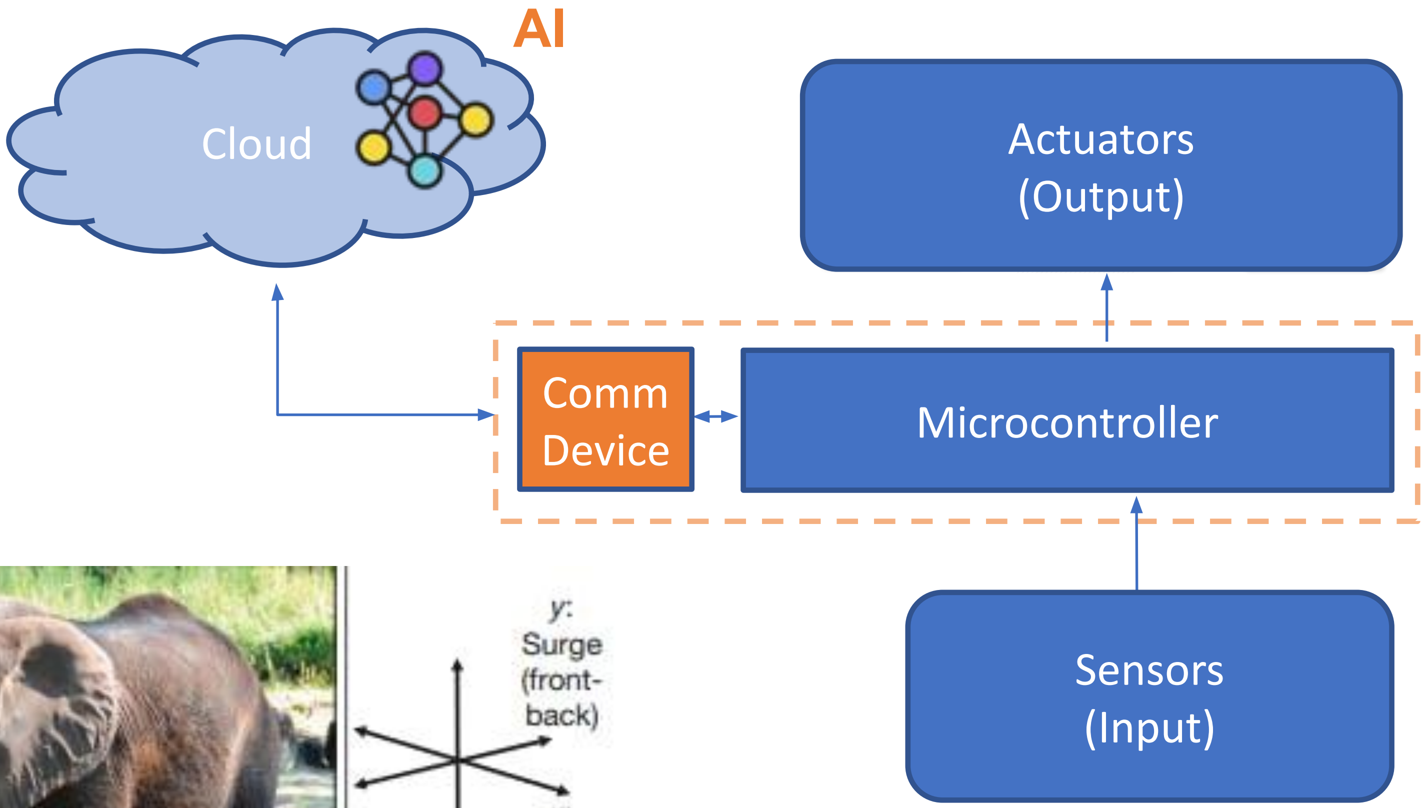
Typical AIoT Project … … Issues

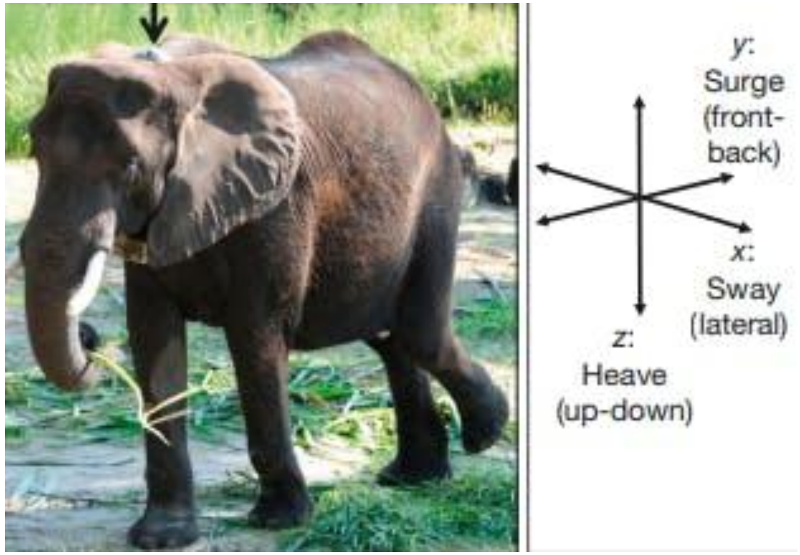


Bandwidth

Latency

Typical AIoT Project ... ... Issues



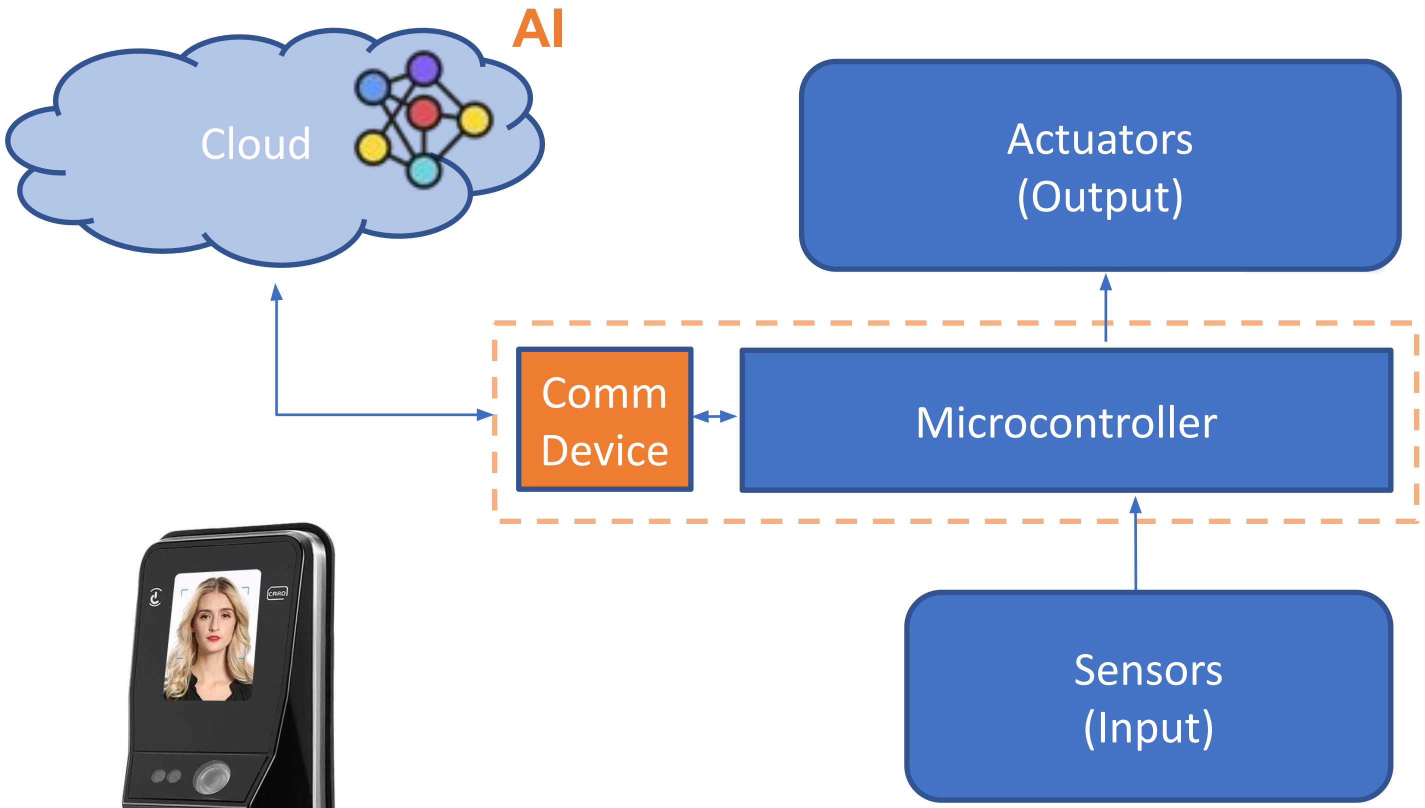


Bandwidth

Latency

Energy

Typical AIoT Project … … Issues



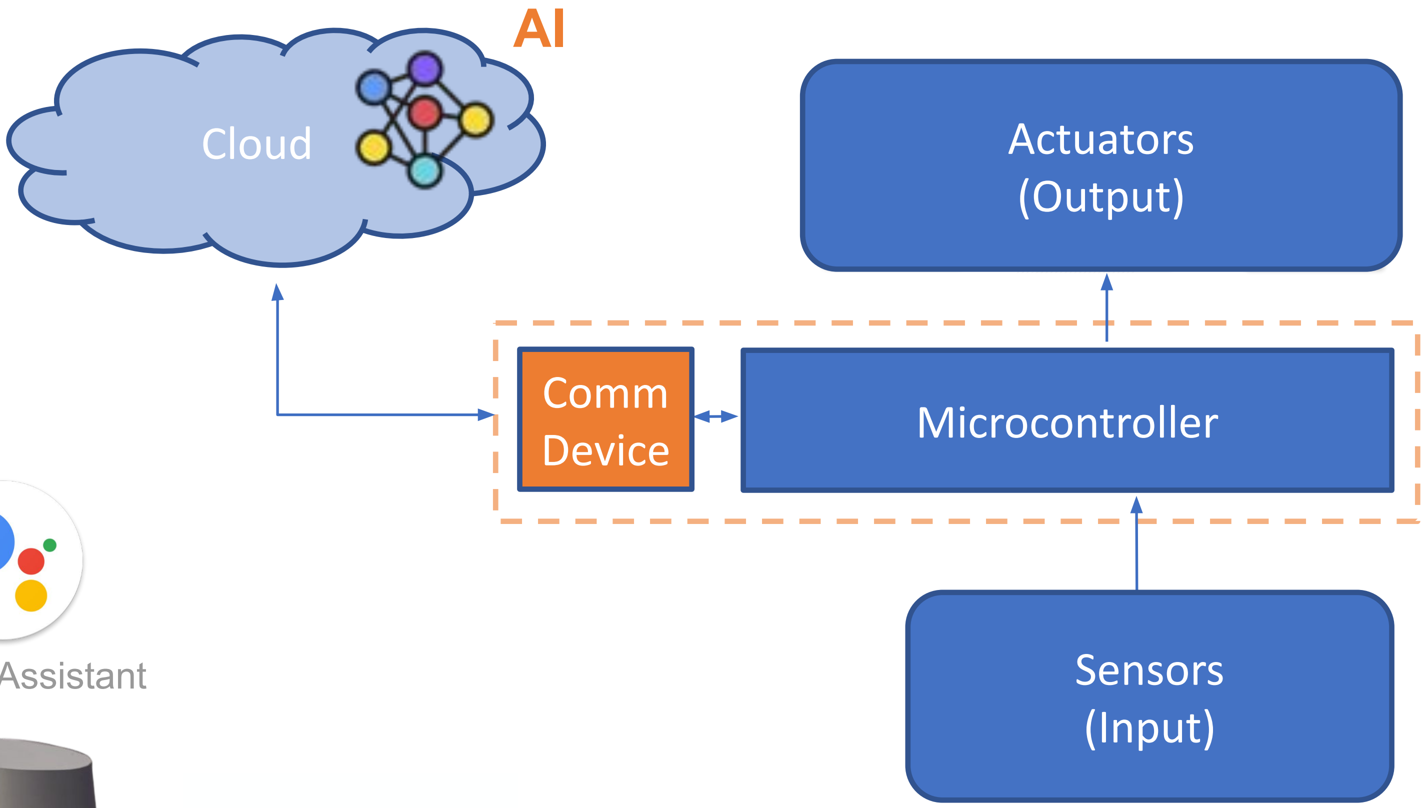
Bandwidth

Latency

Energy

Reliability

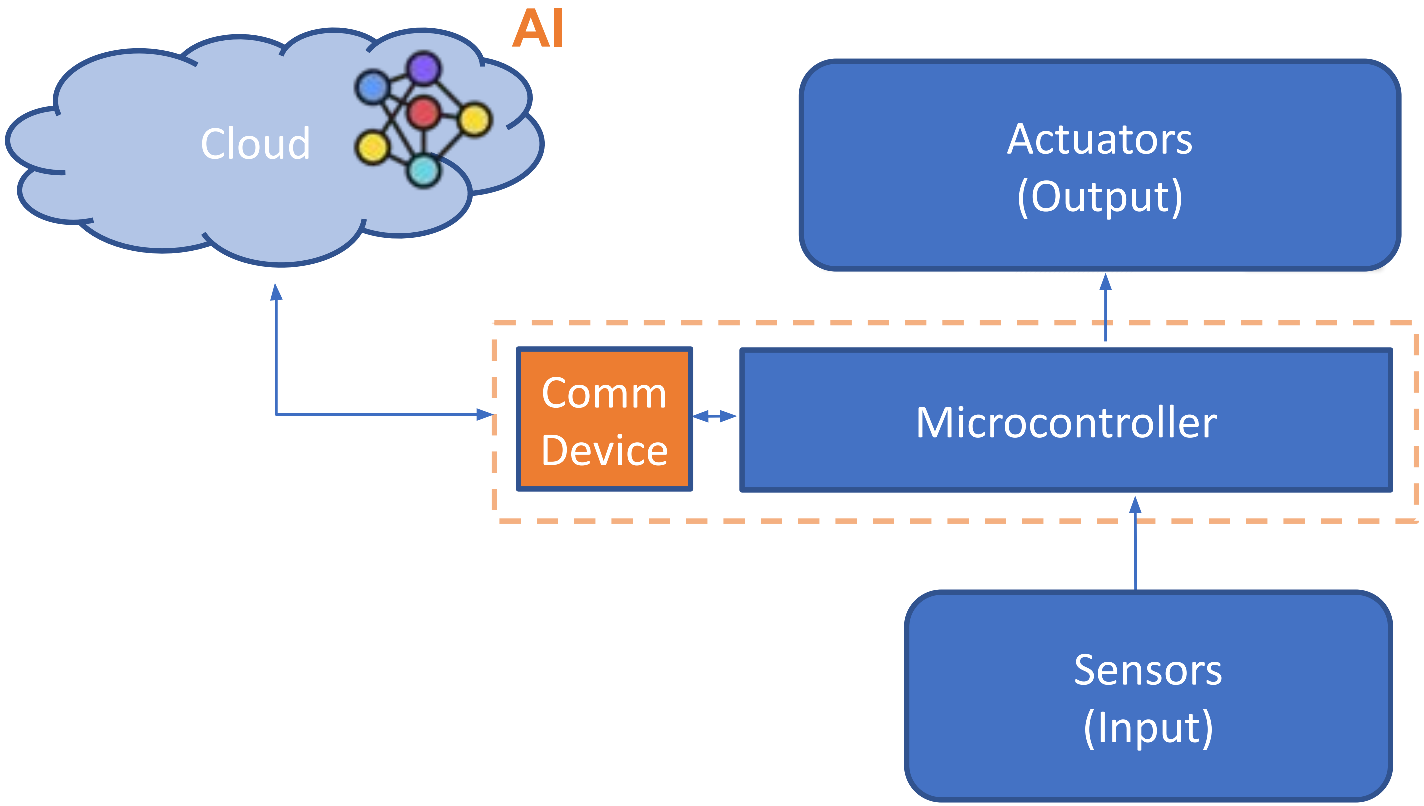
Typical AIoT Project … … Issues



Google Assistant

- **B**andwidth
- **L**atency
- **E**nergy
- **R**eliability
- **P**rivacy

Typical AIoT Project … … Issues



Bandwidth

Latency

Energy

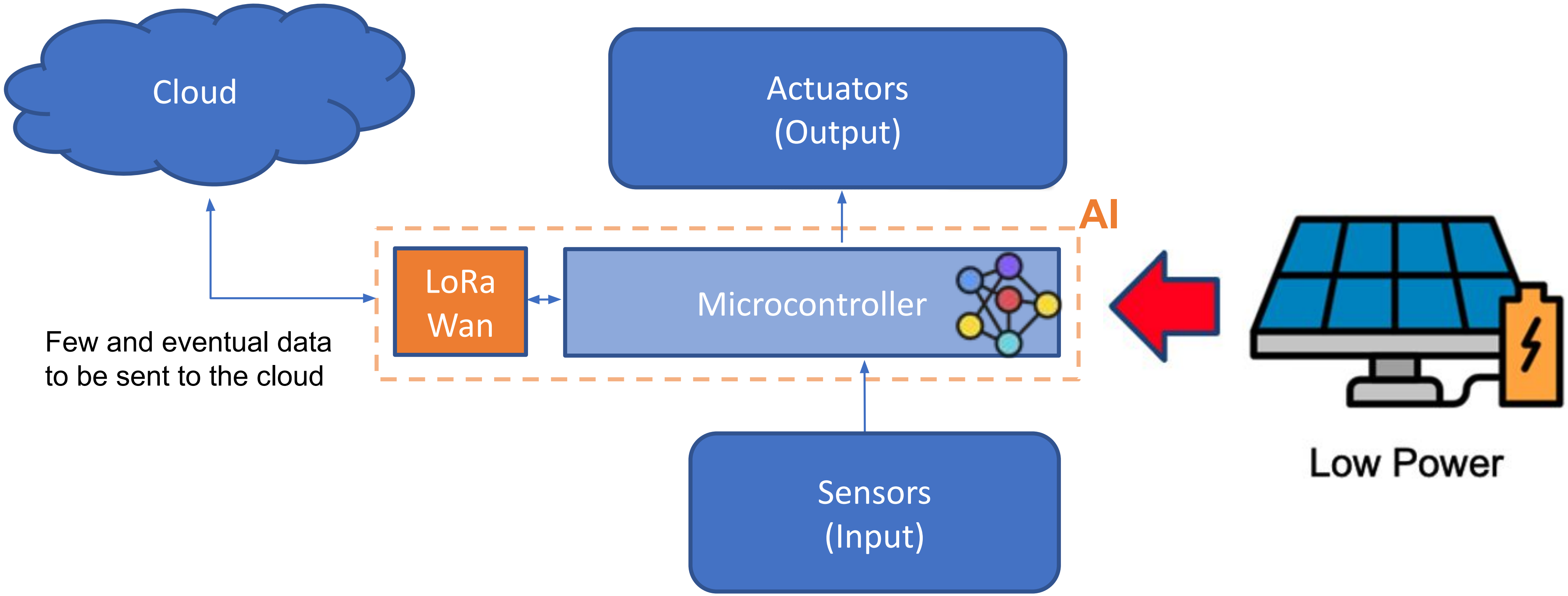
Reliability

Privacy

… **Solution ?**

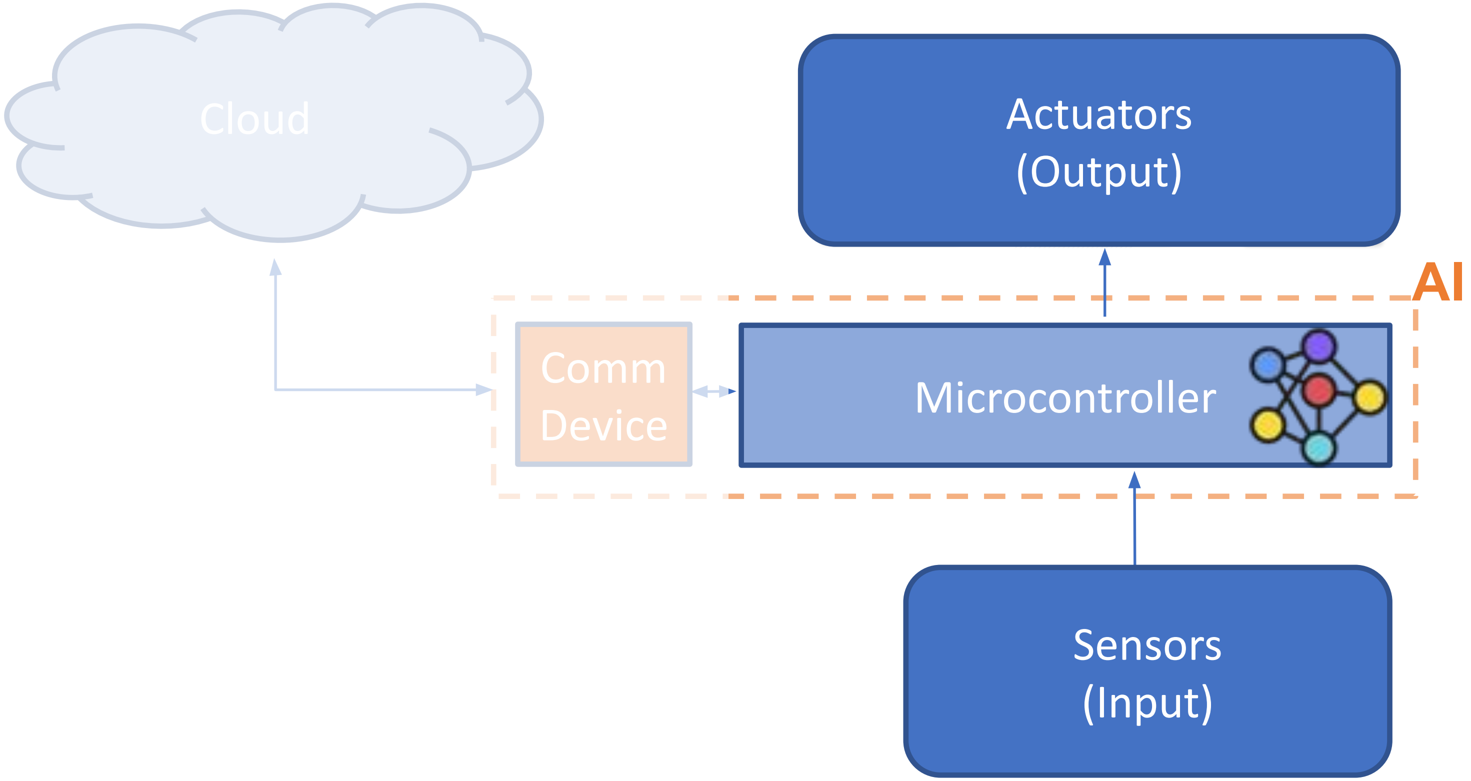
IoT 2.0 * – Edge AI/ML

* Intelligence of Things

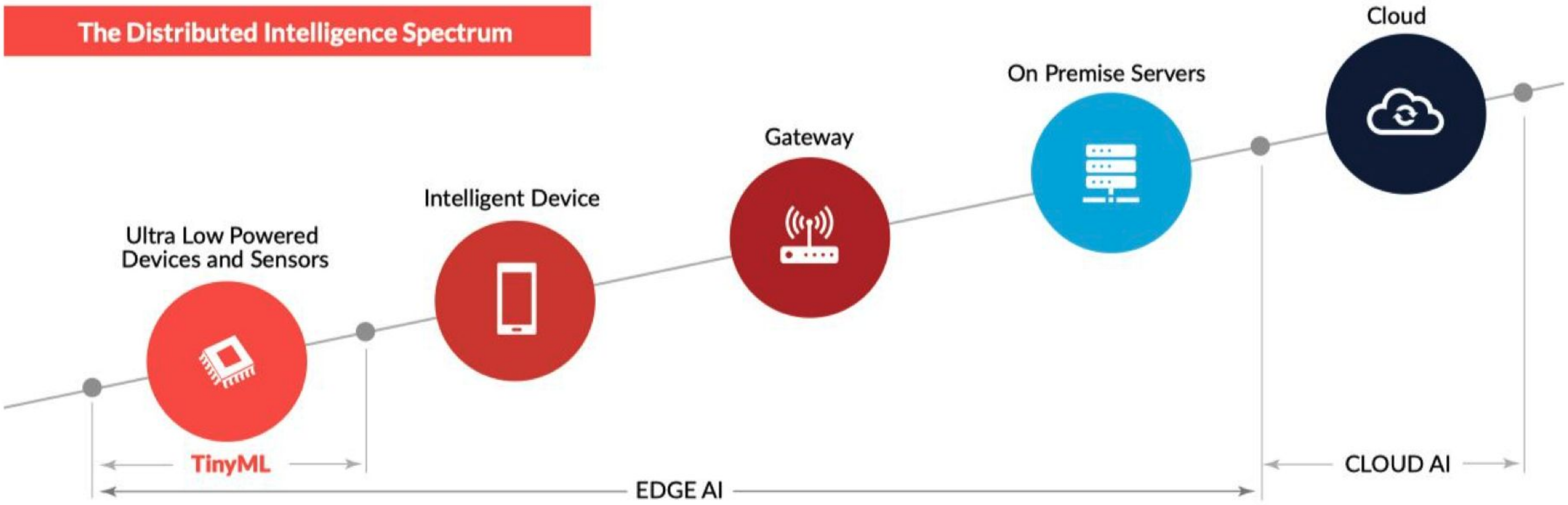


… Solution -> ML goes close to data

When to use an Edge AI/ML approach:

