Patrocinado por

MOUSER ELECTRONICS

Revolucionando

O Caminho do TinyML para a IA Generativa na Borda

Marcelo Rovai
Professor na UNIFEI e
Co-Diretor do TinyML4D

WEBINAR EMBARCADOS

EMBARCADOS

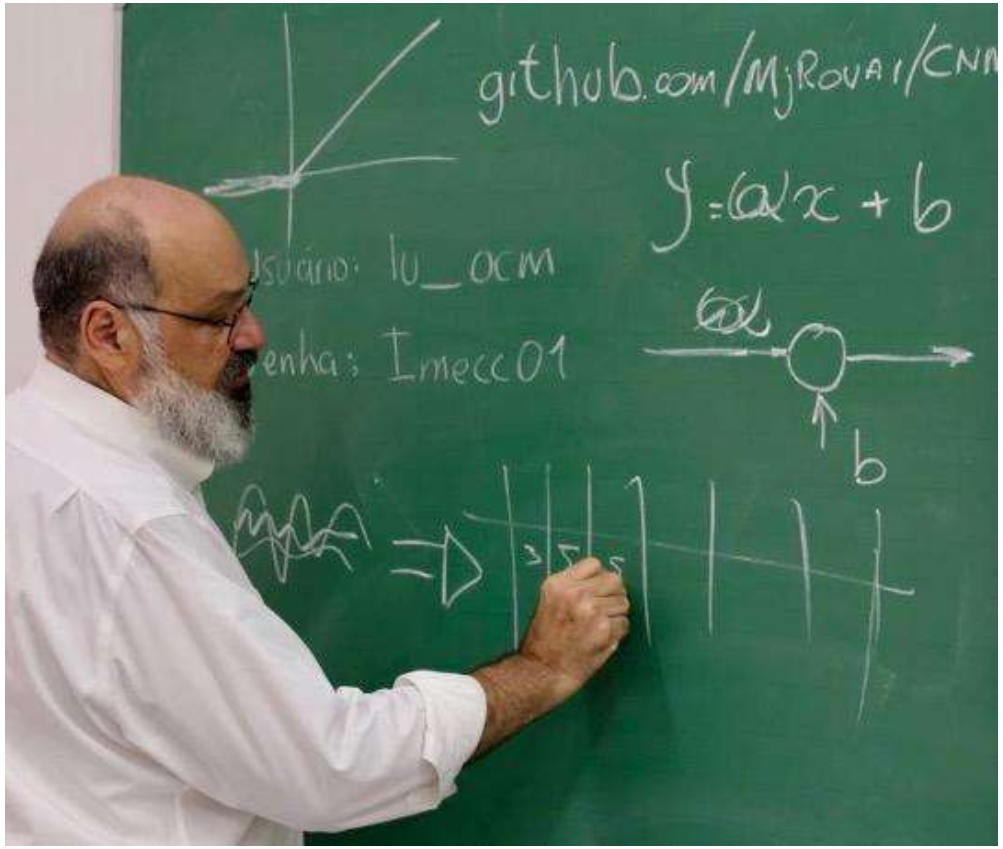
Patrocinado por

MOUSER ELECTRONICS

Agenda

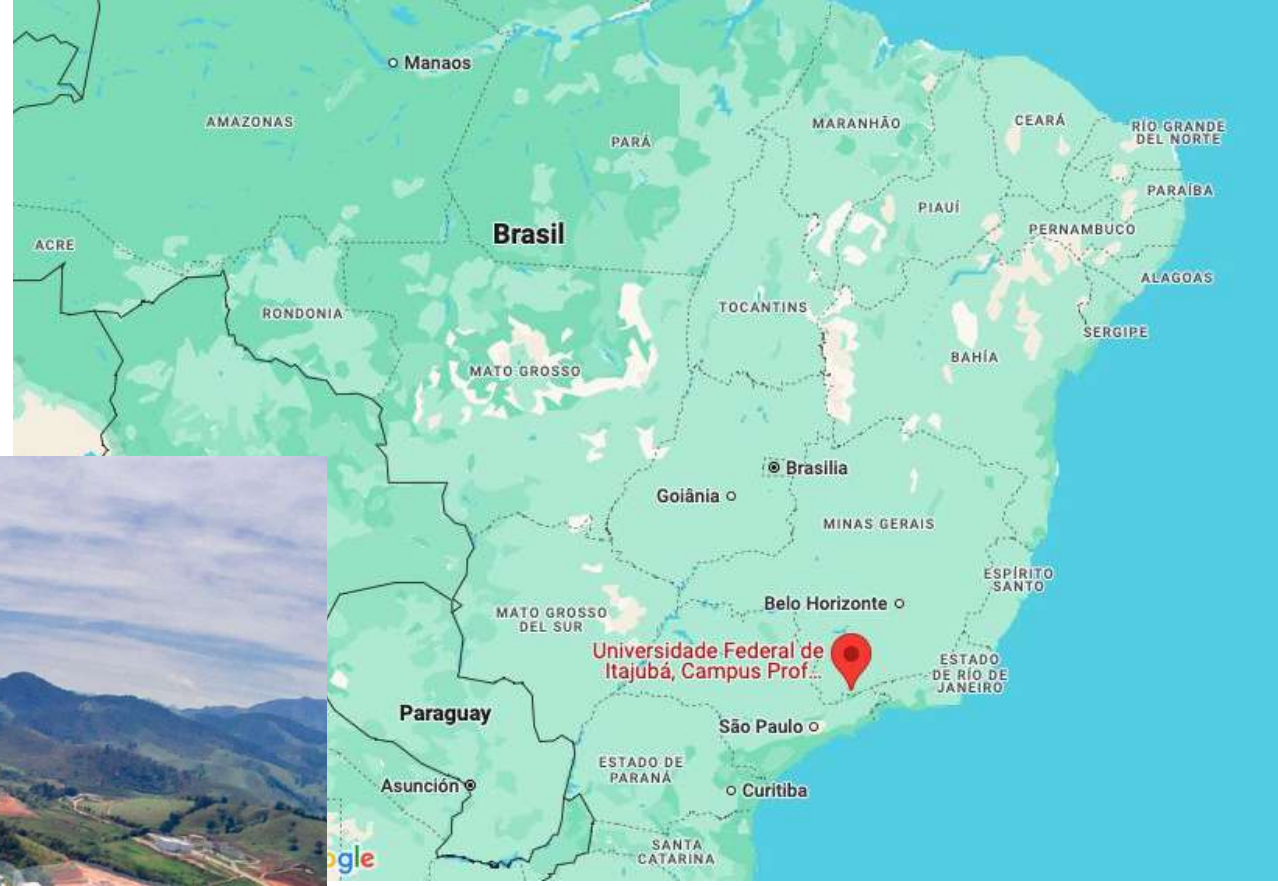
1. Internet of Things to Intelligence of Things
2. Embedded Machine Learning & Challenges
3. Real World Applications
4. Generative AI at the Edge
5. The future
6. TinyML4D and EAIP – Academic & Industry Partnership

Marcelo Rovai is an educator and professional in the field of engineering and technology, holding the title of Professor Honoris Causa from the Federal University of Itajubá, Brazil. His educational background includes an Engineering degree from UNIFEI and a specialization from the Polytechnic School of São Paulo University (POLI/USP). Further enhancing his expertise, he earned an MBA from IBMEC (INSPER) and a Master's in Data Science from the Universidad del Desarrollo (UDD) in Chile.



With a career spanning several high-profile technology companies such as AVIBRAS Airspace, AT&T, NCR, and IGT, where he served as Vice President for Latin America, he brings a wealth of industry experience to his academic endeavors. He is a prolific writer on electronics-related topics and shares his knowledge through open platforms like Hackster.io.

