Embedded ML (TinyML) Intro & Applications

Prof. Marcelo J. Rovai
rovai@unifei.edu.br

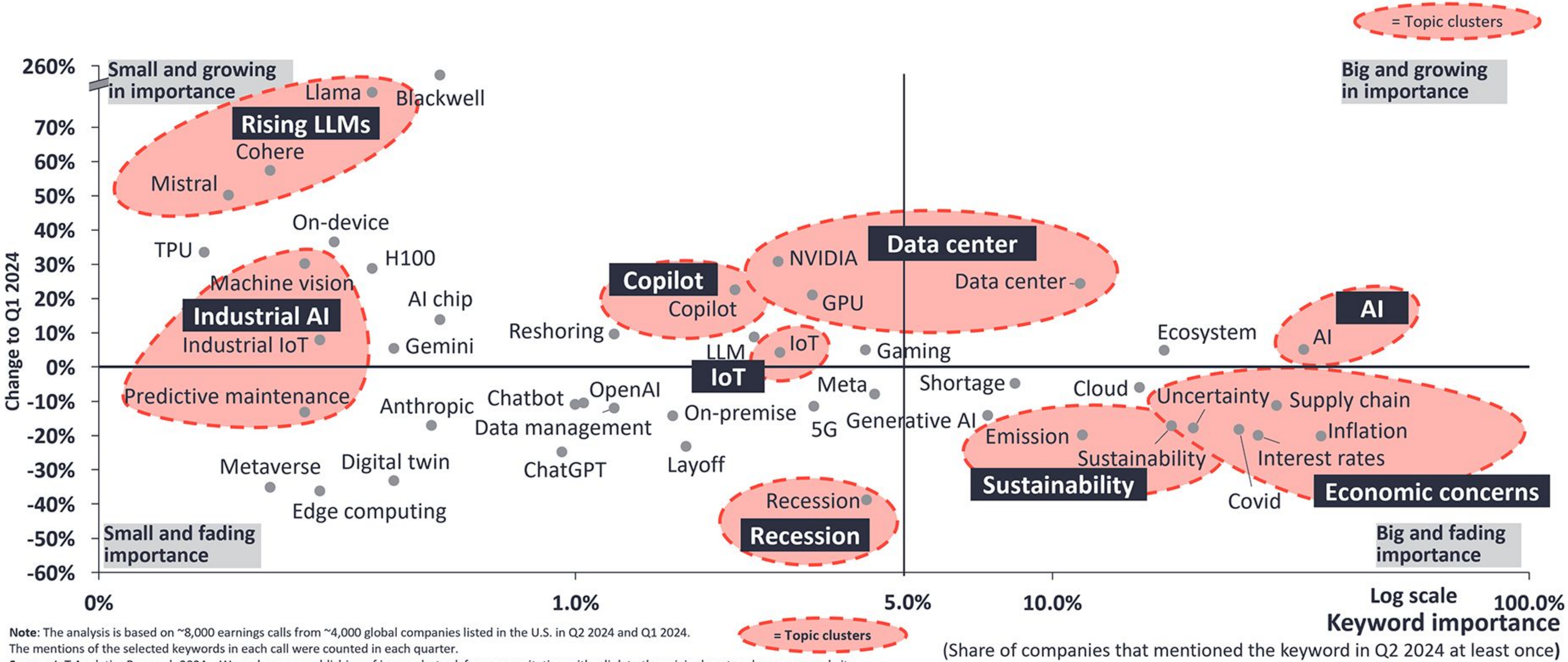
UNIFEI - Federal University of Itajuba, Brazil
TinyML4D Academic Network Co-Chair

Internet of Things (IoT)

IOT ANALYTICS

This analysis is part of the **CEO Insights** series

What CEOs talked about in Q2 2024 (vs. Q1 2024)



Note: The analysis is based on ~8,000 earnings calls from ~4,000 global companies listed in the U.S. in Q2 2024 and Q1 2024. The mentions of the selected keywords in each call were counted in each quarter.

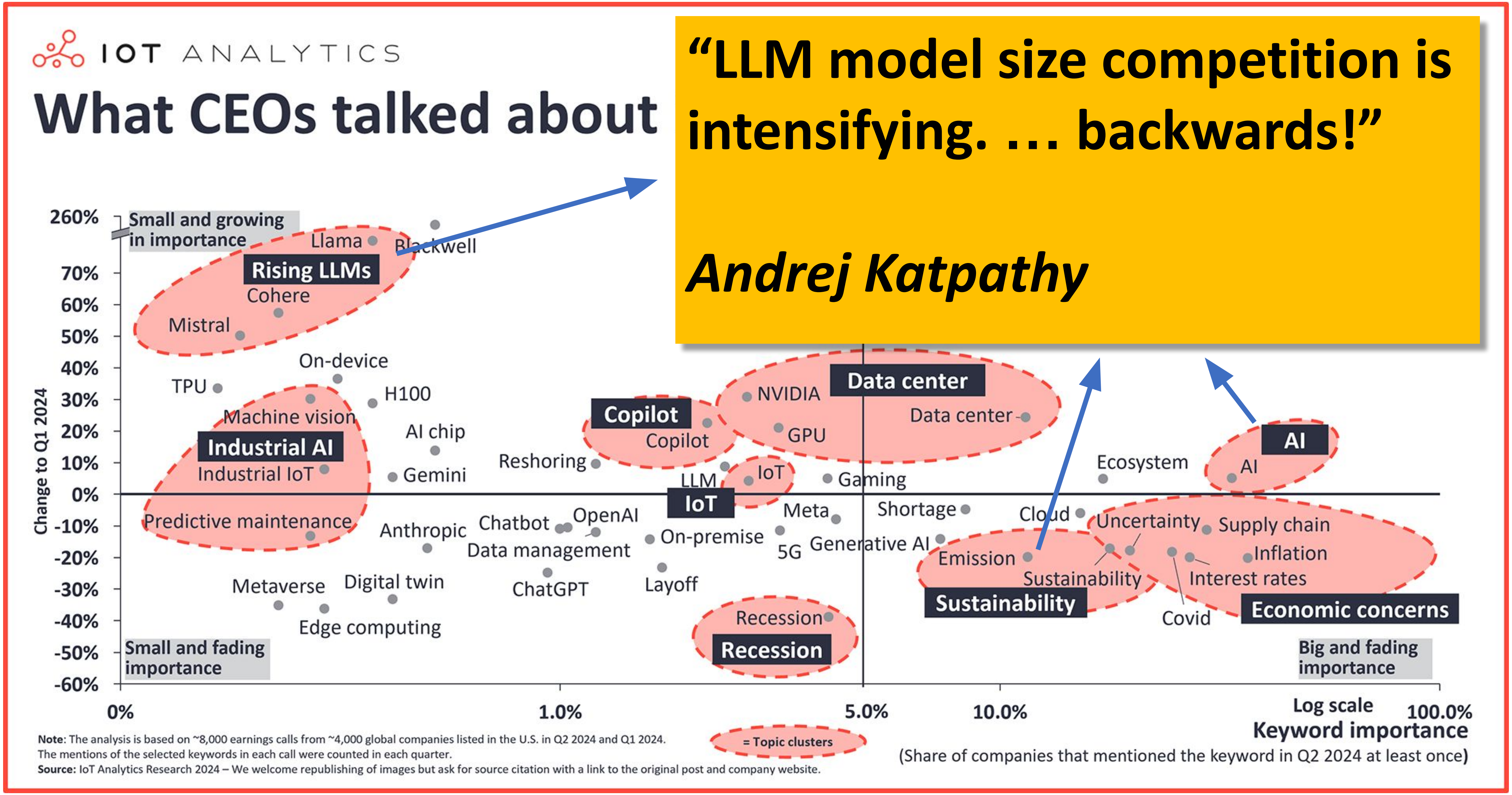
Source: IoT Analytics Research 2024 – We welcome republishing of images but ask for source citation with a link to the original post and company website.

= Topic clusters

https://iot-analytics.com/what-ceos-talked-about-in-q2-2024/

“LLM model size competition is intensifying. … backwards!”

Andrej Katpathy



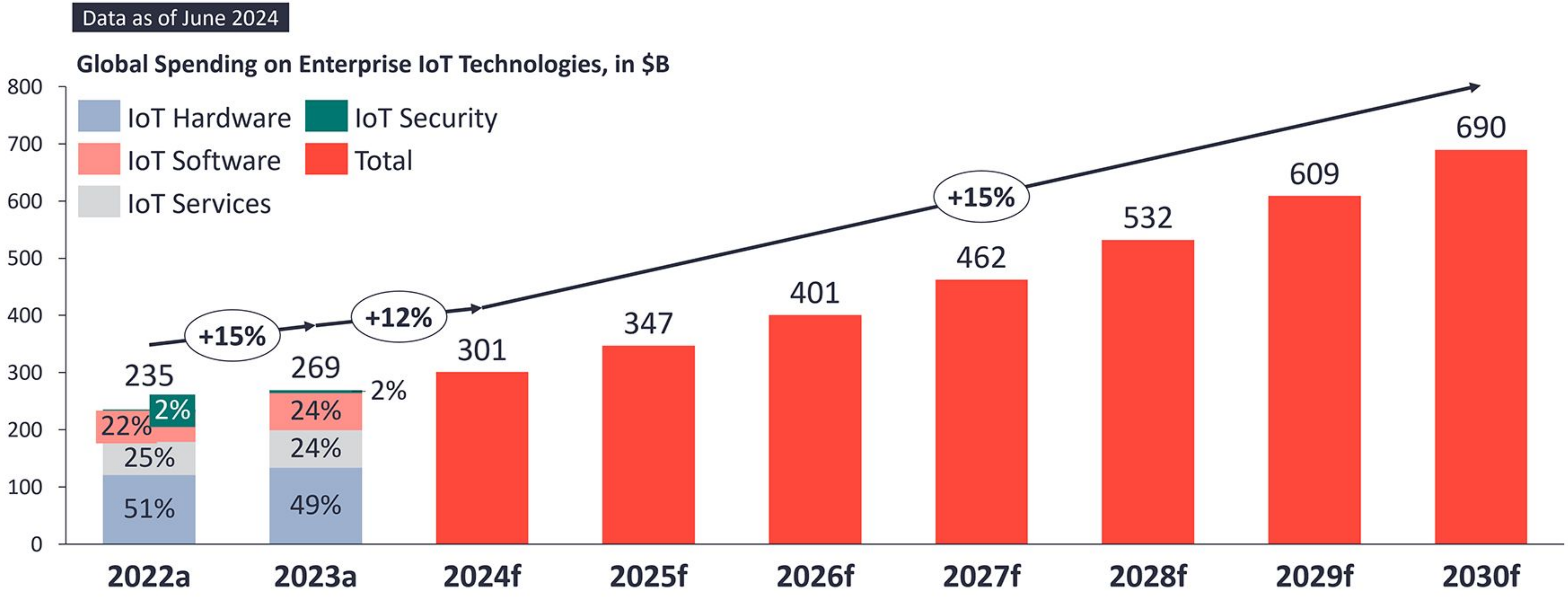
https://iot-analytics.com/what-ceos-talked-about-in-q2-2024/

IOT ANALYTICS

July 2024

Your Global IoT Market Research Partner

The enterprise IoT market by technology 2023–2030



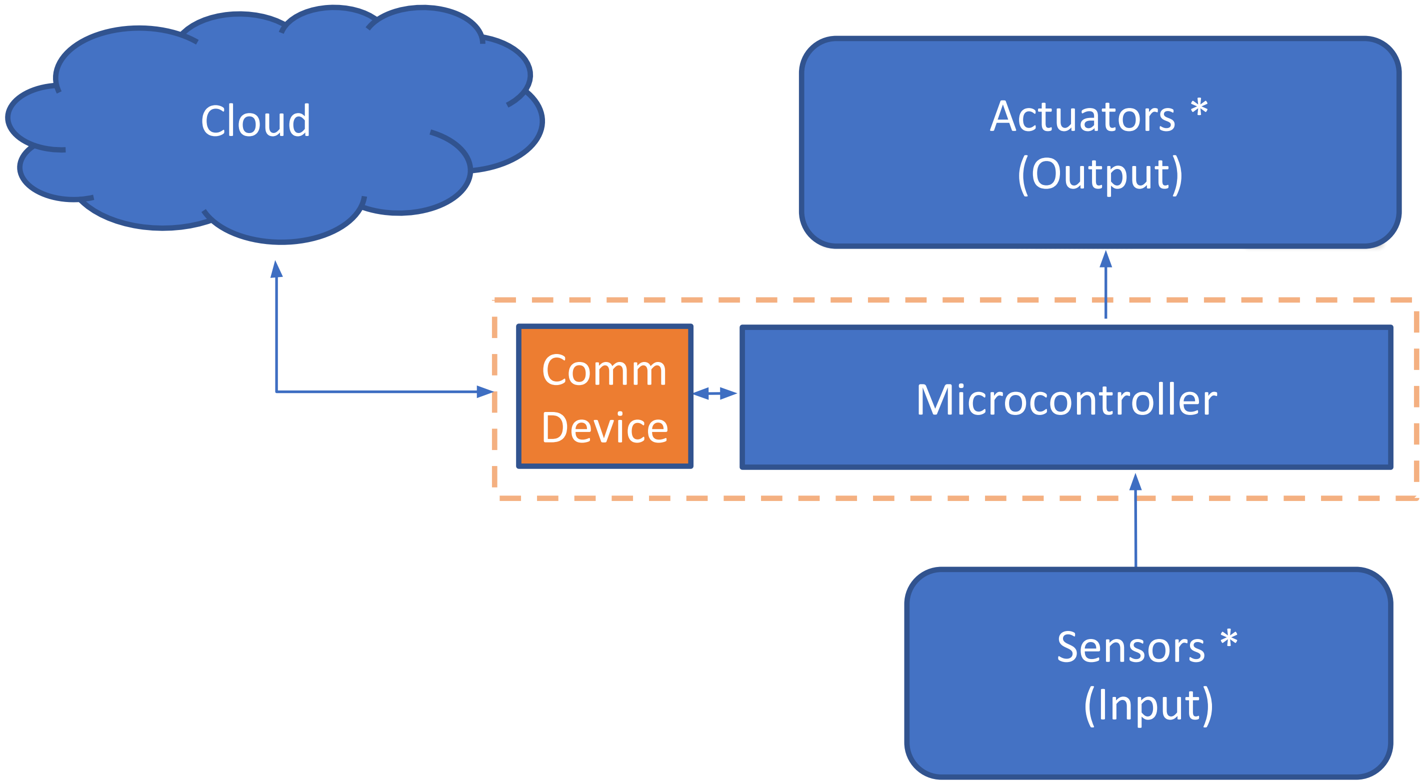
Note: IoT Analytics defines IoT as a network of internet-enabled physical objects. Objects that become internet-enabled (IoT devices) typically interact via embedded systems, some form of network communication, or a combination of edge and cloud computing. The data from IoT-connected devices is often used to create novel end-user applications. Connected personal computers, tablets, and smartphones are not considered IoT, although these may be part of the solution setup. Devices connected via extremely simple connectivity methods, such as radio frequency identification or quick response codes, are not considered IoT devices. Since the last update in 2023 our definition of the enterprise IoT tech stack slightly changed.

a: Actuals, f: Forecast

Source: IoT Analytics Research 2024 – Global IoT Enterprise Spending Dashboard (Q2/2024 update). We welcome republishing of images but ask for source citation with a link to the original post or company website.

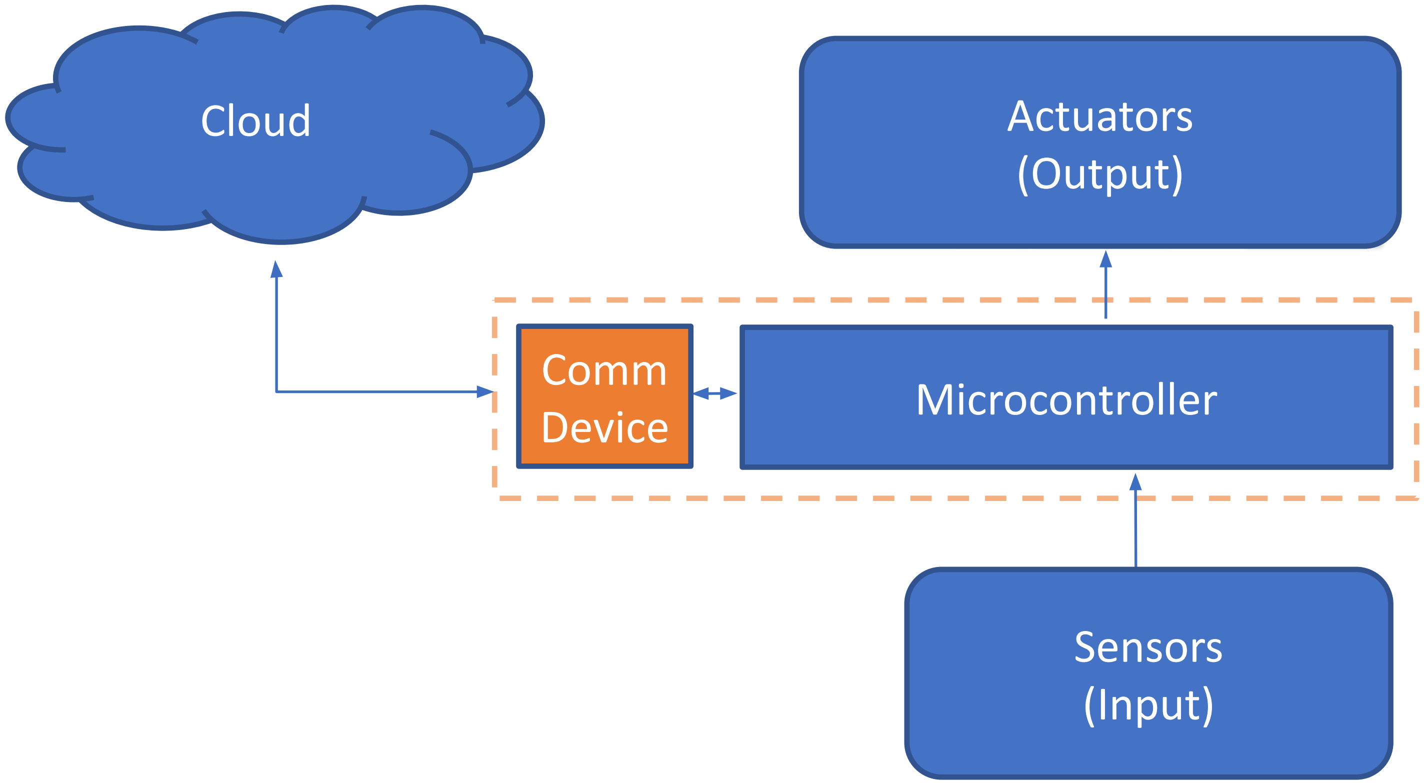
https://iot-analytics.com/iot-market-size/

Typical IoT Project



* “Things”

Typical IoT Project



5 Quintillion

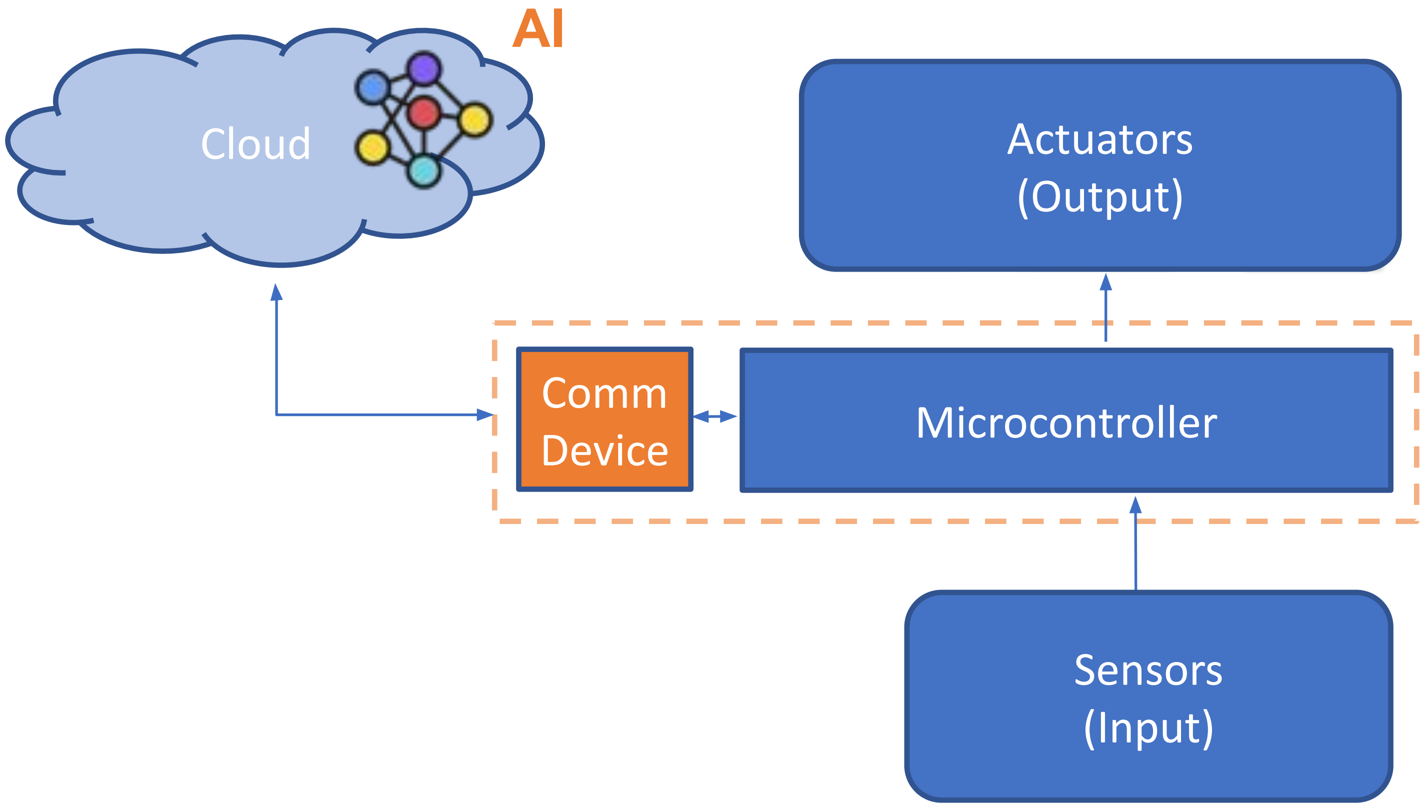
bytes of data produced every day by IoT

<1%

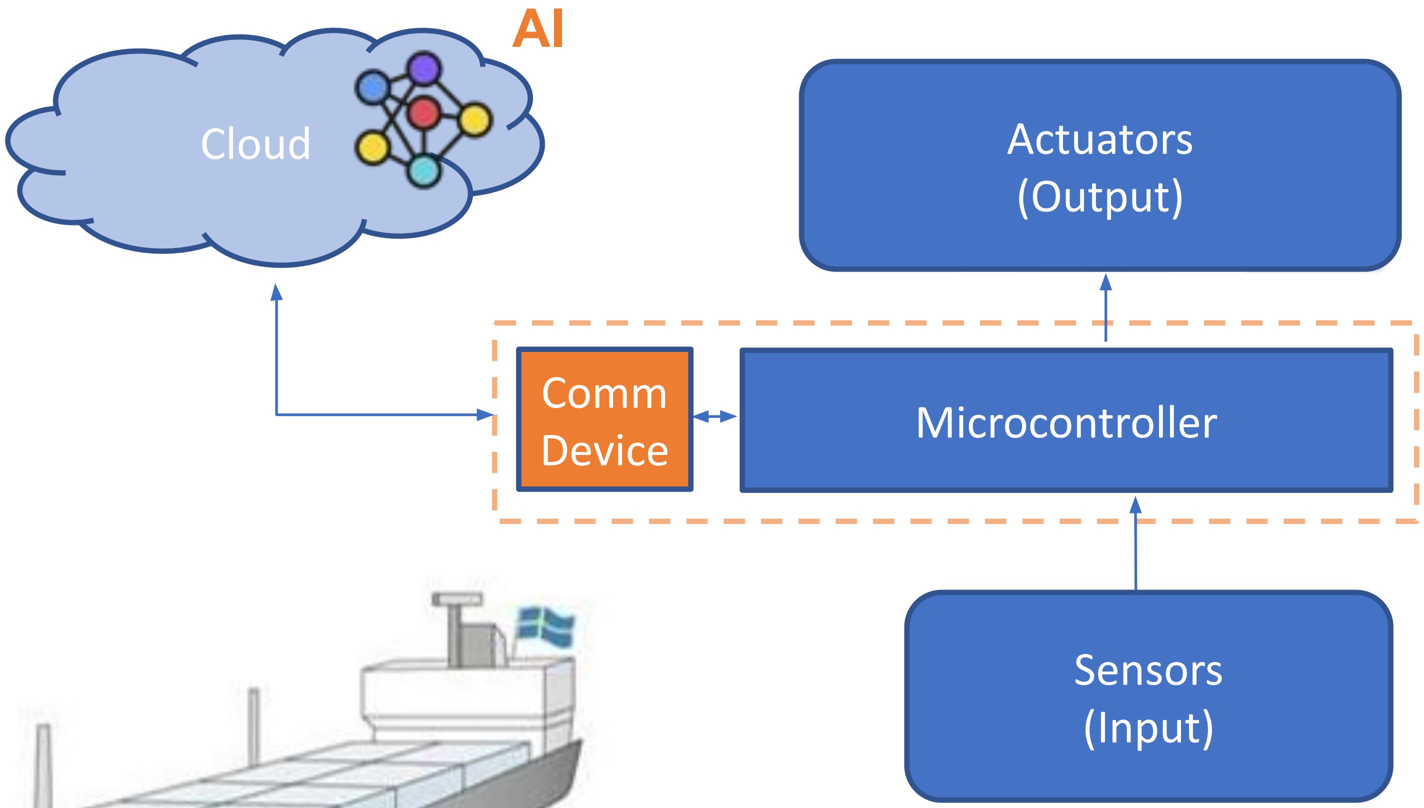
of unstructured data is analyzed or used at all