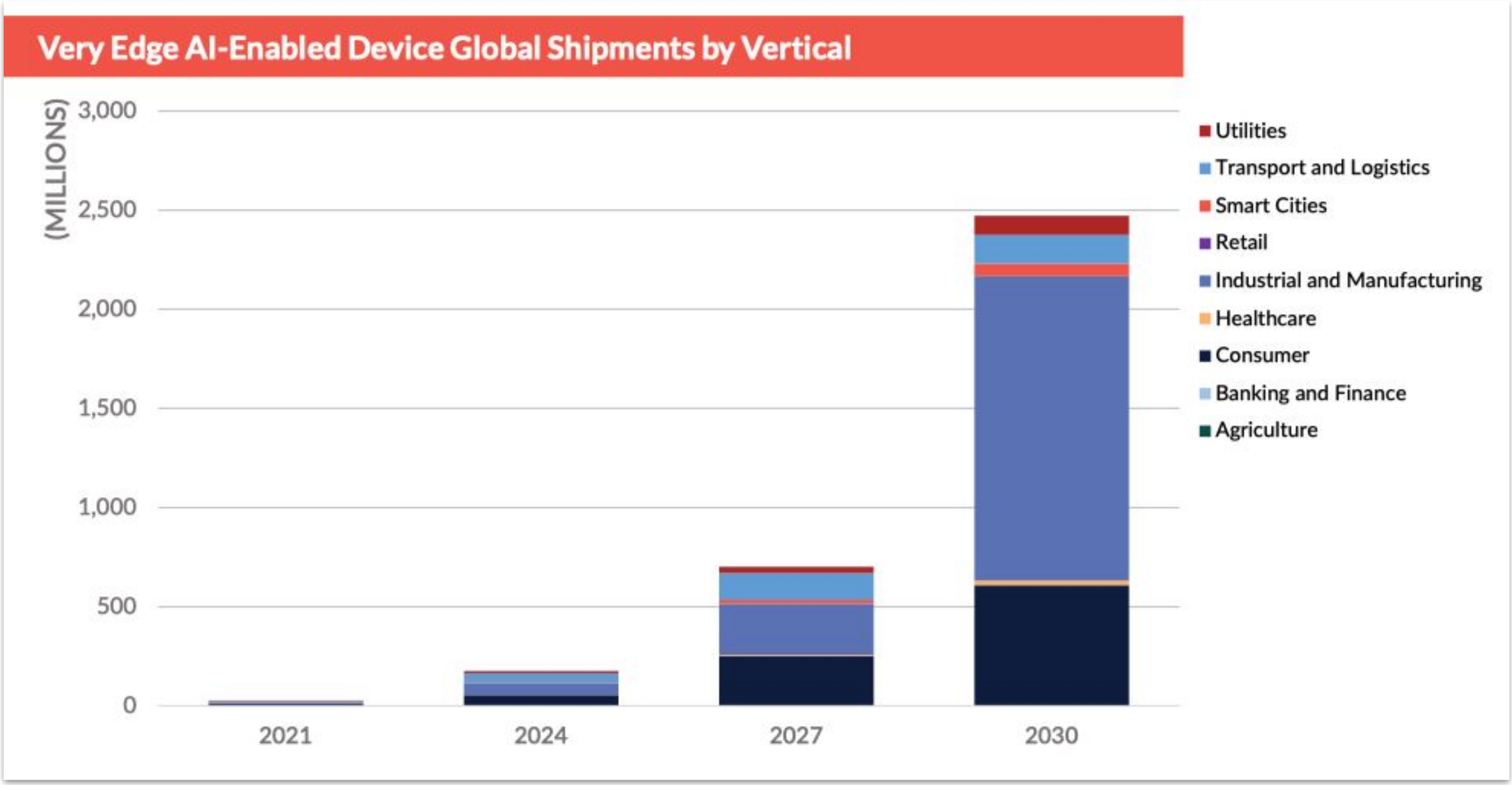


Bandwidth
Latency
Energy
Reliability
Privacy



Source: ABI Research: TinyML

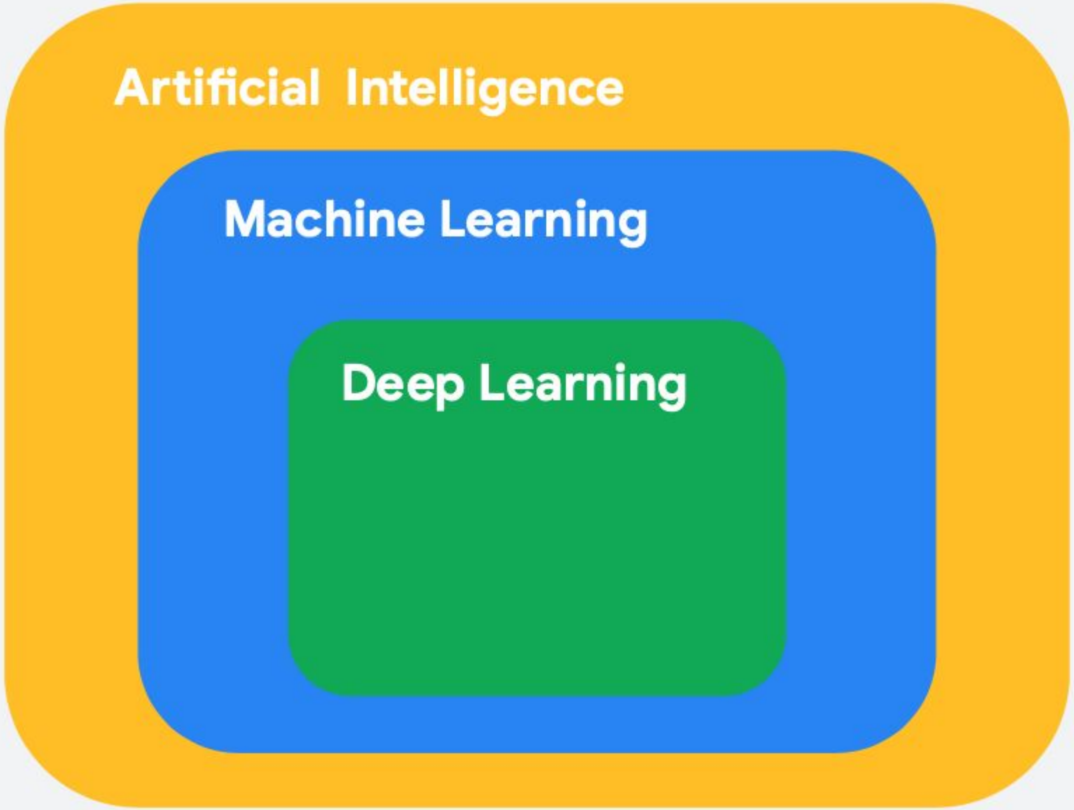
Market Forecast



Source: ABI Research: TinyML

Embedded ML (TinyML)

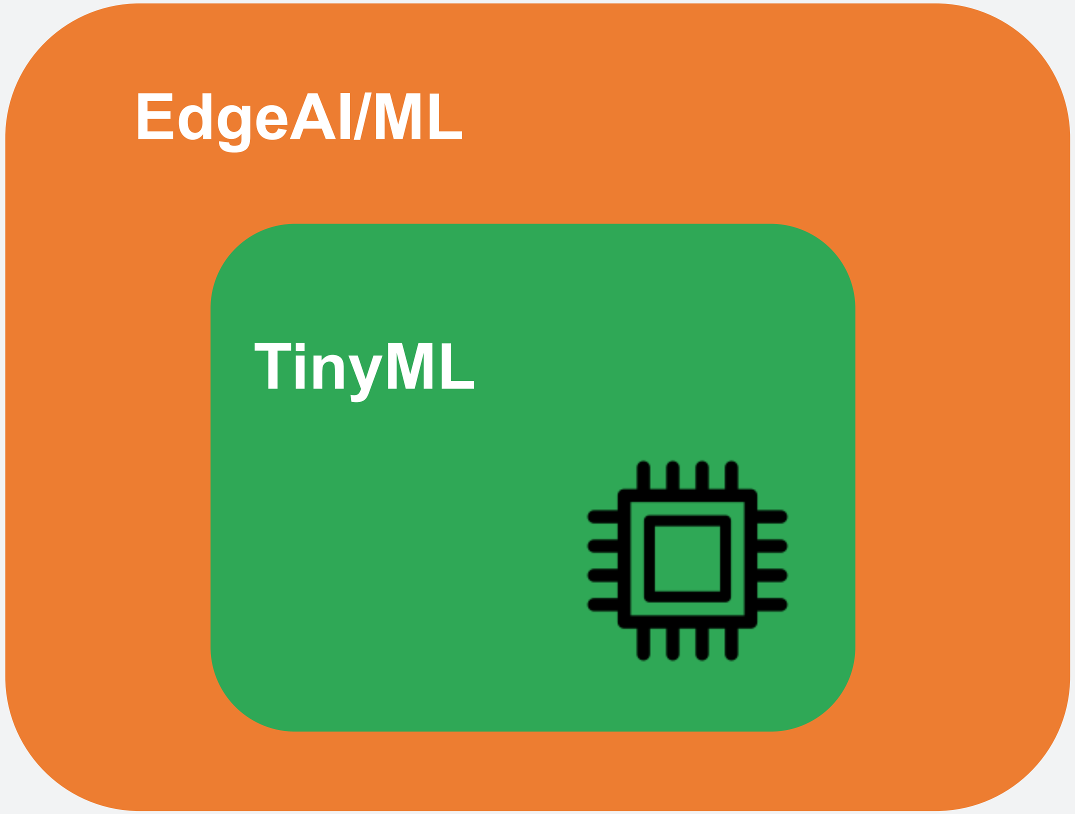
Introduction



AI: Any technique that enables computers to mimic human behavior

ML: Ability to learn without explicitly being programed

DL: Extract patterns from data using neural networks



Edge AI (or Edge ML) is the processing of Artificial Intelligence algorithms on edge, that is, on users' devices. The concept derives from **Edge Computing**, which starts from the same premise: data is stored, processed, and managed directly at the Internet of Things (IoT) endpoints.

TinyML is a subset of EdgeML, where sensors are generating data with ultra-low power consumption (batteries), so that we can ultimately deploy machine learning continuously ("always on devices")

What is Tiny Machine Learning (TinyML)?

TinyML

Fastest-growing field of **ML**

22

What is Tiny Machine Learning (TinyML)?

TinyML

Fastest-growing field of **ML**

Algorithms, hardware, software

23

What is Tiny Machine Learning (TinyML)?

TinyML

Fastest-growing field of **ML**

Algorithms, hardware, software

On-device sensor analytics

 24

What is Tiny Machine Learning (TinyML)?

TinyML

Fastest-growing field of ML

Algorithms, hardware, software

On-device sensor analytics

Low power consumption

25

What is Tiny Machine Learning (**TinyML**)?

TinyML

Fastest-growing field of **ML**

Algorithms, hardware, software

On-device sensor analytics

Low power consumption

Battery-operated

26

What is Tiny Machine Learning (TinyML)?

TinyML

Fastest-growing field of **ML**

Algorithms, hardware, software

On-device sensor analytics

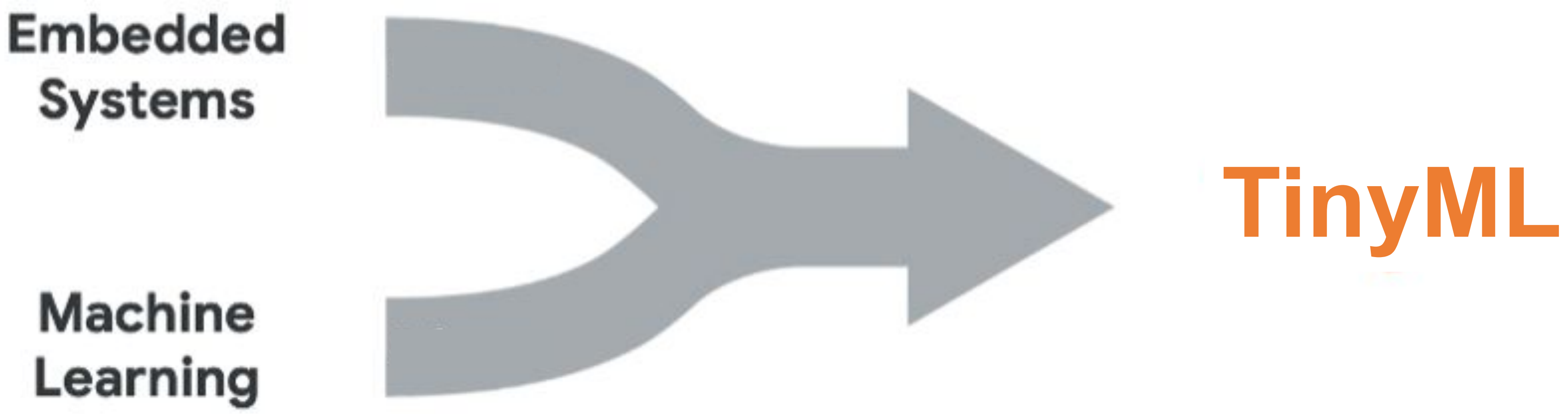
Low power consumption

Battery-operated

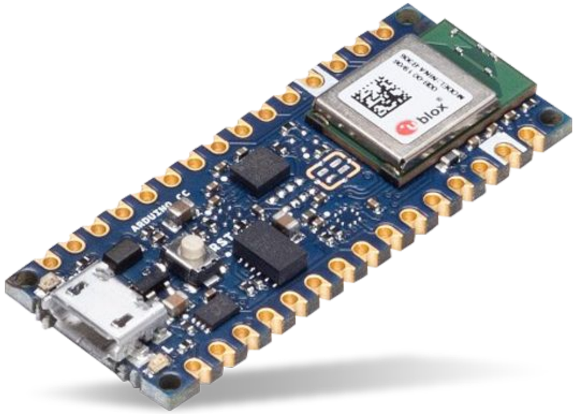
Always-on ML

27

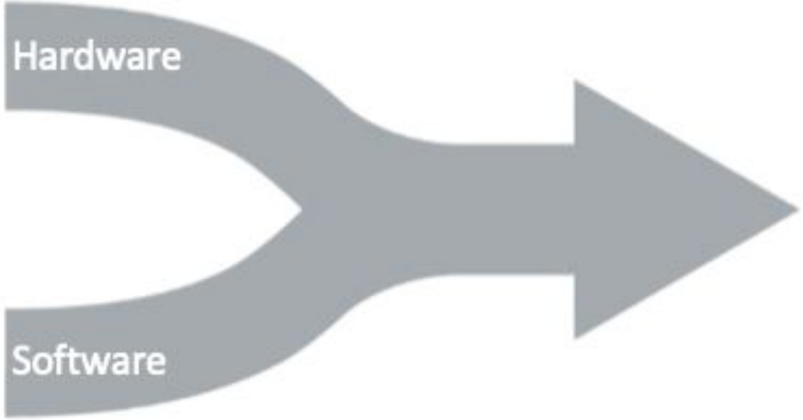
What Makes **TinyML** ?



What Makes TinyML ?

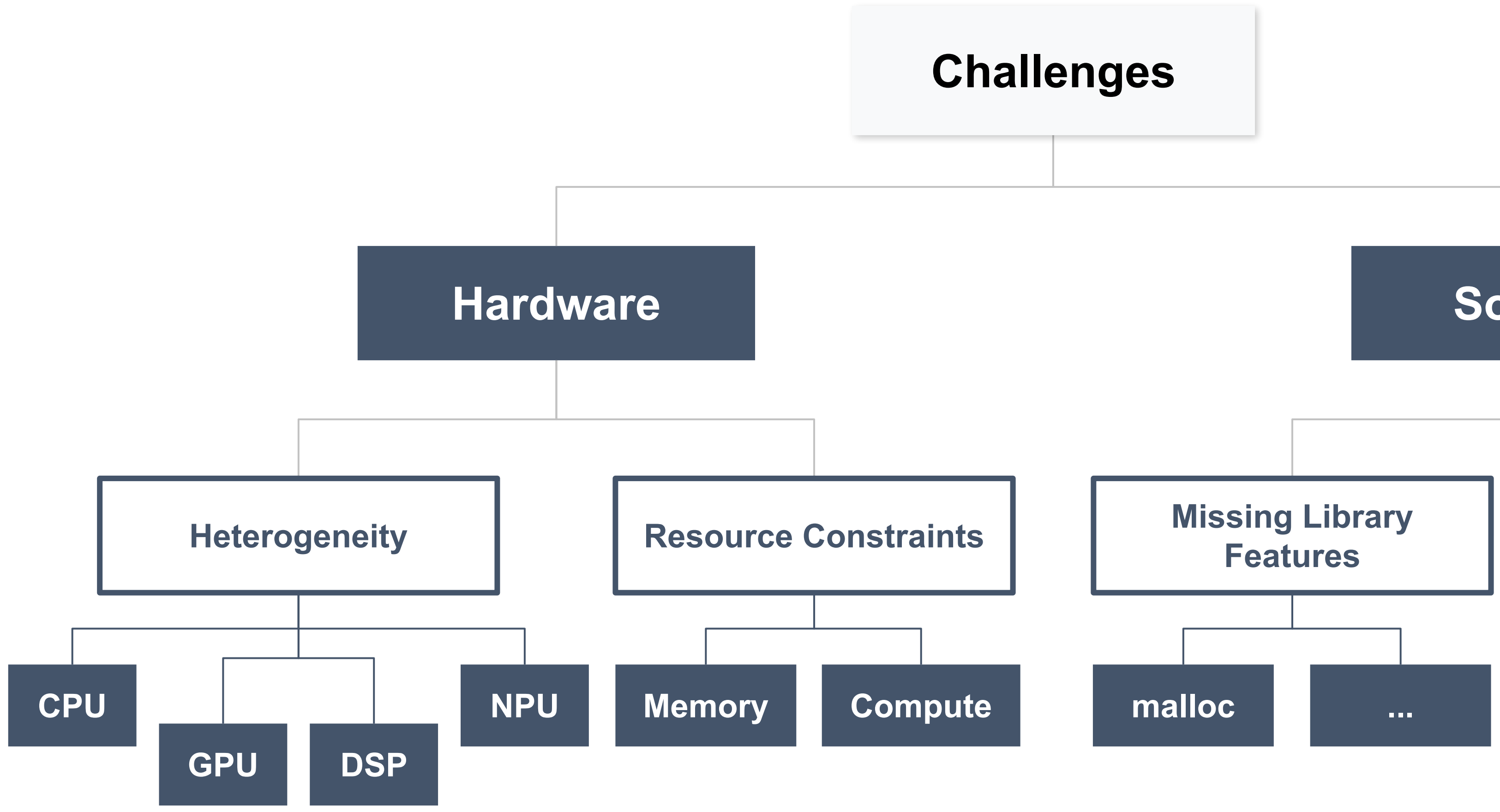


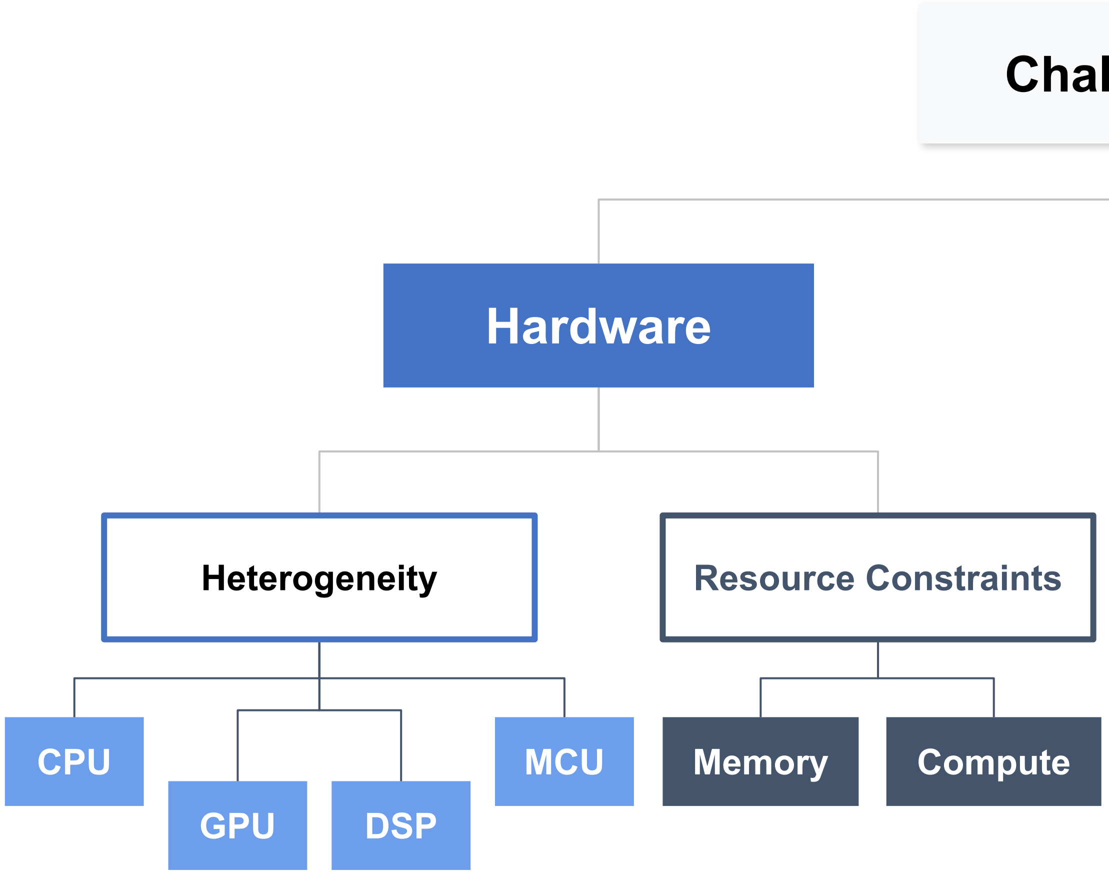
TensorFlow Lite



TinyML

TinyML Challenges


Challenges
Hardware
Software
Heterogeneity
Resource Constraints
Missing Library Features
Limited Operating System Support
CPU
GPU
DSP
NPU
Memory
Compute
malloc
...



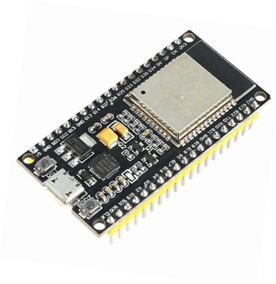
Challenges
Hardware
Software
Heterogeneity
Resource Constraints
Missing Library Features
Limited Operating System Support
CPU
GPU
DSP
MCU
Memory
Compute
malloc
...

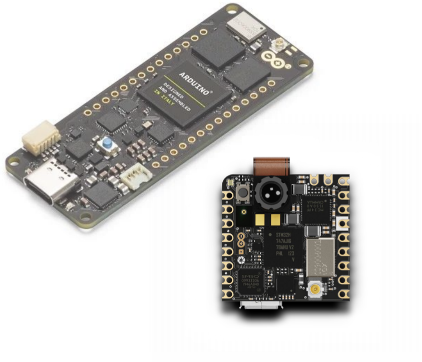
250 Billion

MCUs today

33

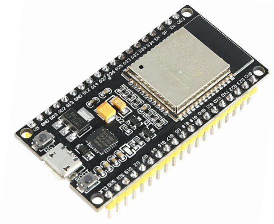
Hardware

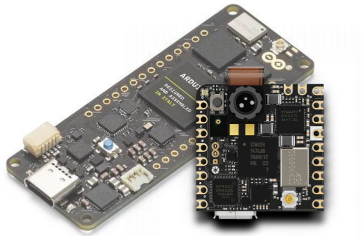




Hardware

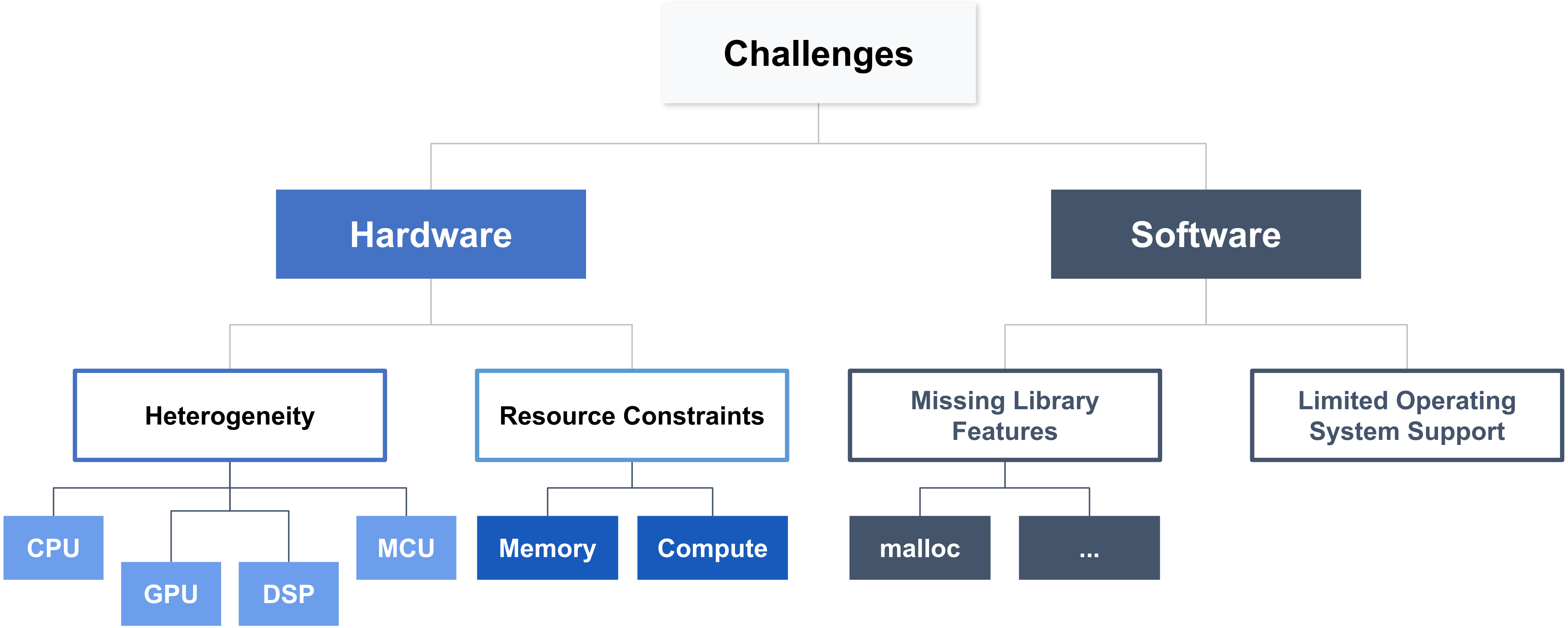




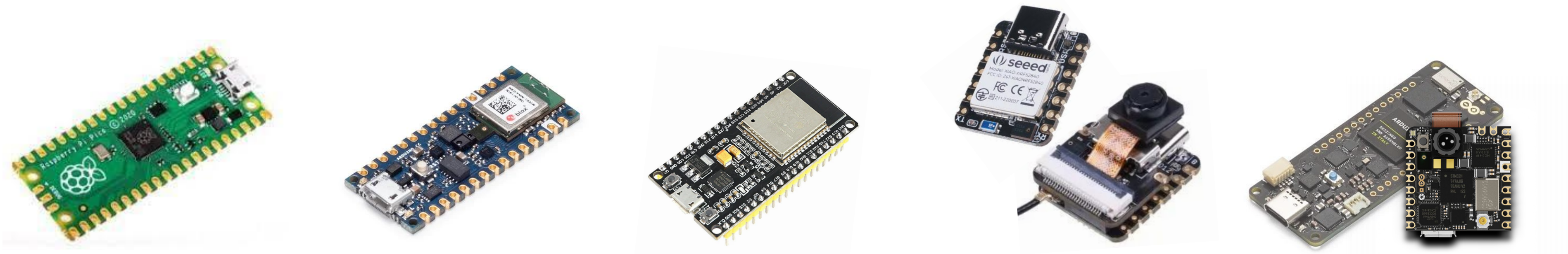


	Raspberry Pico (W)	Arduino Nano Sense	ESP 32	Seeed XIAO Sense / ESP32S3	Arduino Pro
32Bits CPU	Dual-core Arm Cortex-M0+	Arm Cortex-M4F	Xtensa LX6 Dual Core	Arm Cortex-M4F (BLE) Xtensa LX7 Dual Core	Dual Core Arm Cortex M7/M4
CLOCK	133MHz	64MHz	240MHz	64 / 240MHz	480/240MHz
RAM	264KB	256KB	520KB (part available)	256KB / 8MB	1MB
ROM	2MB	1MB	2MB	2MB / 8MB	2MB
Radio	(Yes for W)	BLE	BLE/WiFi	BLE / WiFi (ESP32S3)	BLE/WiFi
Sensors	No	Yes	No	Yes (Sense)	Yes (Nicla)
Bat. Power Manag.	No	No	No	Yes	Yes
Price	$	$$$	$	$$	$$$$$

https://media.digikey.com/Resources/Maker/the-original-guide-to-boards-2022.pdf


Challenges
Hardware
Software
Heterogeneity
Resource Constraints
Missing Library Features
Limited Operating System Support
CPU
GPU
DSP
MCU
Memory
Compute
malloc
...