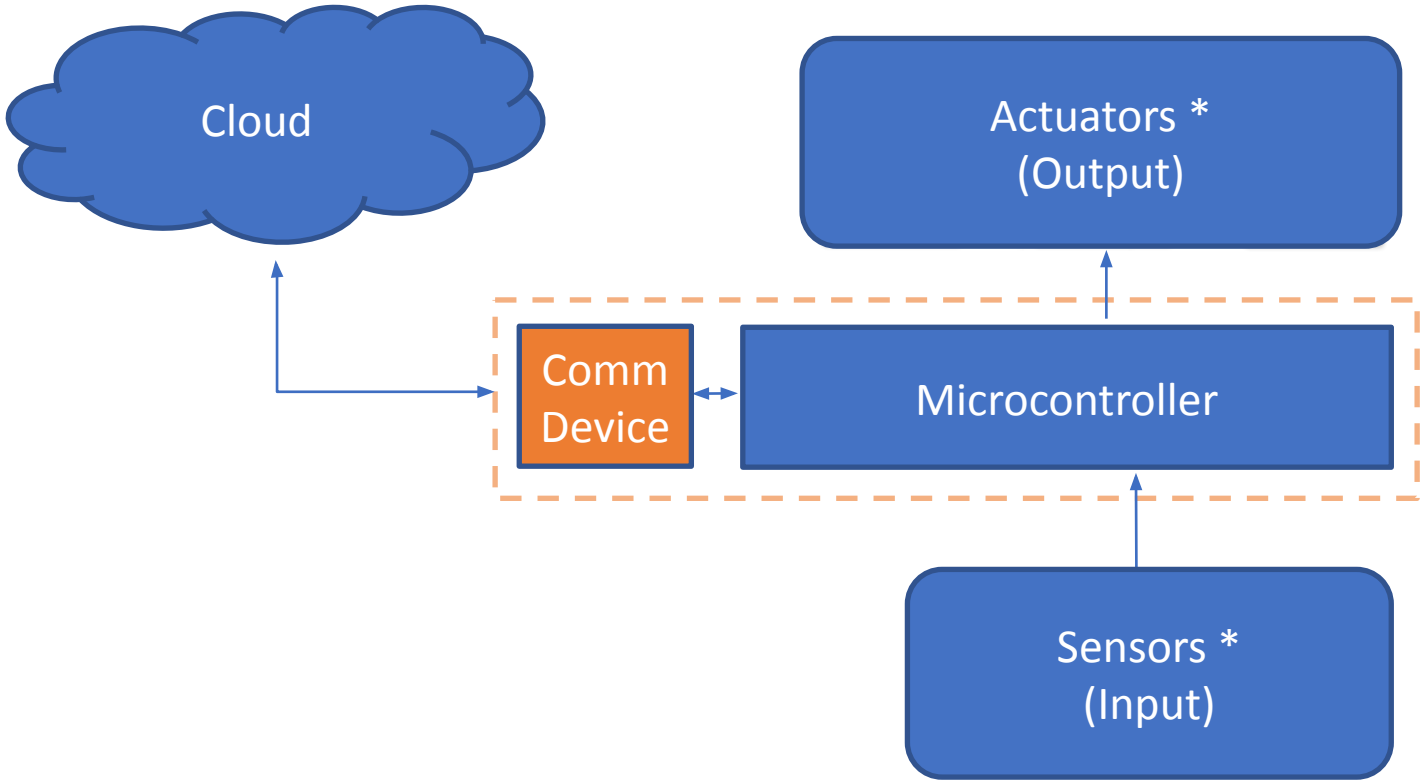
In addition to his professional pursuits, he is dedicated to educational outreach, serving as a volunteer professor at the IESTI (UNIFEI) and engaging with the TinyML4D group and the EDGE AIP – the Academia-Industry Partnership of EDGEAI Foundation as a Co-Chair, promoting EdgeAI education in developing countries. His work underscores a commitment to leveraging technology for societal advancement.