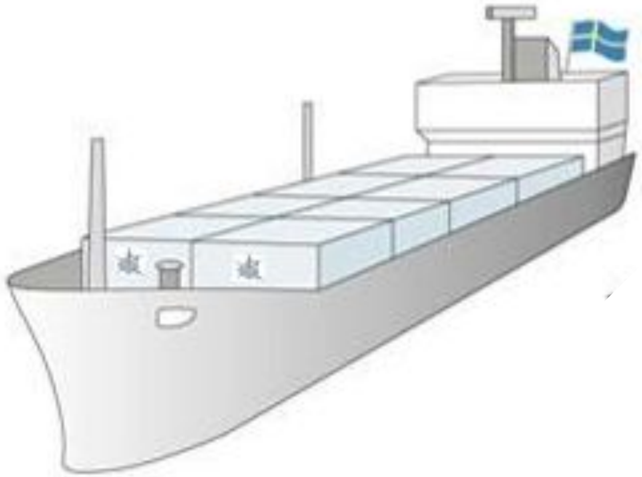
Source: Harvard Business Review, [What's Your Data Strategy?](), April 18, 2017
Cisco, [Internet of Things (IoT) Data Continues to Explode Exponentially. Who Is Using That Data and How?](), Feb 5, 2018

Typical AIoT Project



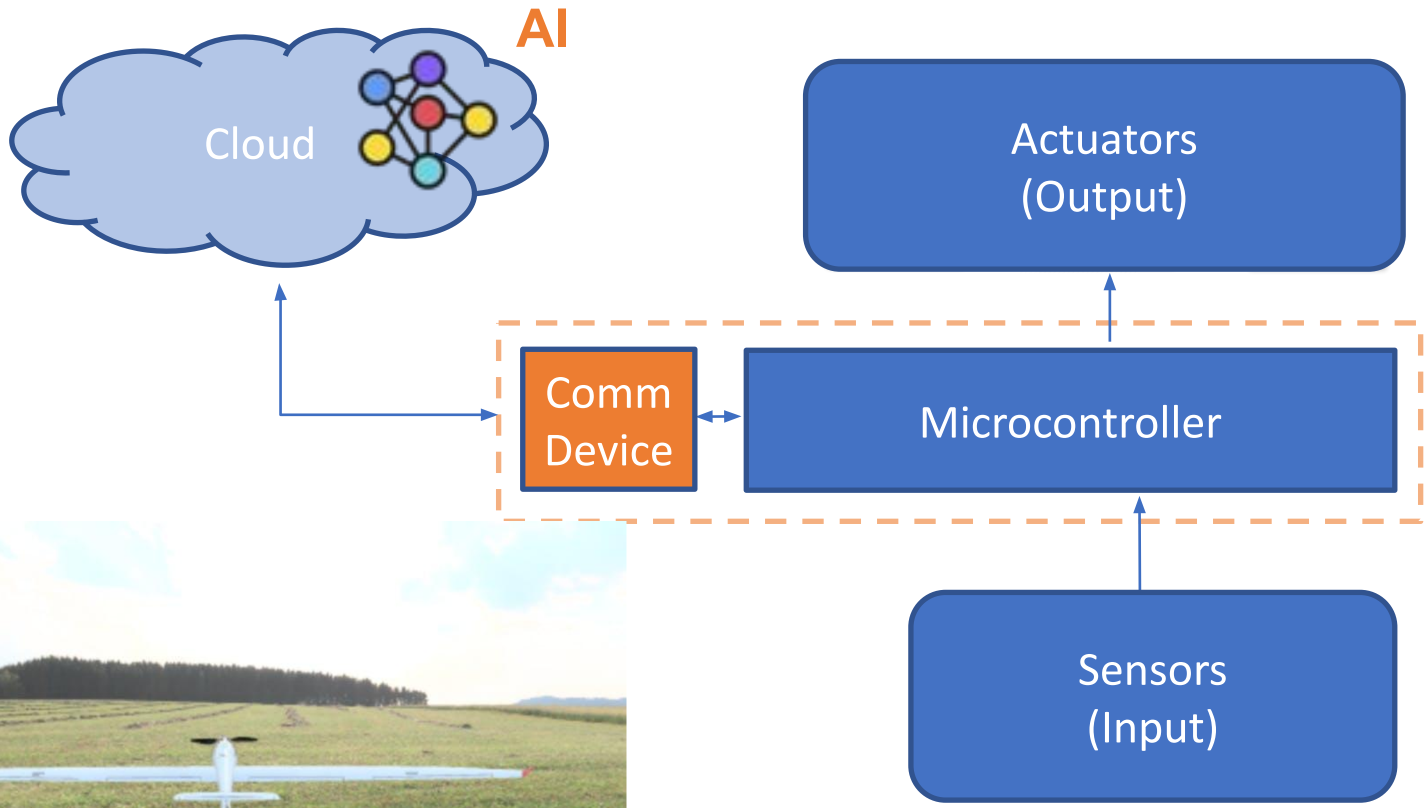
Typical AIoT Project … … Issues





Bandwidth

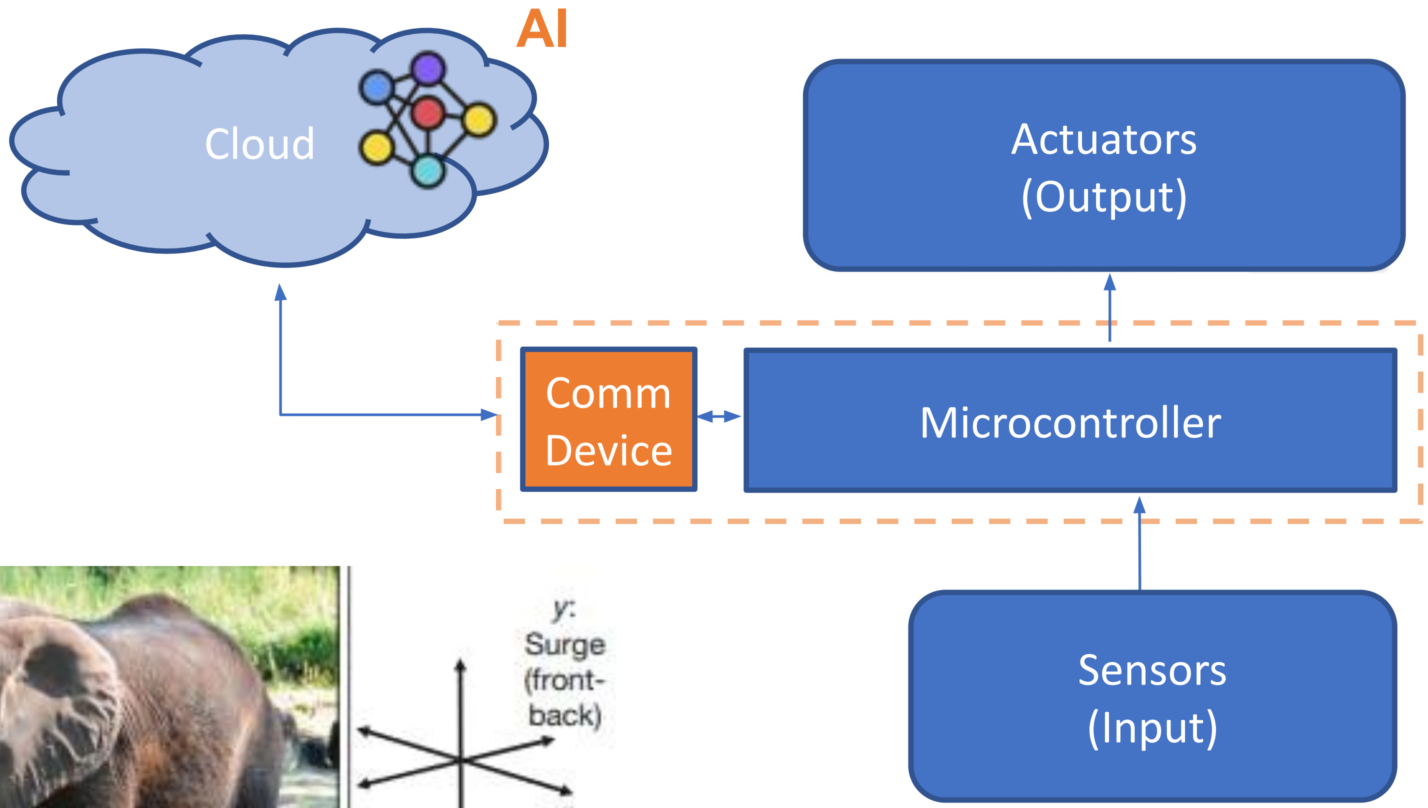
Typical AIoT Project … … Issues

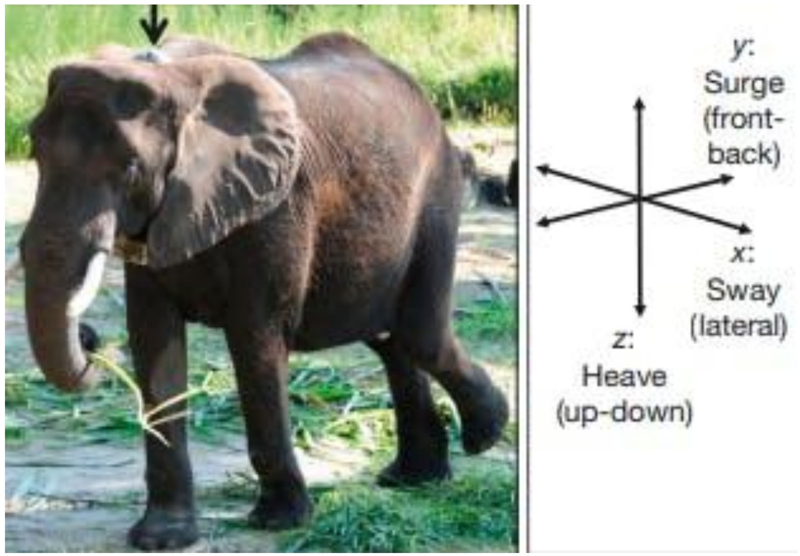


Bandwidth

Latency

Typical AIoT Project … … Issues





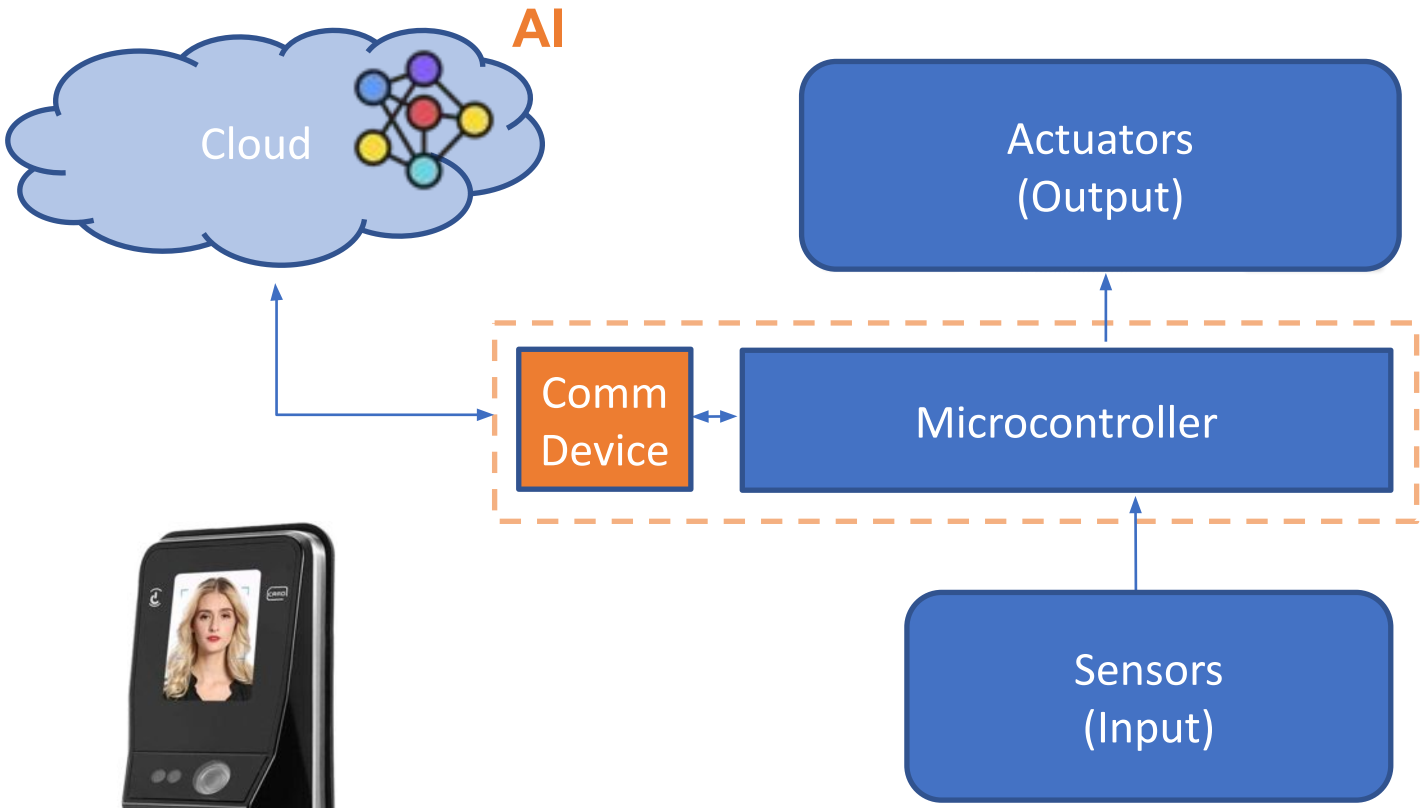
Bandwidth

Latency

Energy

Typical AIoT Project . . .

. . . Issues



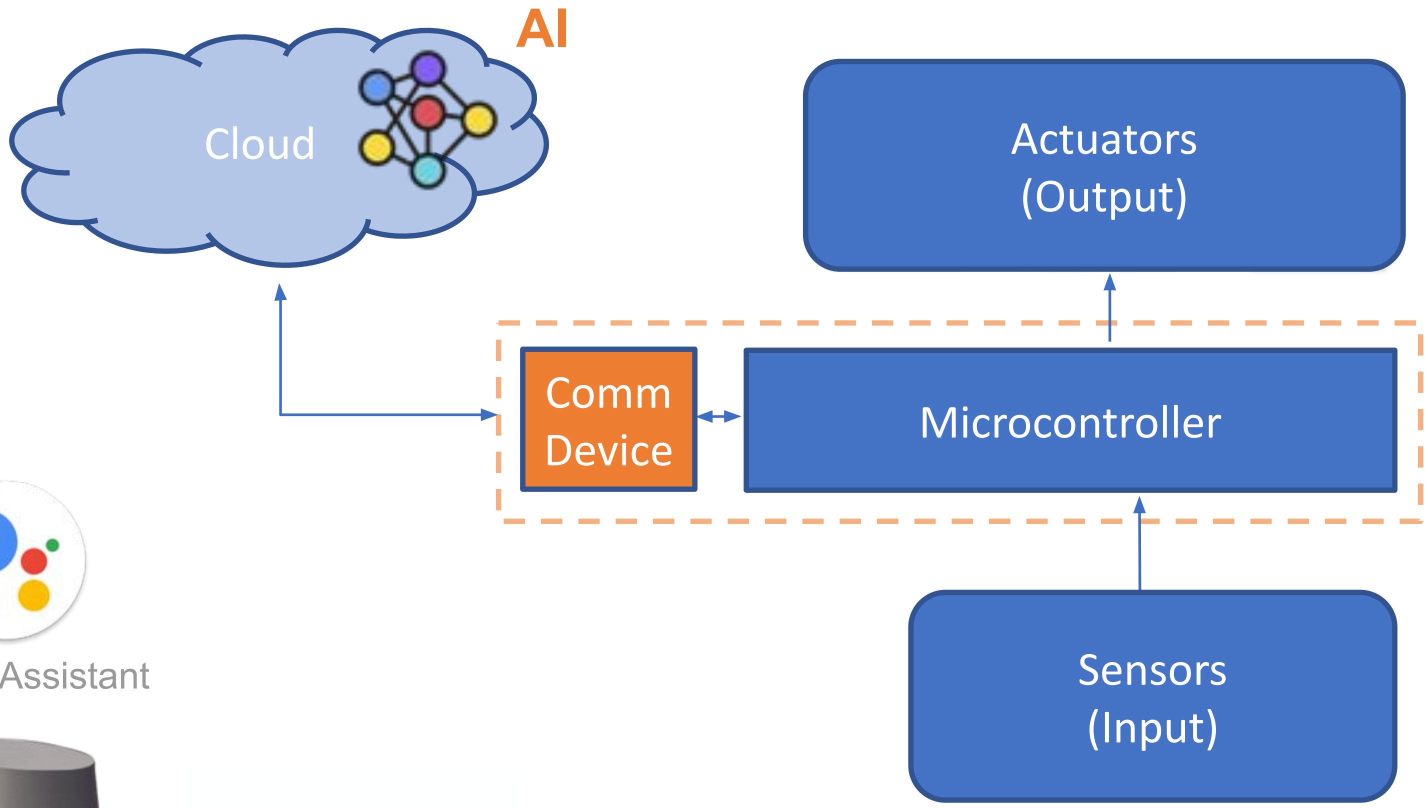
Bandwidth

Latency

Energy

Reliability

Typical AIoT Project … … Issues



Google Assistant

Bandwidth

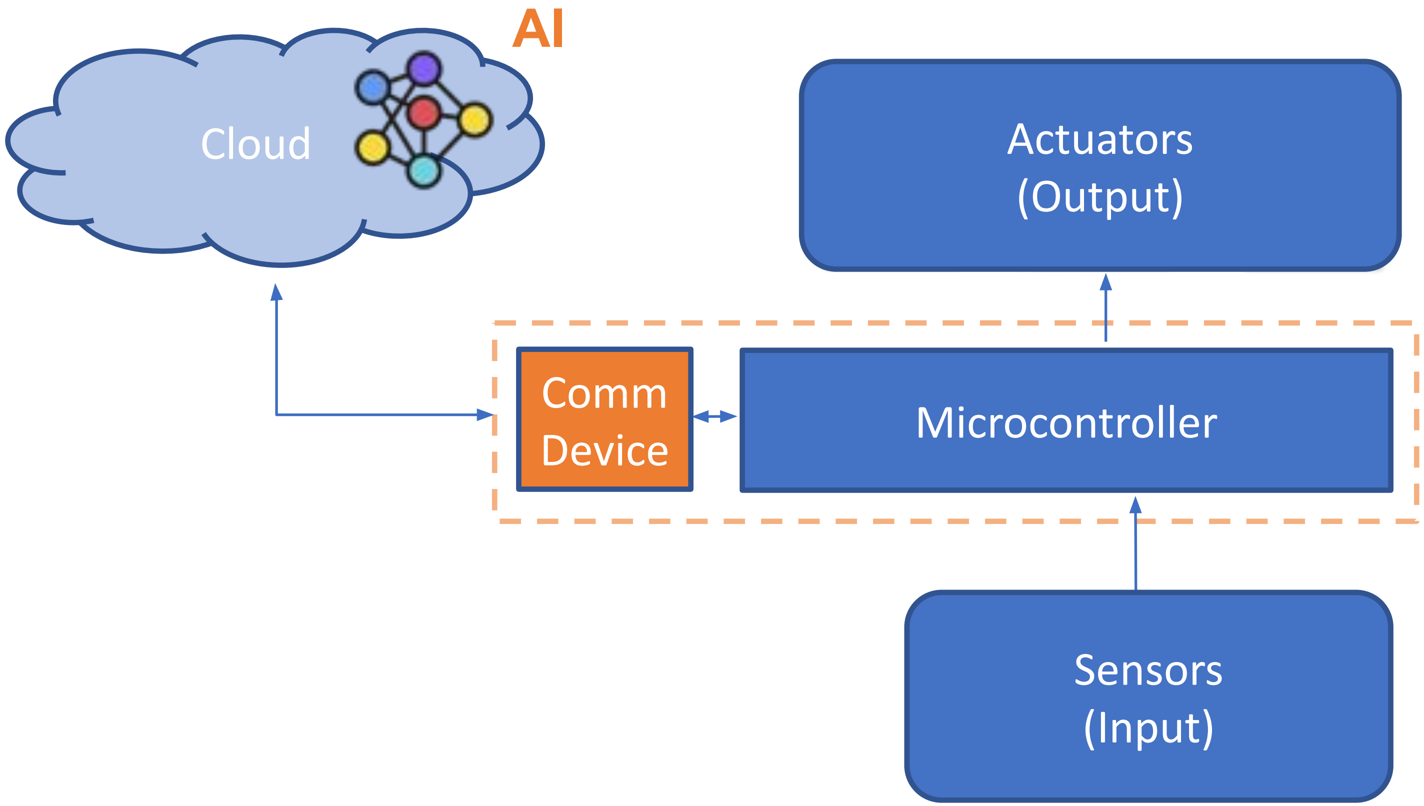
Latency

Energy

Reliability

Privacy

Typical AIoT Project …



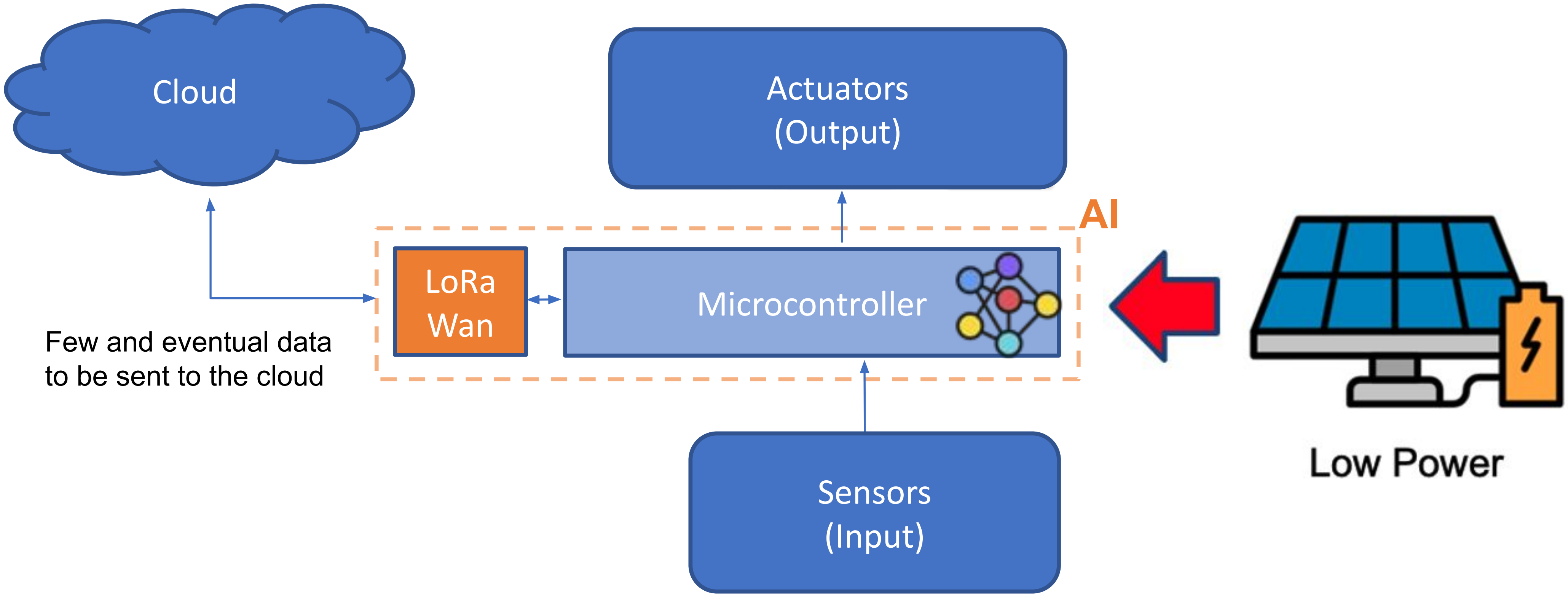
… Issues

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

… Solution ?