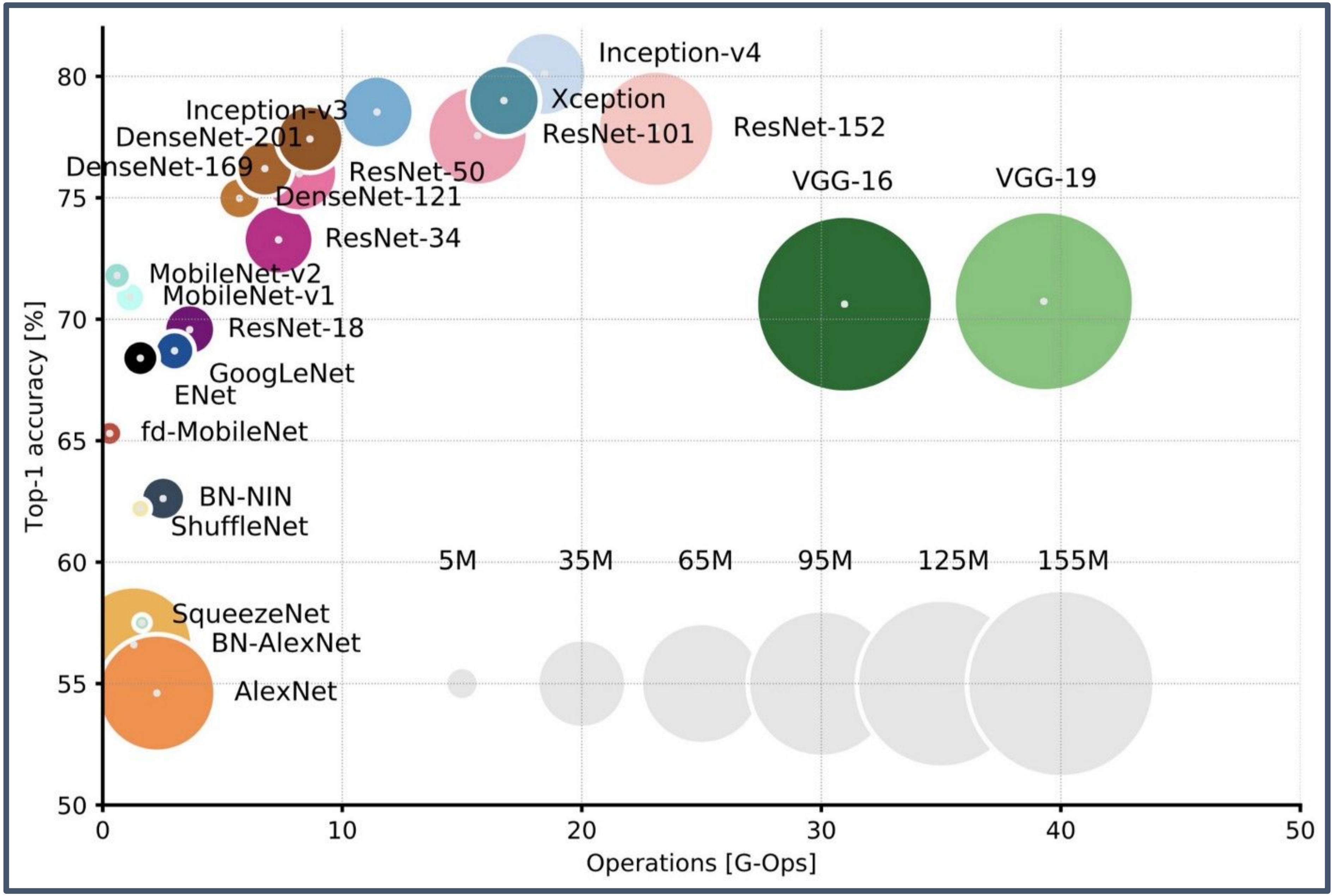
Hardware

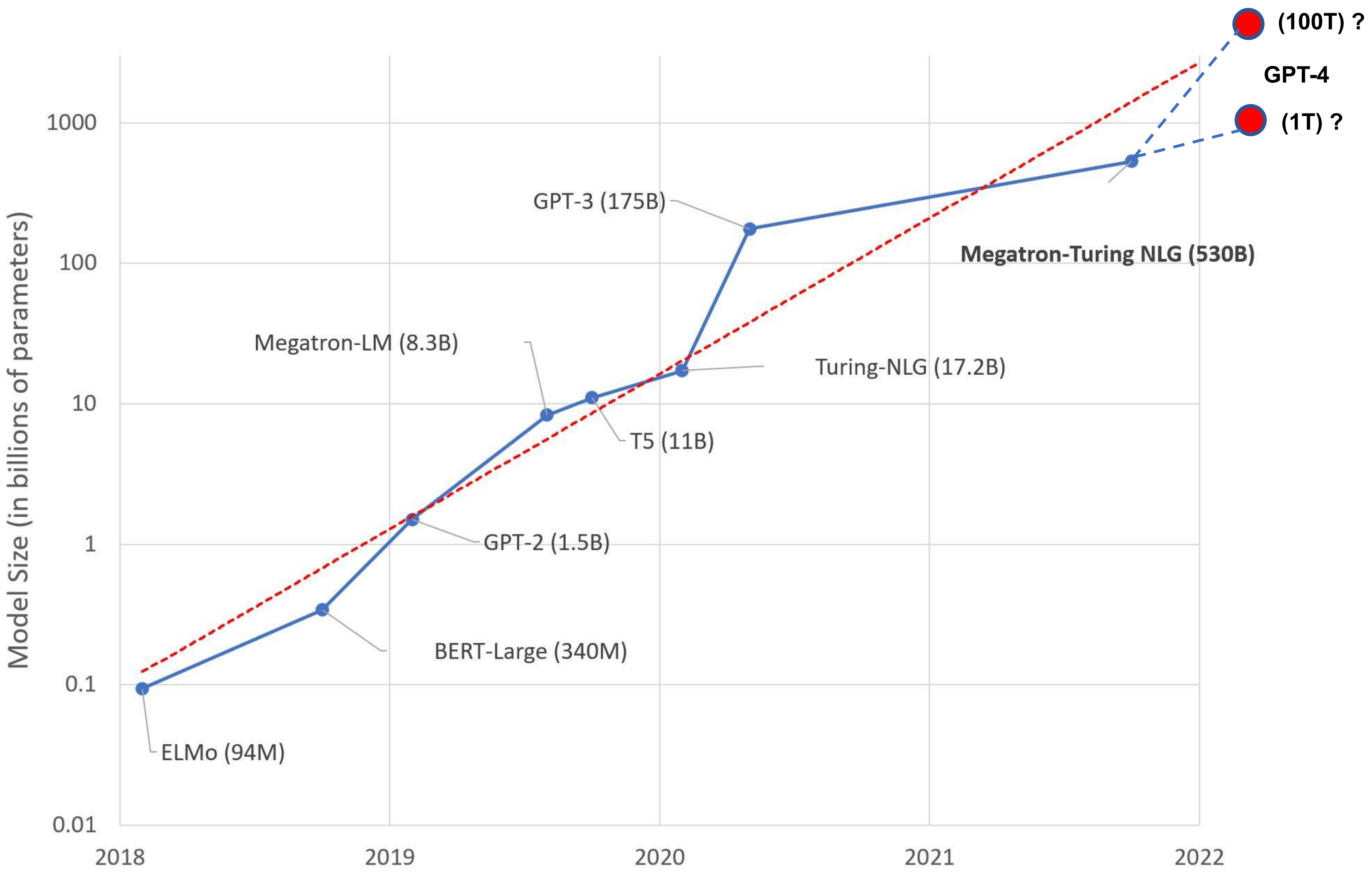


	Raspberry Pico (W)	Arduino Nano Sense	ESP 32	Seeed XIAO Sense / ESP32S3	Arduino Pro
32Bits CPU	Dual-core Arm Cortex-M0+	Arm Cortex-M4F	Xtensa LX6 Dual Core	Arm Cortex-M4F (BLE) Xtensa LX7 Dual Core	Dual Core Arm Cortex M7/M4
CLOCK	133MHz	64MHz	240MHz	64 / 240MHz	480/240MHz
RAM	264KB	256KB	520KB (part available)	256KB / 8MB	1MB
ROM	2MB	1MB	2MB	2MB / 8MB	2MB
Radio	(Yes for W)	BLE	BLE/WiFi	BLE / WiFi (ESP32S3)	BLE/WiFi
Sensors	No	Yes	No	Yes (Sense)	Yes (Nicla)
Bat. Power Manag.	No	No	No	Yes	Yes
Price	$	$$$	$	$$	$$$$$

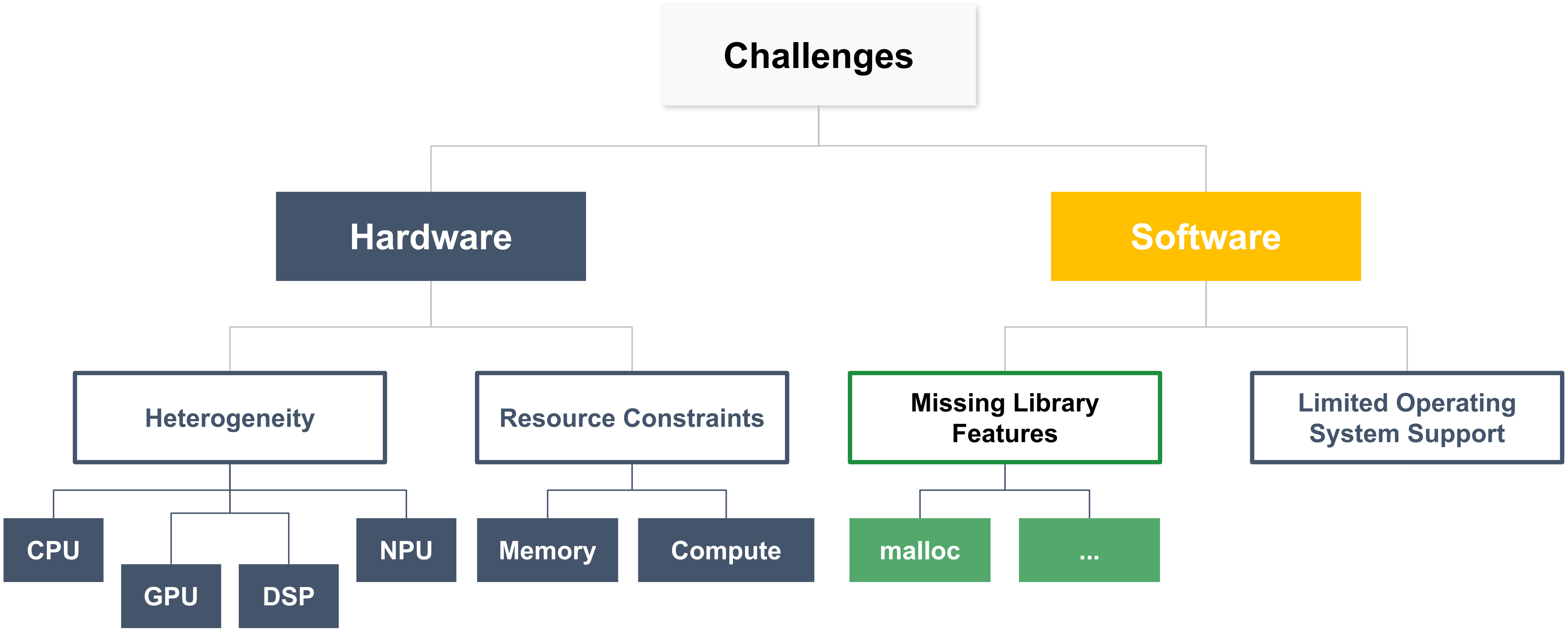
https://media.digikey.com/Resources/Maker/the-original-guide-to-boards-2022.pdf

https://arxiv.org/pdf/1910.01108.pdf
https://towardsdatascience.com/neural-network-architectures-156e5bad51ba





https://huggingface.co/blog/large-language-models https://arxiv.org/abs/2303.08774


Challenges
Hardware
Software
Heterogeneity
Resource Constraints
Missing Library Features
Limited Operating System Support
CPU
GPU
DSP
NPU
Memory
Compute
malloc
...

Datasets
Preprocessing

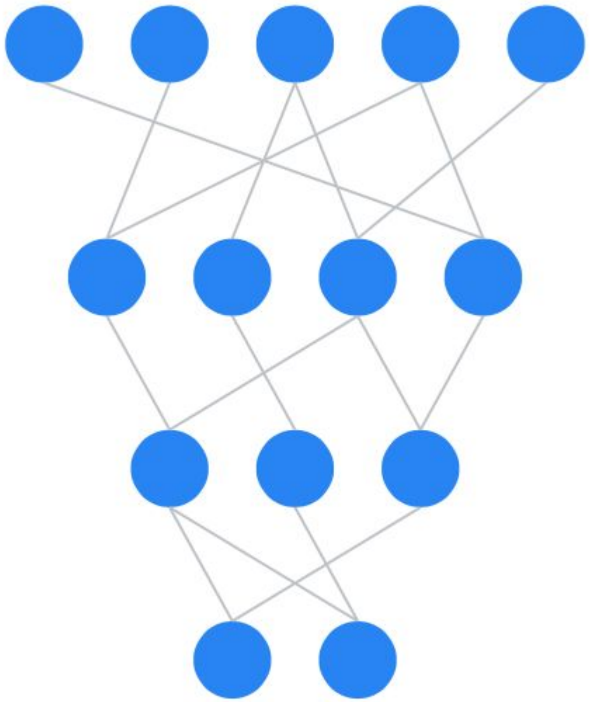
Quantization
Pruning

Resource
constraints

Sound

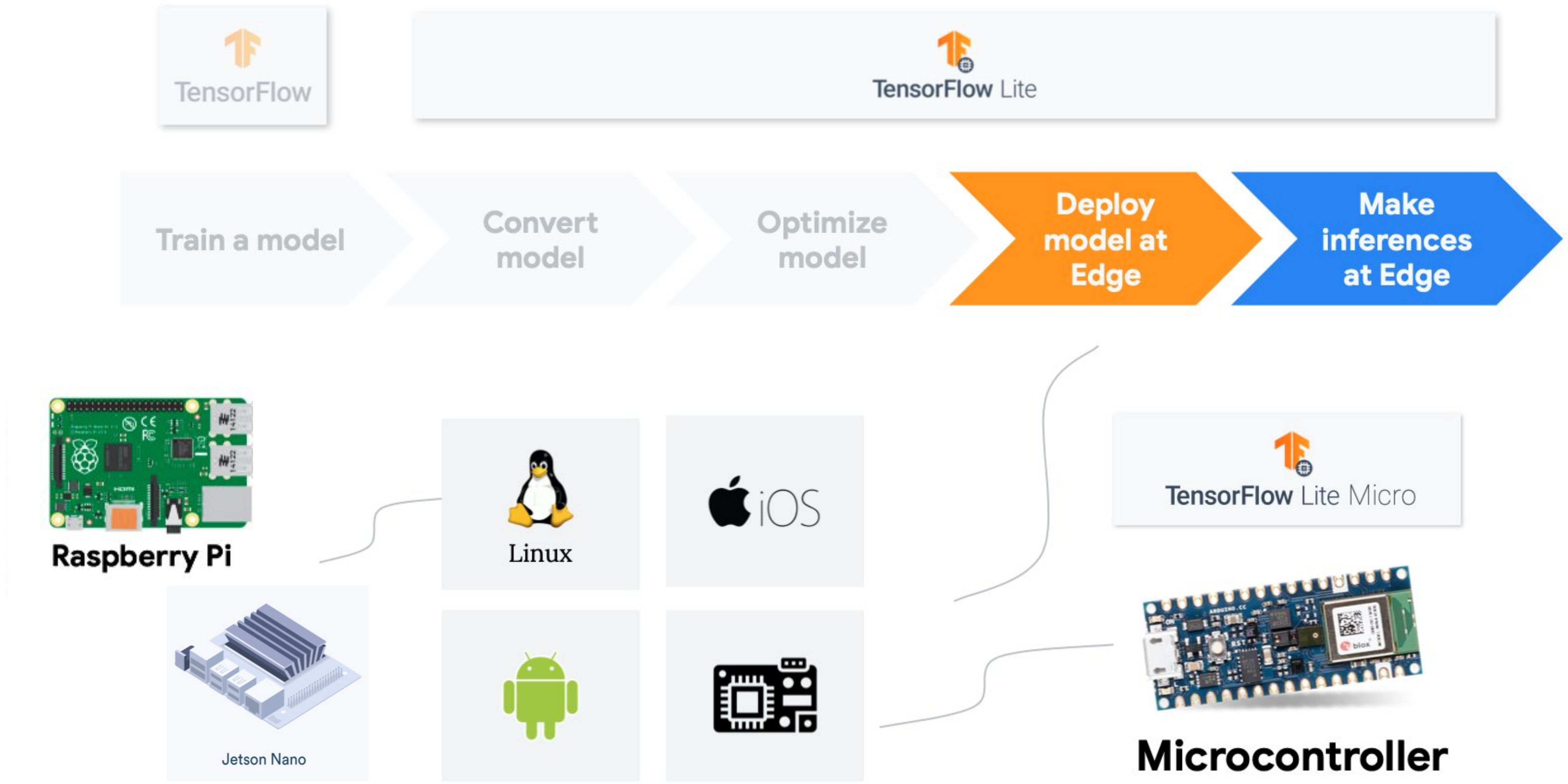
Vision

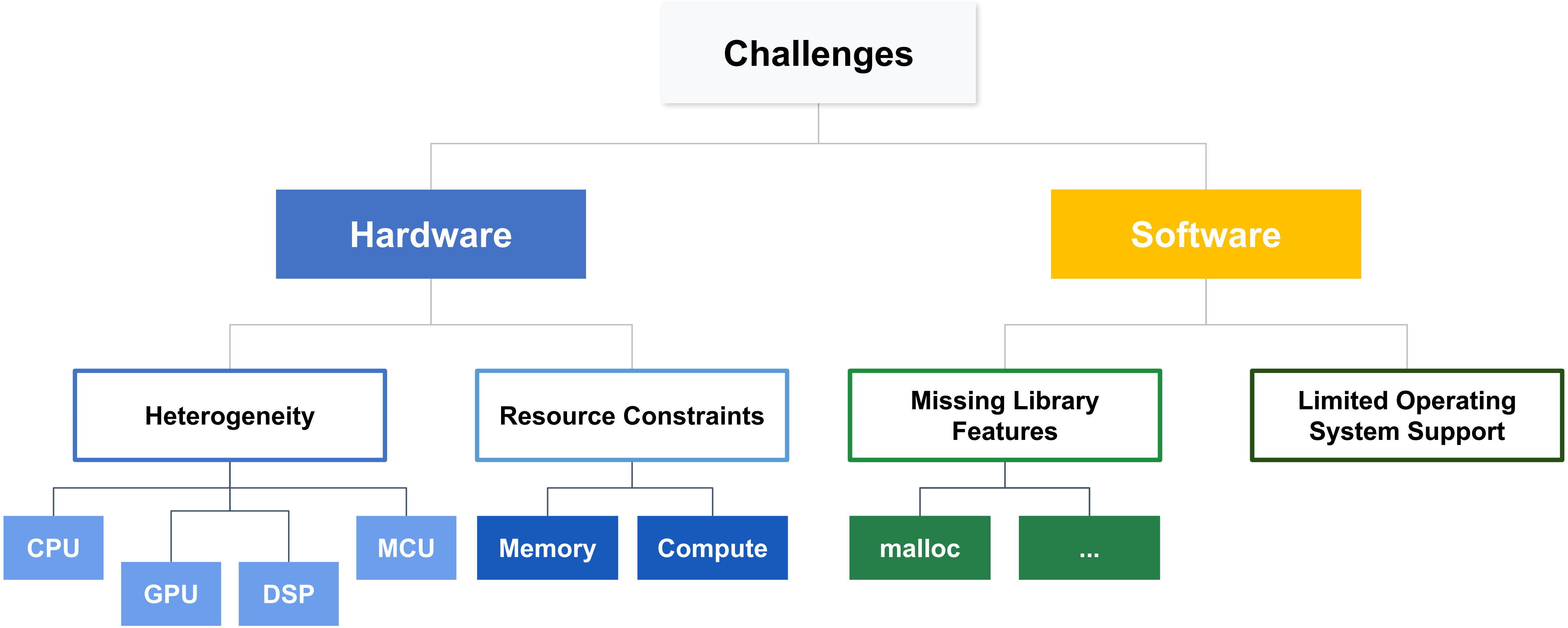
Vibration



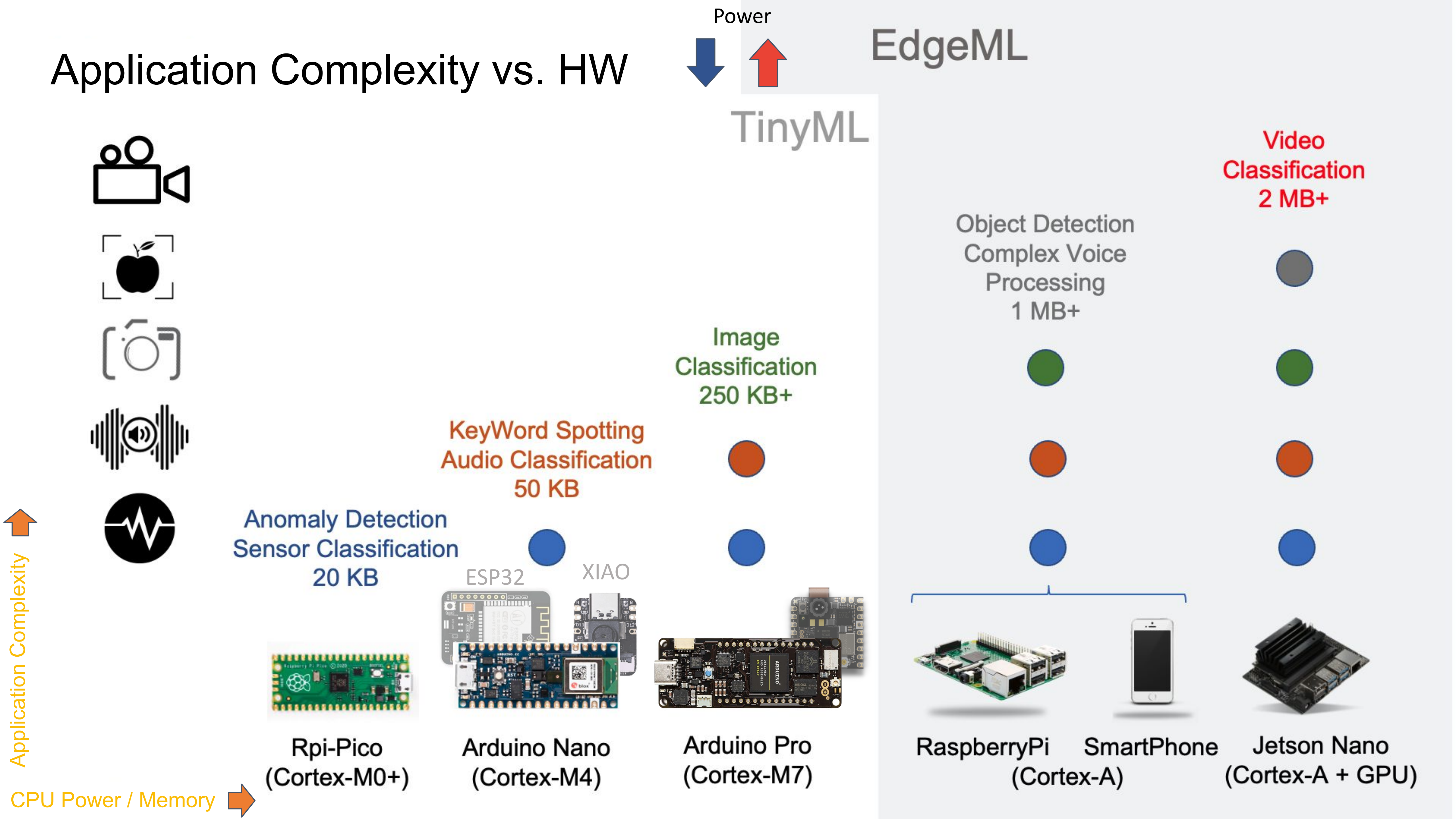
End-to-end **TinyML** application design

Software


TensorFlow
TensorFlow Lite
Train a model
Convert model
Optimize model
Deploy model at Edge
Make inferences at Edge
Raspberry Pi
Jetson Nano
Linux
iOS
TensorFlow Lite Micro
Microcontroller

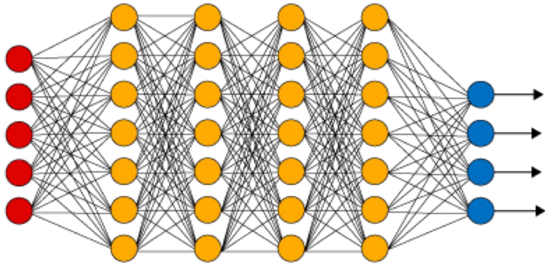

Challenges
Hardware
Software
Heterogeneity
Resource Constraints
Missing Library Features
Limited Operating System Support
CPU
GPU
DSP
MCU
Memory
Compute
malloc
...

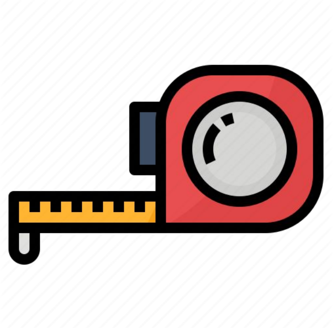
Application Complexity vs. HW



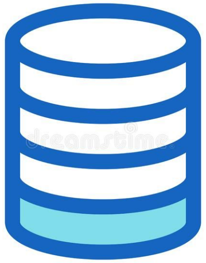
How to Train a ML Model?