Itajubá,
Minas Gerais,
Brazil

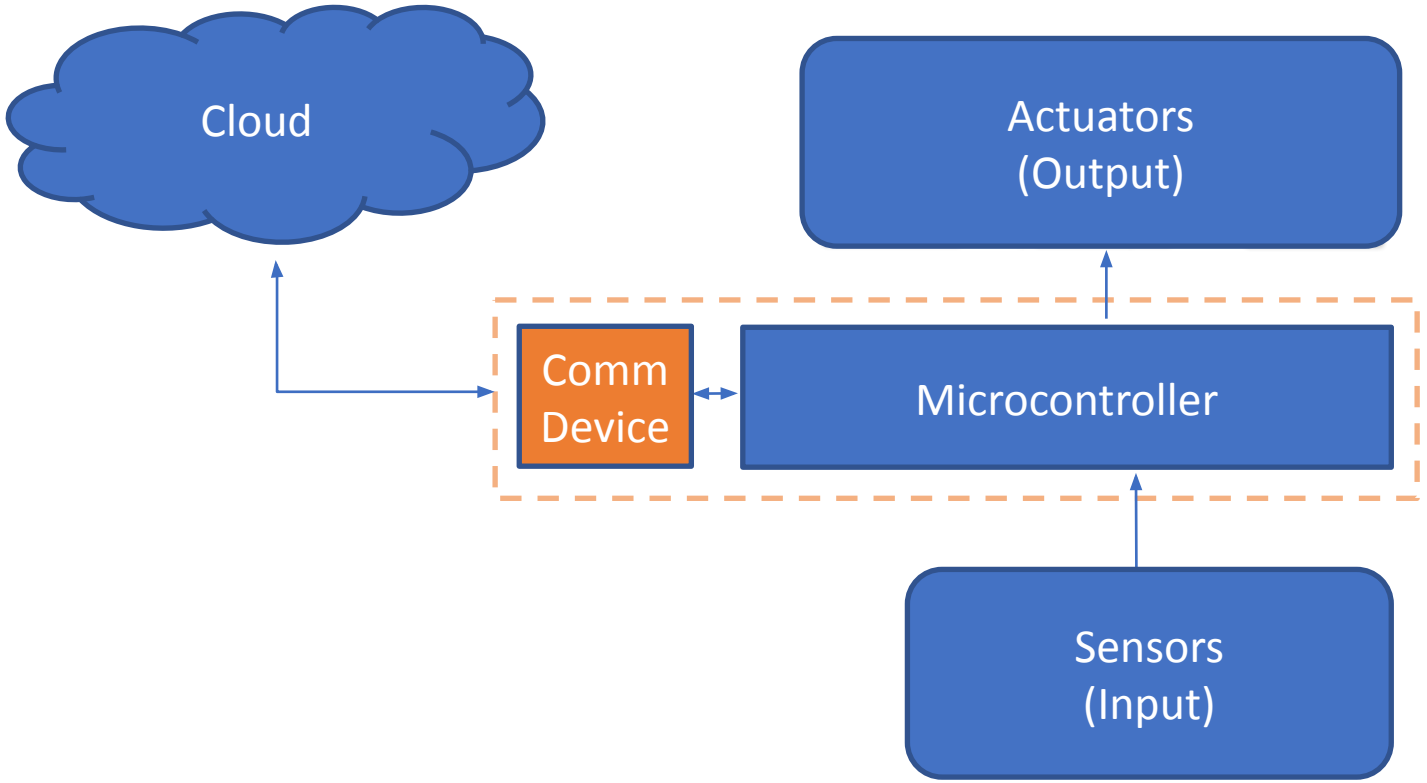
Internet of Things (IoT)

Typical IoT Project



* “Things”