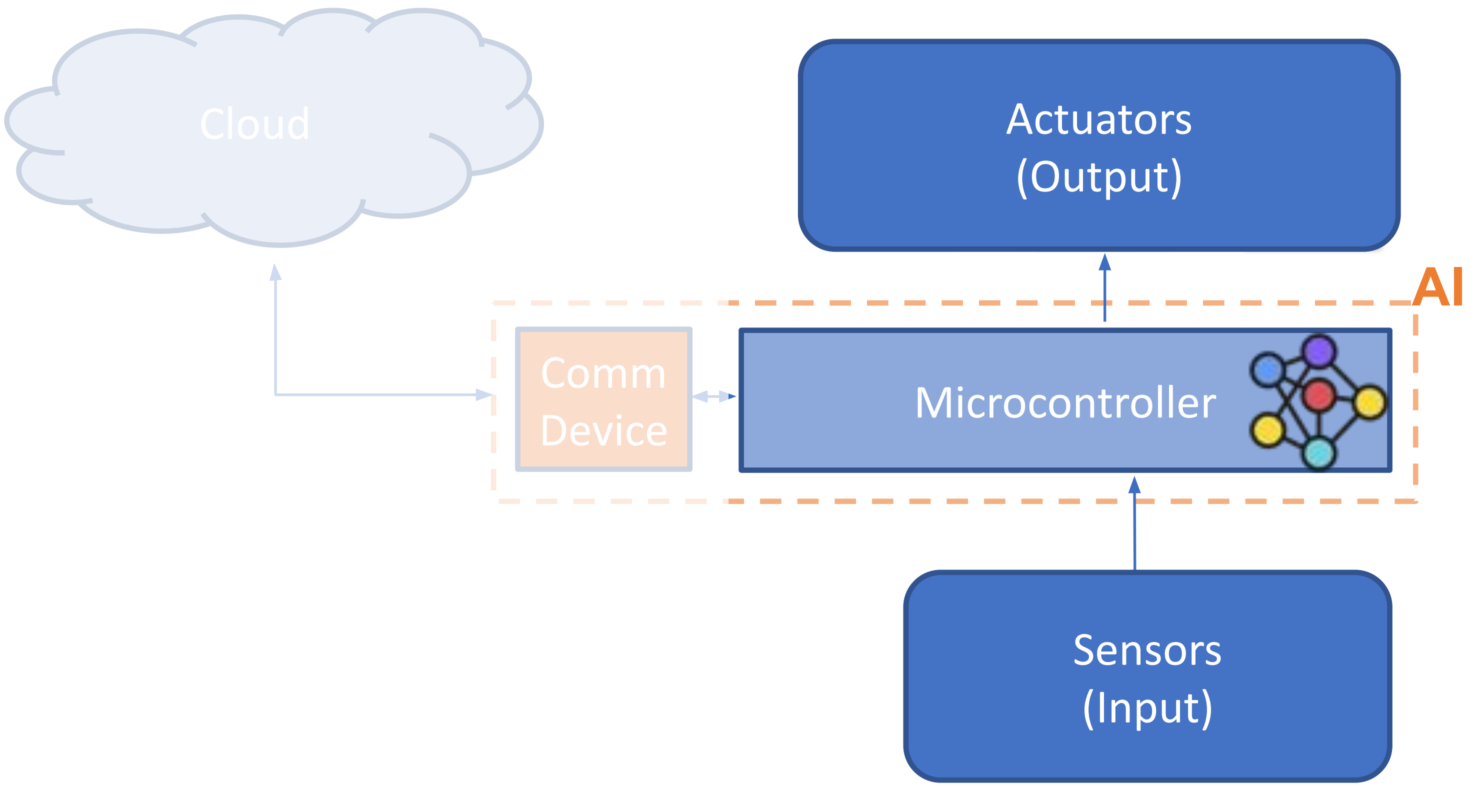
IoT 2.0 * – Edge AI/ML

* Intelligence of Things



… Solution -> ML goes close to data

When to use an Edge AI/ML approach:



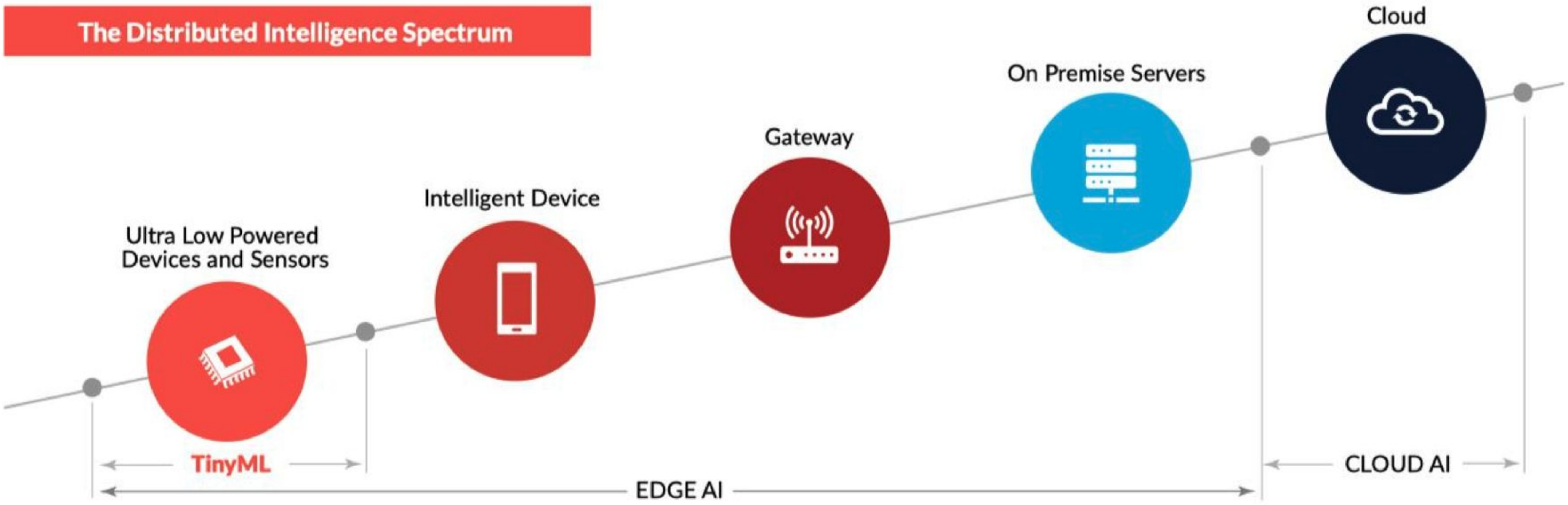
Bandwidth

Latency

Energy

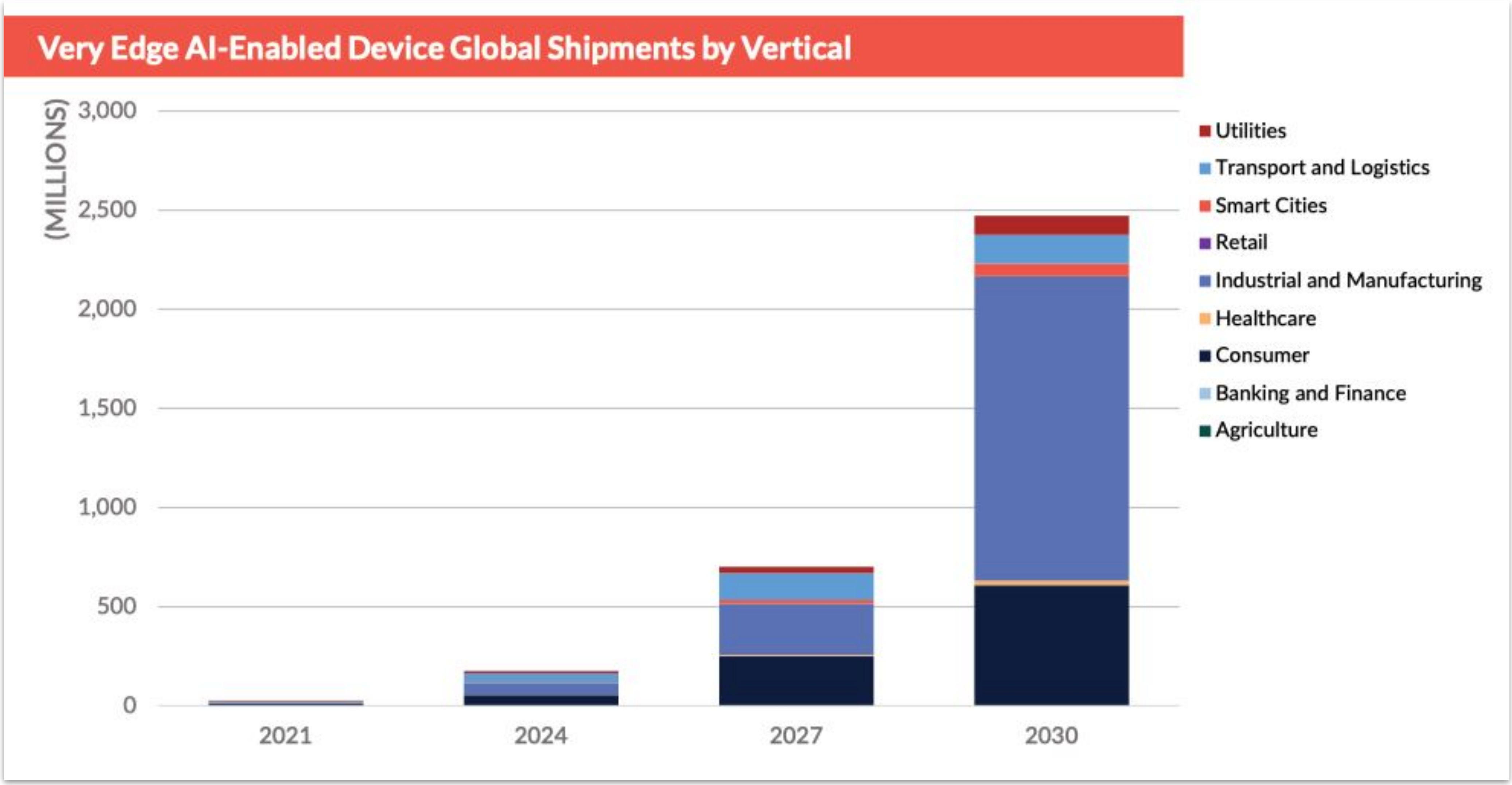
Reliability

Privacy



Source: ABI Research: TinyML

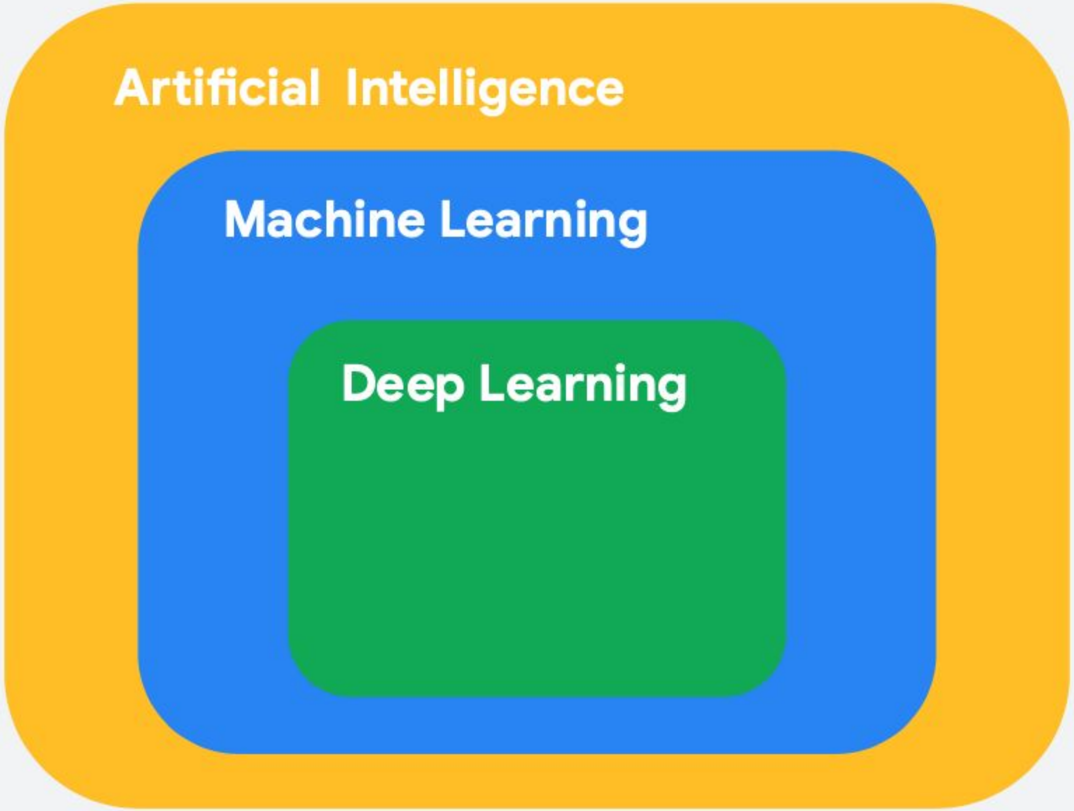
Market Forecast



Source: ABI Research: TinyML

Embedded ML (TinyML)

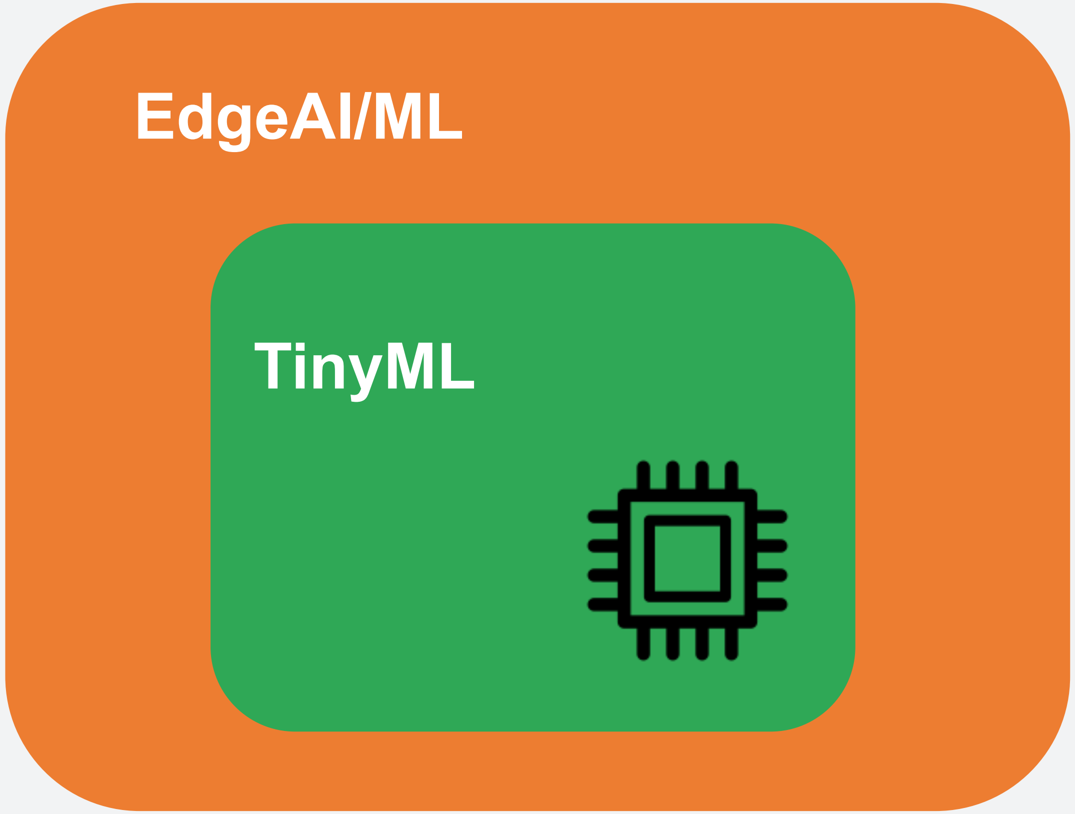
Introduction



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

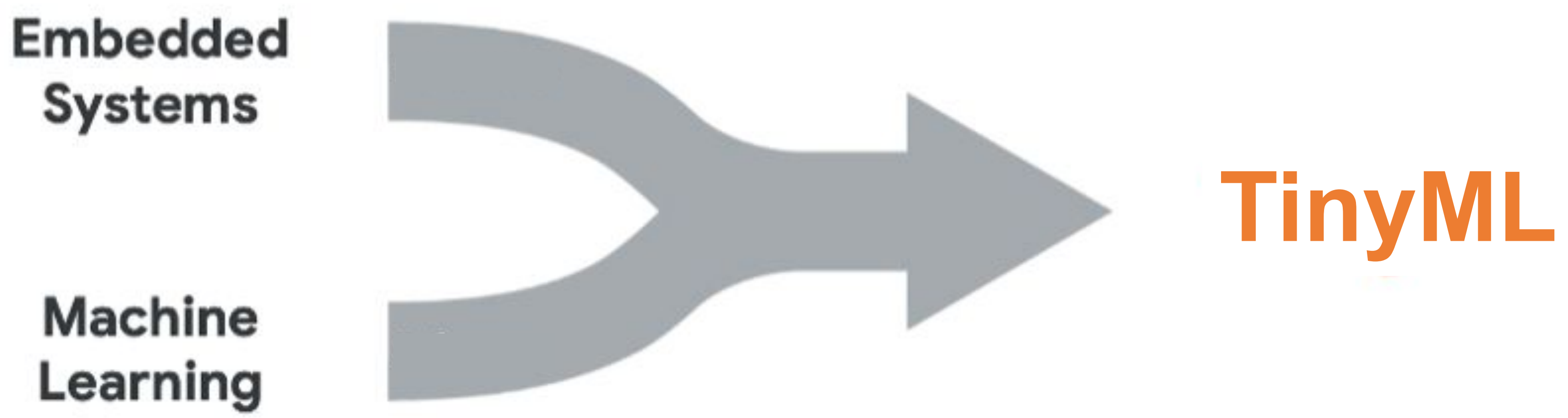
Low power consumption

Battery-operated

Always-on ML

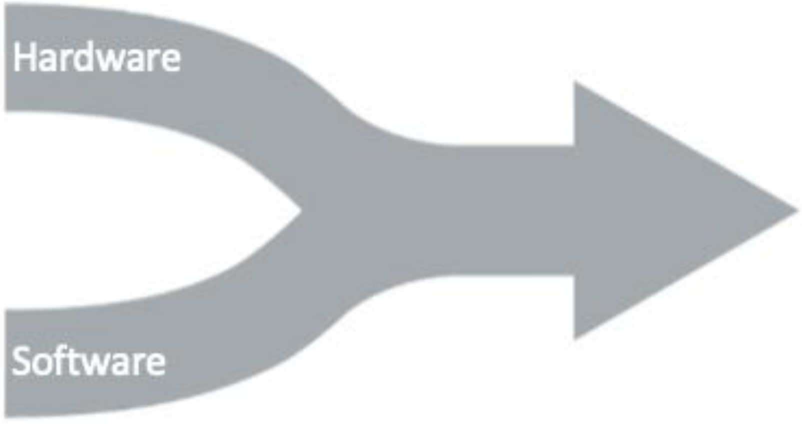
27

What Makes **TinyML** ?



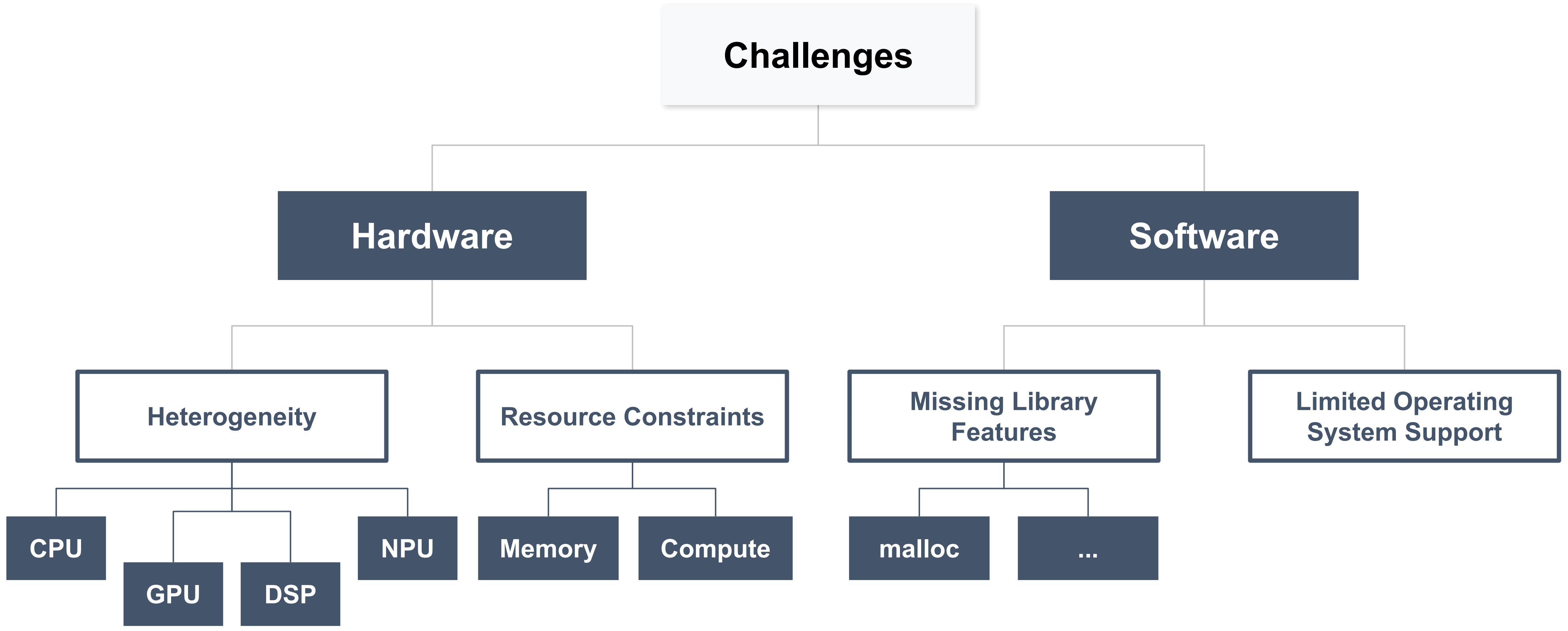
What Makes TinyML ?