Machine Learning Workflow ("What")

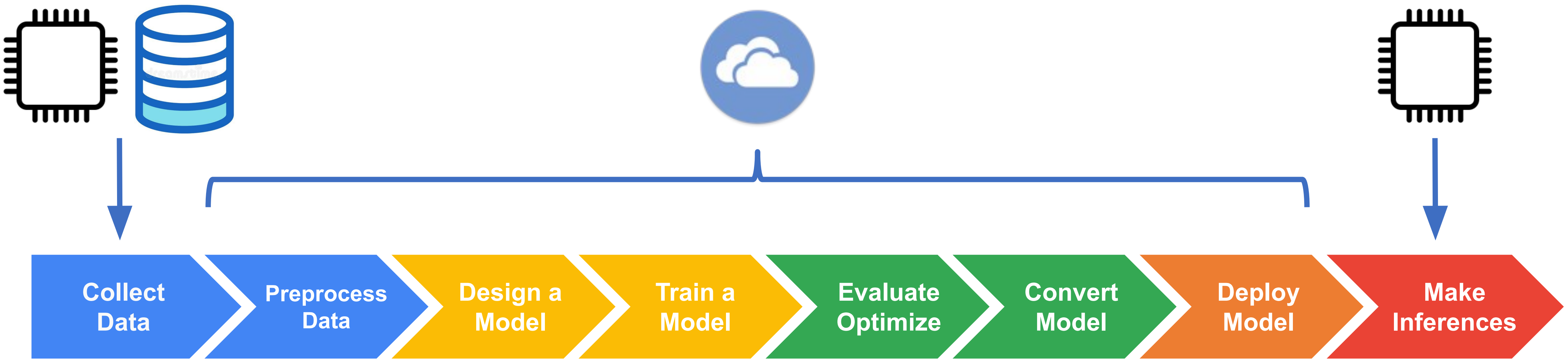




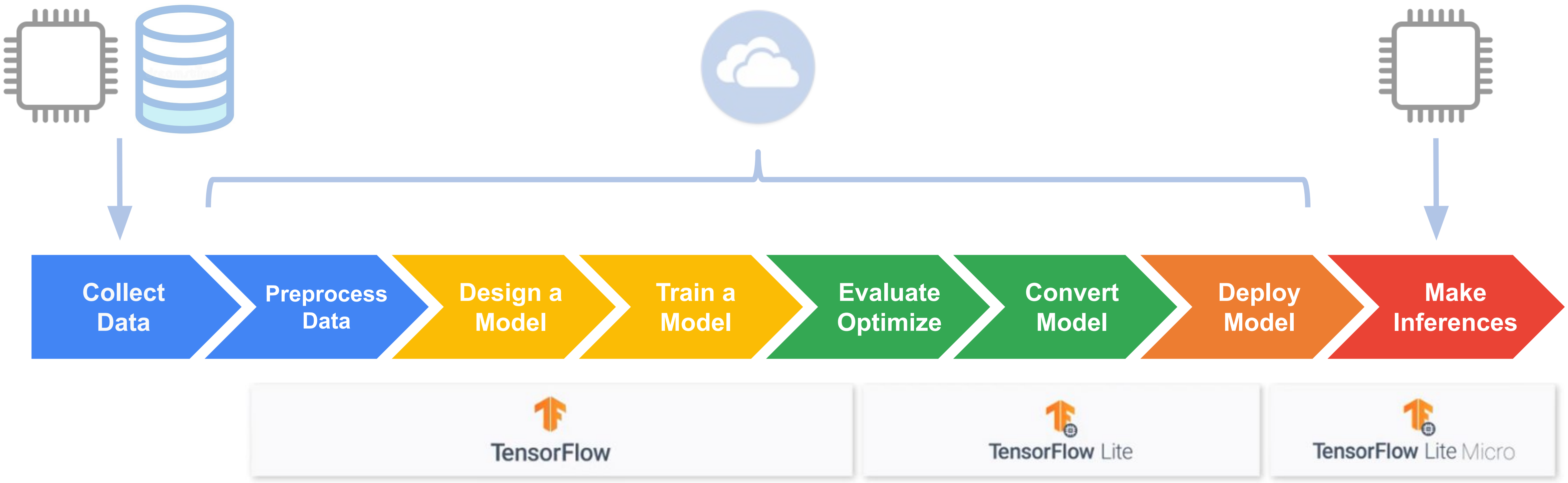
Collect Data → Preprocess Data → Design a Model → Train a Model → Evaluate Optimize → Convert Model → Deploy Model → Make Inferences



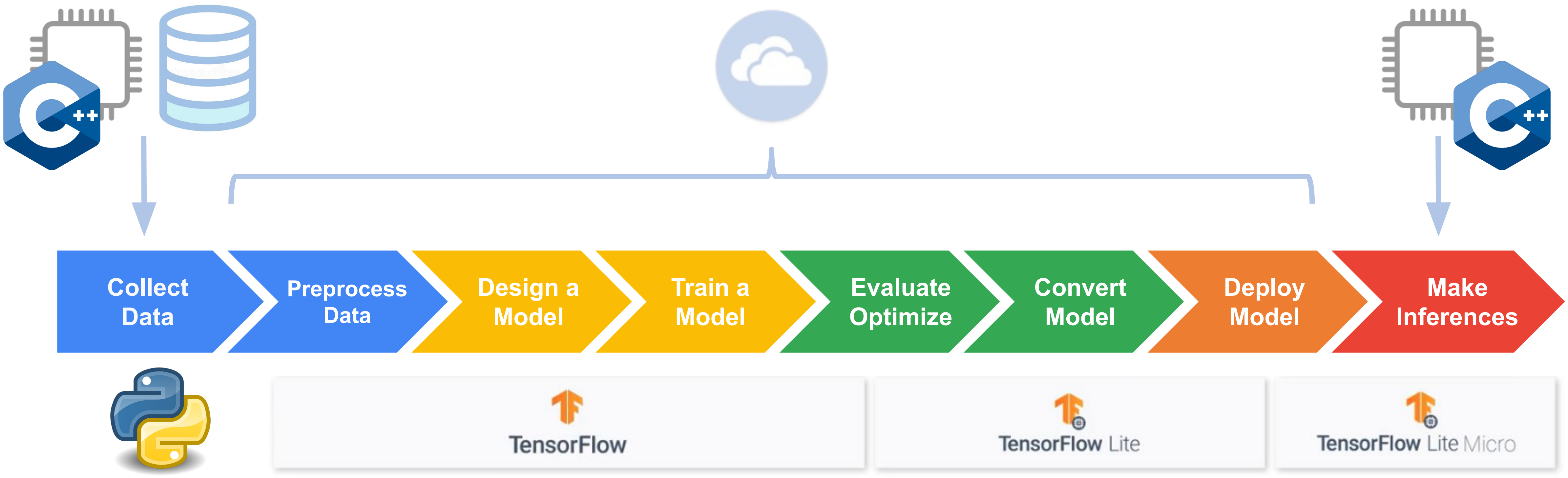
Machine Learning Workflow ("Where")



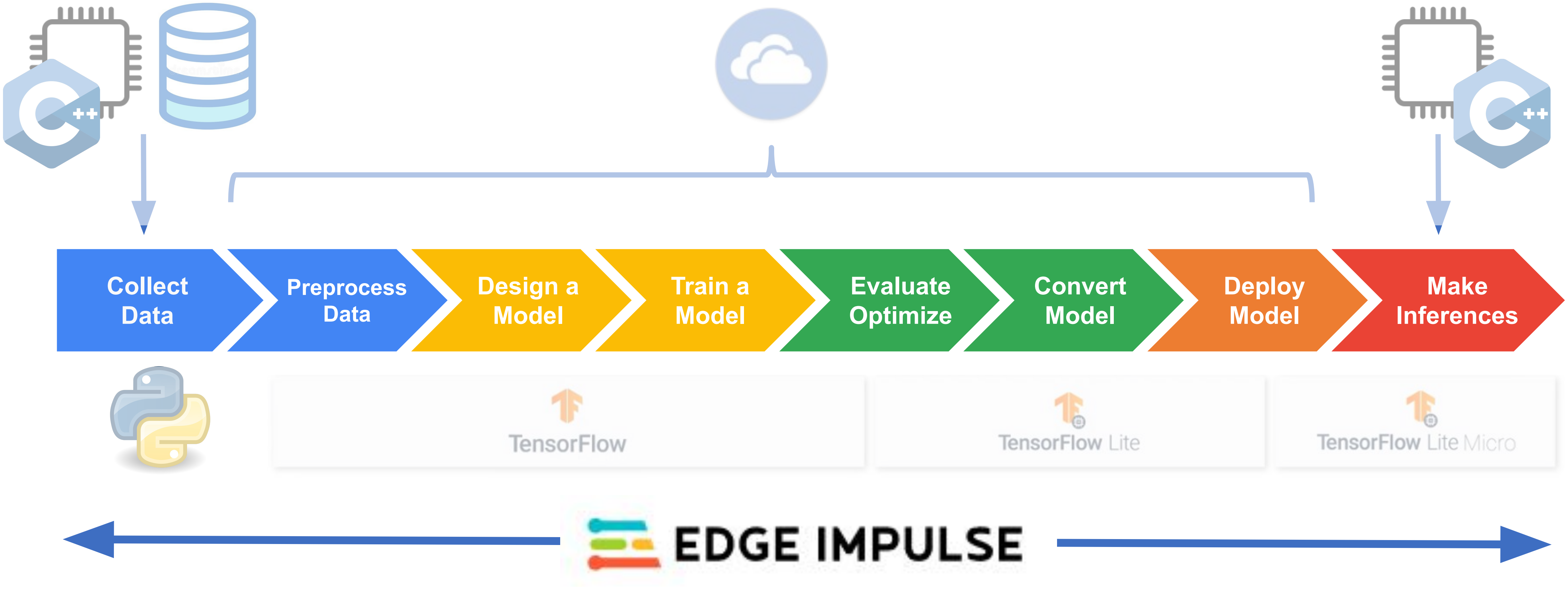
Machine Learning Workflow ("How")

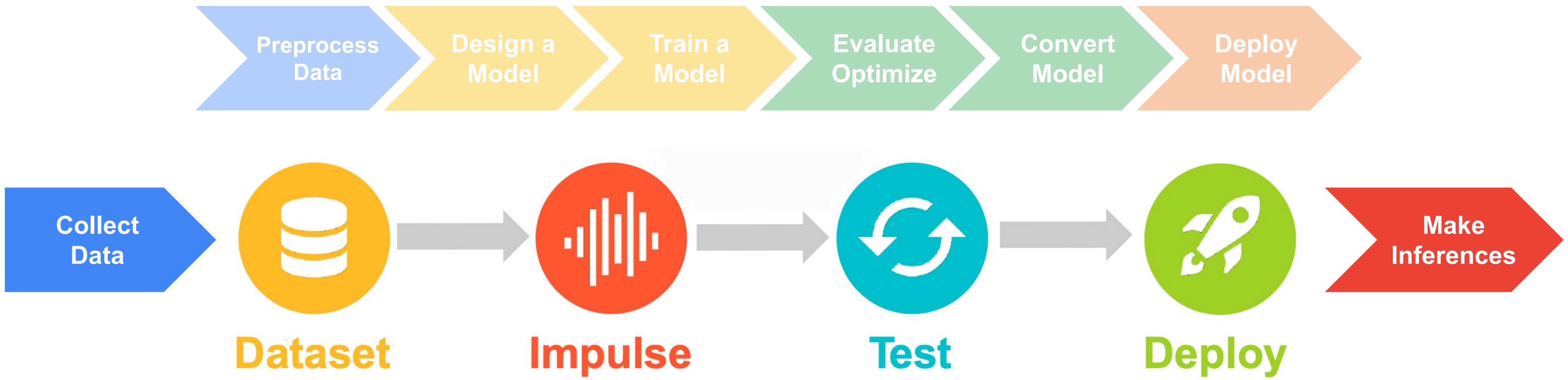


Machine Learning Workflow ("How")

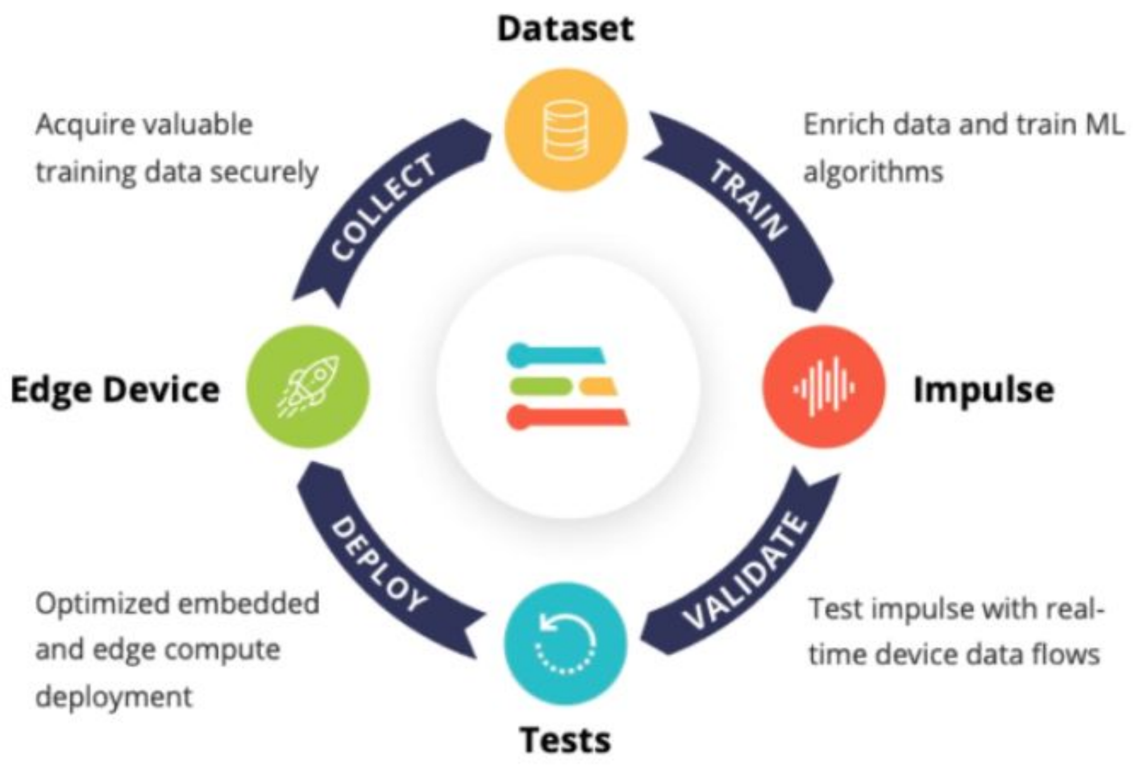


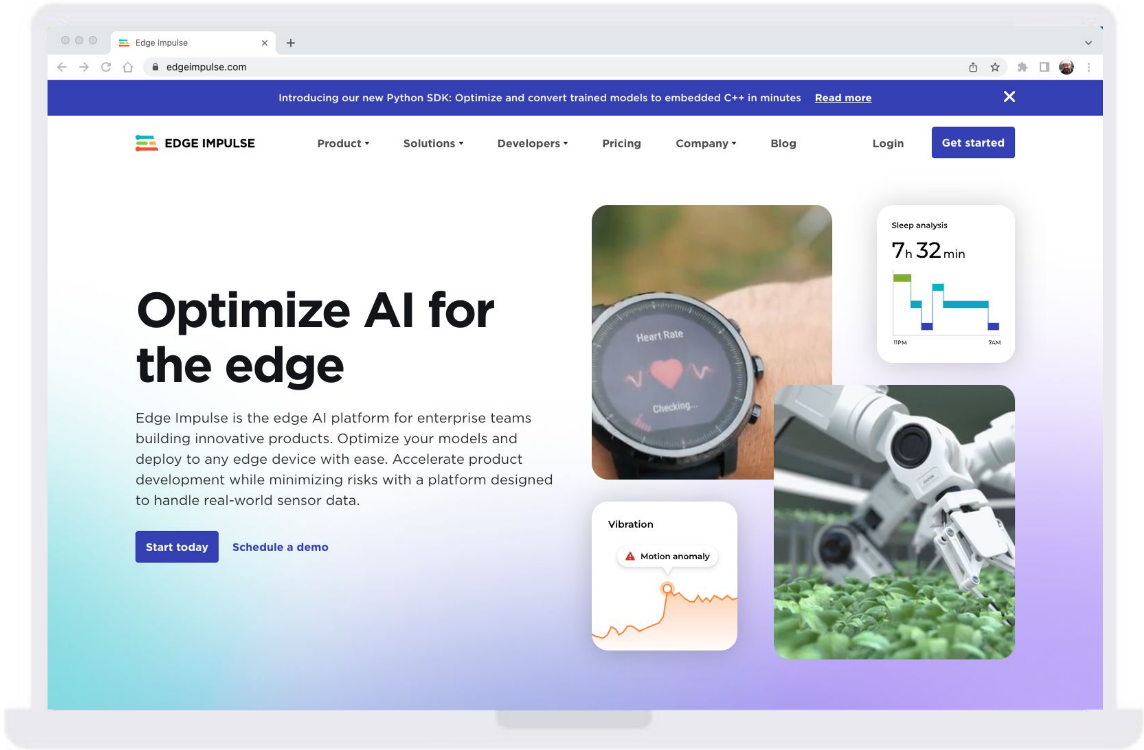
Machine Learning Workflow (“How”)




Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Collect Data
Dataset
Impulse
Test
Deploy
Make Inferences

EI Studio - Embedded ML platform ("AutoML")



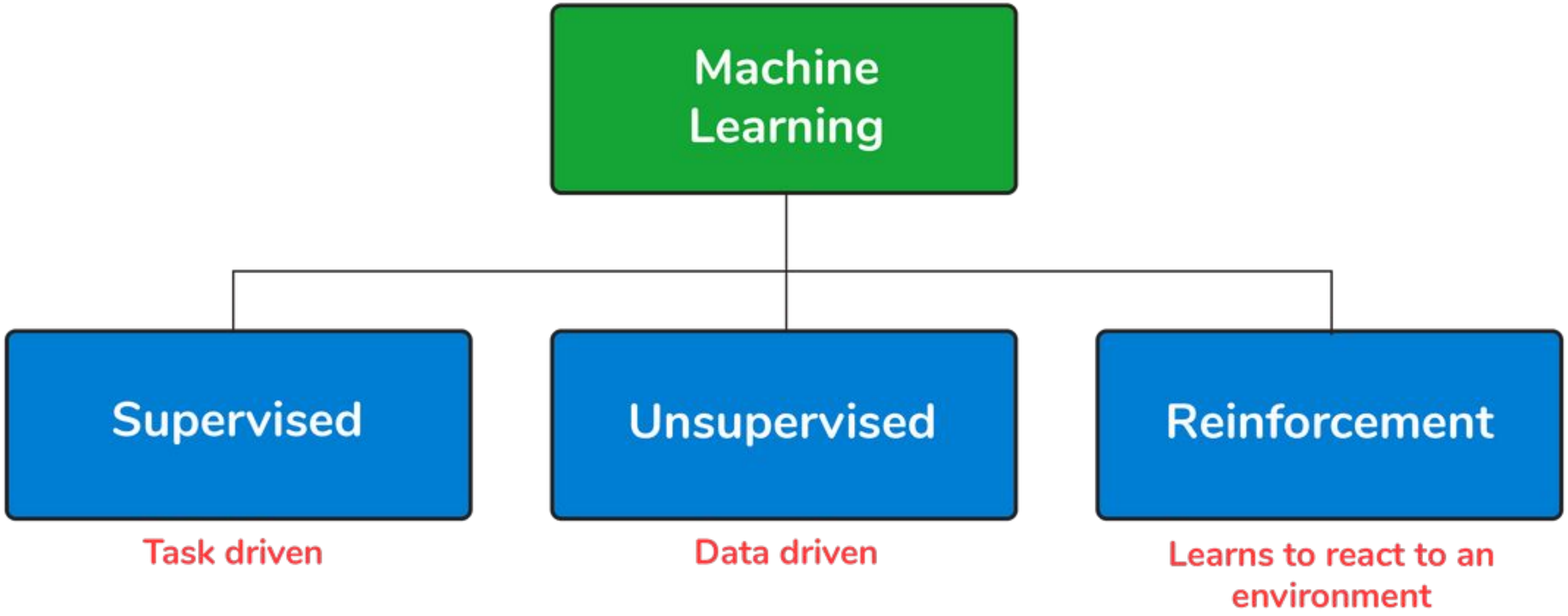


Learn more at http://edgeimpulse.com

[2023 SciTinyML Workshop - Edge Impulse Overview by Shawn Hymel]