Typical IoT Project



5 Quintillion

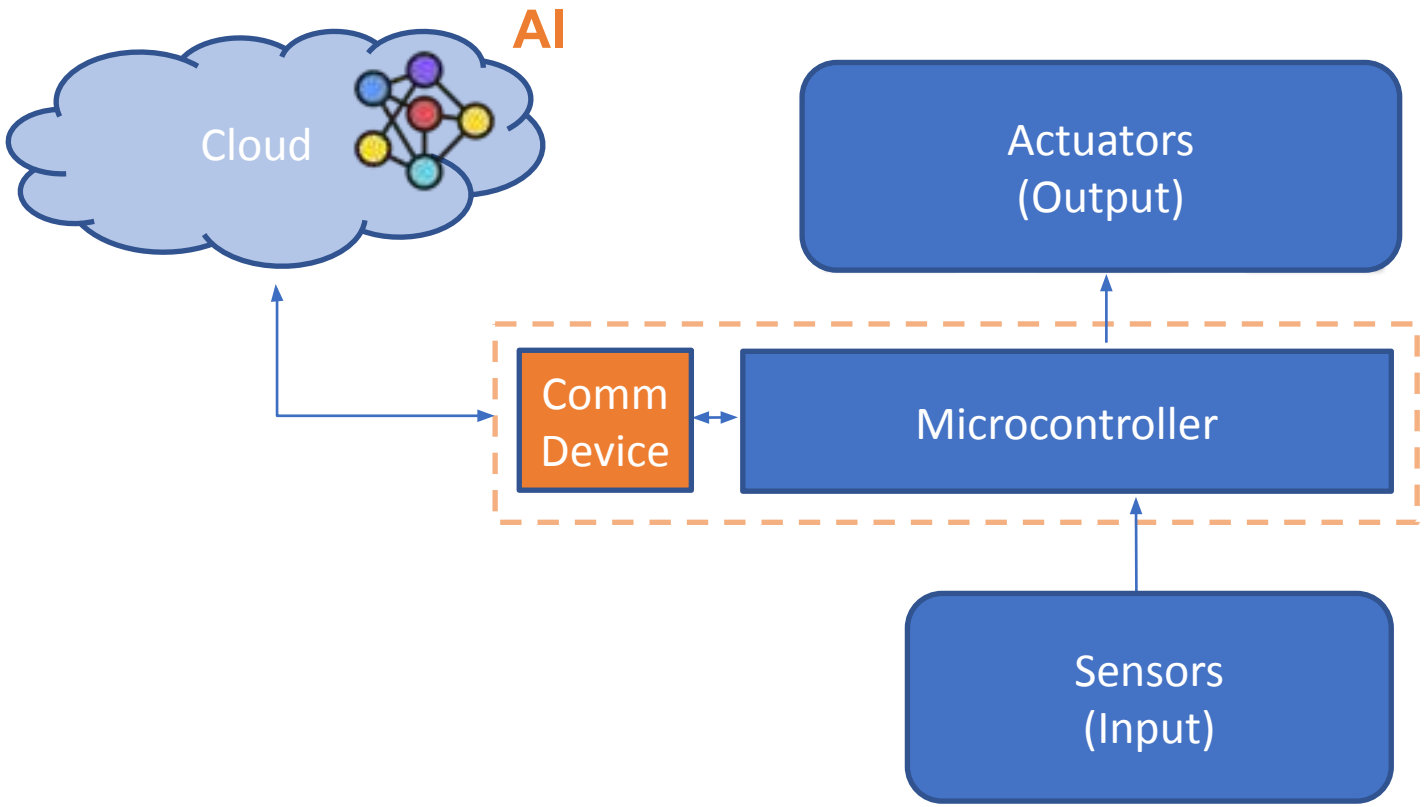
bytes of data produced every day by IoT

<1%

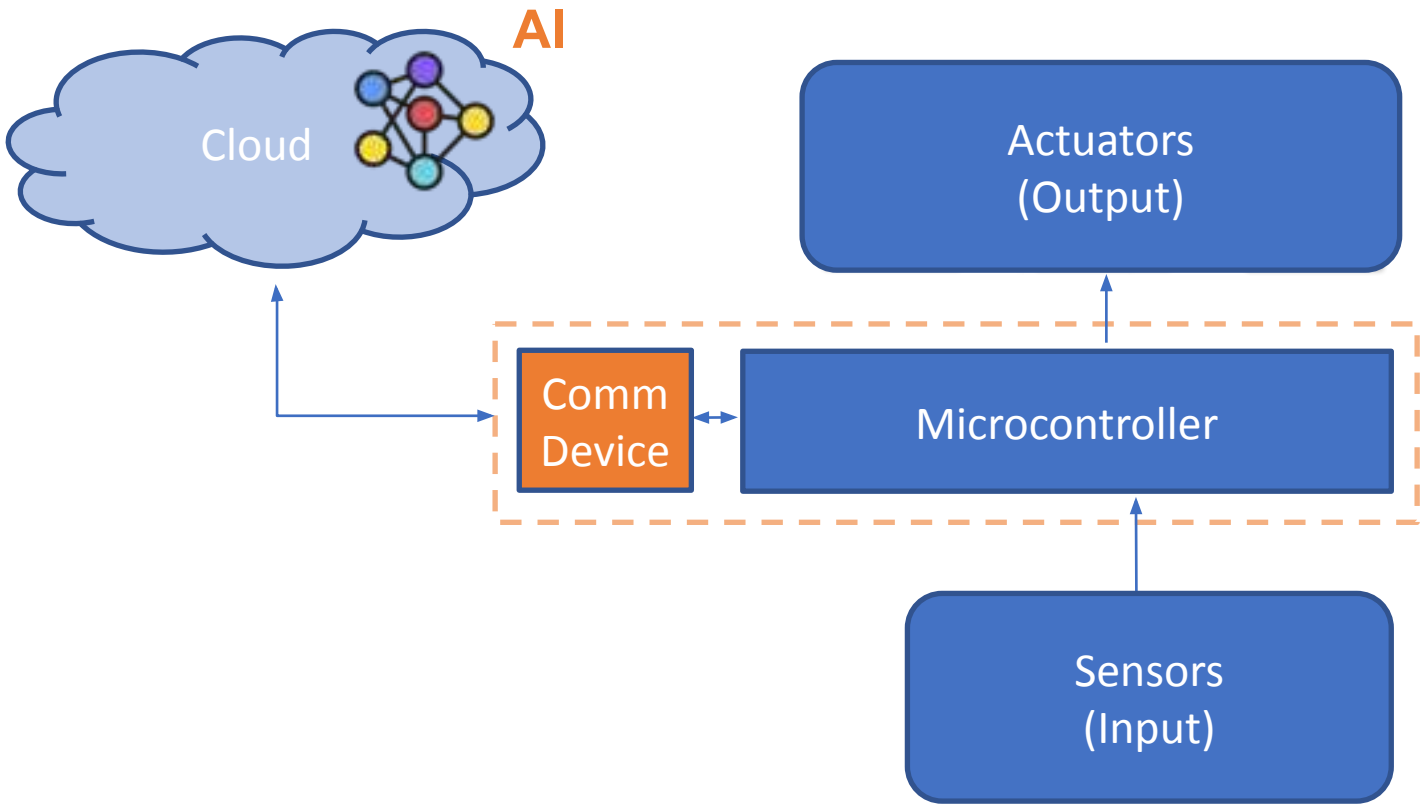
of unstructured data is analyzed or used at all