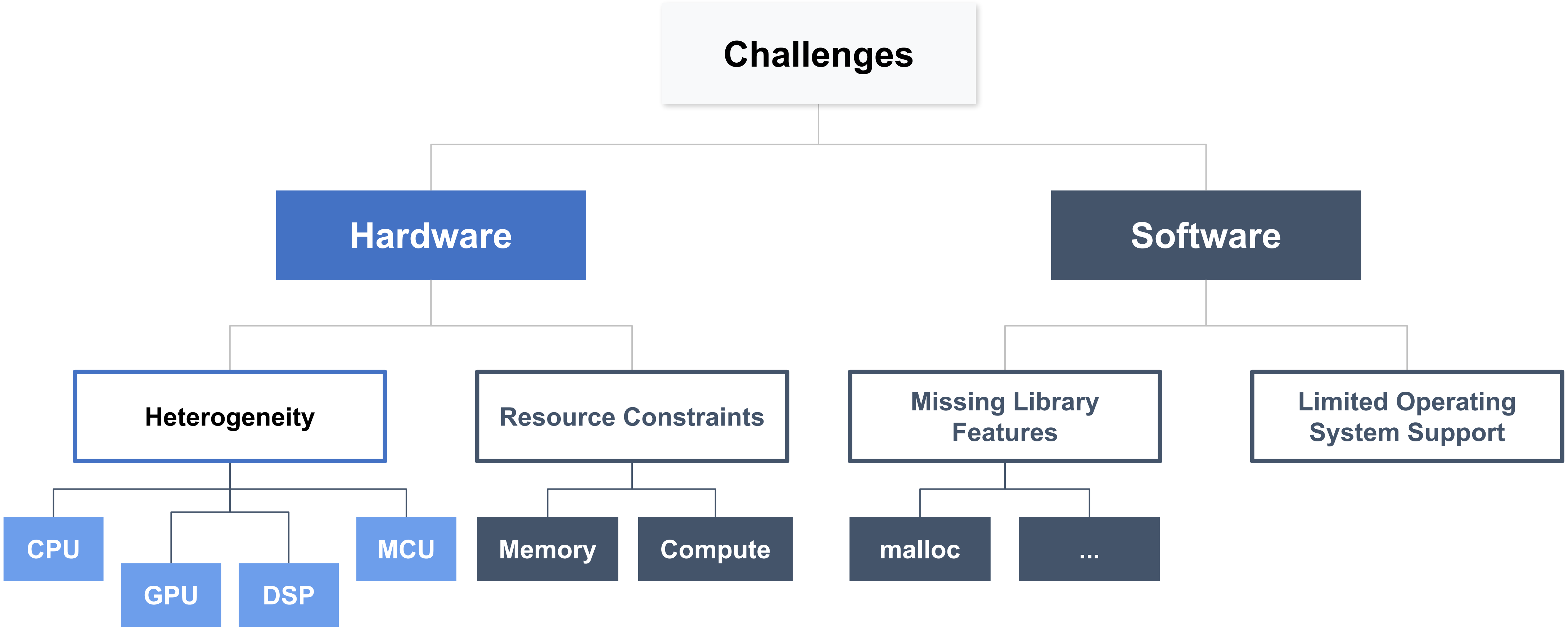
TensorFlow Lite



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

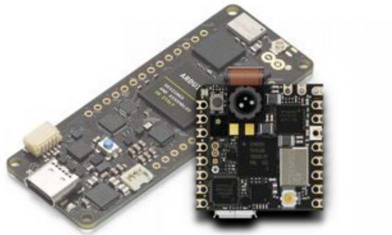
33

Hardware

Hardware

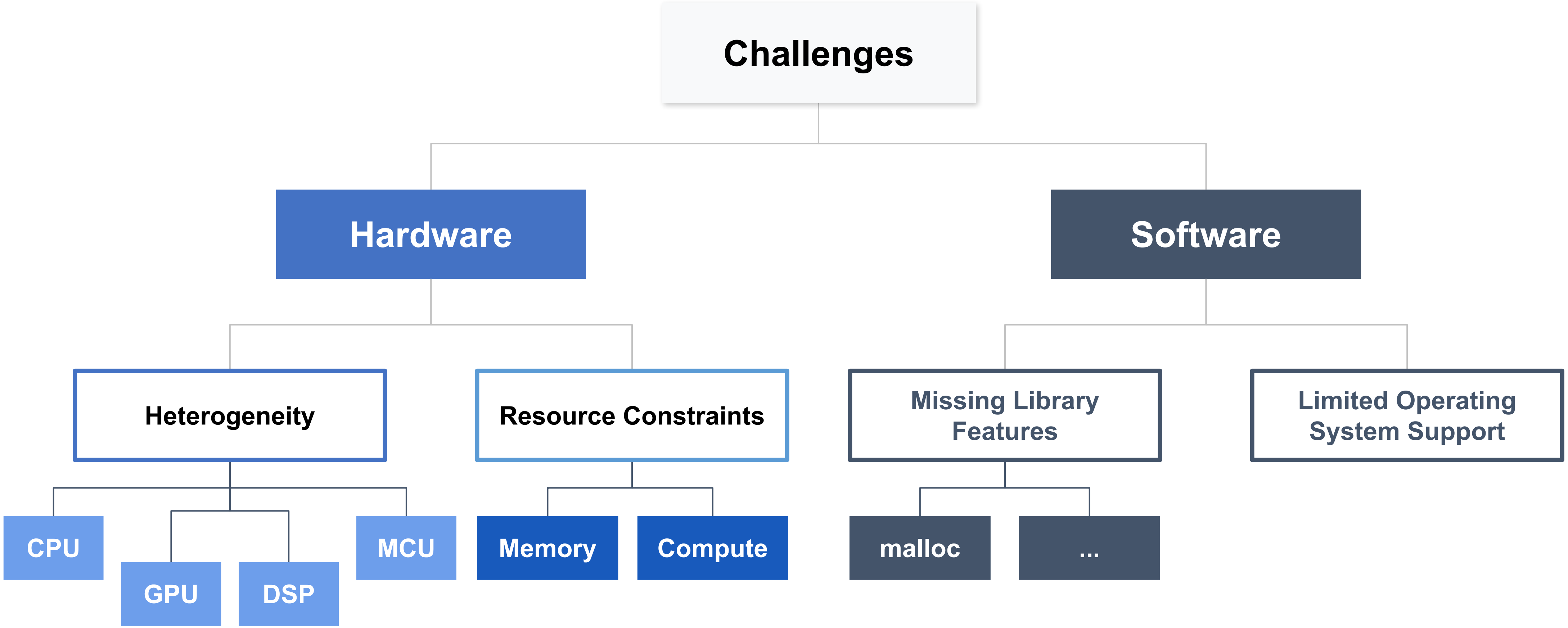






	Raspberry Pico (W)	Arduino Nano Sense	ESP 32	Seeed XIAO Sense / ESP32S3	Arduino Pro
32Bits CPU	Dual-core Arm Cortex-M0+	Arm Cortex-M4F	Xtensa LX6 Dual Core	Arm Cortex-M4F (BLE) Xtensa LX7 Dual Core	Dual Core Arm Cortex M7/M4
CLOCK	133MHz	64MHz	240MHz	64 / 240MHz	480/240MHz
RAM	264KB	256KB	520KB (part 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	$	$$$	$	$$	$$$$

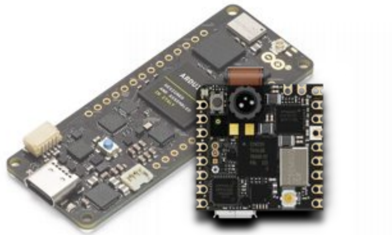
https://media.digikey.com/Resources/Maker/the-original-guide-to-boards-2022.pdf


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

Hardware



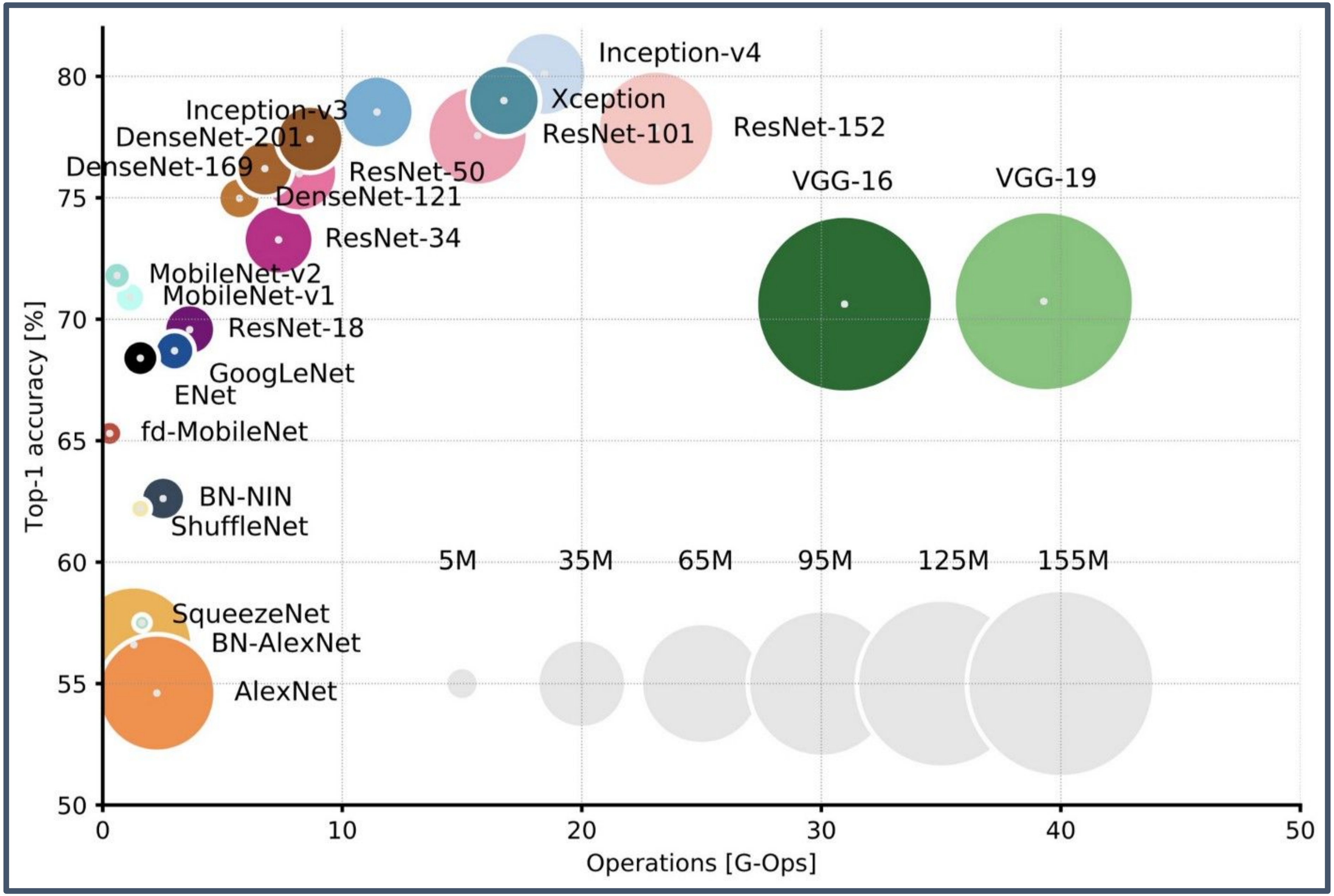


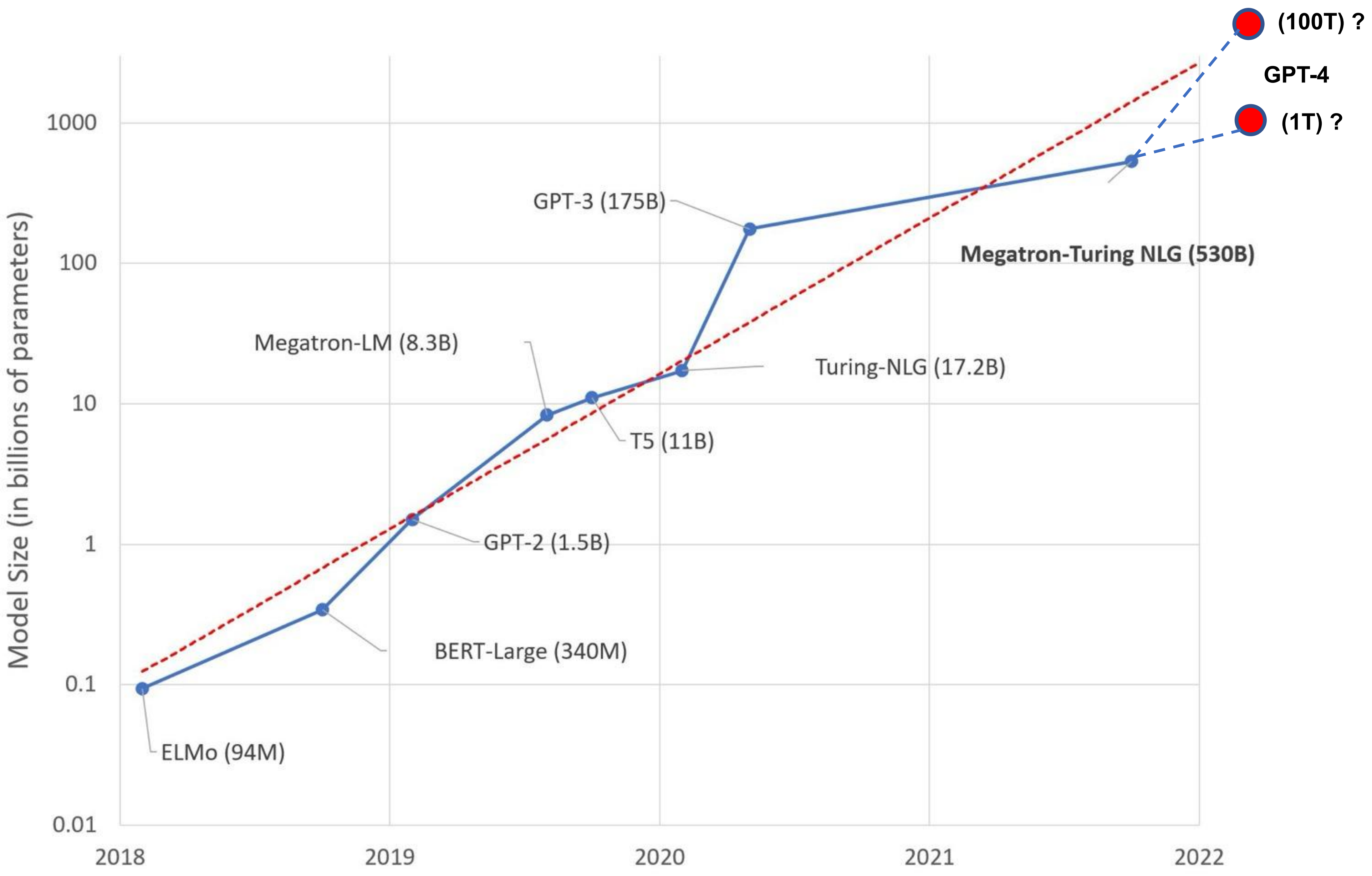


	Raspberry Pico (W)	Arduino Nano Sense	ESP 32	Seeed XIAO Sense / ESP32S3	Arduino Pro
32Bits CPU	Dual-core Arm Cortex-M0+	Arm Cortex-M4F	Xtensa LX6 Dual Core	Arm Cortex-M4F (BLE) Xtensa LX7 Dual Core	Dual Core Arm Cortex M7/M4
CLOCK	133MHz	64MHz	240MHz	64 / 240MHz	480/240MHz
RAM	264KB	256KB	520KB (part 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	$	$$$	$	$$	$$$$

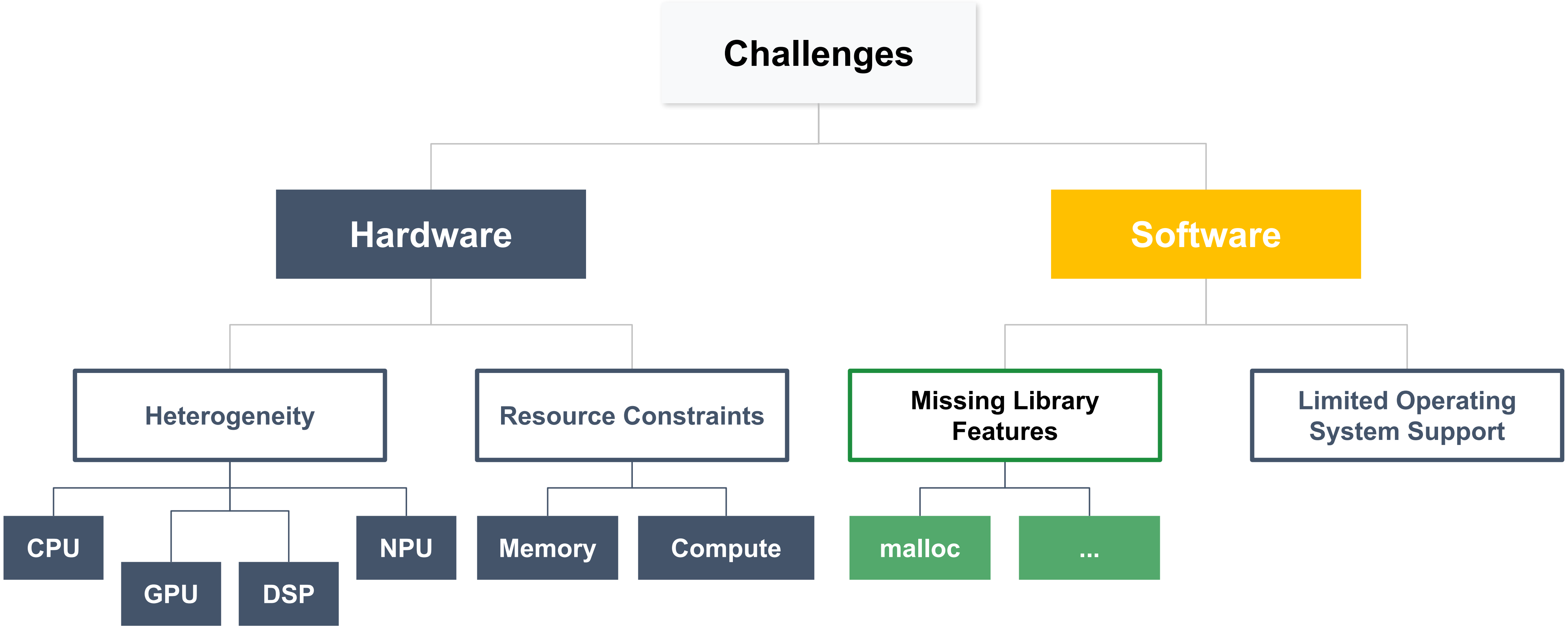
https://media.digikey.com/Resources/Maker/the-original-guide-to-boards-2022.pdf

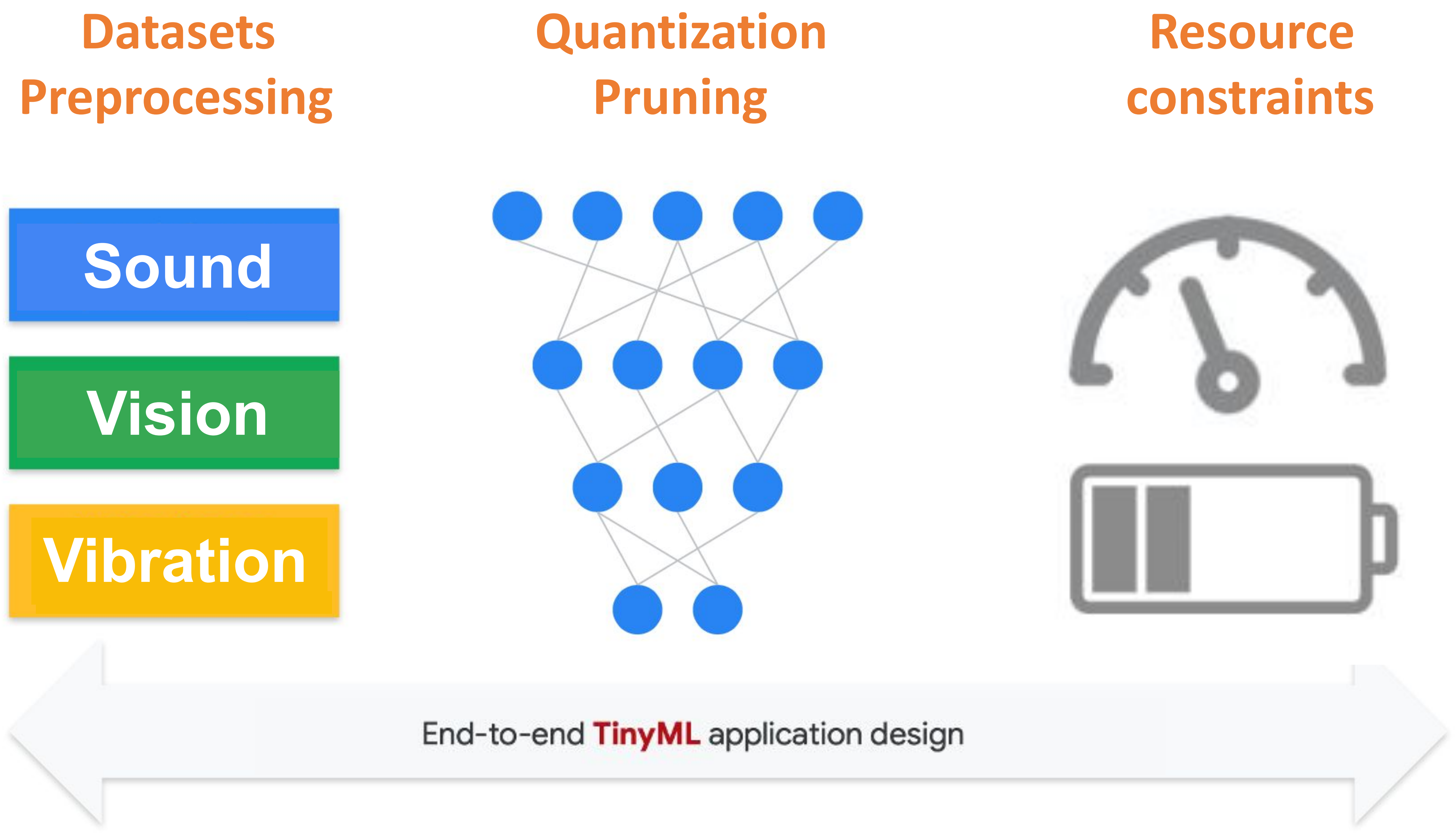
https://arxiv.org/pdf/1910.01108.pdf
https://towardsdatascience.com/neural-network-architectures-156e5bad51ba



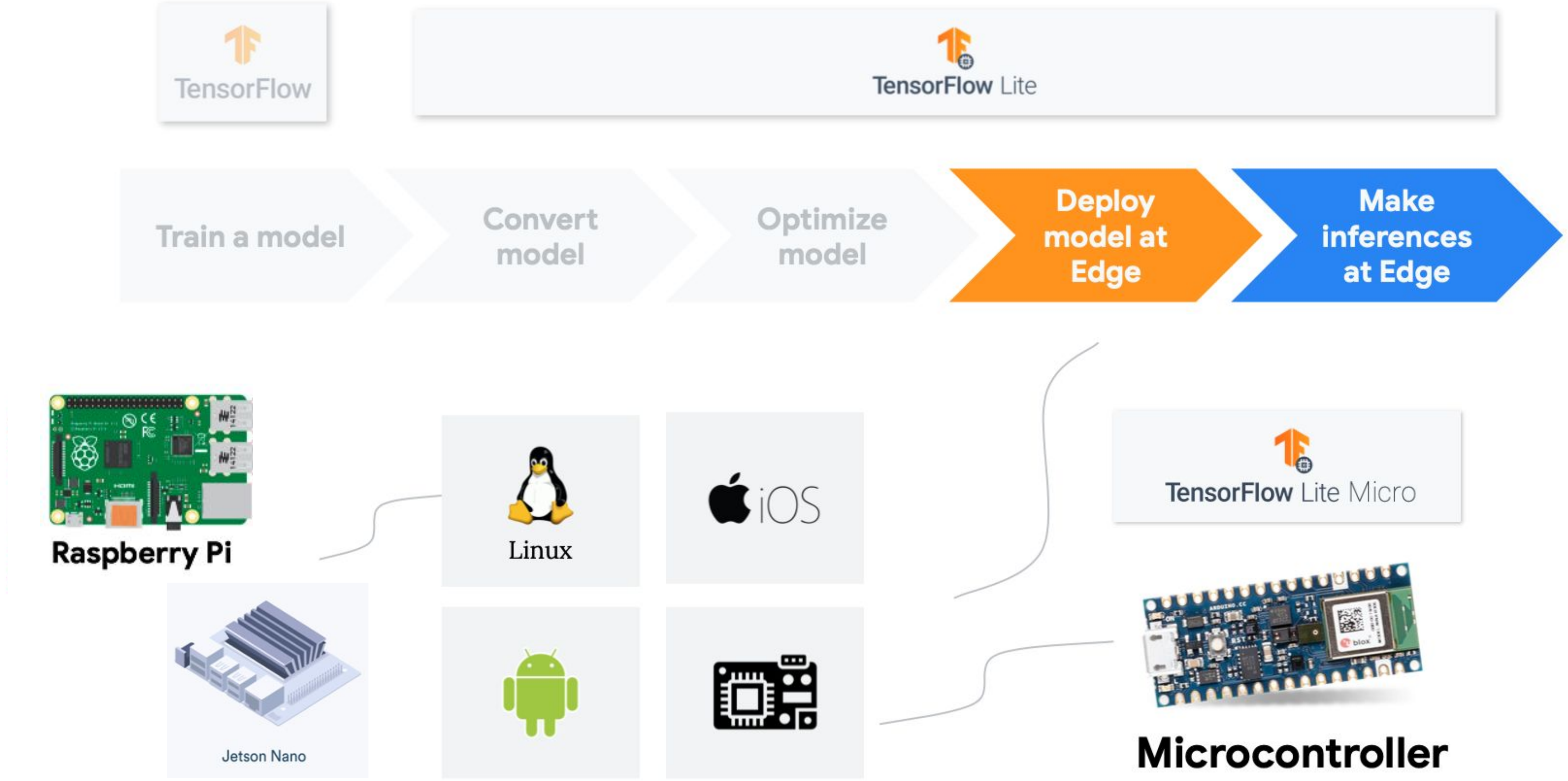


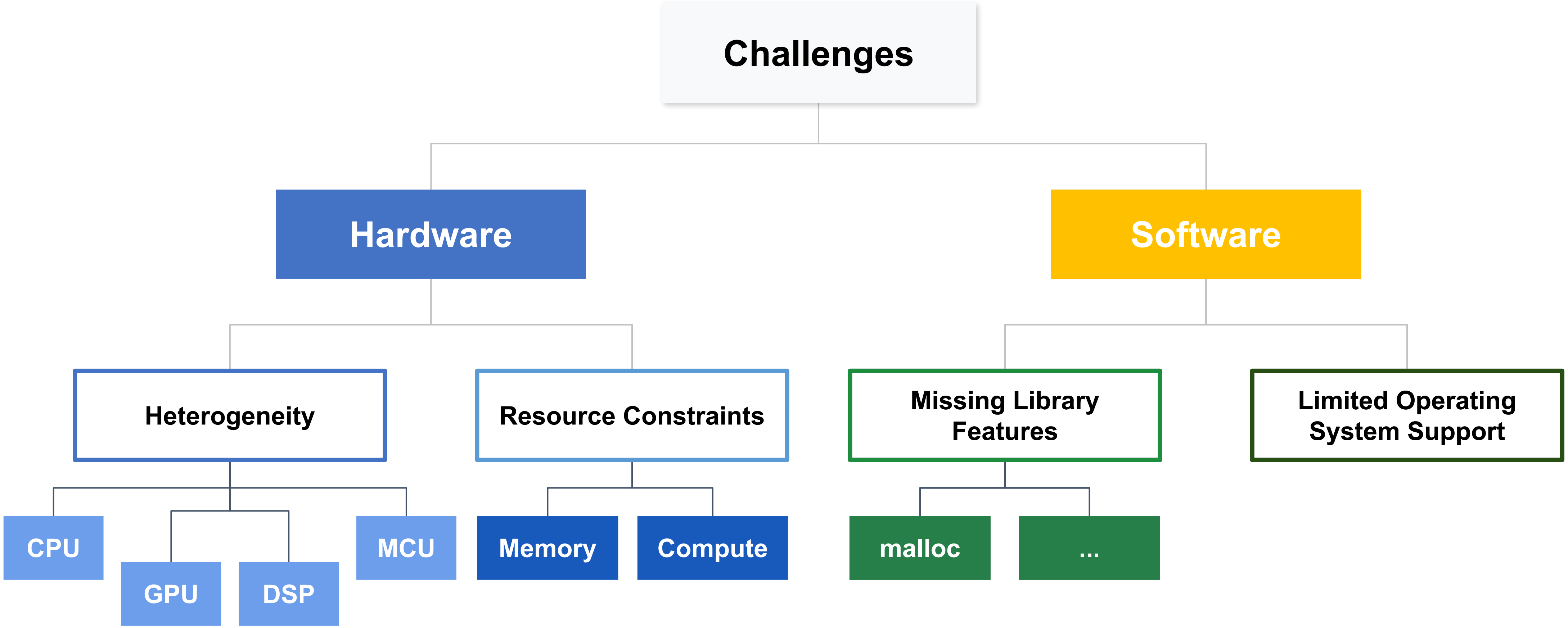
https://huggingface.co/blog/large-language-models https://arxiv.org/abs/2303.08774