TinyML Application

Examples



- Regression
- Classification
- Object Detection

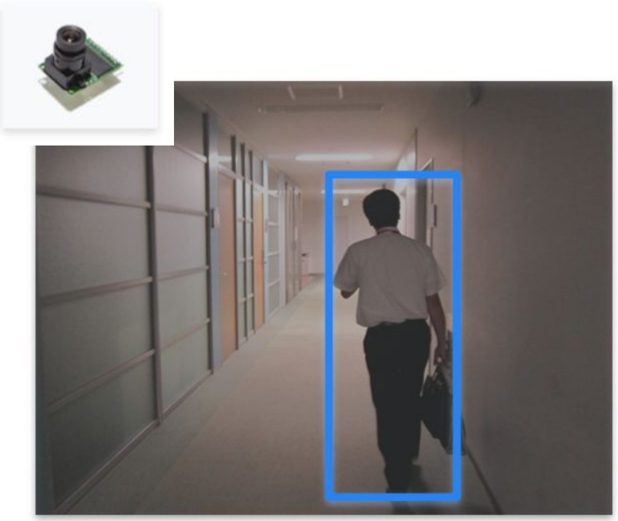
- Anomaly Detection

- Autonomous Navigation

Sound

Vibration

Vision



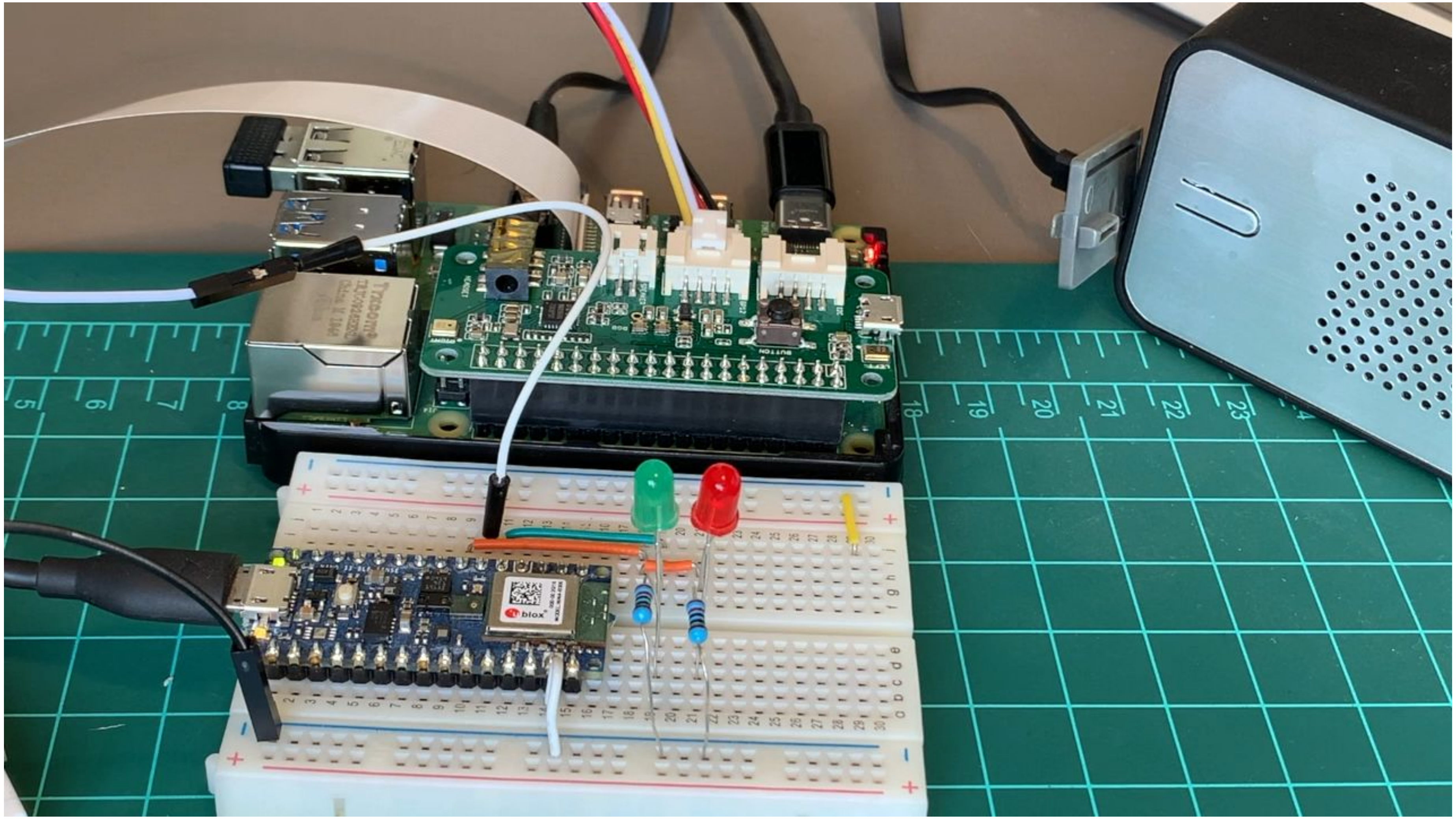
Sound

Vibration

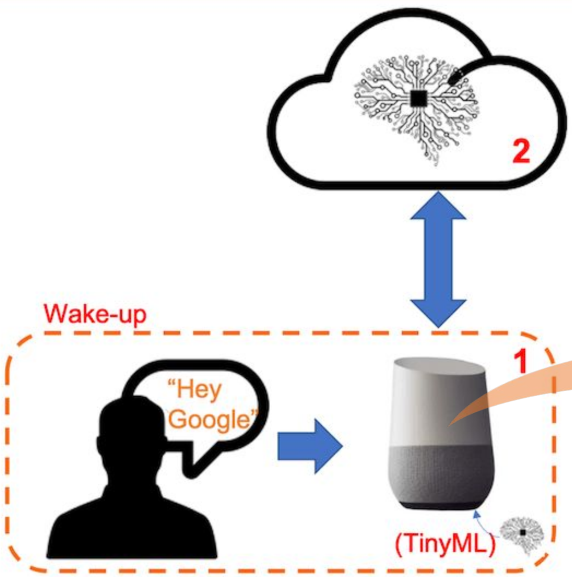
Vision

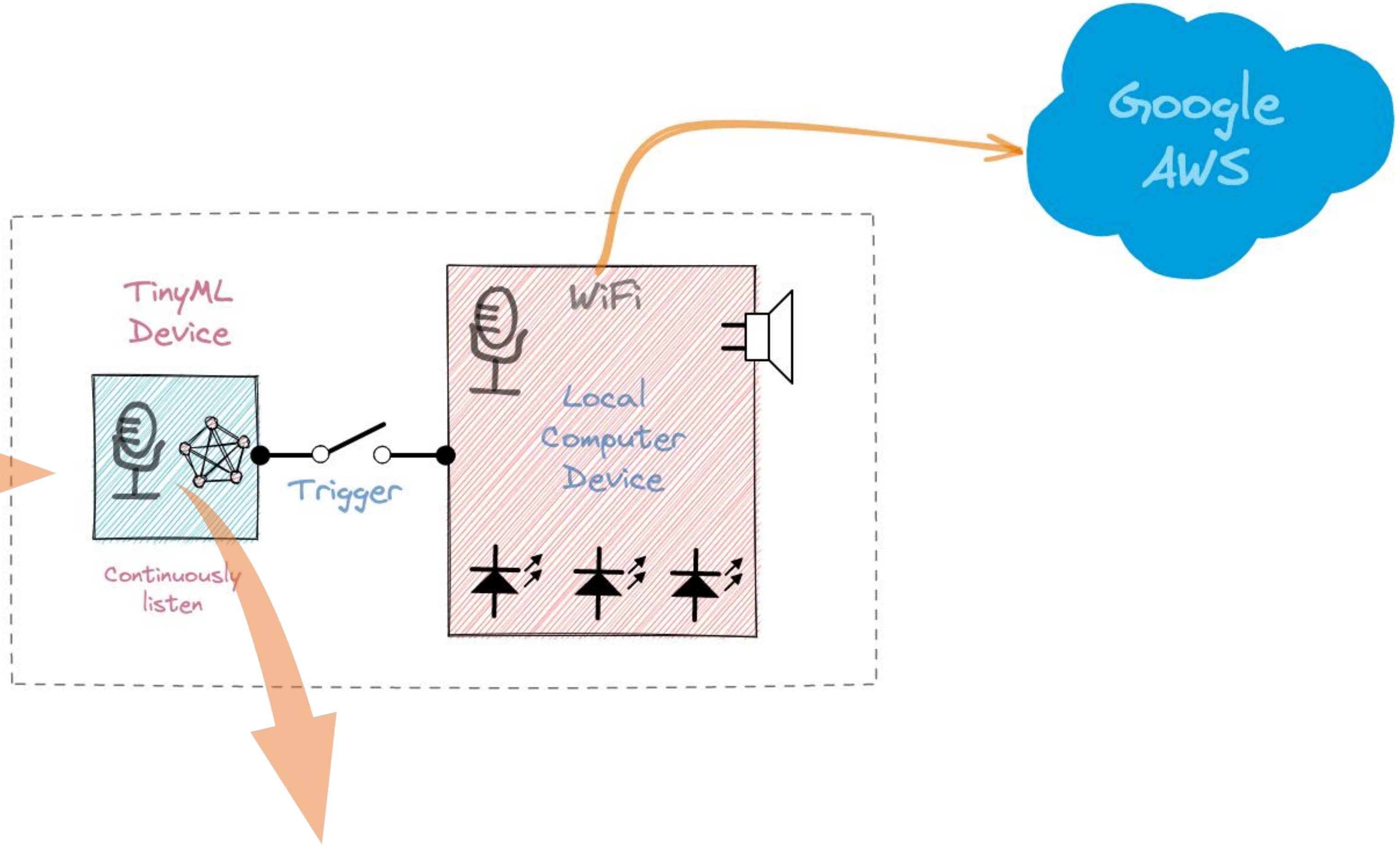
Personal Assistant

Personal Assistant



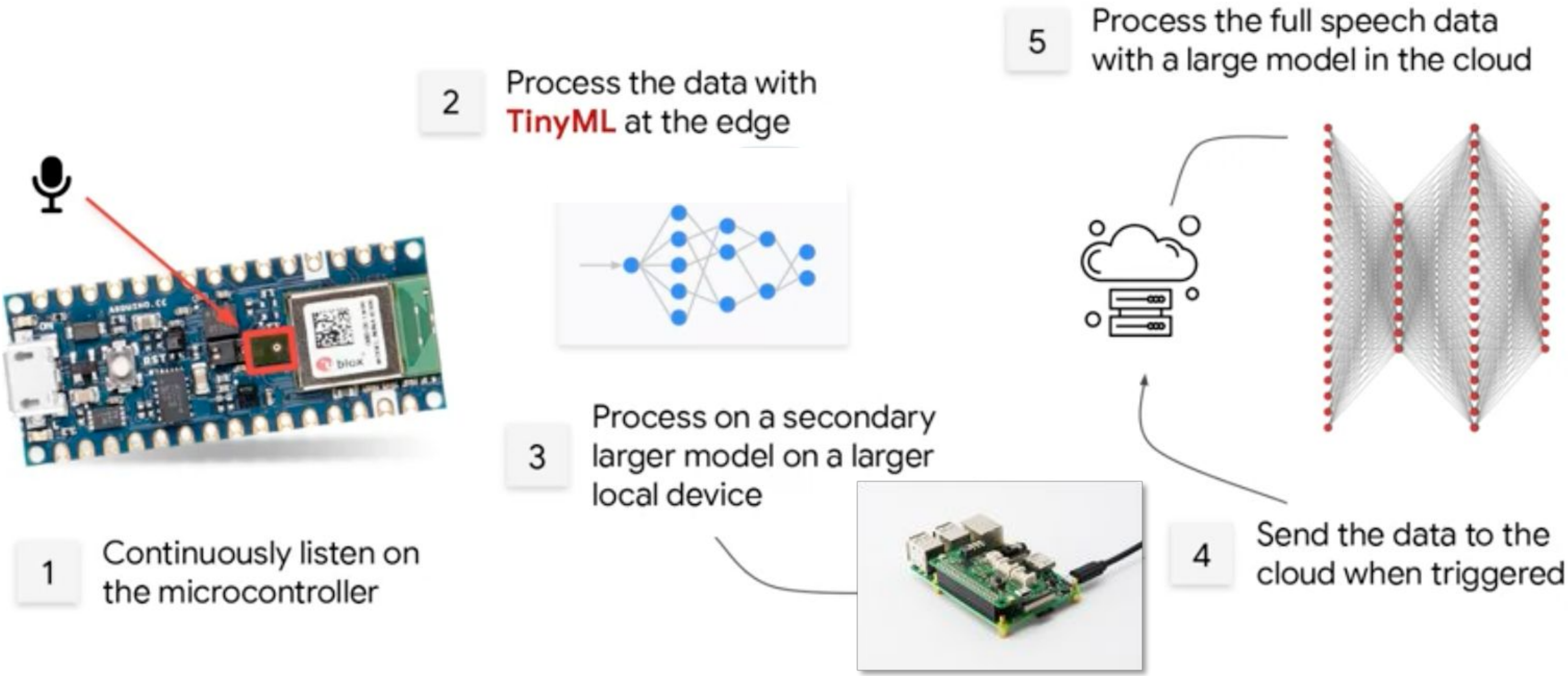
Personal Assistant



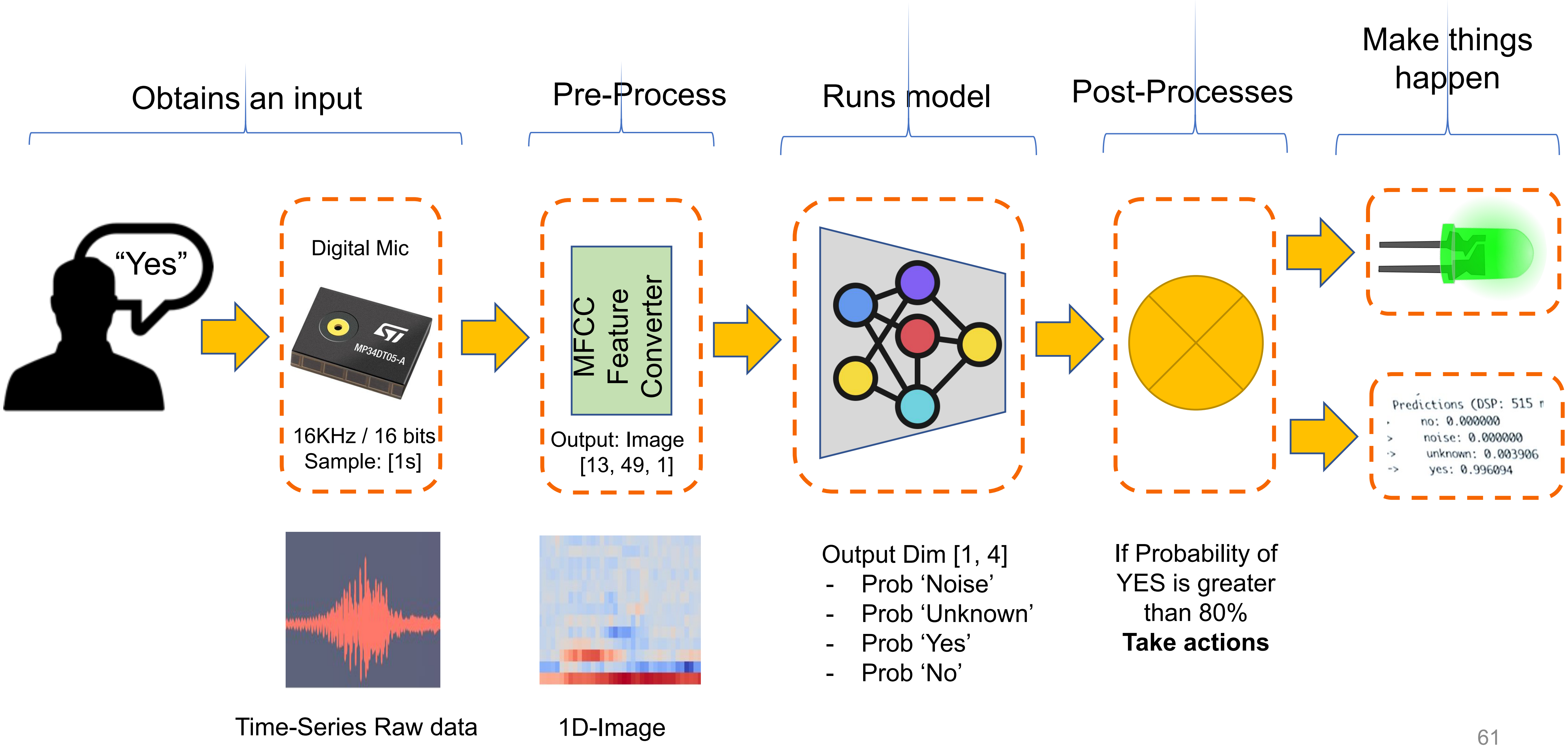


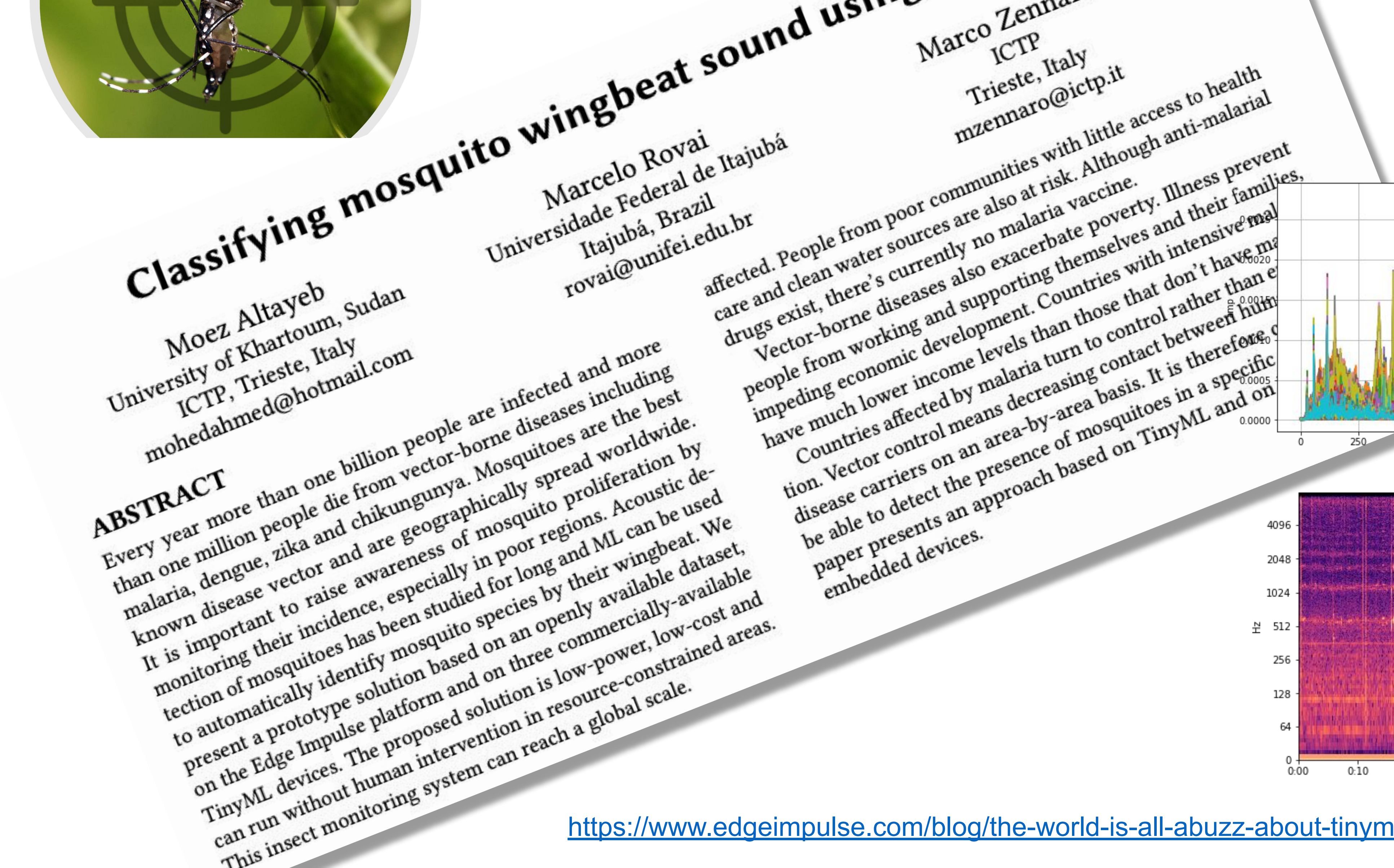
KWS

“Cascade” Detection: multi-stage model



KeyWord Spotting (KWS) - Inference





Classifying mosquito wingbeat sound using TinyML

Moez Altayeb
University of Khartoum, Sudan
ICTP, Trieste, Italy
mohedahmed@hotmail.com

Marcelo Rovai
Universidade Federal de Itajubá
Itajubá, Brazil
rovai@unifei.edu.br

Marco Zennaro
ICTP
Trieste, Italy
mzennaro@ictp.it

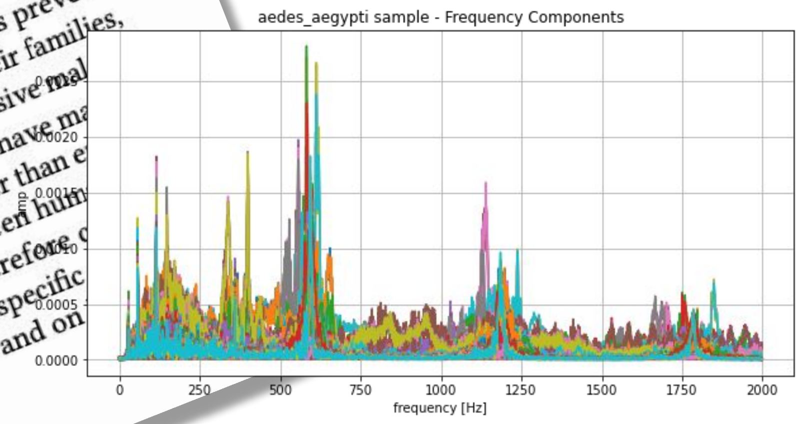
ABSTRACT

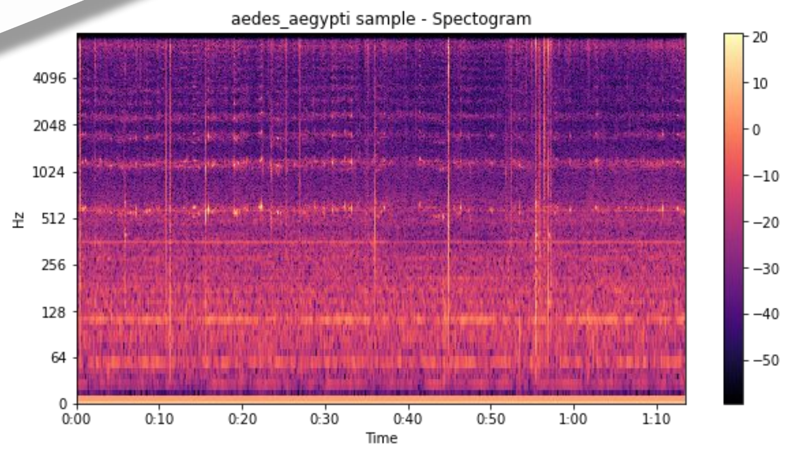
Every year more than one billion people are infected and more than one million people die from vector-borne diseases including malaria, dengue, zika and chikungunya. Mosquitoes are the best known disease vector and are geographically spread worldwide. It is important to raise awareness of mosquito proliferation by monitoring their incidence, especially in poor regions. Acoustic detection of mosquitoes has been studied for long and ML can be used to automatically identify mosquito species by their wingbeat. We present a prototype solution based on an openly available dataset, on the Edge Impulse platform and on three commercially-available TinyML devices. The proposed solution is low-power, low-cost and can run without human intervention in resource-constrained areas. This insect monitoring system can reach a global scale.

affected. People from poor communities with little access to health care and clean water sources are also at risk. Although anti-malarial drugs exist, there's currently no malaria vaccine.

Vector-borne diseases also exacerbate poverty. Illness prevent people from working and supporting themselves and their families, impeding economic development. Countries with intensive mal have much lower income levels than those that don't have ma

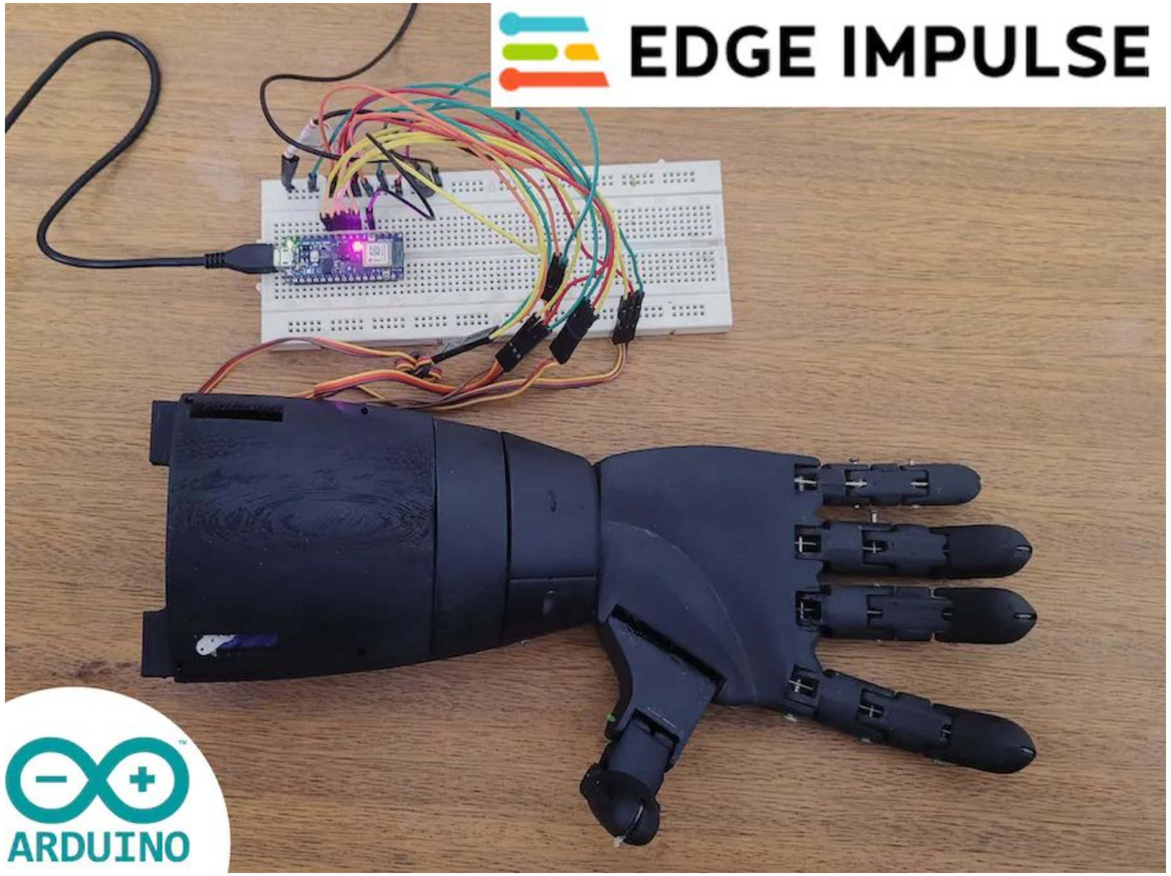
Countries affected by malaria turn to control rather than e tion. Vector control means decreasing contact between hum disease carriers on an area-by-area basis. It is therefore c be able to detect the presence of mosquitoes in a specific paper presents an approach based on TinyML and on embedded devices.





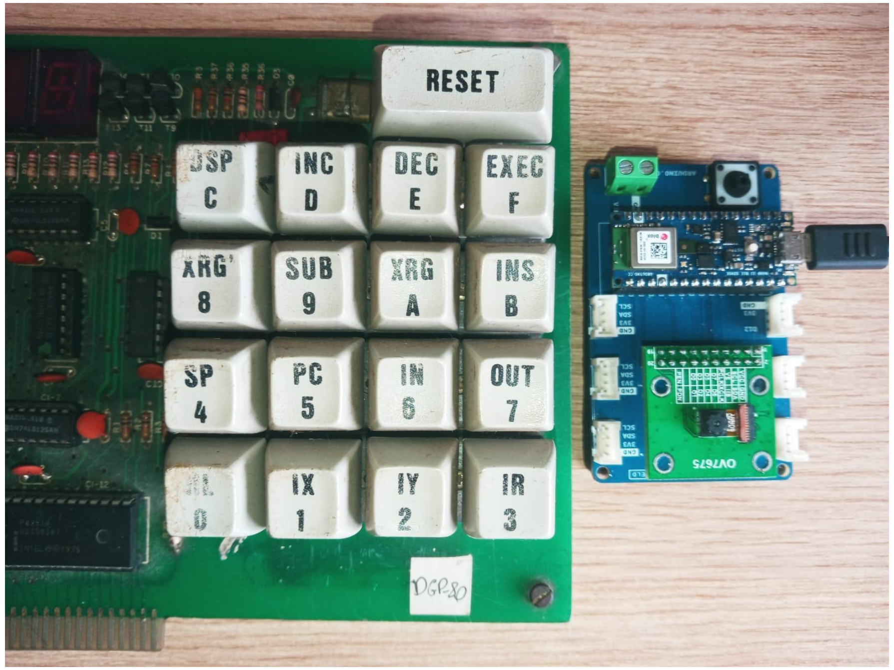
https://www.edgeimpulse.com/blog/the-world-is-all-abuzz-about-tinyml

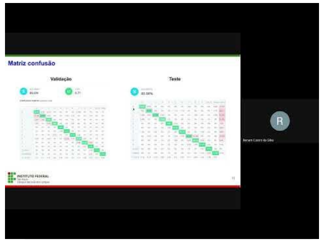
Bionic Hand Voice Commands Module



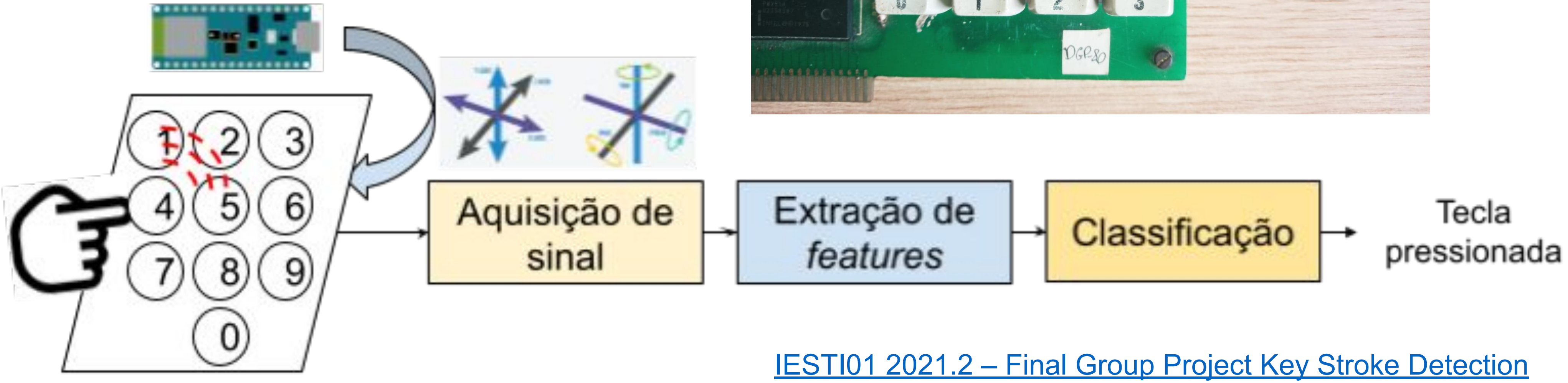
https://www.hackster.io/ex-machina/bionic-hand-voice-commands-module-w-edge-impulse-arduino-aa97e3

Keystroke **Sound** Detection





Renam Castro
Professor IFESP



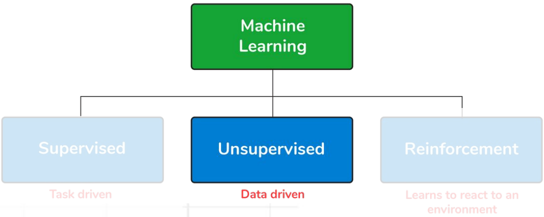
[IESTI01 2021.2 – Final Group Project Key Stroke Detection]()

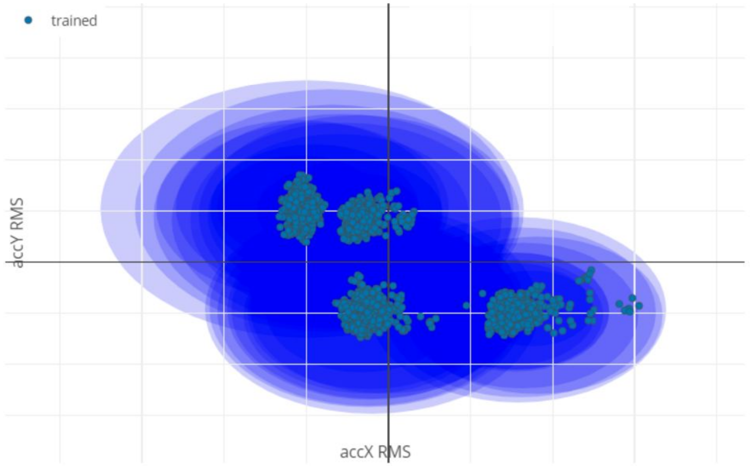
Sound

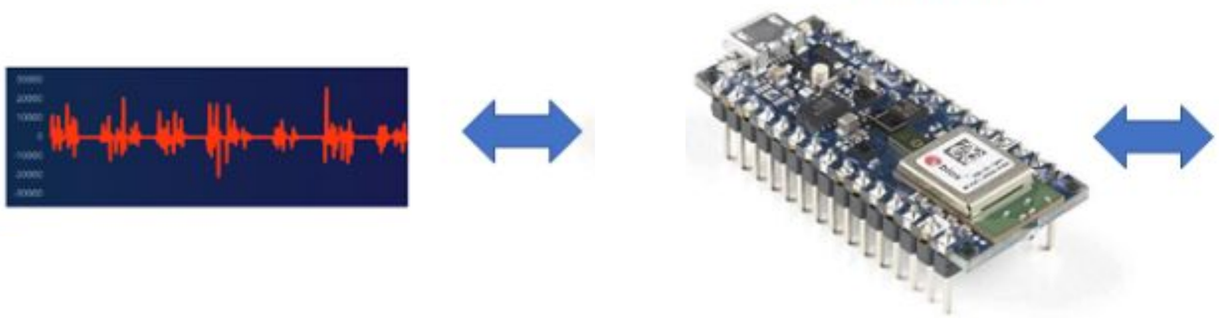
Vibration

Vision

Industrial – Anomaly Detection

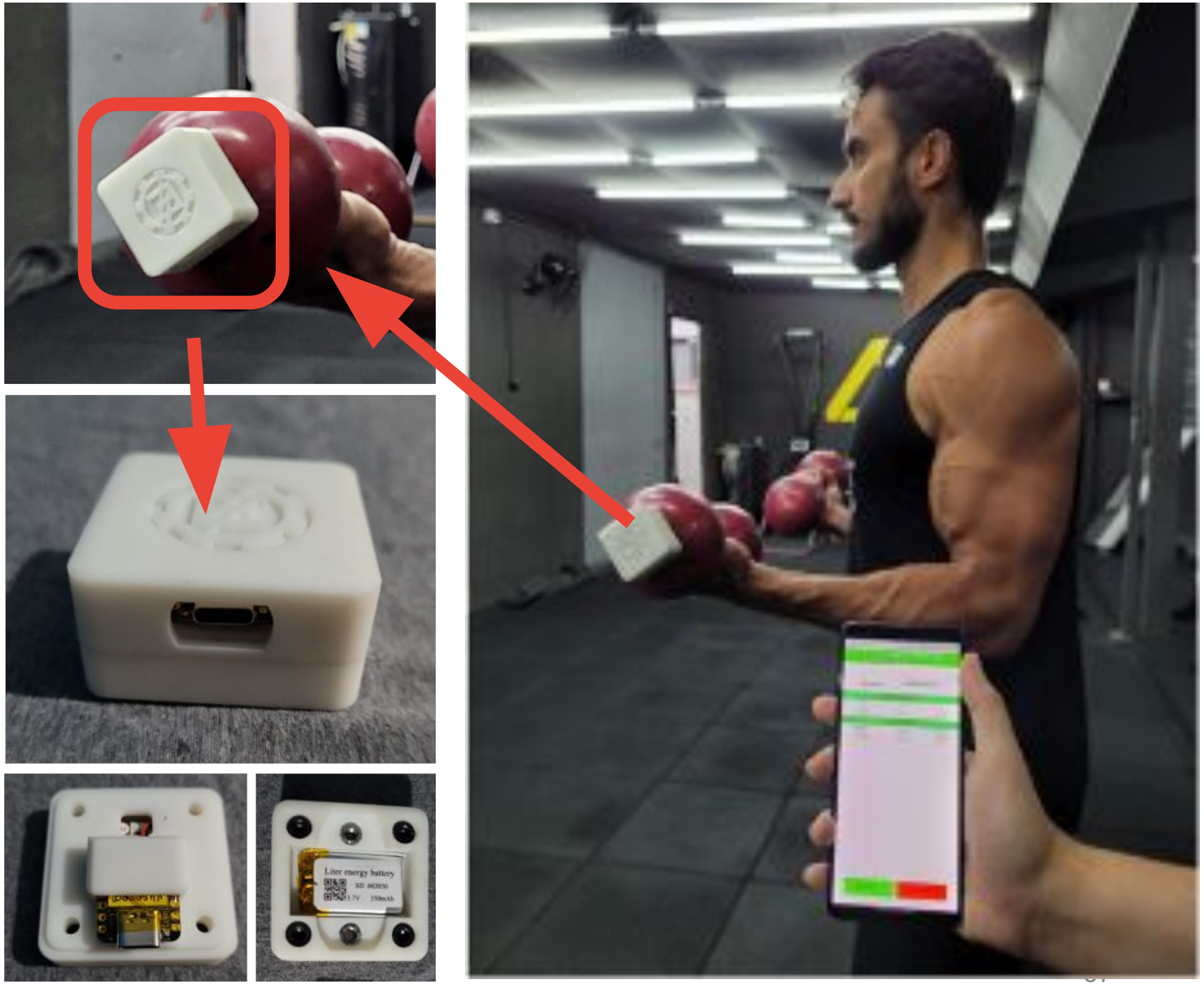






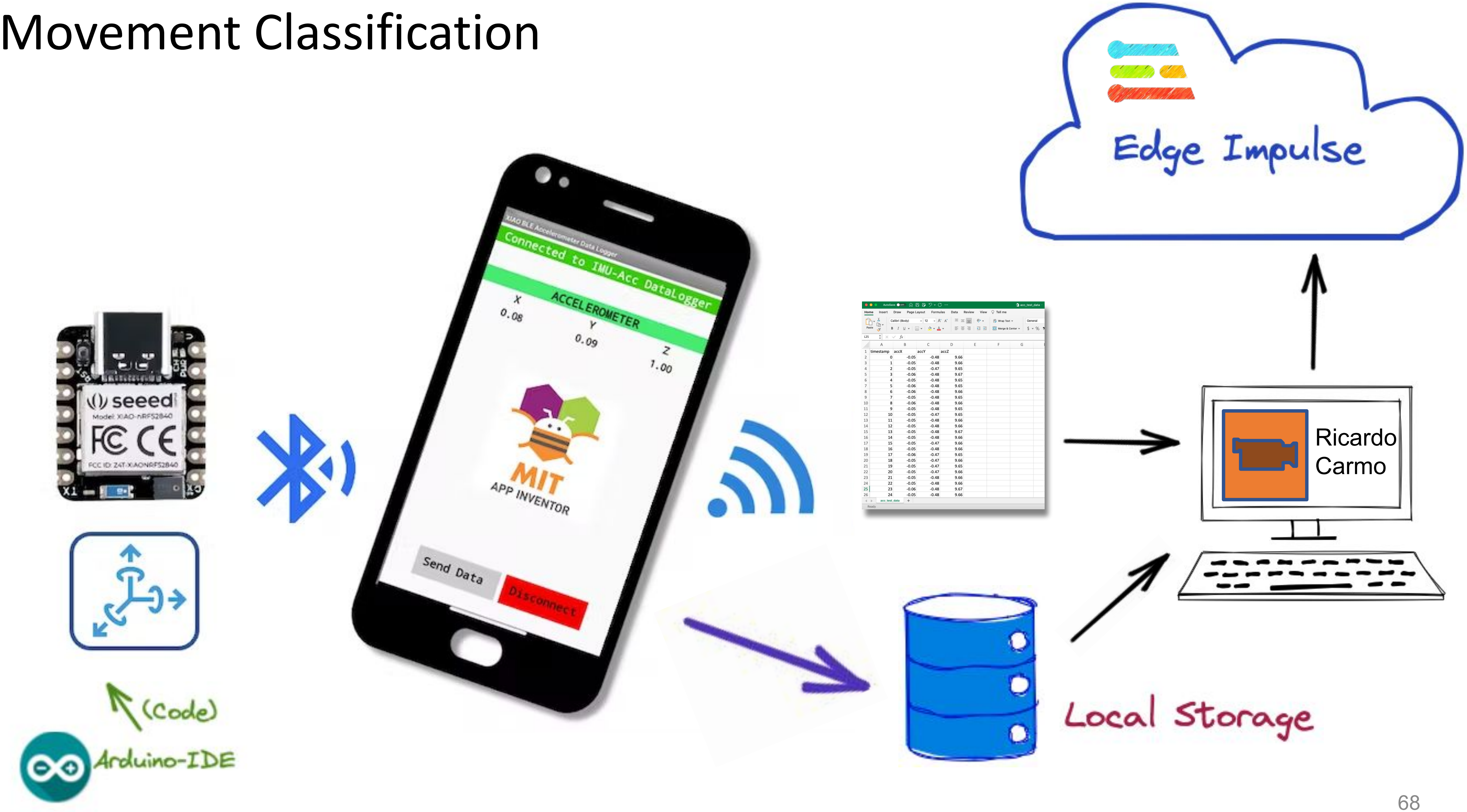
[IESTI01 2021.2 - Final Group Project: Bearing Failure Detection]

Movement Classification



CNPq UNIFEI

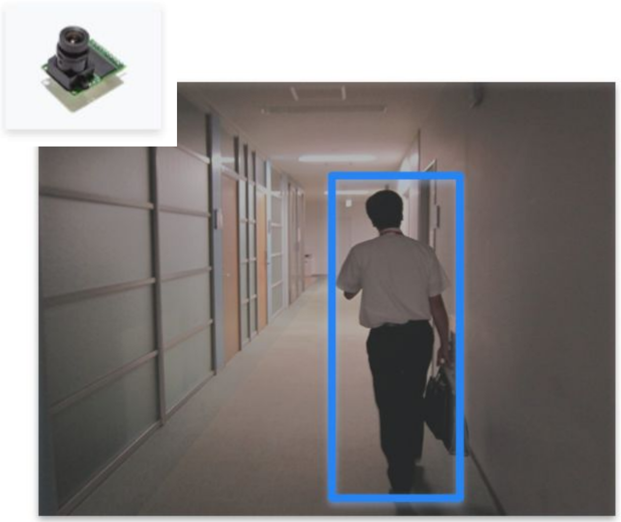
Movement Classification



Sound

Vibration

Vision



Computer Vision Main Types

Image Classification

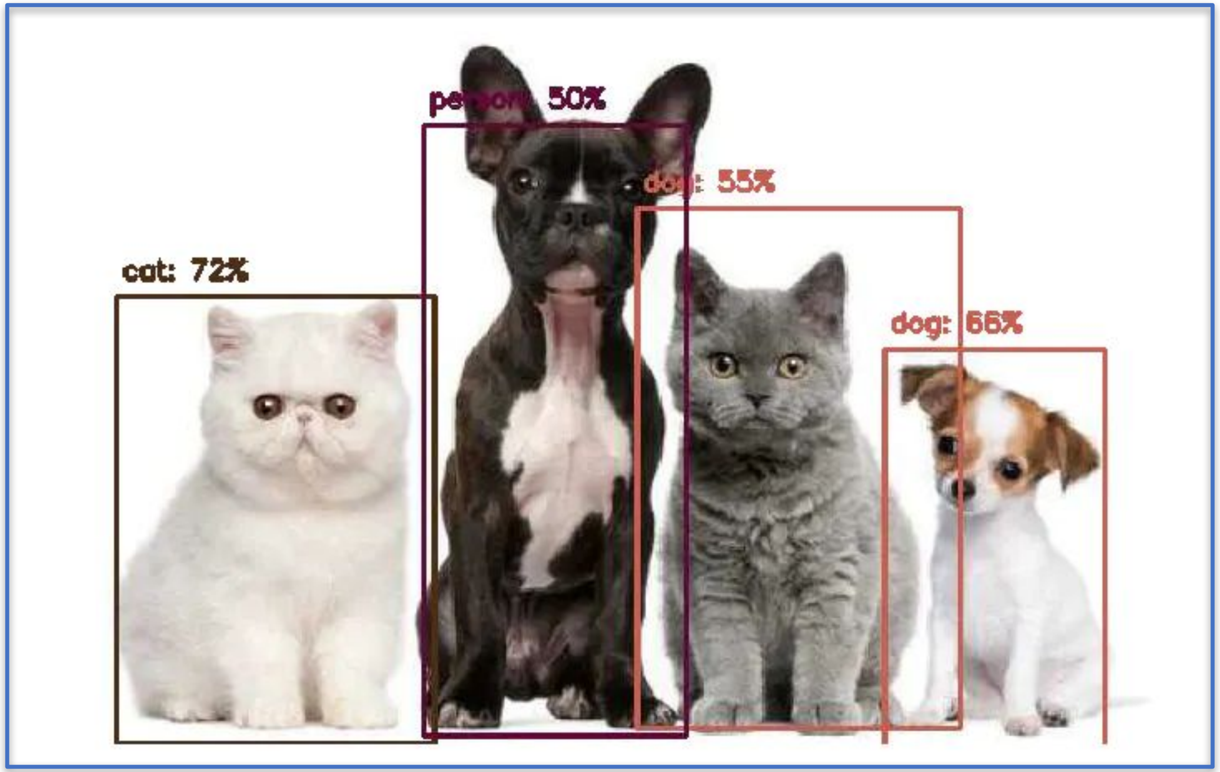
(**Multi-Class** Classification)

Cat: 70%

Dog: 80%

Object Detection

Multi-Label Classification + Object **Localization**



Computer Vision Main Types

Image Classification

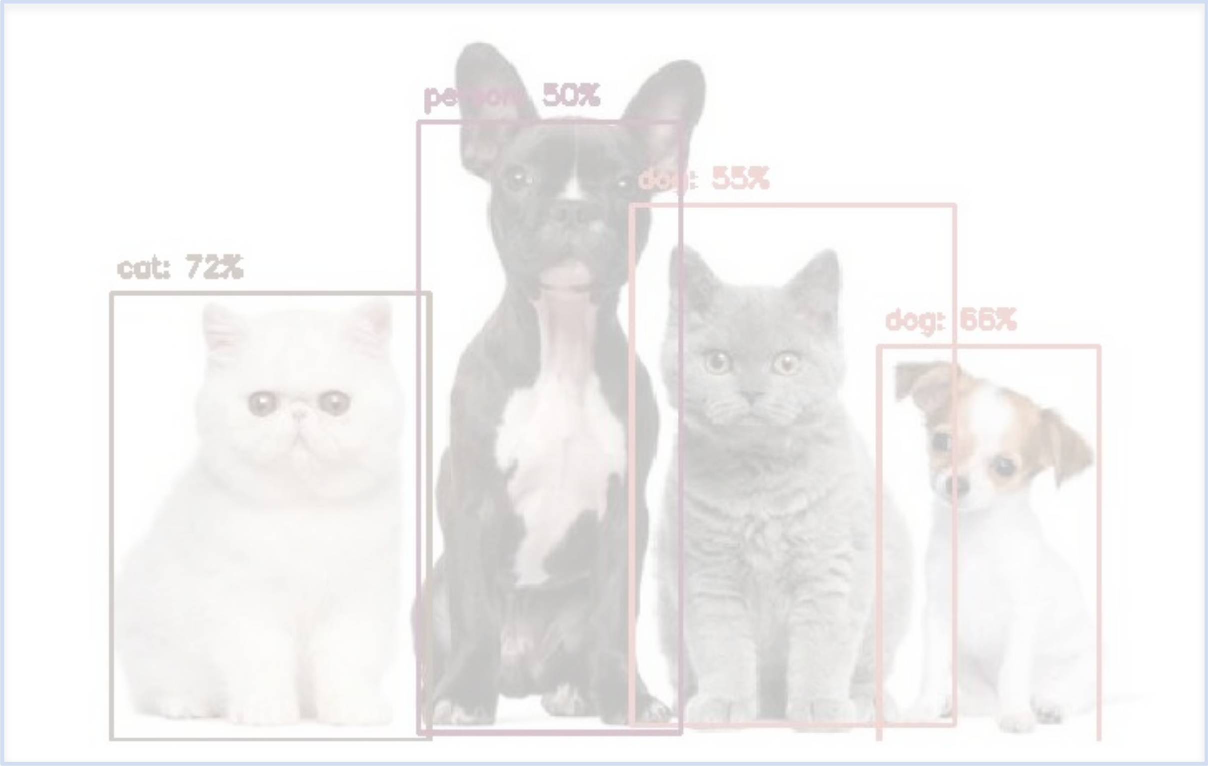
(**Multi-Class** Classification)

Cat: 70%

Dog: 80%

Object Detection

Multi-Label Classification + Object **Localization**



Forest Fire Detection

[TinyML Aerial Forest Fire Detection]

[IESTI01 - Forest Fire Detection – Proof of Concept]

Coffee Disease Classification

João Vitor Yukio Bordin Yamashita
Graduando em Engenharia Eletrônica pela UNIFEI

https://www.hackster.io/Yukio/coffee-disease-classification-with-ml-b0a3fc

Computer Vision Main Types

Image Classification

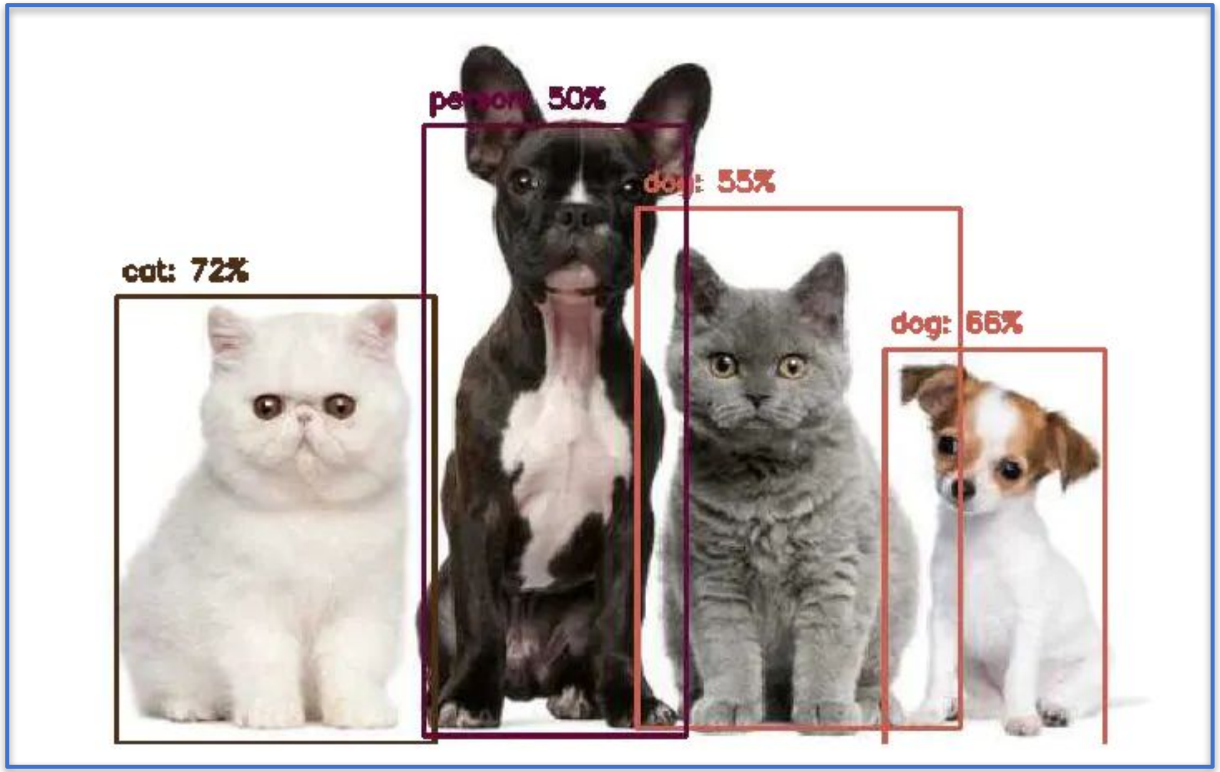
(**Multi-Class** Classification)

Cat: 70%

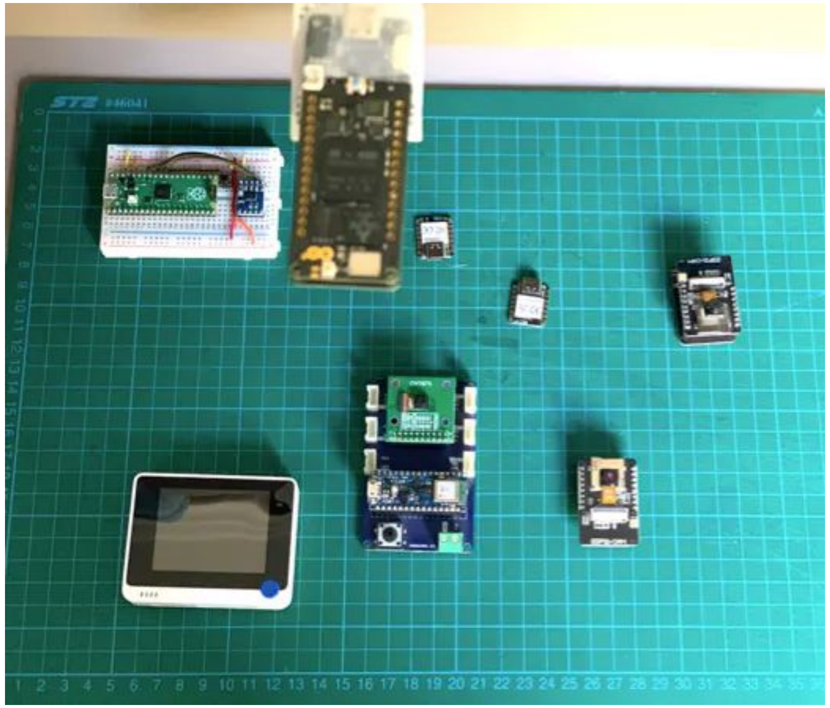
Dog: 80%

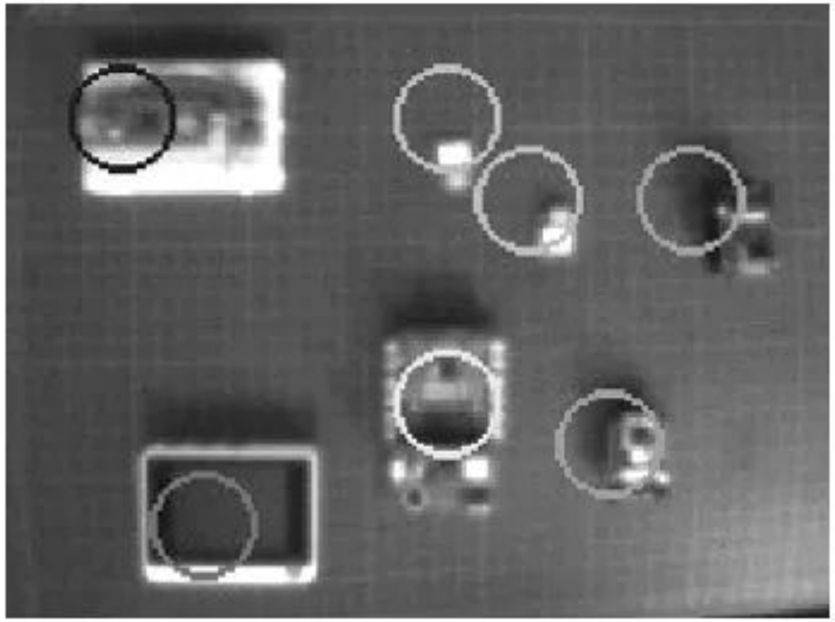
Object Detection

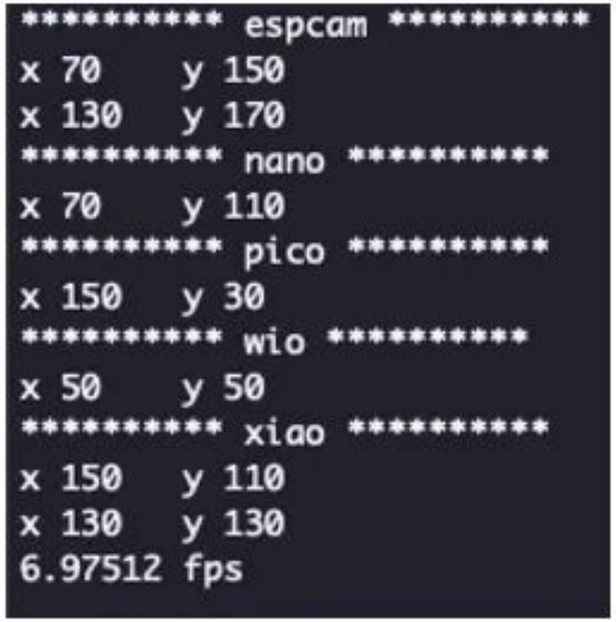
Multi-Label Classification + Object **Localization**



Detecting Objects using TinyML (FOMO)

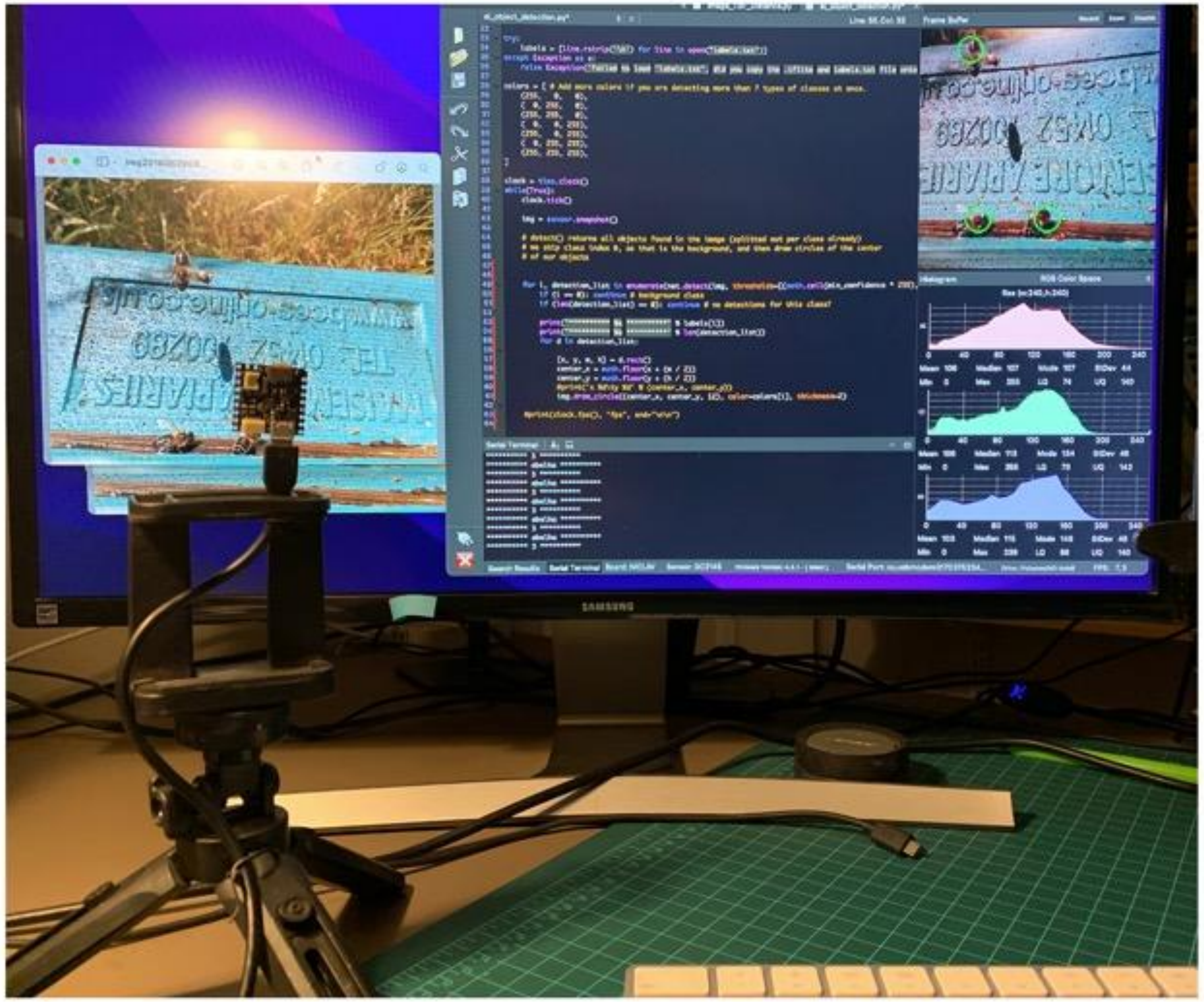






[EdgeAI made simple - Exploring Image Processing (Object Detection) on microcontrollers with Arduino Portenta, Edge Impulse FOMO, and OpenMV]

Detecting Objects using TinyML (FOMO)