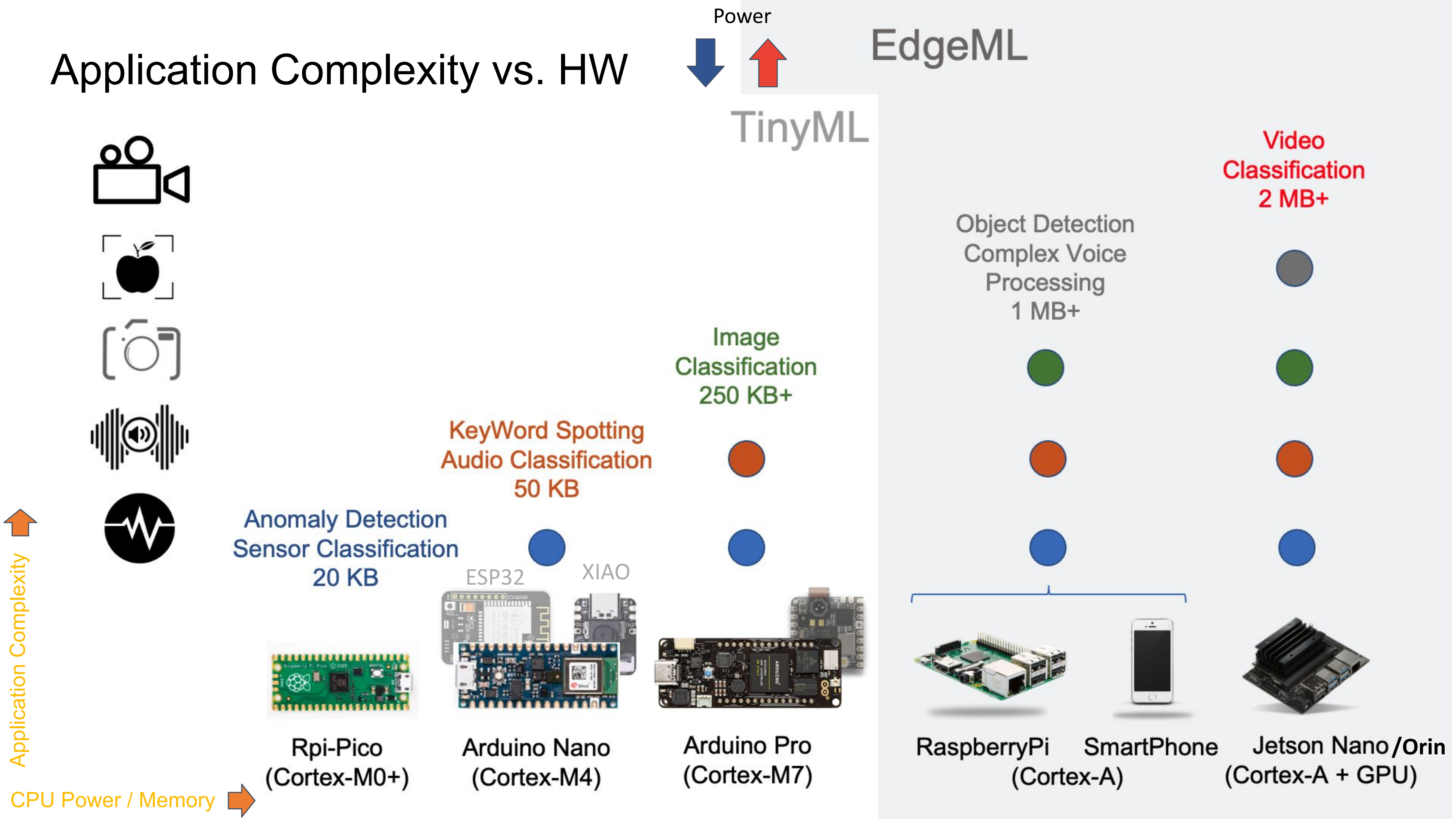

Challenges
Hardware
Software
Heterogeneity
Resource Constraints
Missing Library Features
Limited Operating System Support
CPU
GPU
DSP
NPU
Memory
Compute
malloc
...


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
...

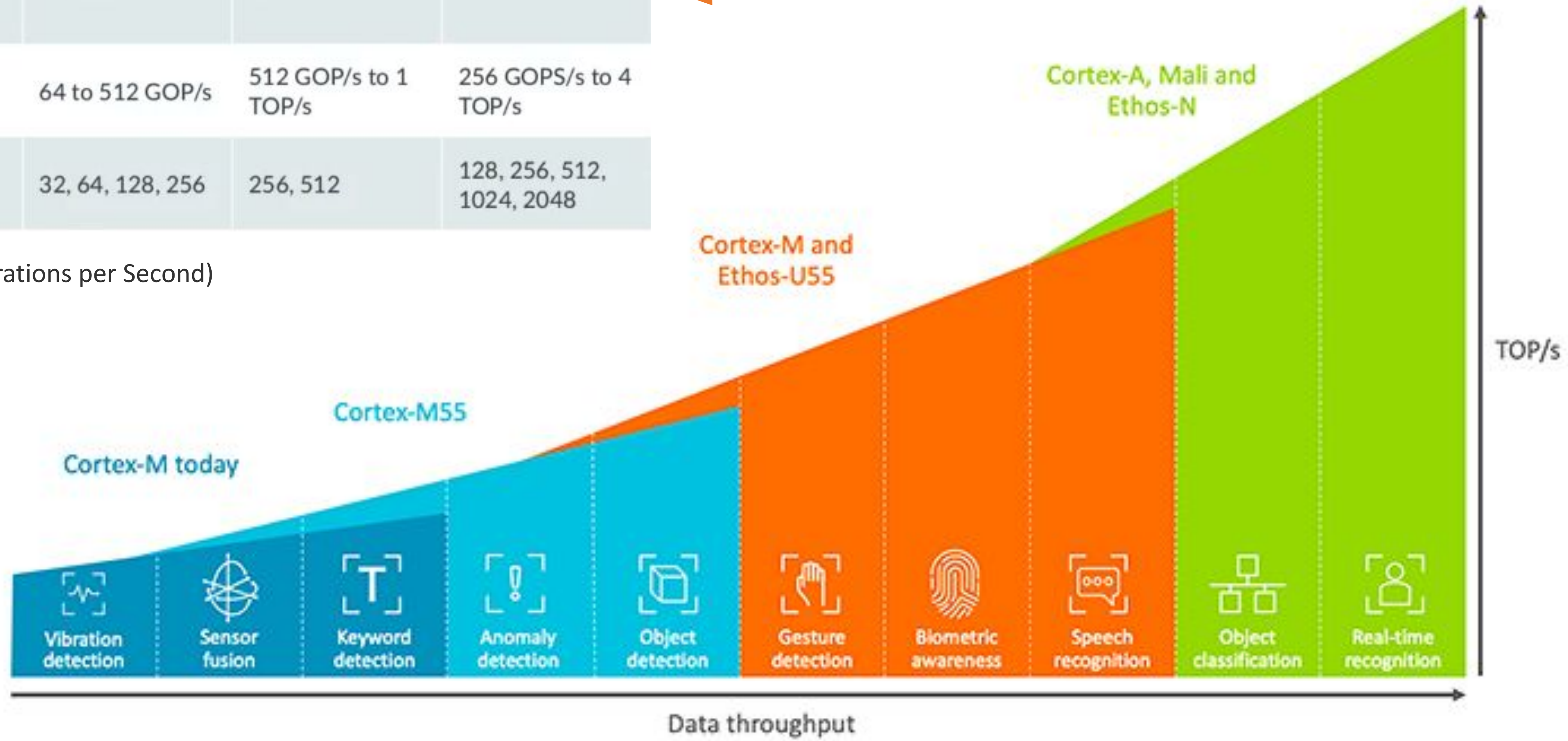


Application Complexity vs. HW
Power
EdgeML
TinyML
Video Classification 2 MB+
Object Detection Complex Voice Processing 1 MB+
Image Classification 250 KB+
KeyWord Spotting Audio Classification 50 KB
Anomaly Detection Sensor Classification 20 KB
ESP32
XIAO
Rpi-Pico (Cortex-M0+)
Arduino Nano (Cortex-M4)
Arduino Pro (Cortex-M7)
RaspberryPi
SmartPhone
(Cortex-A)
Jetson Nano /Orin (Cortex-A + GPU)
Application Complexity
CPU Power / Memory

ML- optimized Solutions (w/microNPUs)

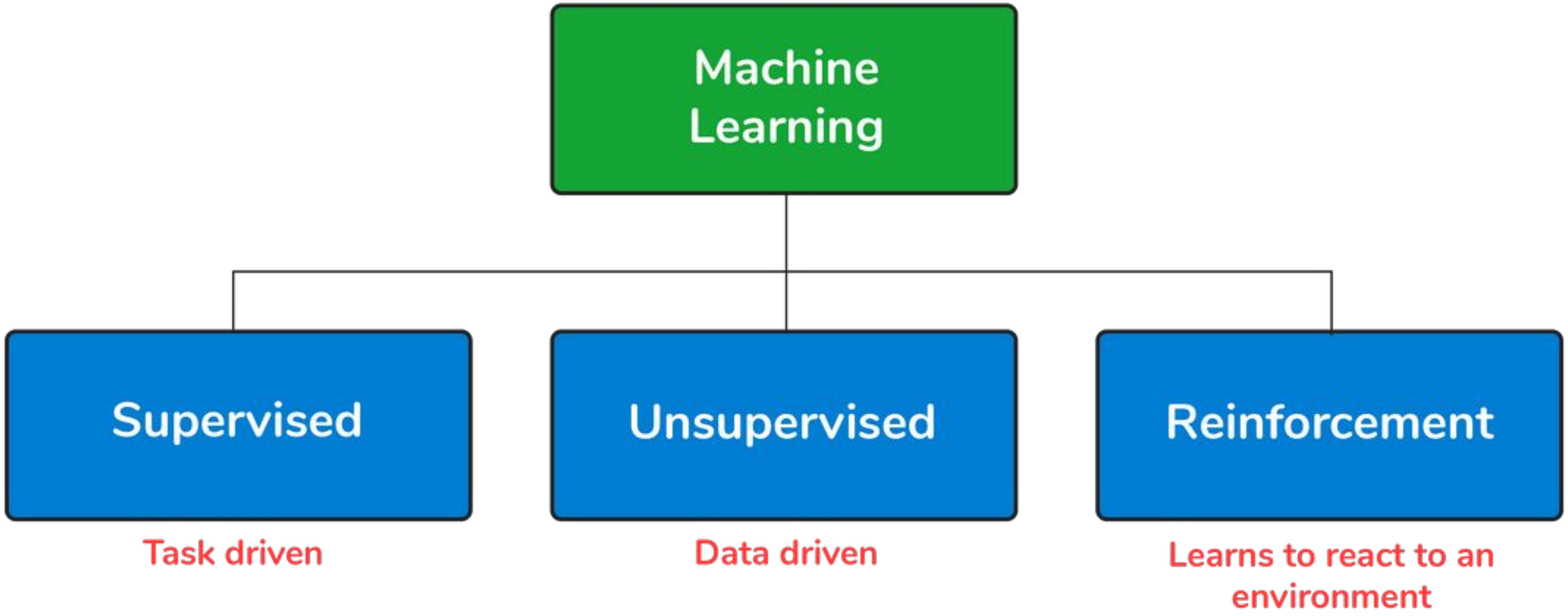
	Ethos-U55	Ethos-U65	Ethos-U85
Performance (At 1 GHz)	64 to 512 GOP/s	512 GOP/s to 1 TOP/s	256 GOPS/s to 4 TOP/s
MACs (8x8)	32, 64, 128, 256	256, 512	128, 256, 512, 1024, 2048

TOPS (Tera Operations per Second)



TinyML Application

Examples



- Regression
- Classification
- Object Detection
- Time Series Forecasting

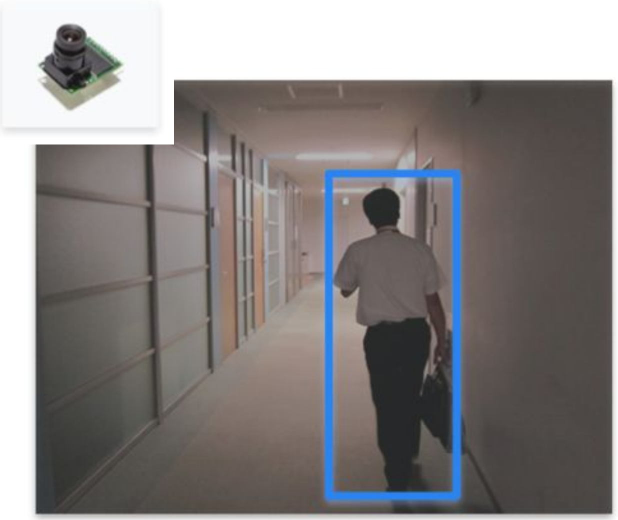
- Anomaly Detection
- Clustering
- Dimensionality Reduction