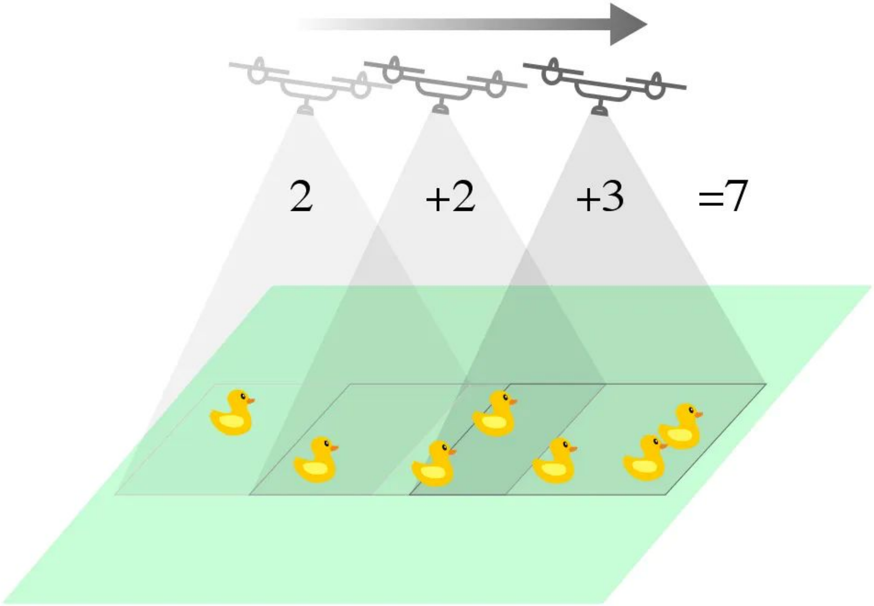


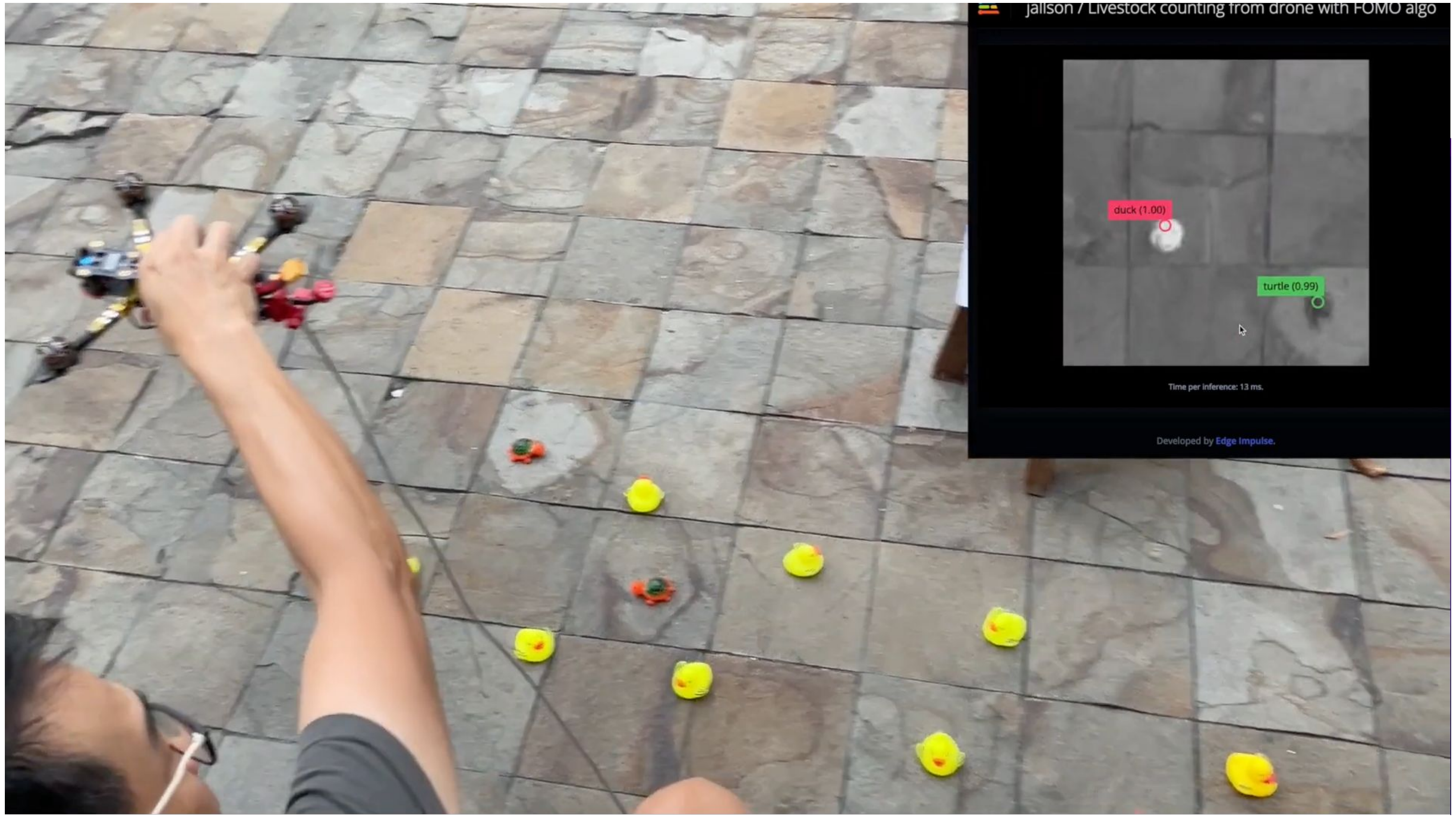
OpenMV

MicroPython

EDGE IMPULSE

Livestock / Wildlife Counting from Drone with FOMO





https://www.hackster.io/jallsonsuryo/livestock-wildlife-counting-from-drone-with-fomo-algorithm-a2f734

Other TinyML / MCUs Project Examples

Vision

- Image Classification with ESP32-CAM [Doc]
- Image Classification with Portenta H7 [Doc]
- Object Detection with Portenta H7 [Doc]

Sound

- Listening Temperature with Nano 33 [Doc]
- COPD Detection with Nano 33 [Doc]
- Sound Classification with XIAO BLE Sense [Doc]

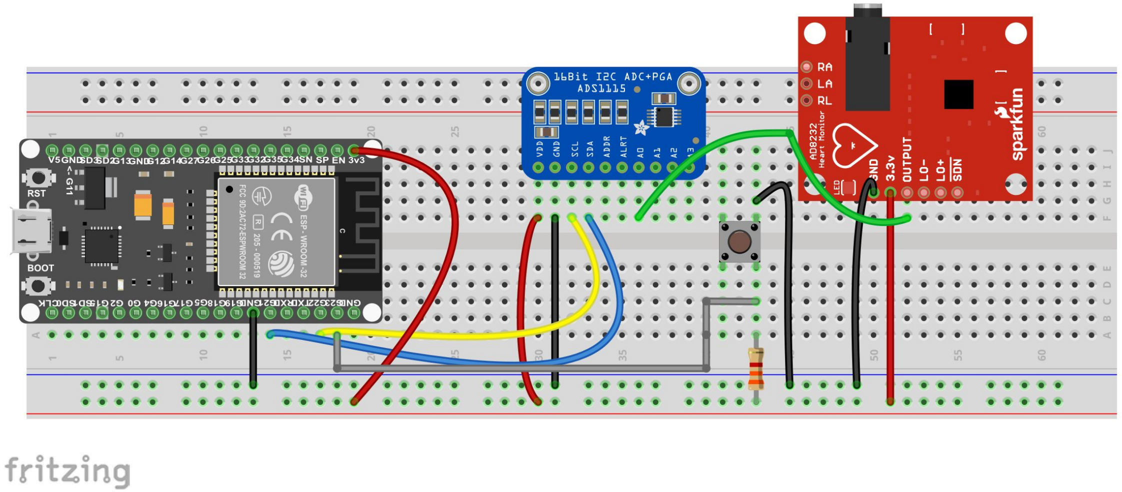
Vibration

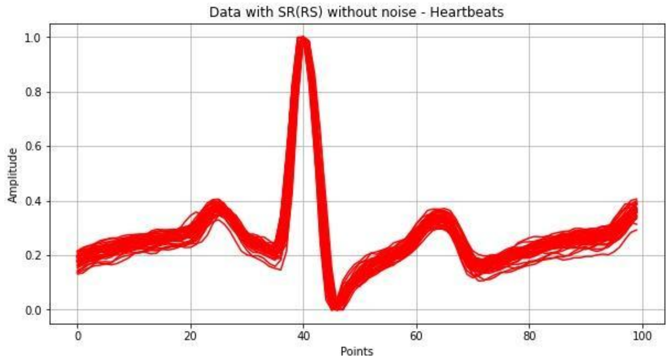
- Motion Recognition with RPi Pico [Doc]
- Gesture Recognition with Wio Terminal [Doc]
- Anomaly Detection with XIAO BLE Sense [Doc]

Other Sensors / MCUs / Models

Examples

AD8232 - Single Lead Heart Rate Monitor

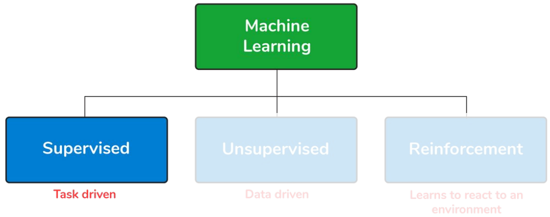


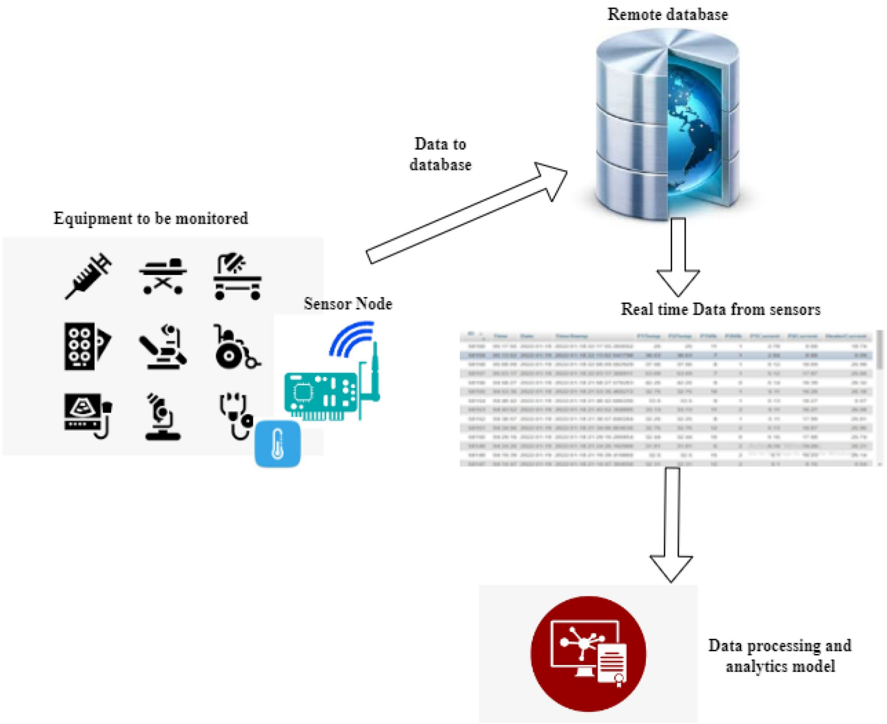


Guilherme Silva
Engenheiro - UNIFEI

[Atrial Fibrillation Detection on ECG using TinyML
Silva et al. UNIFEI 2021]

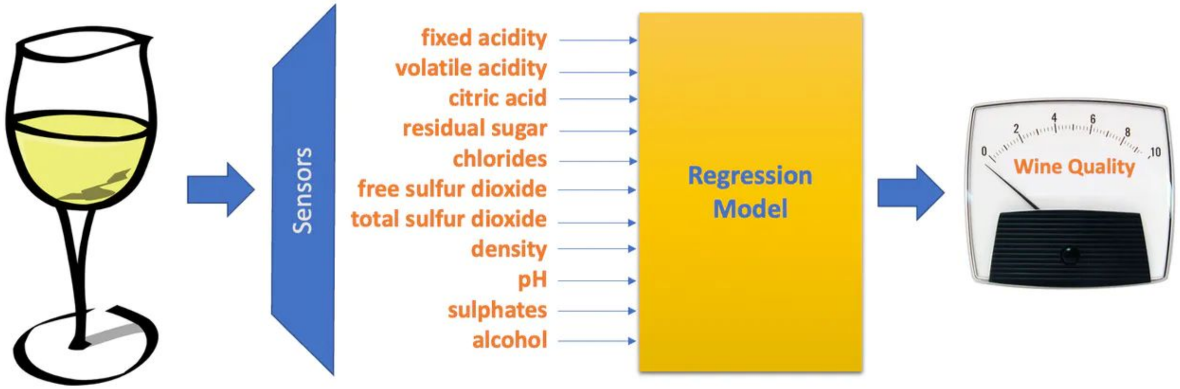
Regression on TinyML





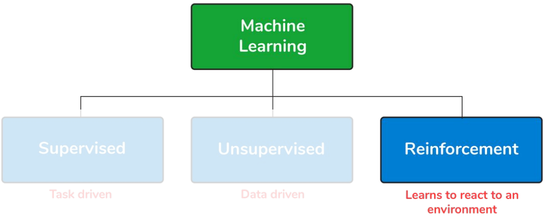
[On-Device IoT-Based Predictive Maintenance Analytics Model: Comparing TinyLSTM and TinyModel from Edge Impulse]

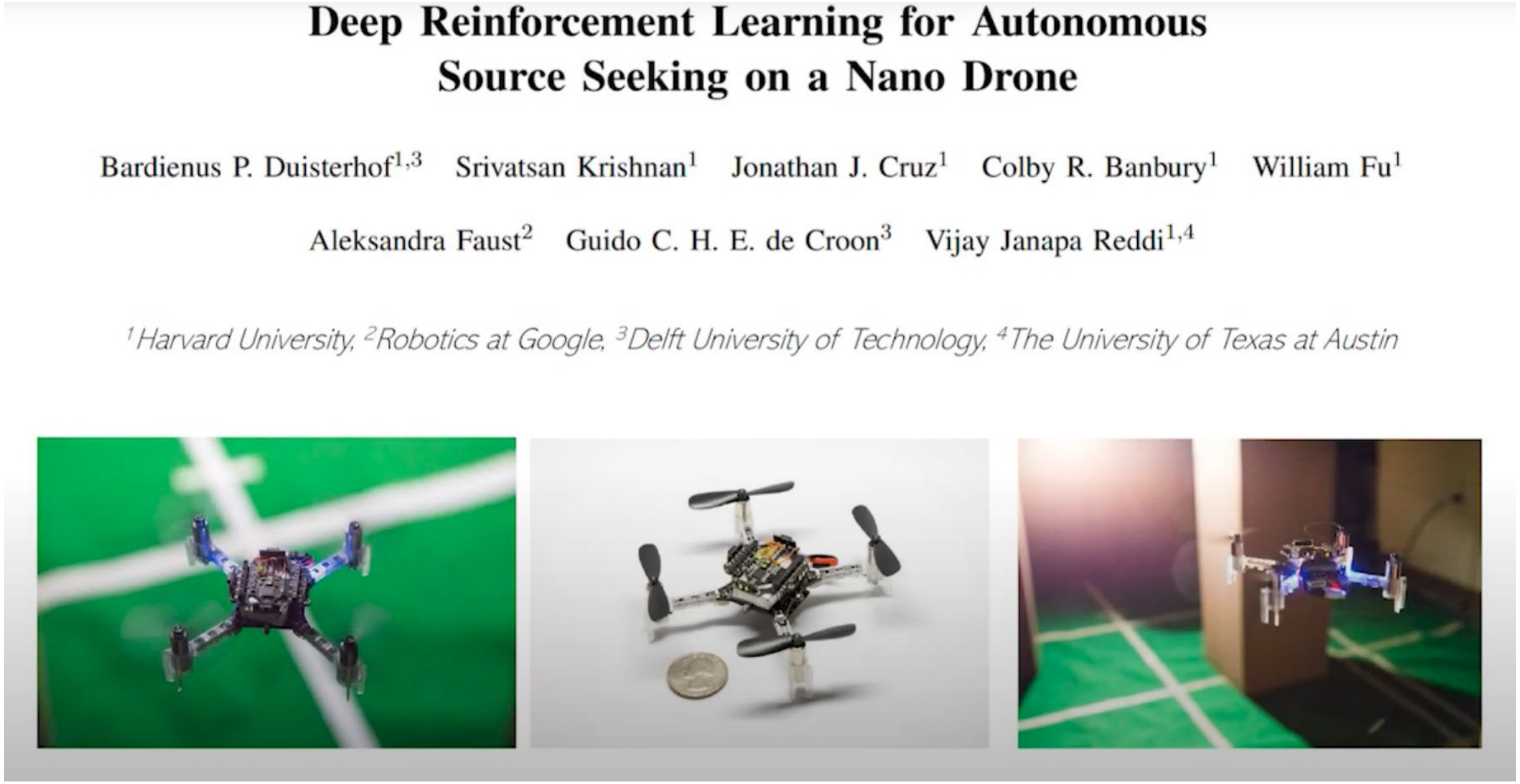
Sensor fusion



[TinyML Made Easy: Exploring Regression - White Wine Quality]

Reinforcement on TinyML





Deep Reinforcement Learning for Autonomous Source Seeking on a Nano Drone

Bardienus P. Duisterhof[1,3] Srivatsan Krishnan[1] Jonathan J. Cruz[1] Colby R. Banbury[1] William Fu[1]

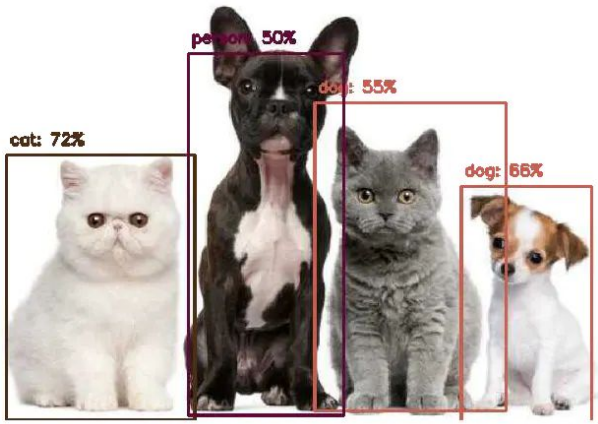
Aleksandra Faust[2] Guido C. H. E. de Croon[3] Vijay Janapa Reddi[1,4]

[1]Harvard University, [2]Robotics at Google, [3]Delft University of Technology, [4]The University of Texas at Austin

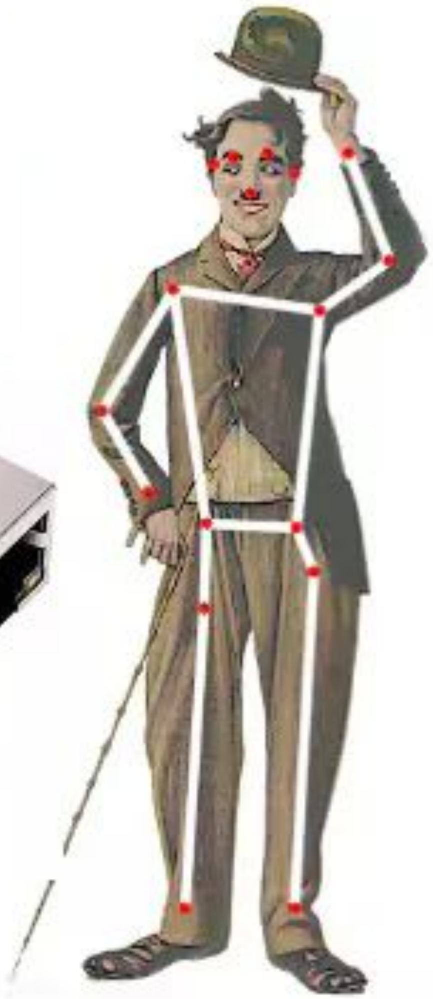
https://arxiv.org/abs/1909.11236

https://youtu.be/wmVKbX7MOnU

Exploring AI at the edge (Computer Vision)



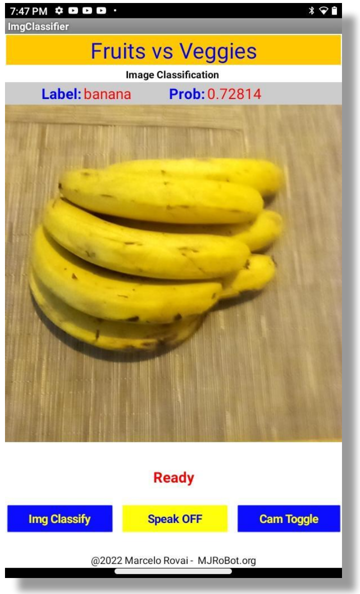
TensorFlow Lite

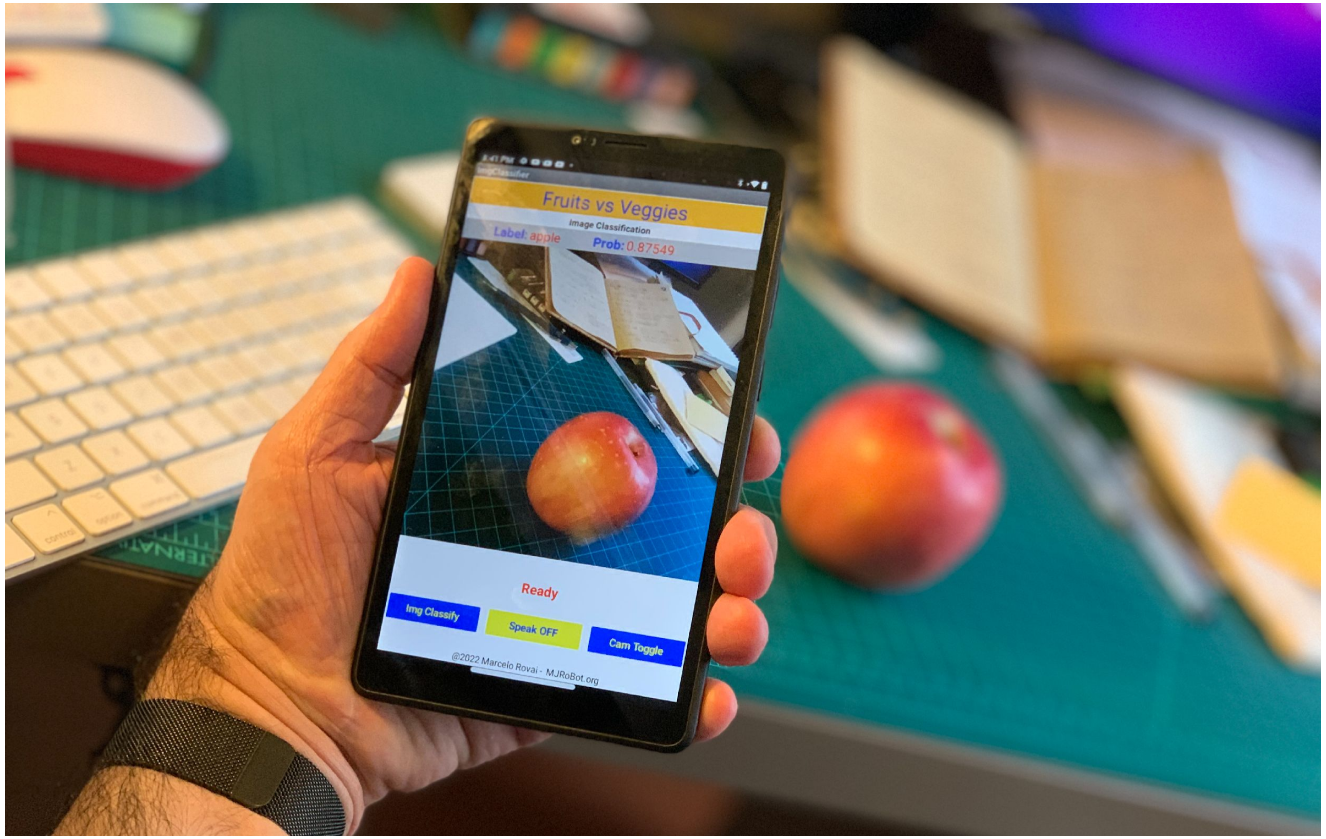


[Exploring AI at the Edge!]

Classifying Images using Smartphones

AI WITH APP INVENTOR



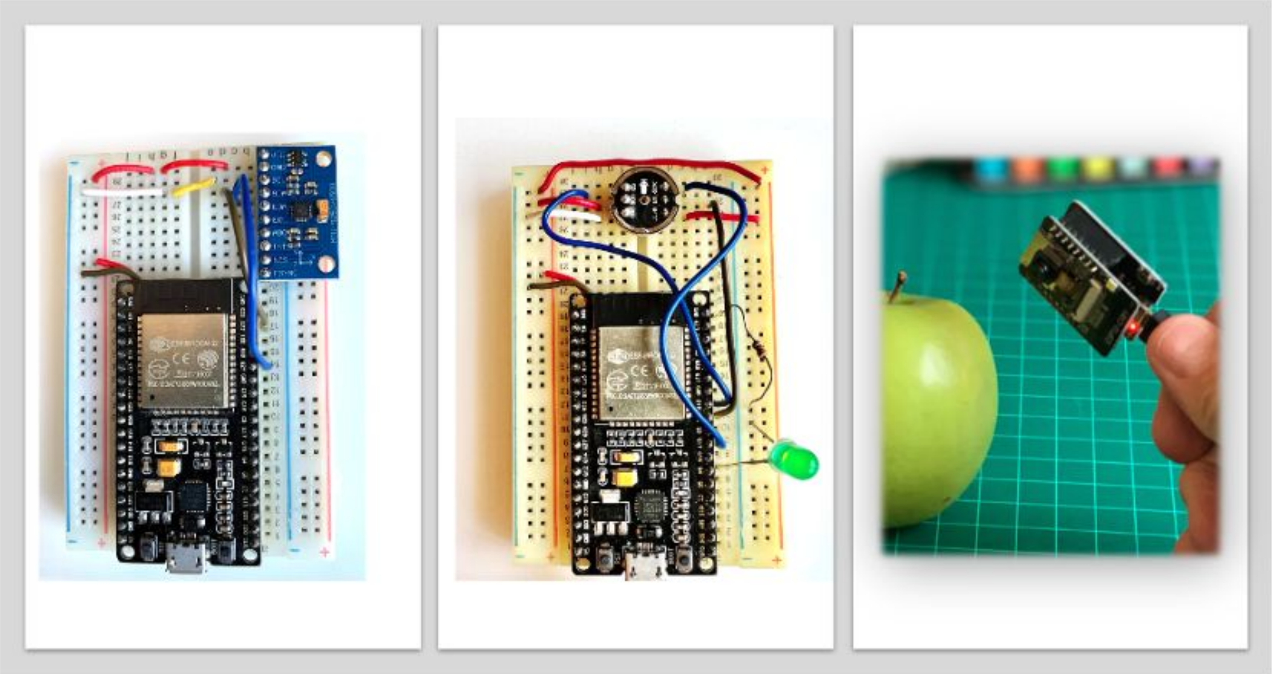


https://www.hackster.io/mjrobot/app-inventor-edgeml-image-classification-fruit-vs-veggies-b671da

To learn more ...

ESP32-TinyML

Exploring TinyML with ESP32 MCUs.



Seeed-XIAO-BLE-Sense

KWS, Anomaly Detection & Motion Classification and Micropython - Exploring the Seeed XIAO BLE Sense.

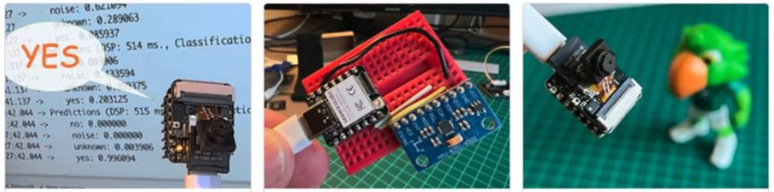
Programming Tiny devices with MicroPython. The easiest way!
MJRoBot (Marcelo Rovai)

Sensor DataLogger
MJRoBot (Marcelo Rovai)

TinyML Made Easy: Anomaly Detection & Motion Classification
MJRoBot (Marcelo Rovai)

TinyML Made Easy: Sound Classification (KWS)
MJRoBot (Marcelo Rovai)

XIAO-ESP32S3-Sense



TinyML Made Easy: KeyWord Spotting (KWS)
MJRoBot (Marcelo Rovai)

Exploring Machine Learning with the new XIAO ESP32S3
MJRoBot (Marcelo Rovai)

TinyML Made Easy: Image Classification
MJRoBot (Marcelo Rovai)

To learn more …

- IESTI01 TinyML - Machine Learning for Embedding Devices (Videos: Pt)
- WALC 22 – Applied AI - TinyML (Videos in Spanish)
- Professional Certificate in Tiny Machine Learning (TinyML) – edX/Harvard
- Introduction to Embedded Machine Learning - Coursera/Edge Impulse
- Computer Vision with Embedded Machine Learning - Coursera/Edge Impulse
- "Deep Learning with Python" book by François Chollet
- "TinyML" book by Pete Warden, Daniel Situnayake
- "TinyML Cookbook" by Gian Marco Iodice
- "AI at the Edge" book by Daniel Situnayake, Jenny Plunkett

On the TinyML4D website, You can find lots of educational materials on TinyML. They are all free and open-source for educational uses – we ask that if you use the material, please cite them! TinyML4D is an initiative to make TinyML education available to everyone globally.