- Autonomous Navigation

Sound

Vibration

Vision



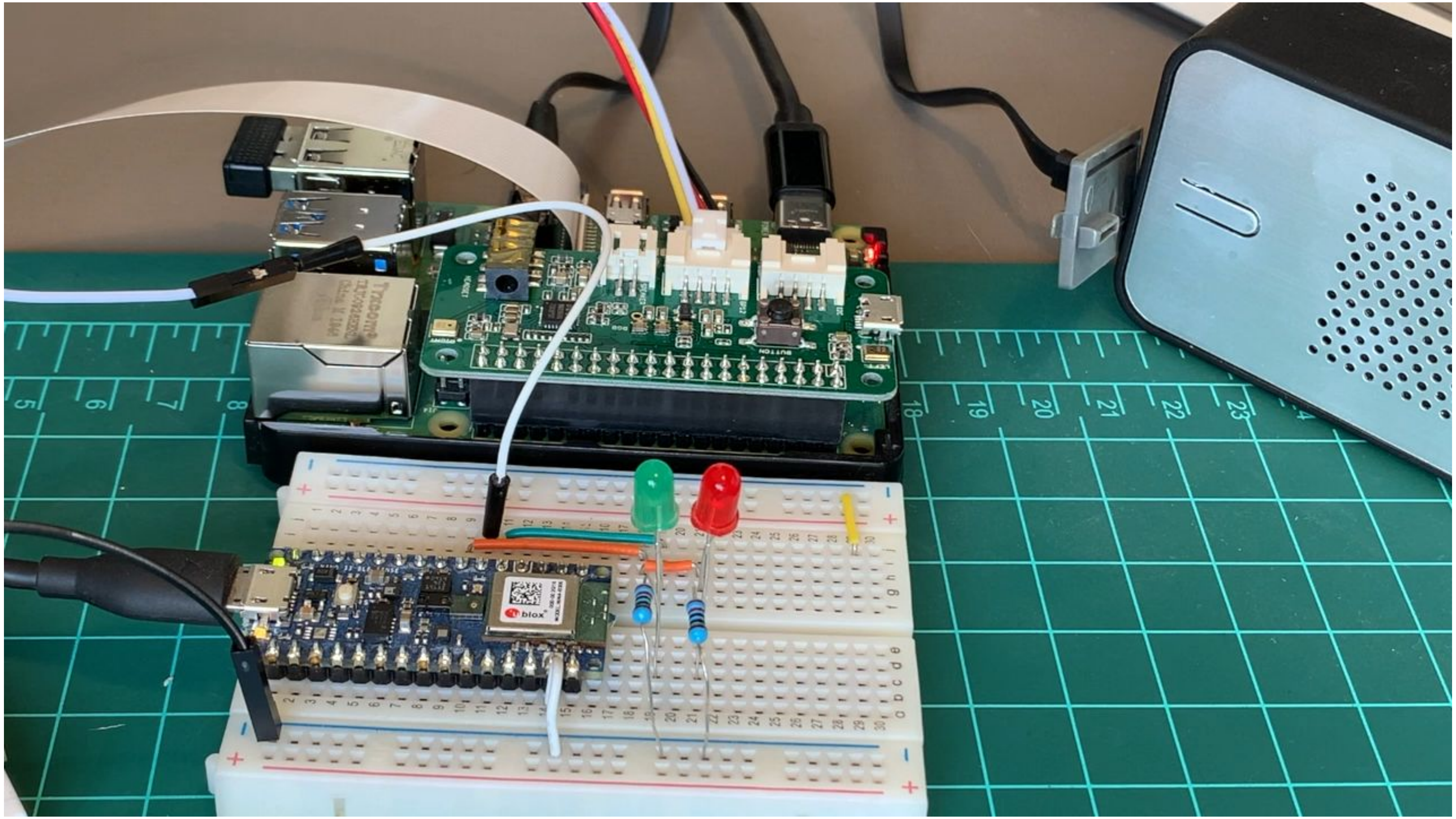
Sound

Vibration

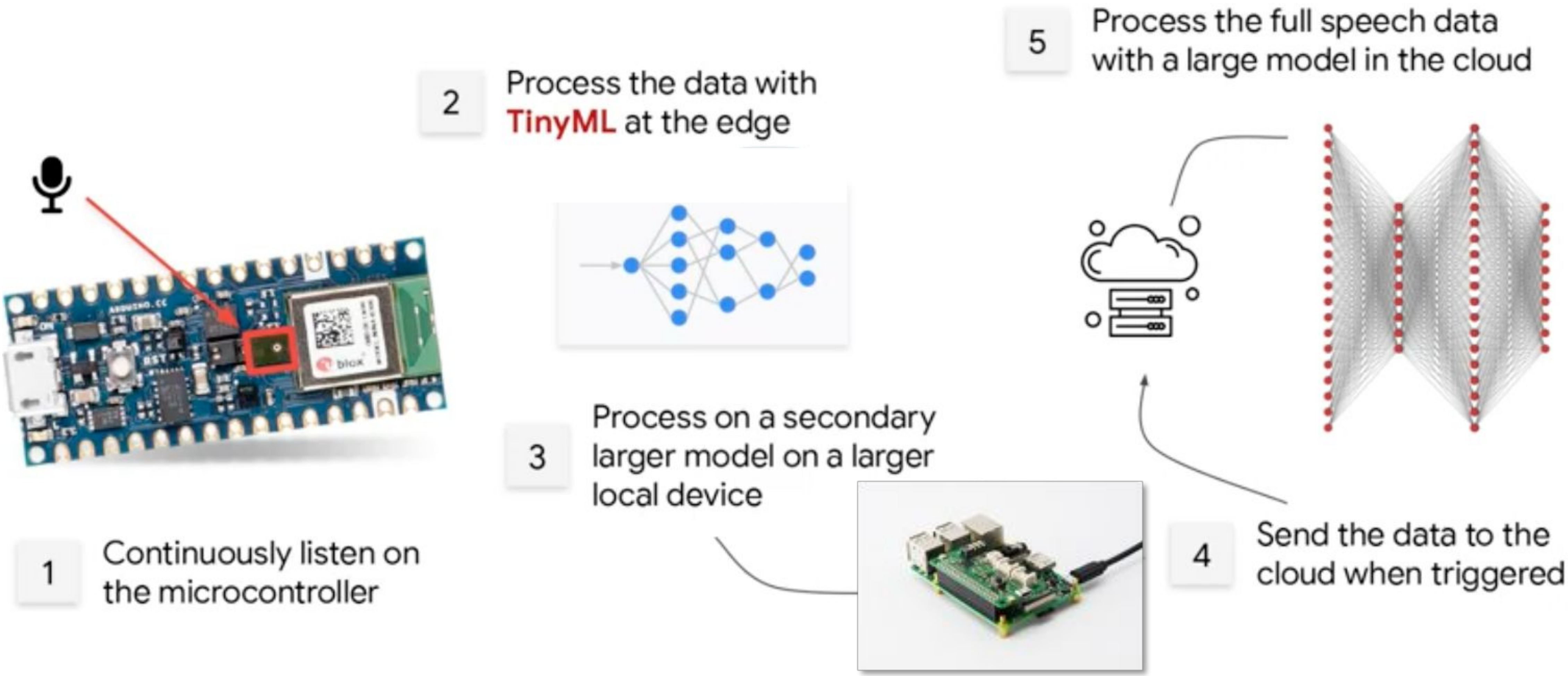
Vision

Personal Assistant

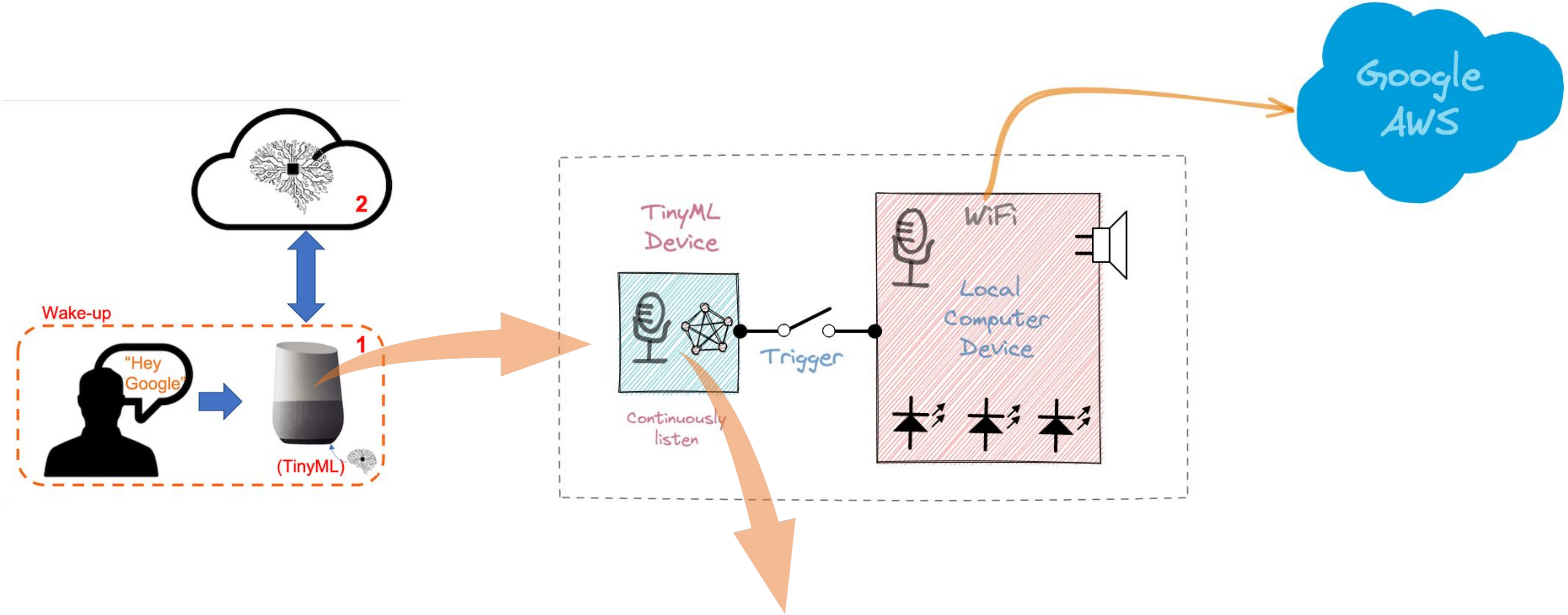
Personal Assistant



"Cascade" Detection: multi-stage model

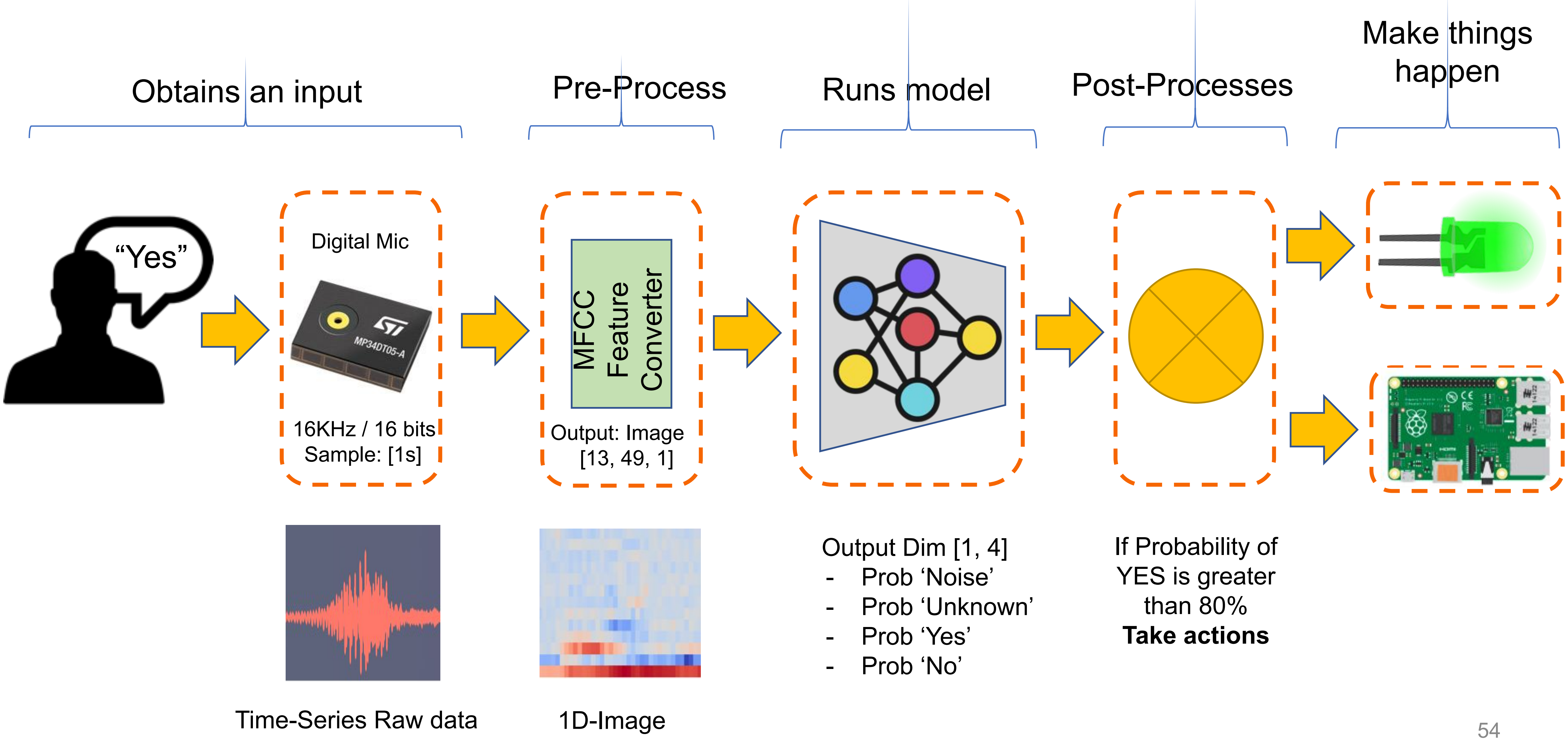


Personal Assistant



KWS

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

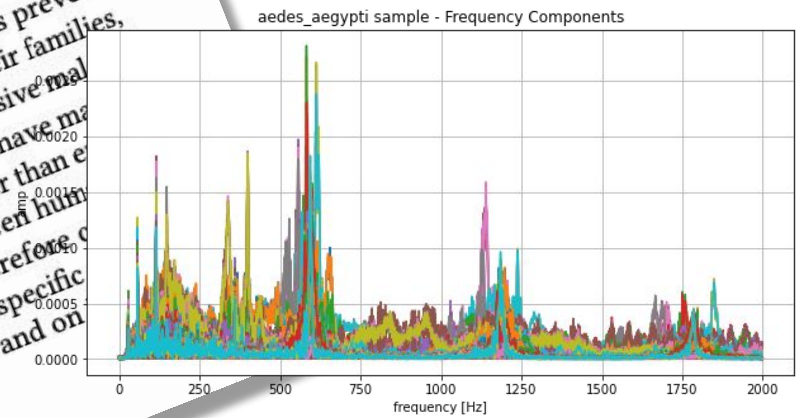
ABSTRACT

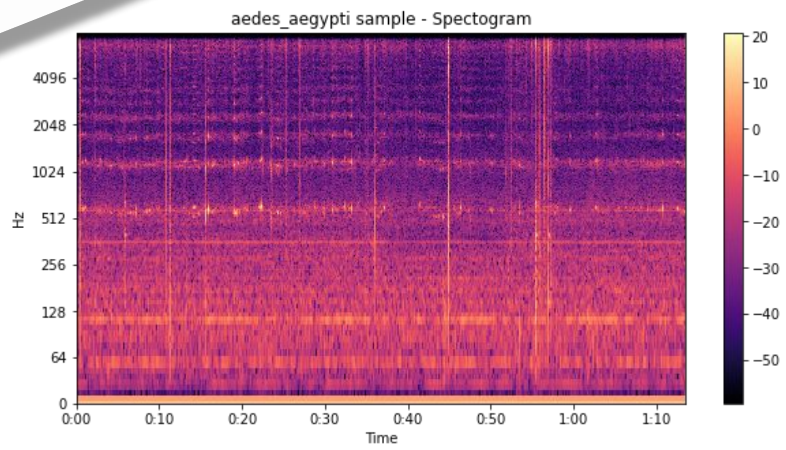
Every year more than one billion people are infected and more than one million people die from vector-borne diseases including malaria, dengue, zika and chikungunya. Mosquitoes are the best known disease vector and are geographically spread worldwide. It is important to raise awareness of mosquito proliferation 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

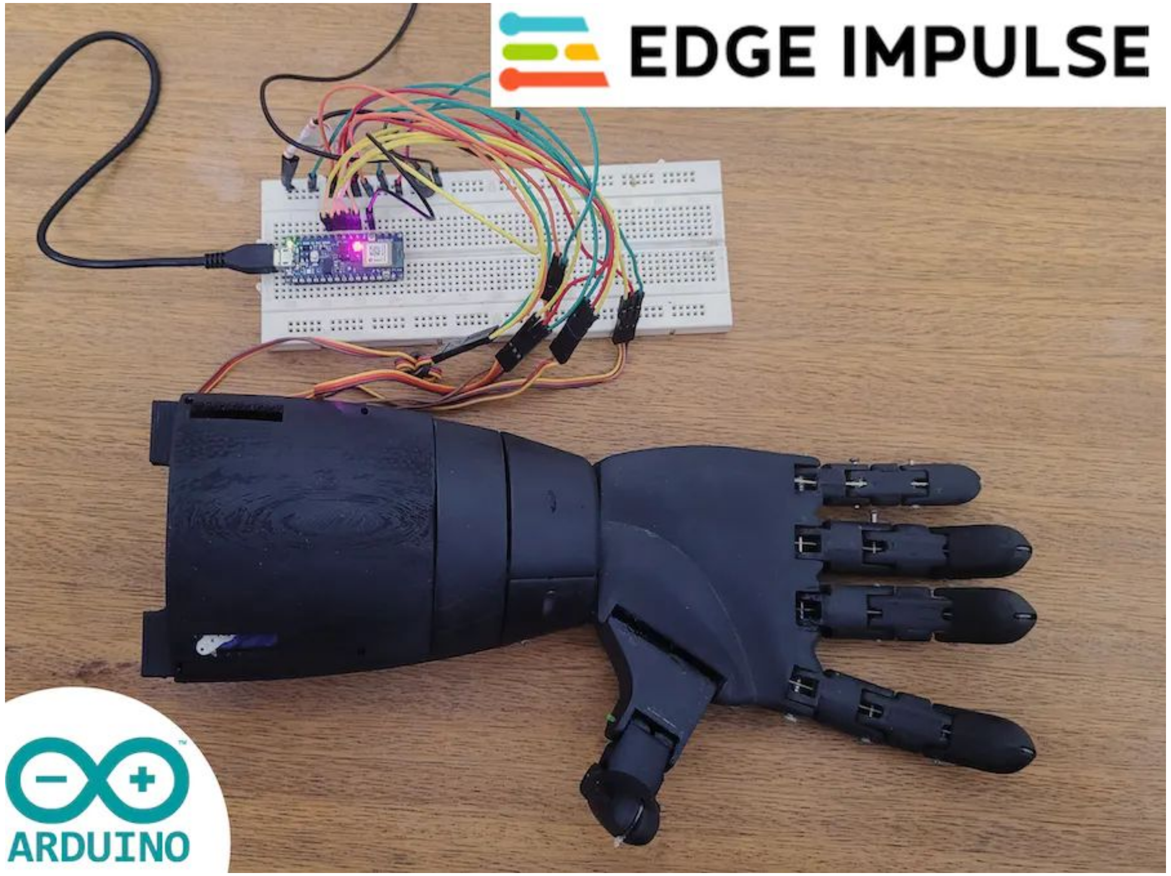
Countries affected by malaria turn to control rather than e tion. Vector control means decreasing contact between hum disease carriers on an area-by-area basis. It is therefore c be able to detect the presence of mosquitoes in a specific paper presents an approach based on TinyML and on embedded devices.





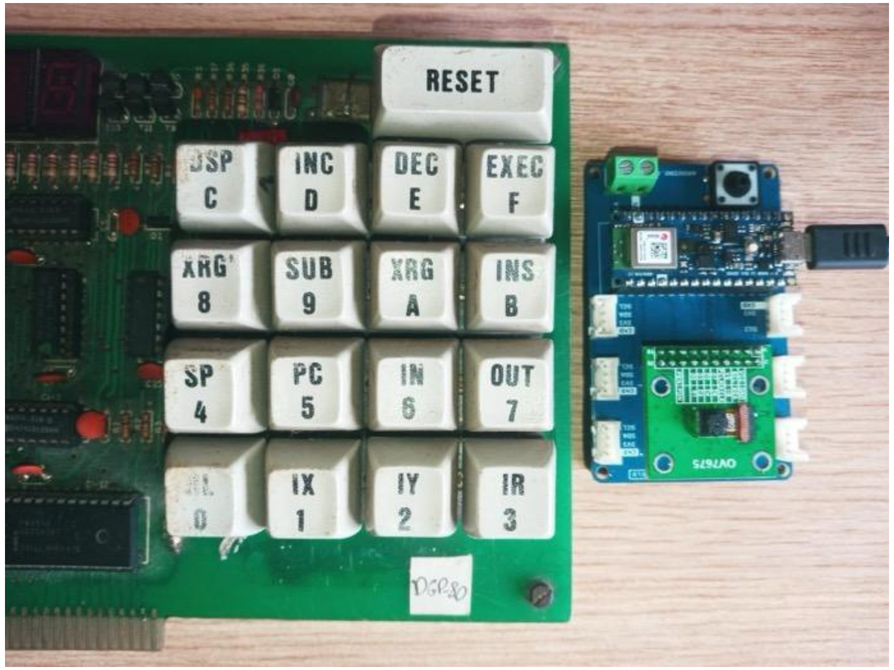
https://www.edgeimpulse.com/blog/the-world-is-all-abuzz-about-tinyml

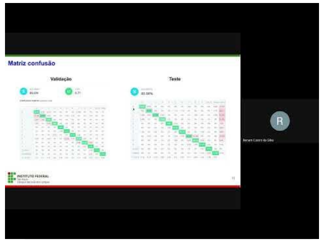
Bionic Hand Voice Commands Module



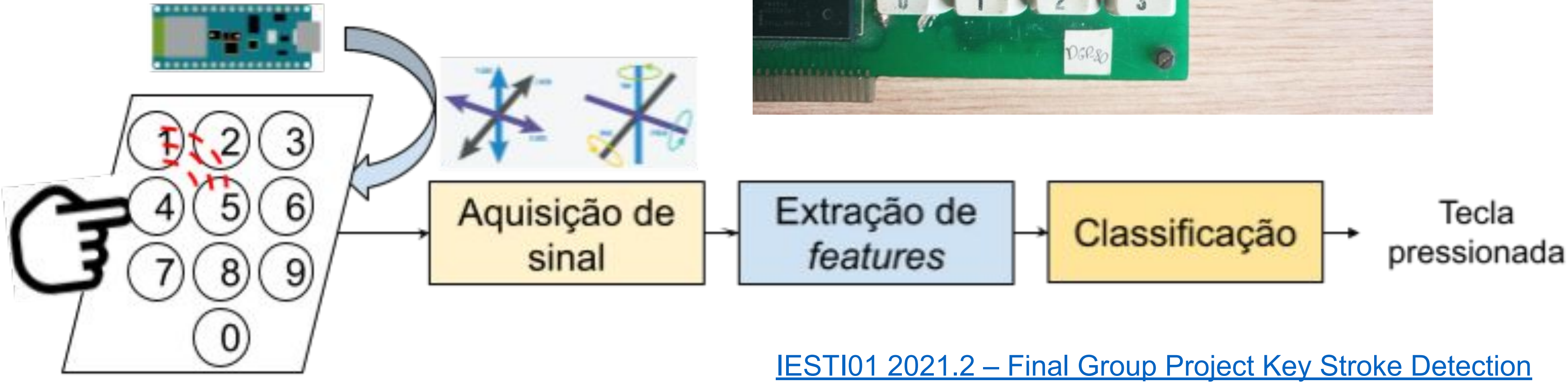
https://www.hackster.io/ex-machina/bionic-hand-voice-commands-module-w-edge-impulse-arduino-aa97e3

Keystroke Sound Detection





Renam Castro
Professor IFESP



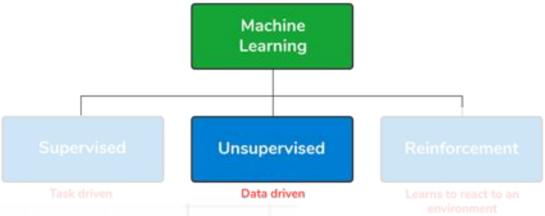
[IESTI01 2021.2 – Final Group Project Key Stroke Detection]

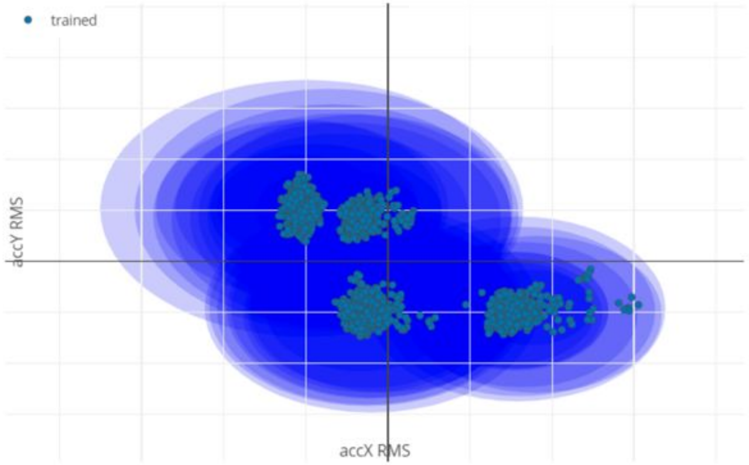
Sound

Vibration

Vision

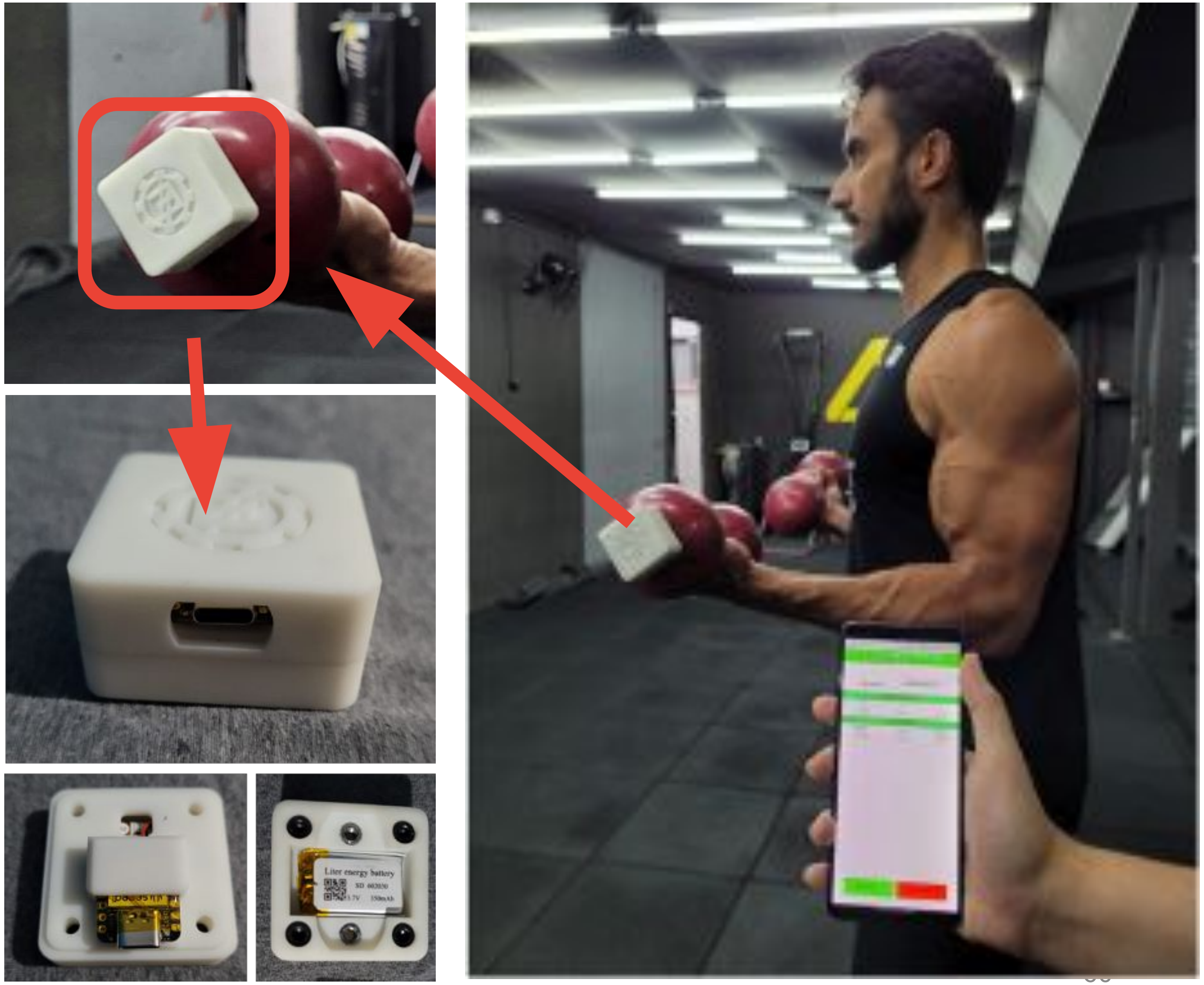
Industrial – Anomaly Detection





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

Movement Classification

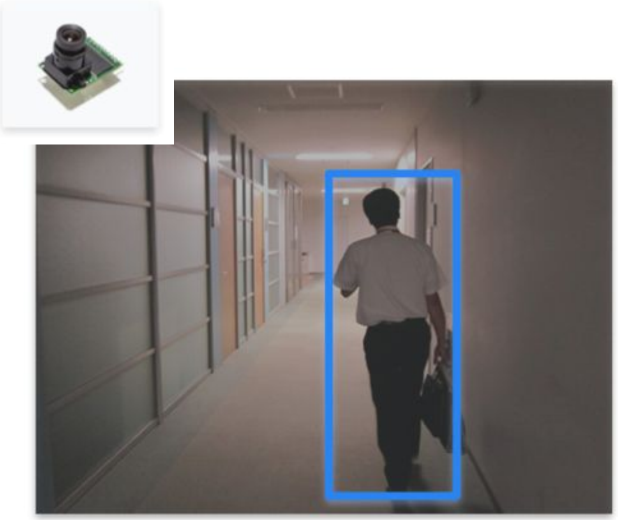


CNPq UNIFEI

Sound

Vibration

Vision



Computer Vision Main Types

Image Classification

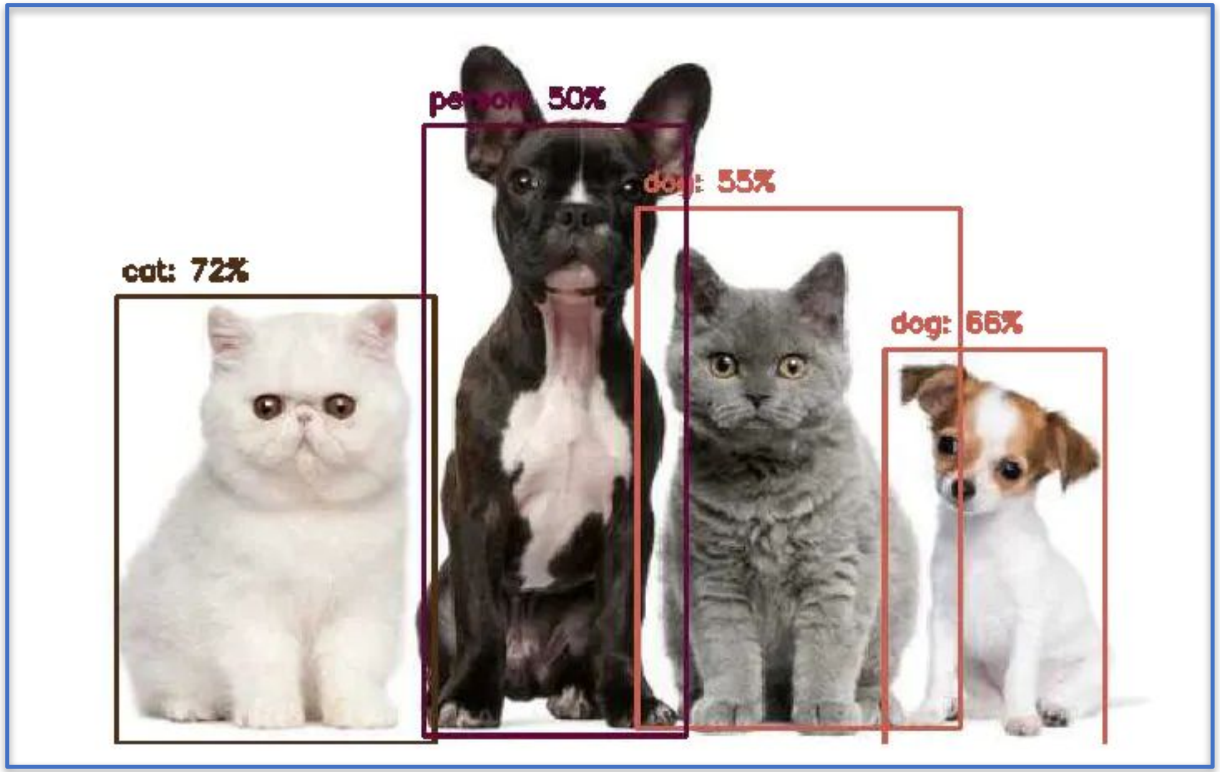
(**Multi-Class** Classification)