TinyML4D Show&Tell Presentations

Date	Thread	Video
August 31st, 2023	TBD	Video here when ready
May 25th, 2023	Thread here	Video here when ready
April 20th, 2023	Thread here	https://youtu.be/uoM_IjXjDFY
March 30th, 2023	thread here	https://youtu.be/UQ0l-SwBwUY
February 23rd, 2023	thread here	https://youtu.be/BAEdil7X68Y
January 26th, 2023	thread here 17	https://youtu.be/-0xRZ-5UYUc 9
December 1st, 2022	thread here 2	https://youtu.be/e49pkjnIMlQ 8
October 27th, 2022	thread here 2	https://youtu.be/s8_hKpOWUwY 1

TinymML4D Academic Network Show and Tell Main Index.

The TinyML4D Academic Network Students should use this form to sign up for the latest presentations.
https://forms.gle/ic52HZMqVv4pBrkP7 2

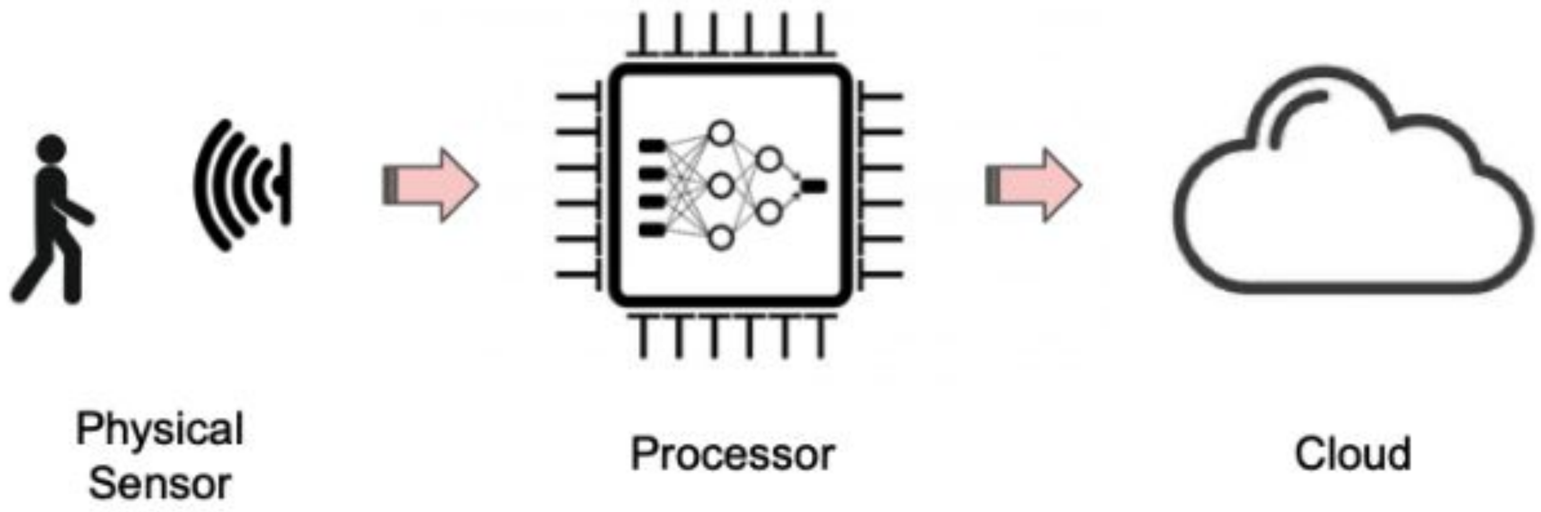
The Show and Tell are typically held at 2 pm UTC on the last Thursday of each month and will take place in this Zoom room.
https://zoom.us/j/95229860797 1
Meeting ID: 952 2986 0797
Passcode: 141278

Projects by Students (UNIFEI – IESTI01)

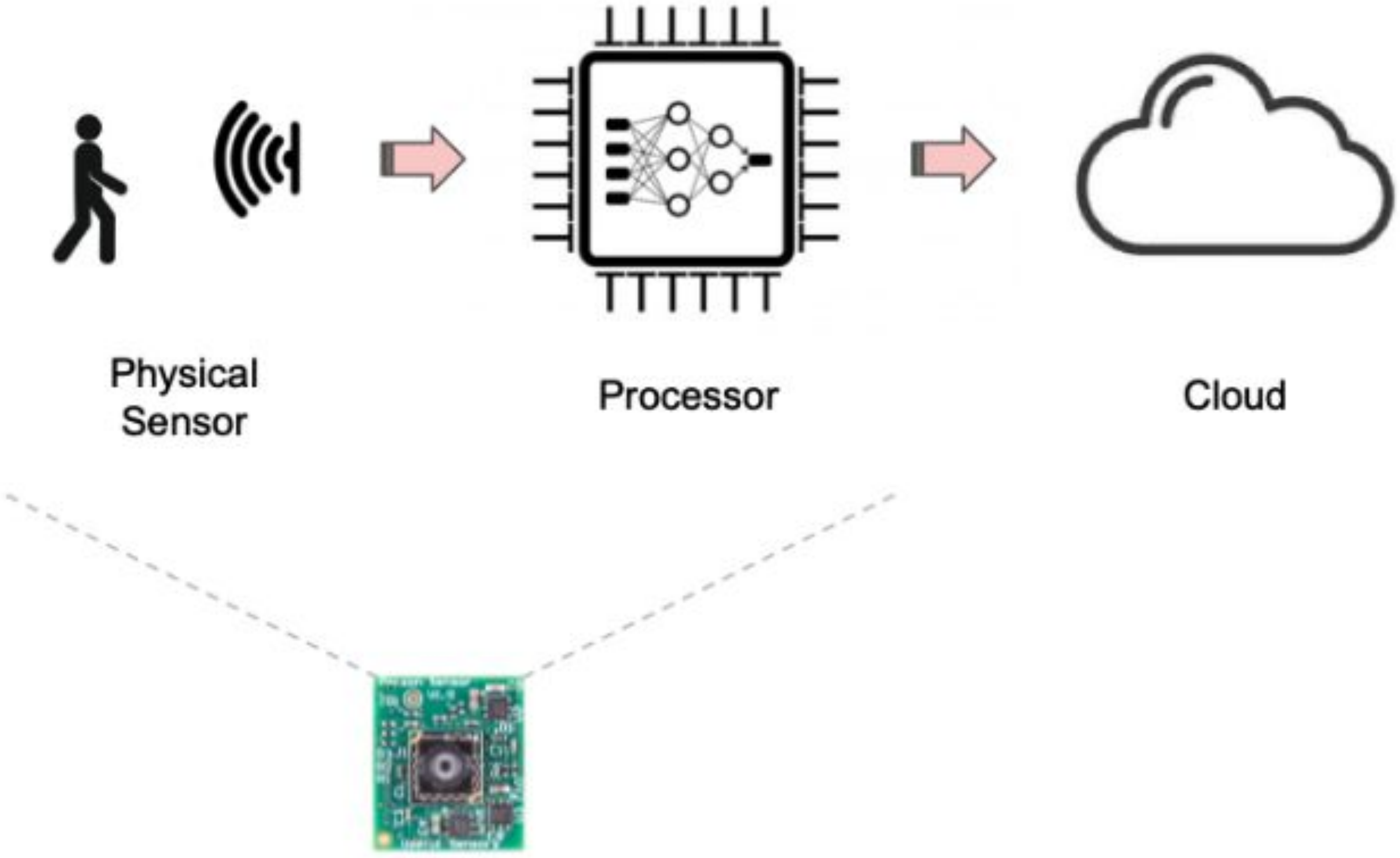
- **Sound:**
  - Earthquake detection
  - Covid Detection (cough)
  - Key Detection
  - Pulmonary Disease
  - Snore Detection
  - Bionic Hand Control
- **Other Sensors:**
  - Bionic Hand – Finger Detection
  - Electric Charges
  - ECG – Fibrial Atrilation detection
- **Image:**
  - Mask Detection
  - Forest Fire Detection
  - Helmet Detection
  - Water Consumption (hydrometer)
  - Sign Language
  - Coffee Disease Classification
  - Bee Counting
- **Vibration:**
  - Personal Trainer
  - Bearing – Anomaly Detection

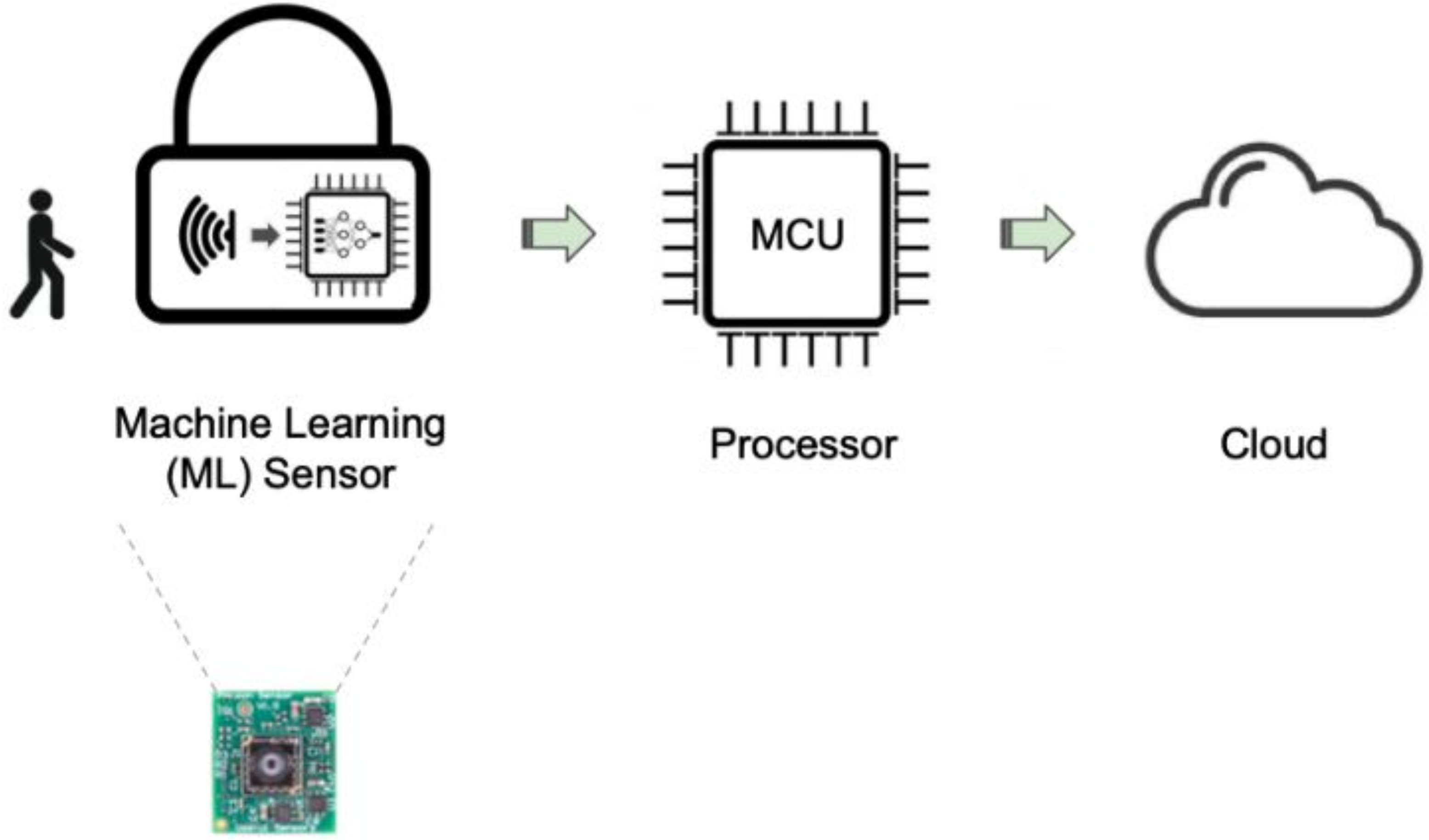
UNIFEI IESTI01 2021.1 – Final Projects   UNIFEI IESTI01 2021.2 – Final Projects   UNIFEI IESTI01 2022.1 – Final Projects

The Future of the EdgeAI



Sensor 1.0


Physical
Sensor
Processor
Cloud



Sensor 2.0

The Person Sensor has built in facial recognition and determines how many people there are, as well as their relative position.

[USD 10 -> https://www.sparkfun.com/products/21231](https://www.sparkfun.com/products/21231)

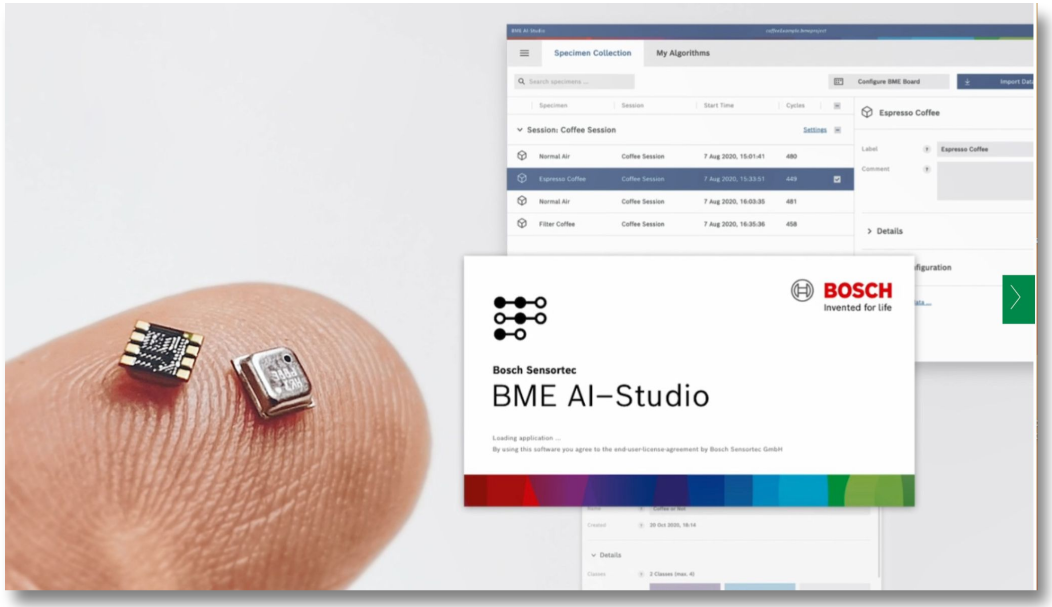
Bosch BME688 - Environmental sensing with AI

Relative humidity barometric pressure

Excellent temperature stability

Humidity

Gas sensing



https://www.bosch-sensortec.com/products/environmental-sensors/gas-sensors/bme688/

mlsensors.org

https://github.com/harvard-edge/ML-Sensors

arXiv:2206.03266v1 [cs.LG] 7 Jun 2022

Machine Learning Sensors

Pete Warden [1] Matthew Stewart [2] Brian Plancher [2] Colby Banbury [2] Shvetank Prakash [2] Emma Chen [2]
Zain Asgar [1] Sachin Katti [1] Vijay Janapa Reddi [2]

[1]Stanford University [2]Harvard University

Abstract

Machine learning sensors represent a paradigm shift for the future of embedded machine learning applications. Current instantiations of embedded machine learning (ML) suffer from complex integration, lack of modularity, and privacy and security concerns from data movement. This article proposes a more data-centric paradigm for embedding sensor intelligence on edge devices to combat these challenges. Our vision for "sensor 2.0" entails segregating sensor input data and ML processing from the wider system at the hardware level and providing a thin interface that mimics traditional sensors in functionality. This separation leads to a modular and easy-to-use ML sensor device. We discuss challenges presented by the standard approach of building ML processing into the software stack of the controlling microprocessor on an embedded system and how the modularity of ML sensors alleviates these problems. ML sensors increase privacy and accuracy while making it easier for system builders to integrate ML into their products as a simple component. We provide examples of prospective ML sensors and an illustrative datasheet as a demonstration and hope that this will build a dialogue to progress us towards sensor 2.0.

1 Introduction

Since the advent of AlexNet [43], deep neural networks have proven to be robust solutions to many challenges that involve making sense of data from the physical world. Machine learning (ML) models can now run on low-cost, low-power hardware capable of deployment as part of an embedded device. Processing data close to the sensor on an embedded device allows for an expansive new variety of always-on ML use-cases that preserve bandwidth, latency, and energy while improving responsiveness and maintaining data privacy. This emerging field, commonly referred to as embedded ML or tiny machine learning (TinyML) [73, 18, 39, 59], is paving the way for a prosperous new array of use-cases, from personalized health initiatives to improving manufacturing productivity and everything in-between.

However, the current practice for combining inference and sensing is cumbersome and raises the barrier of entry to embedded ML. At present, the general design practice is to design or leverage a board with decoupled sensors and compute (in the form of a microcontroller or DSP), and for the developer to figure out how to run ML on these embedded platforms. The developer is expected to train and optimize ML models and fit them within the resource constraints of the embedded device. Once an acceptable prototype implementation is developed, the model is integrated with the rest of the software on the device. Finally, the widget is tethered to the device under test to run inference. The current approach is slow, manual, energy-inefficient, and error-prone.

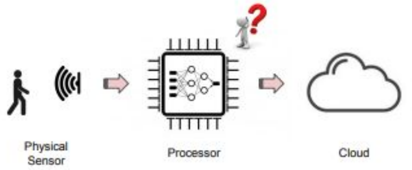


Figure 1. The Sensor 1.0 paradigm tightly couples the ML model with the application processor and logic, making it difficult to provide hard guarantees about the ML sensor's ultimate behavior.

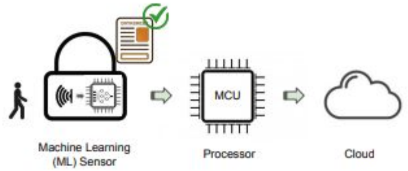
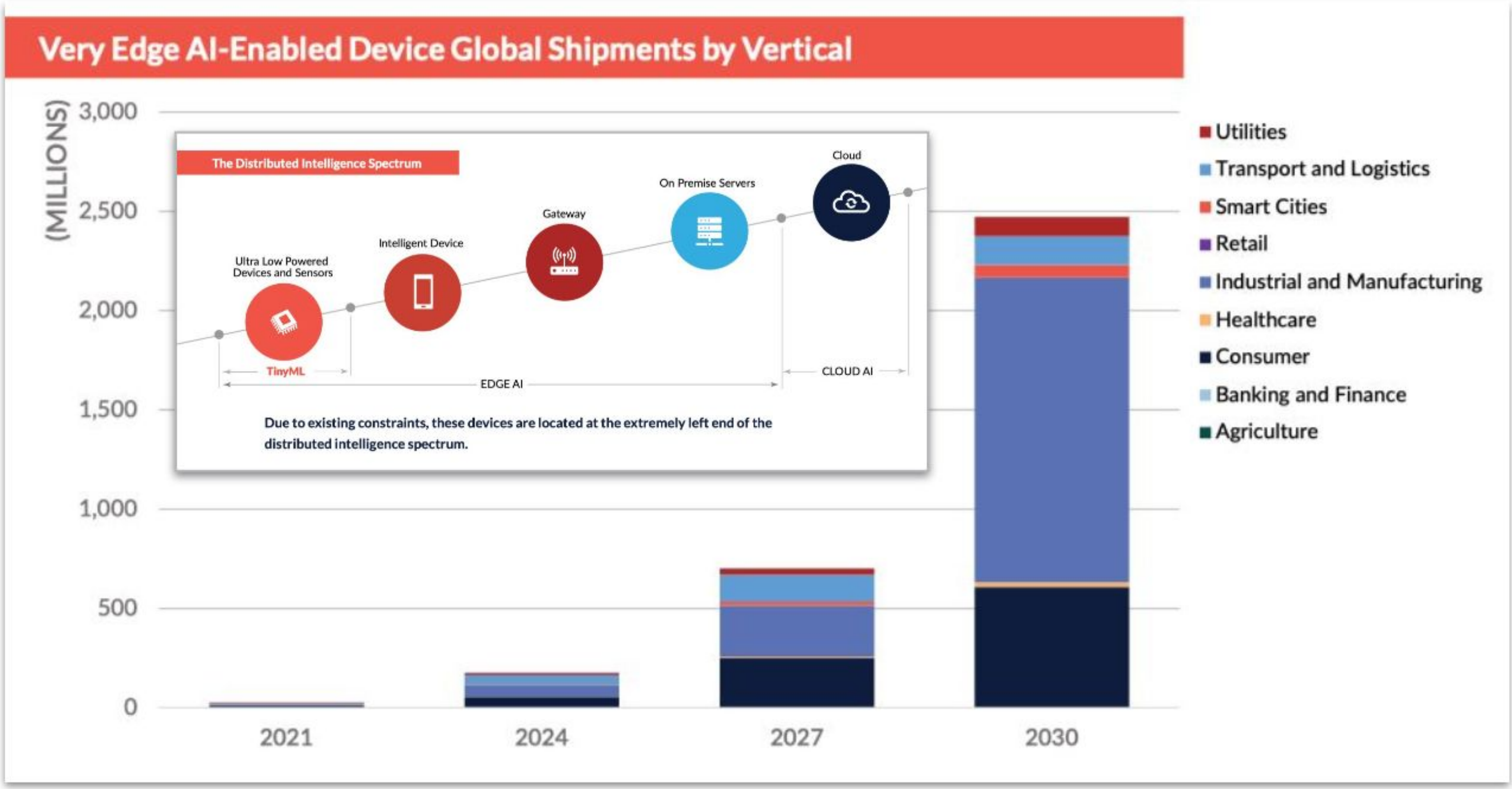


Figure 2. Our proposed Sensor 2.0 paradigm. The ML model is tightly coupled with the physical sensor, separate from the application processor, and comes with an ML sensor datasheet that makes its behavior transparent to the system integrators and developers.

It requires a sophisticated understanding of ML and the intricacies of ML model implementations to optimize and fit a model within the constraints of the embedded device.

Massive Potential for Impact



Source: ABI Research: TinyML

Edge AI has a very high impact potential, and it is for now!

Conclusion

The Future of ML is Tiny and Bright

***Vijay Janapa Reddi**, Ph. D. | Associate Professor |*
John A. Paulson School of Engineering and Applied Sciences | Harvard University |

VE RI TAS
HARVARD

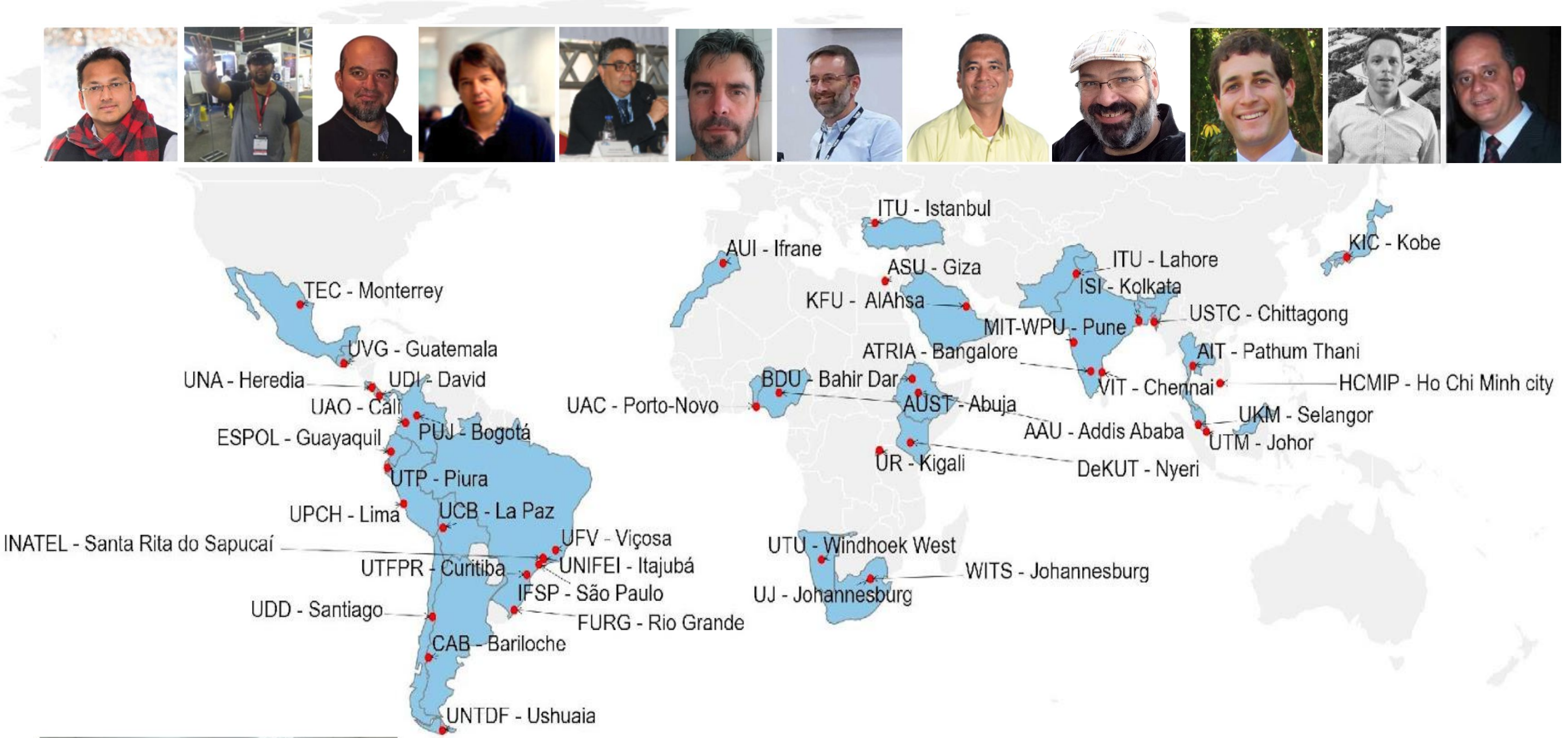
Q & A

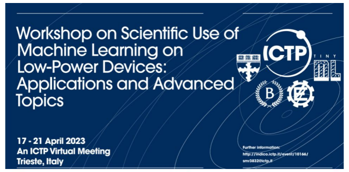
Marcelo Rovai was born in São Paulo and holds a Master's degree in Data Science from the Universidad del Desarrollo (UDD) in Chile and an MBA from IBMEC (INSPER) in Brazil. He graduated in 1982 as an Engineer from UNIFEI, Federal University of Itajubá, with a specialization from Polytechnic School of São Paulo University (USP), both institutions located in Brazil.

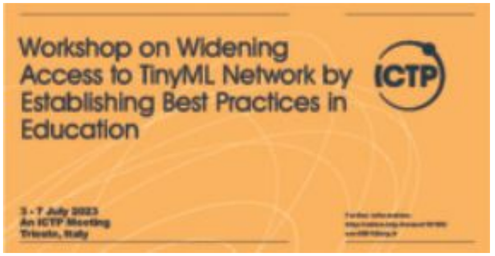
Rovai has experience as a teacher, engineer, and executive in several technology companies such as CDT/ETEP, AVIBRAS Aeroespacial, SID Informática, ATT-GIS, NCR, DELL, COMPAQ (HP), and more recently at IGT as a VP and a Senior Advisor for Latin America.

Marcelo Rovai publishes articles about electronics on websites such as MJRoBot.org, Hackster.io, Instructables.com, and Medium.com. Furthermore, he is a volunteer Professor at the UNIFEI in Brazil and a lecturer at several Congresses and Universities on IoT and TinyML. He is an active member and a Co-Chair of the TinyML4D group, an initiative to bring TinyML education to developing countries.

TinyML4D Academic Network







Thanks

UNIFEI