Cat: 70%

Dog: 80%

Object Detection

Multi-Label Classification + Object **Localization**



Computer Vision Main Types

Image Classification

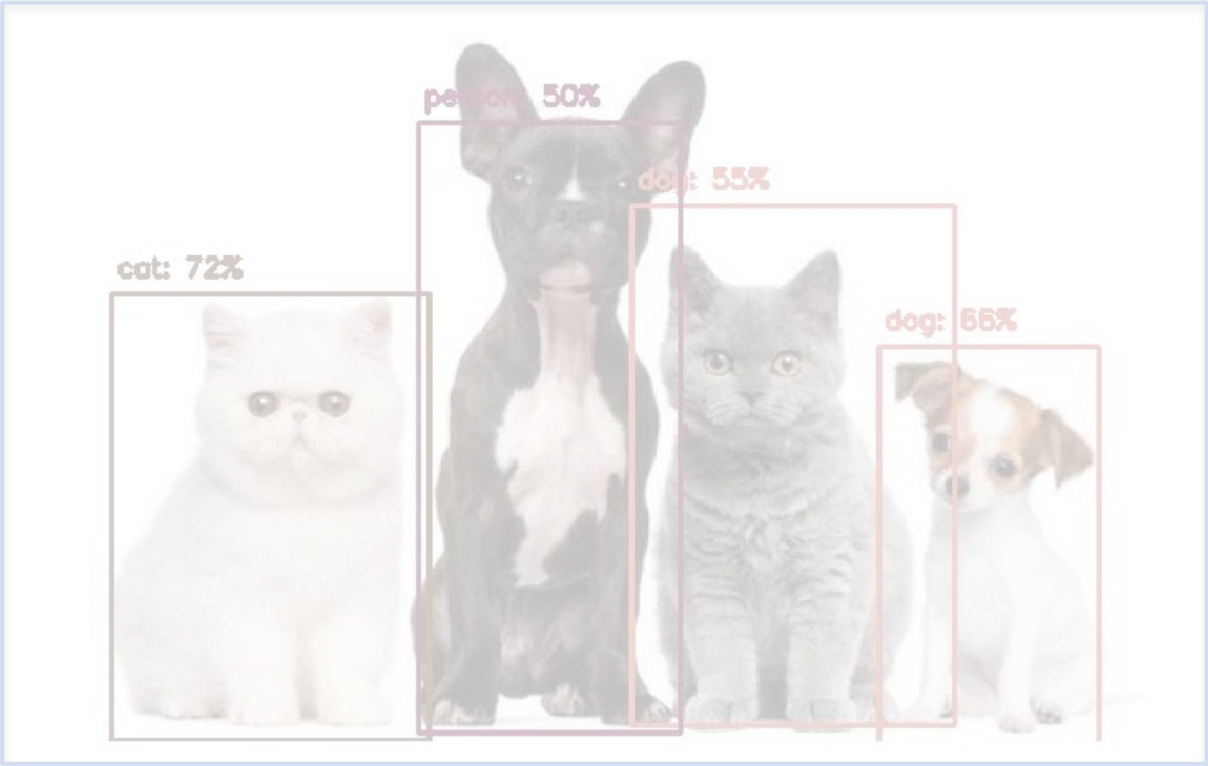
(**Multi-Class** Classification)

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

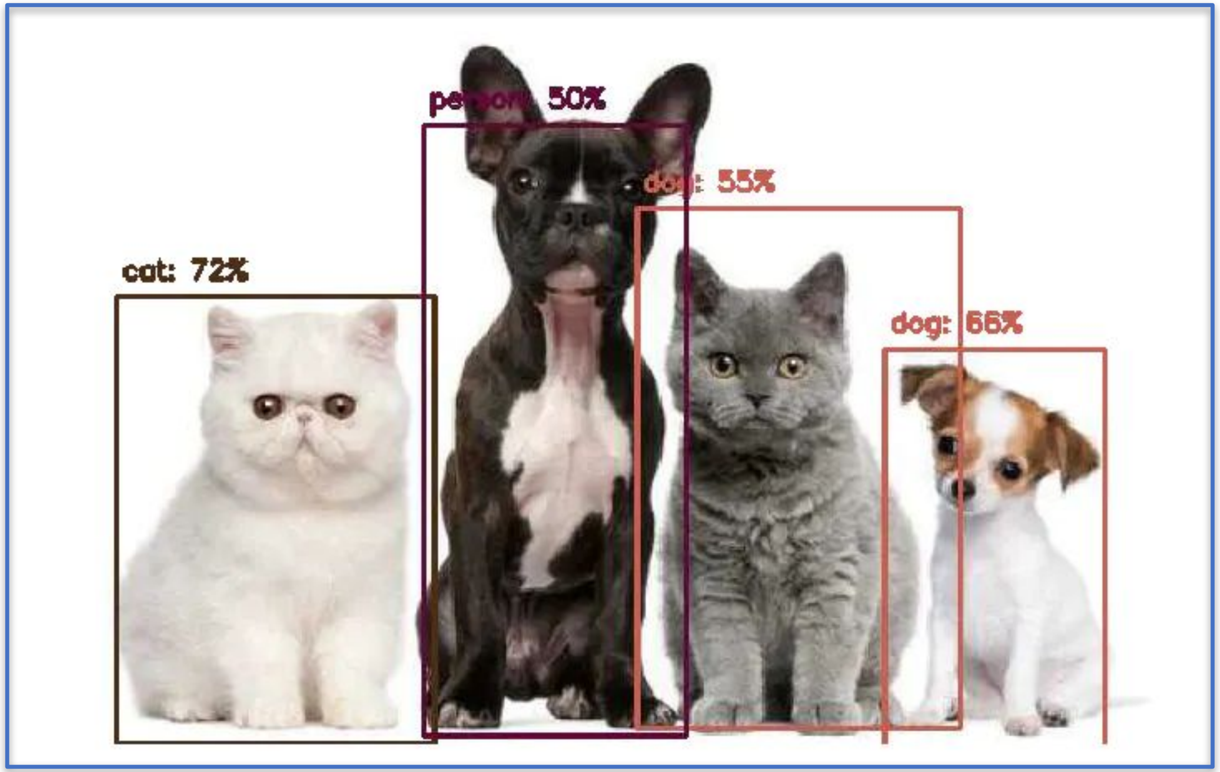
(**Multi-Class** Classification)

Cat: 70%

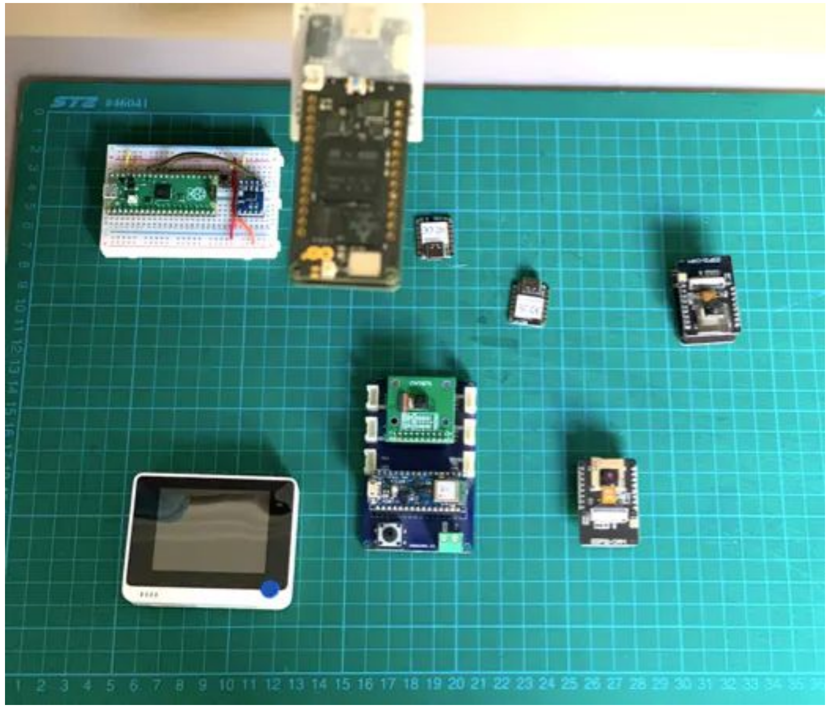
Dog: 80%

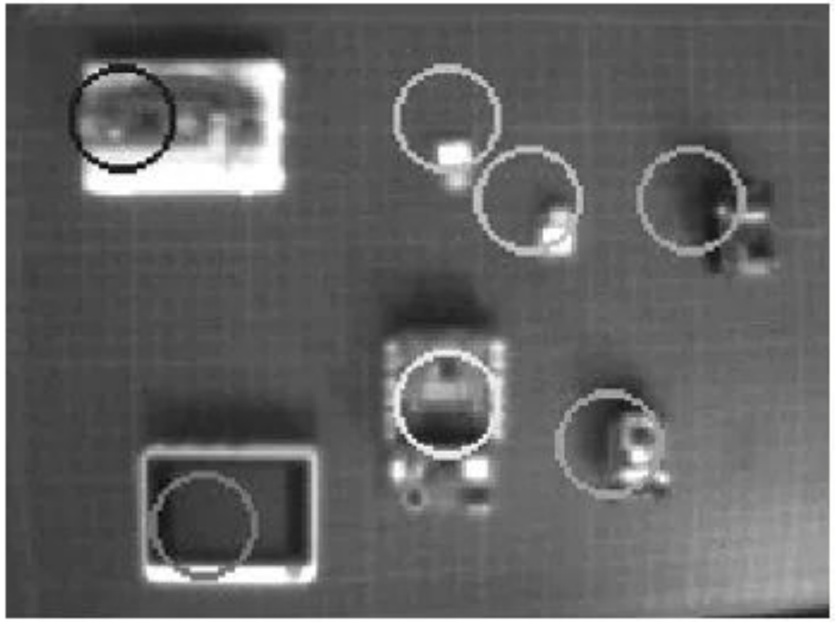
Object Detection

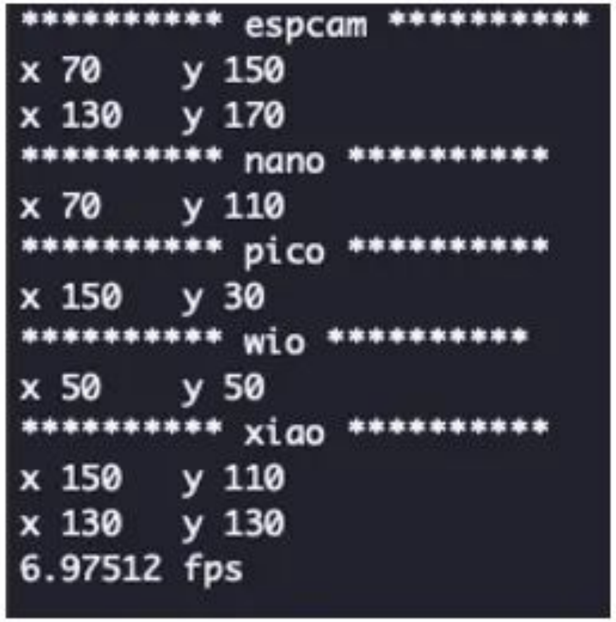
Multi-Label Classification + Object **Localization**



Detecting Objects using TinyML (FOMO)

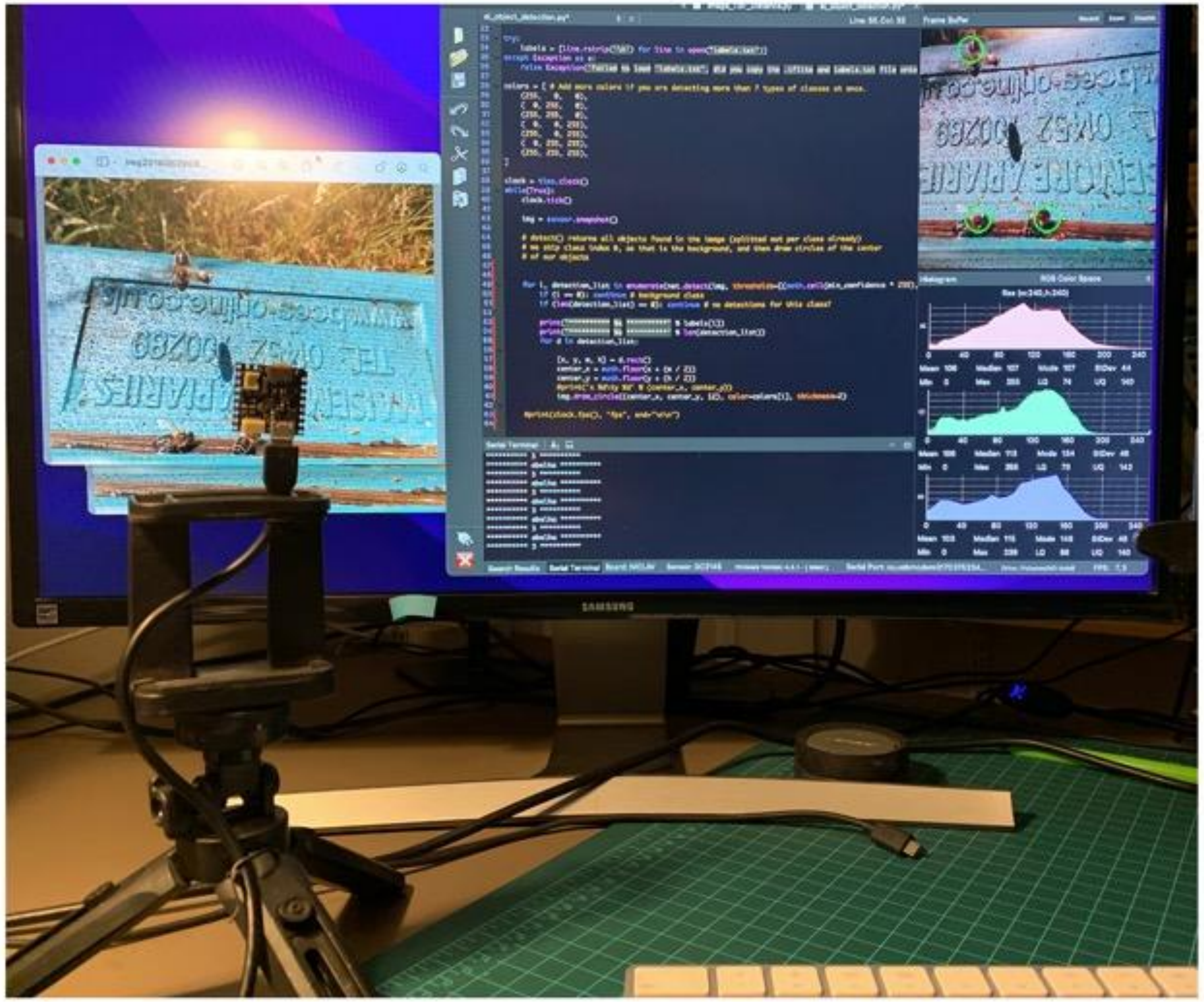






[EdgeAI made simple - Exploring Image Processing (Object Detection) on microcontrollers with Arduino Portenta, Edge Impulse FOMO, and OpenMV]

Detecting Objects using TinyML (FOMO)

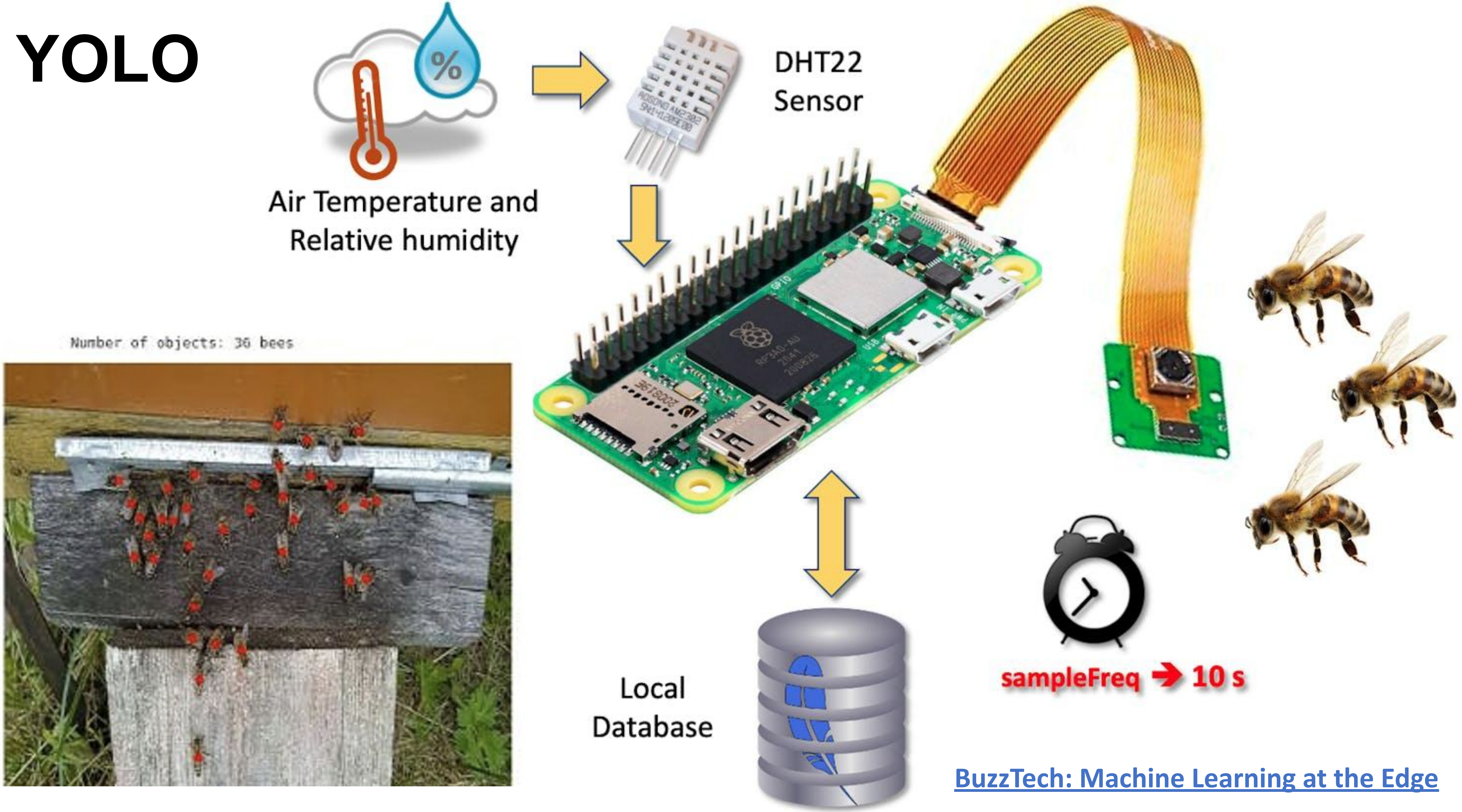


OpenMV

MicroPython

EDGE IMPULSE

YOLO



[BuzzTech: Machine Learning at the Edge]()

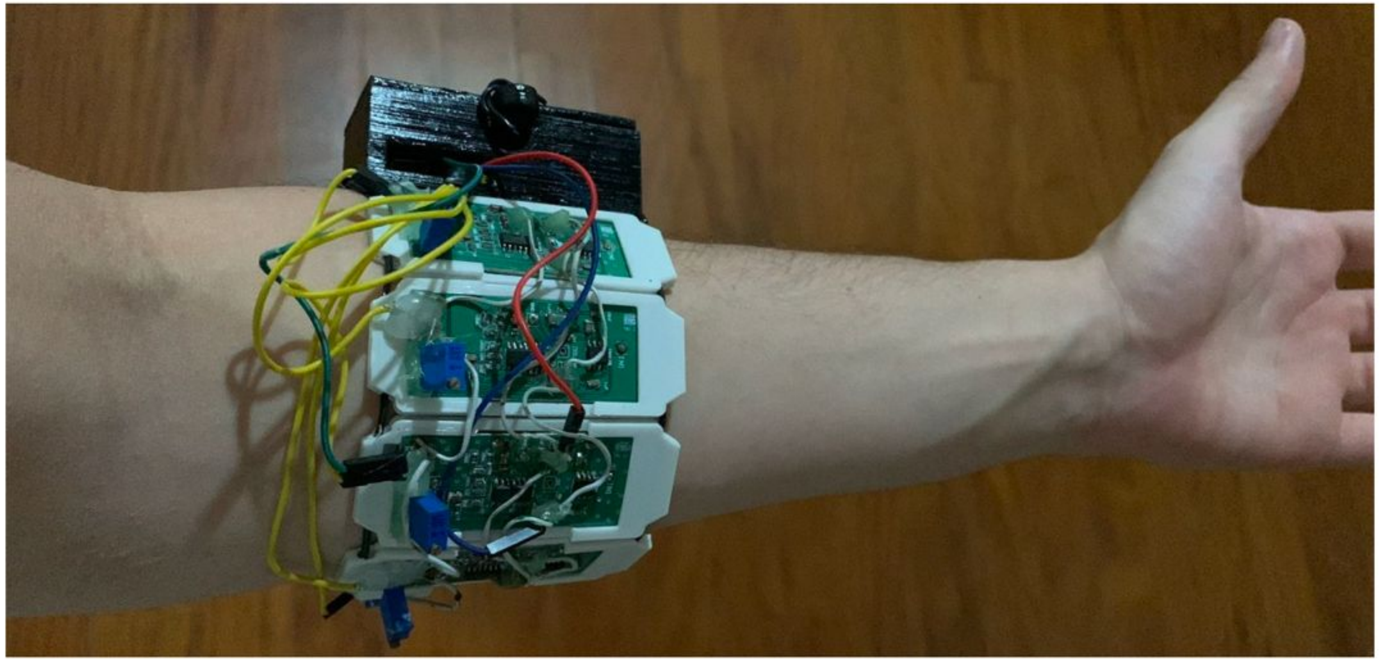
YOLO

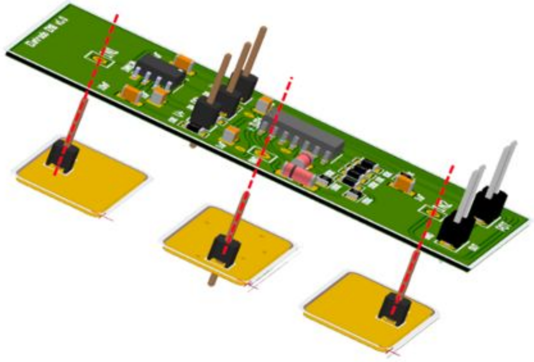
Ant Detection

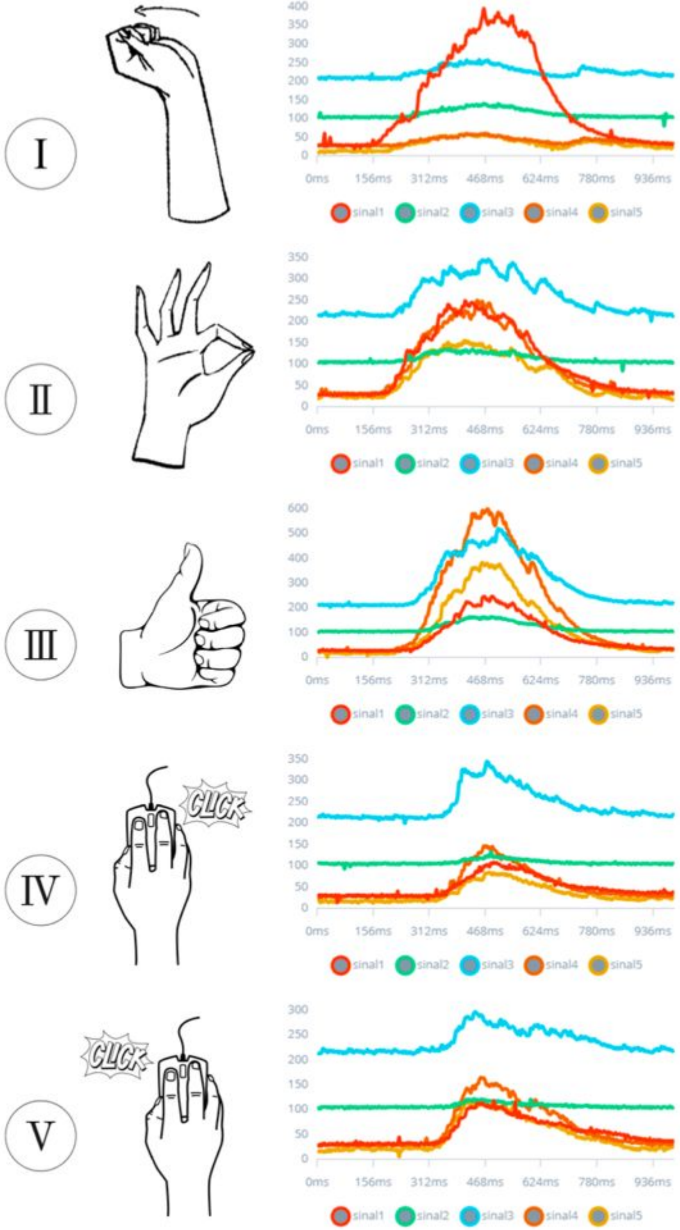
Other Sensors / MCUs / Models

Examples

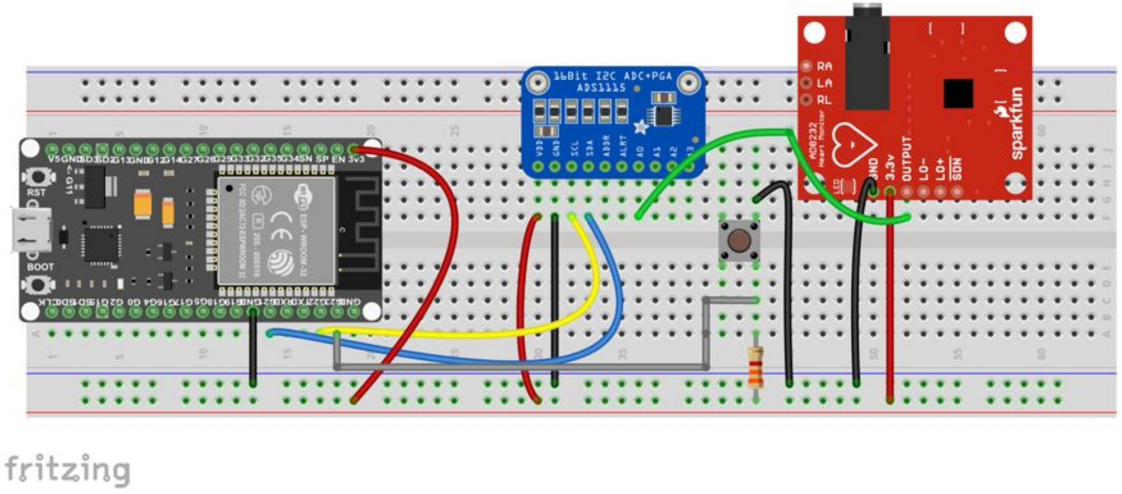
Surface electromyography

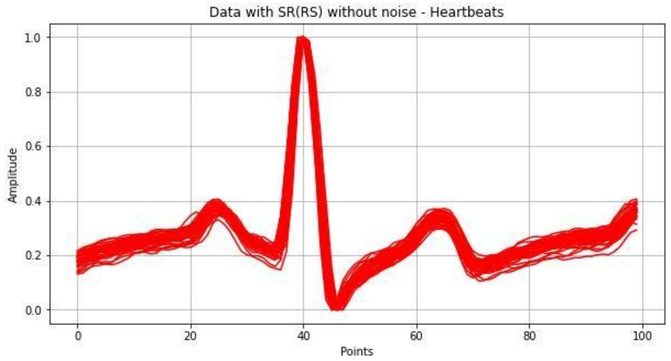






AD8232 - Single Lead Heart Rate Monitor

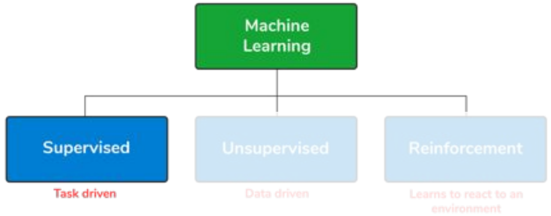


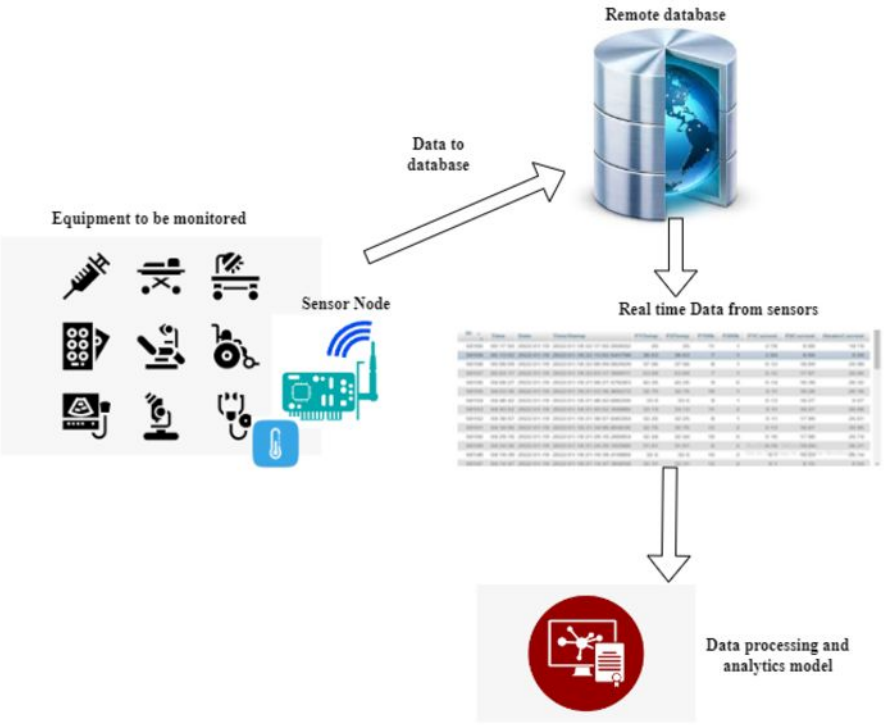


Guilherme Silva
Engenheiro - UNIFEI

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

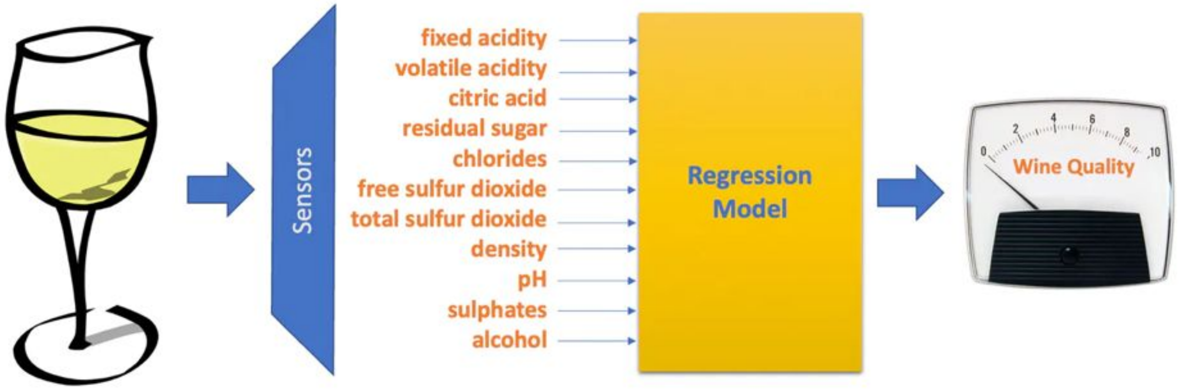
Regression on TinyML





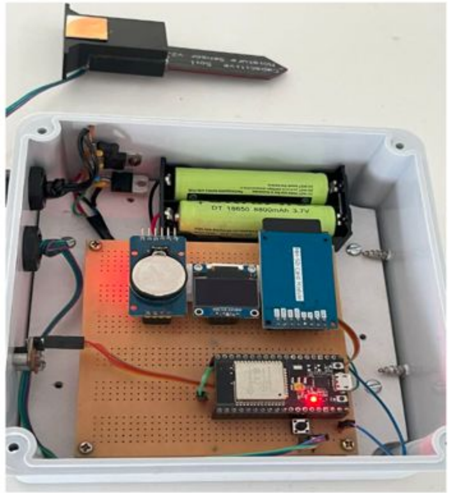
[On-Device IoT-Based Predictive Maintenance Analytics Model: Comparing TinyLSTM and TinyModel from Edge Impulse]

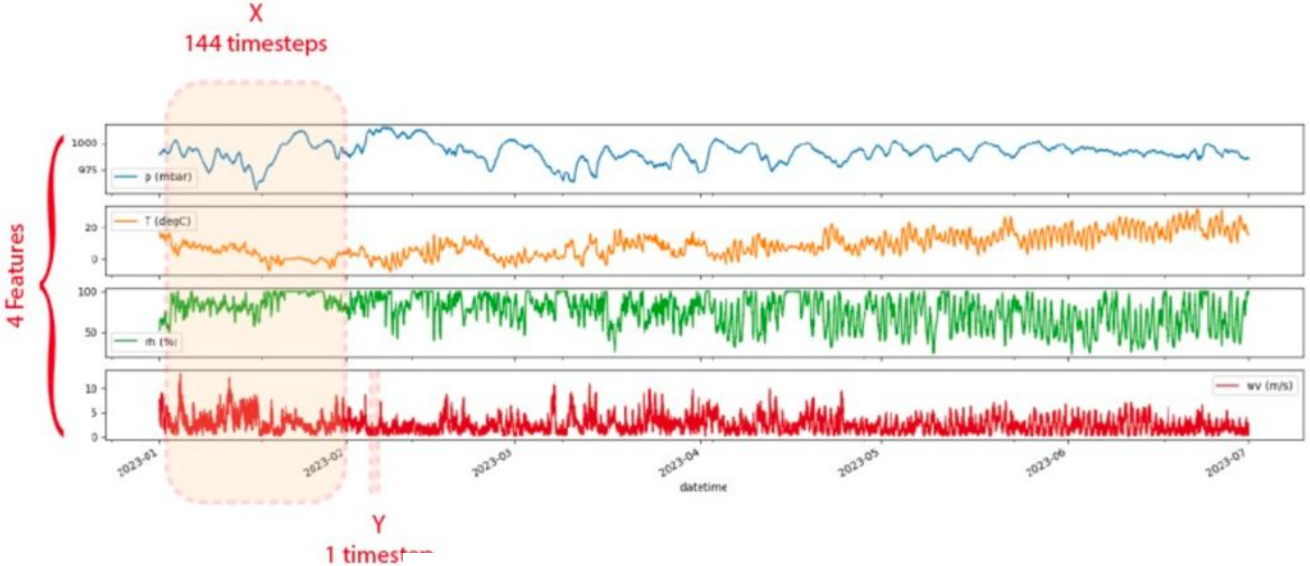
Sensor fusion



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

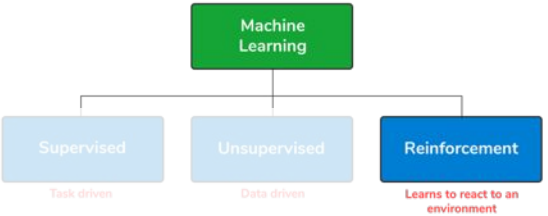
LSTM





<u>ESP32 LSTM Phenolic Sponge Moisture</u>

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

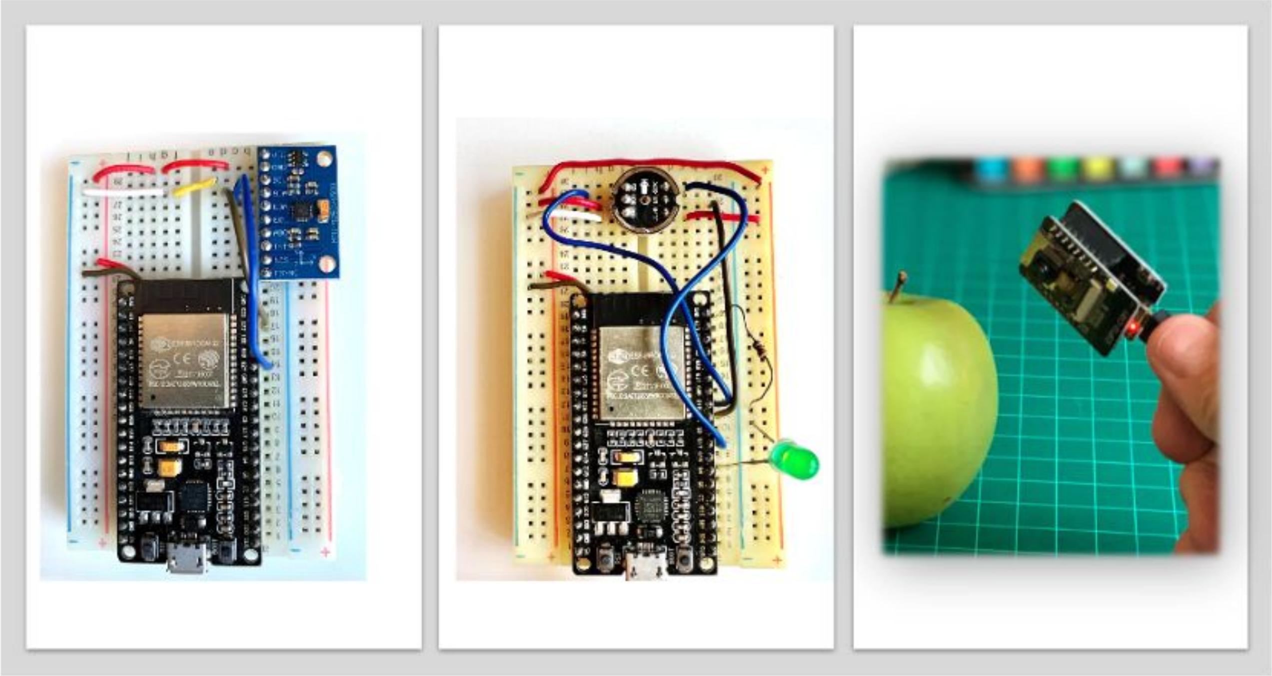
https://arxiv.org/abs/1909.11236

https://youtu.be/wmVKbX7MOnU

More MCUs…

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)

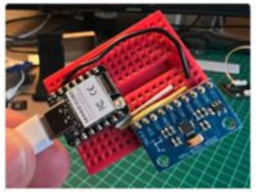
Sensor DataLogger
MJRoBot (Marcelo Rovai)

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

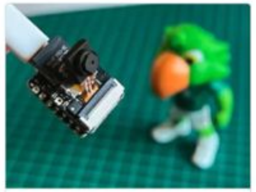
TinyML Made Easy: Sound Classification (KWS)
MJRoBot (Marcelo Rovai)

XIAO-ESP32S3-Sense

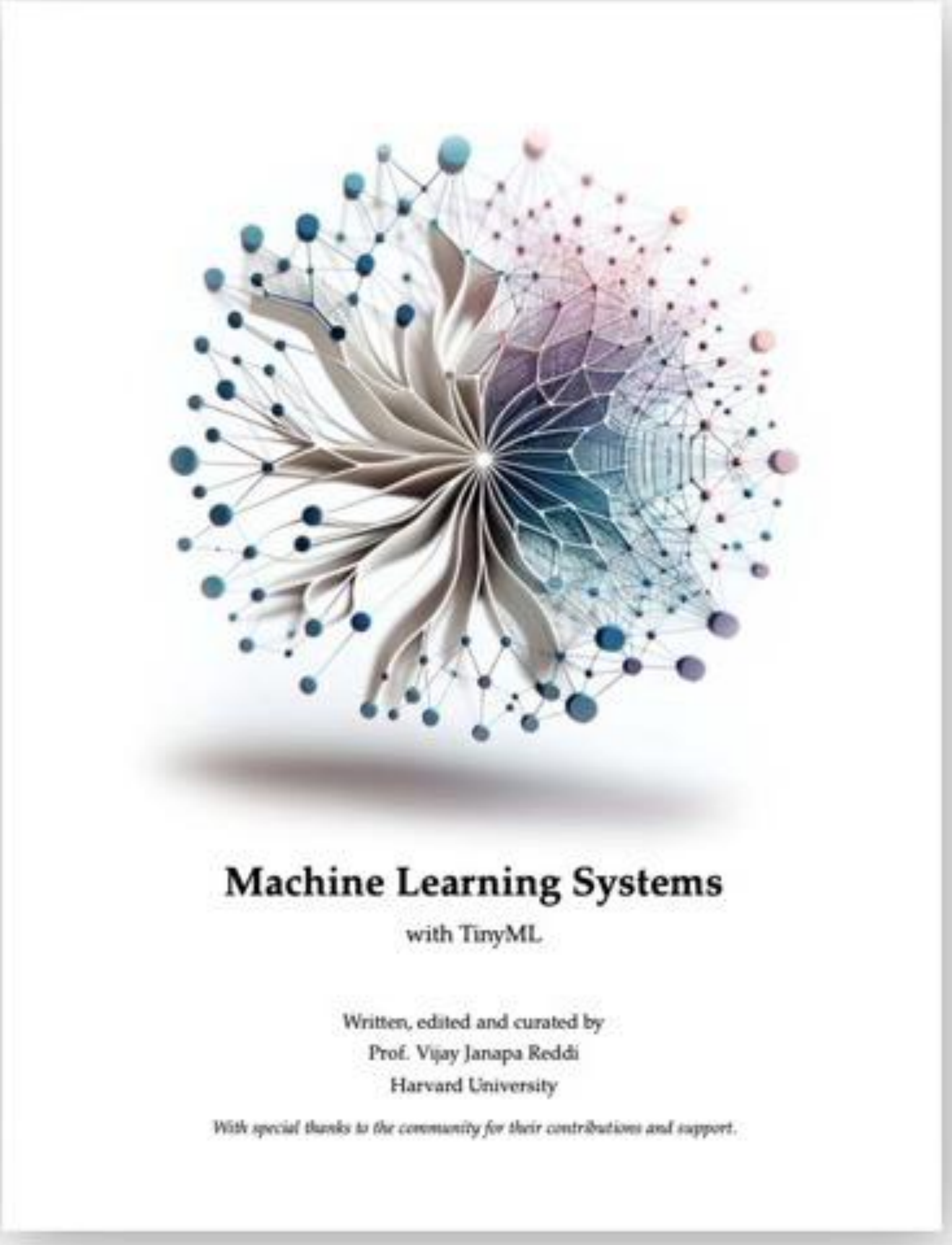
TinyML Made Easy: KeyWord Spotting (KWS)
MJRoBot (Marcelo Rovai)



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

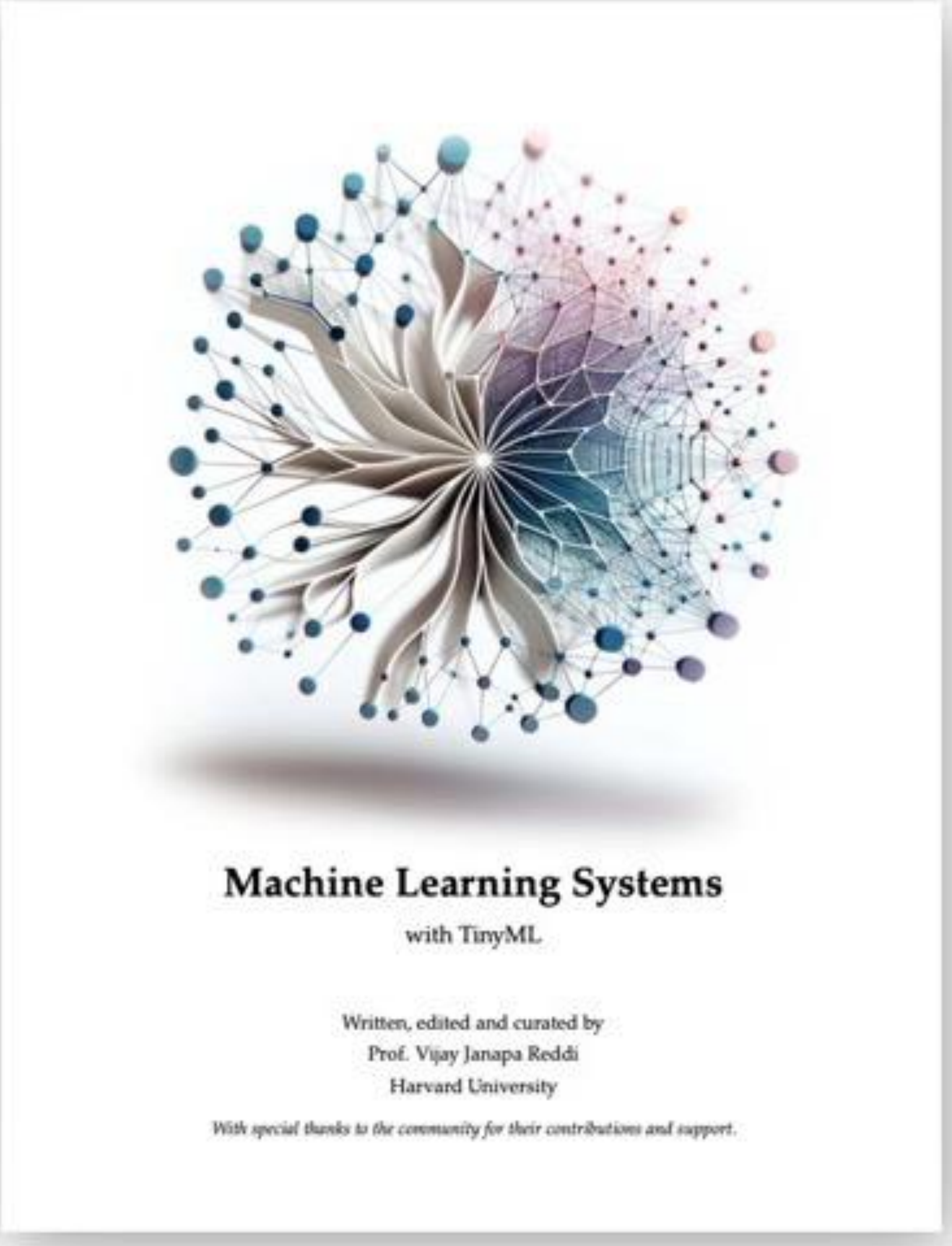


TinyML Made Easy: Image Classification
MJRoBot (Marcelo Rovai)

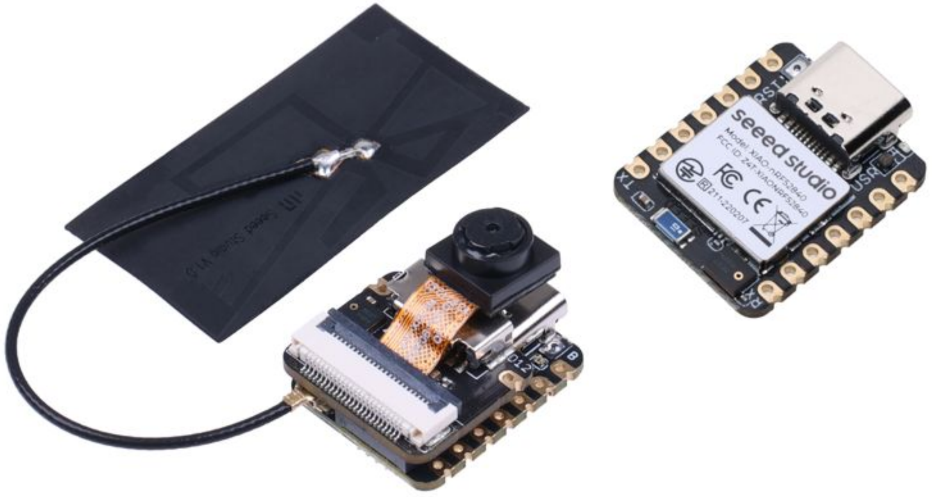


Nicla Vision

XIAO ESP32S3


Machine Learning Systems
with TinyML
Written, edited and curated by
Prof. Vijay Janapa Reddi
Harvard University
With special thanks to the community for their contributions and support.

Seeed Studio XIAO



To learn more …

Online Courses

Harvard School of Engineering and Applied Sciences - CS249r: Tiny Machine Learning
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
UNIFEI-IESTI01 TinyML: “Machine Learning for Embedding Devices”

Books

“Python for Data Analysis” by Wes McKinney
“Deep Learning with Python” by François Chollet - GitHub Notebooks
“TinyML” by Pete Warden and Daniel Situnayake
“TinyML Cookbook 2nd Edition” by Gian Marco Iodice
“Technical Strategy for AI Engineers, In the Era of Deep Learning” by Andrew Ng
“AI at the Edge” book by Daniel Situnayake and Jenny Plunkett
“XIAO: Big Power, Small Board” by Lei Feng and Marcelo Rovai
“MACHINE LEARNING SYSTEMS for TinyML” by a collaborative effort

Projects Repository

Edge Impulse Expert Network

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

[TinymML4D Academic Network Show and Tell Main Index.]

The TinyML4D Academic Network Students should use this form to propose 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 Meet link:
https://meet.google.com/rns-yyrx-ggw

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 |

Thanks

TINYML4D