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

Latency

Energy

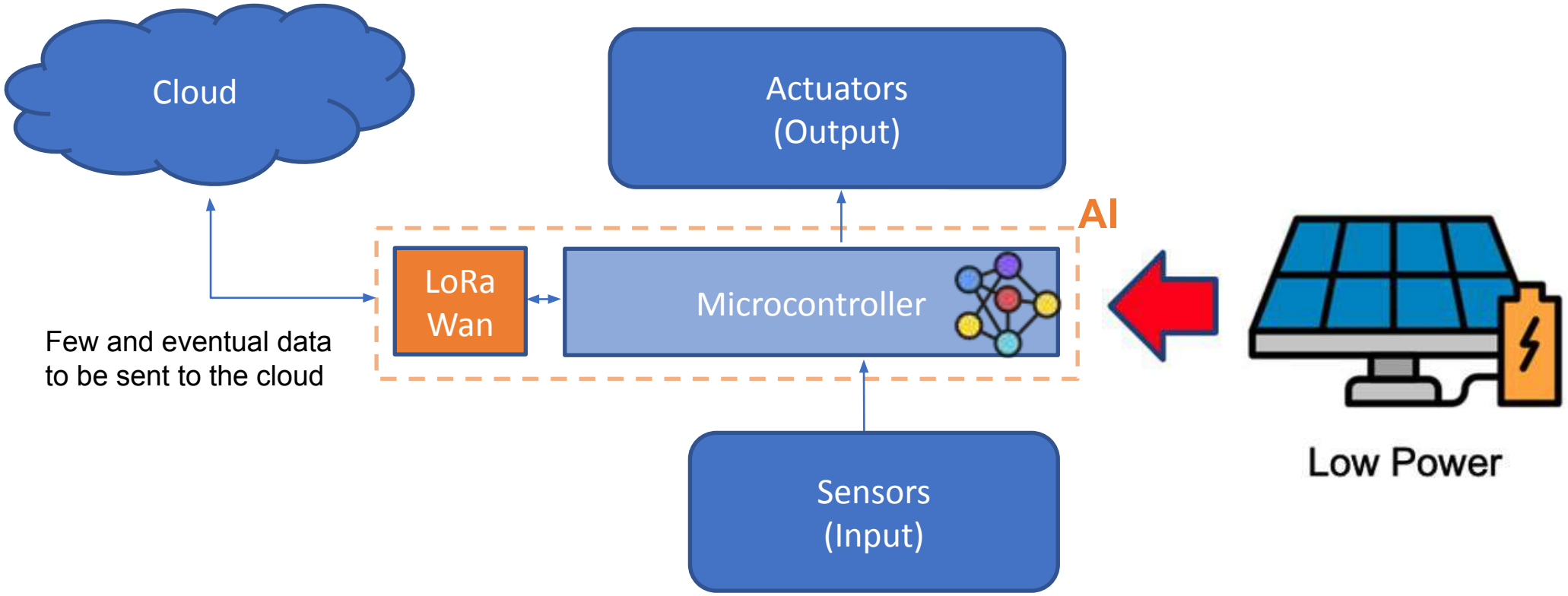
Reliability

Privacy

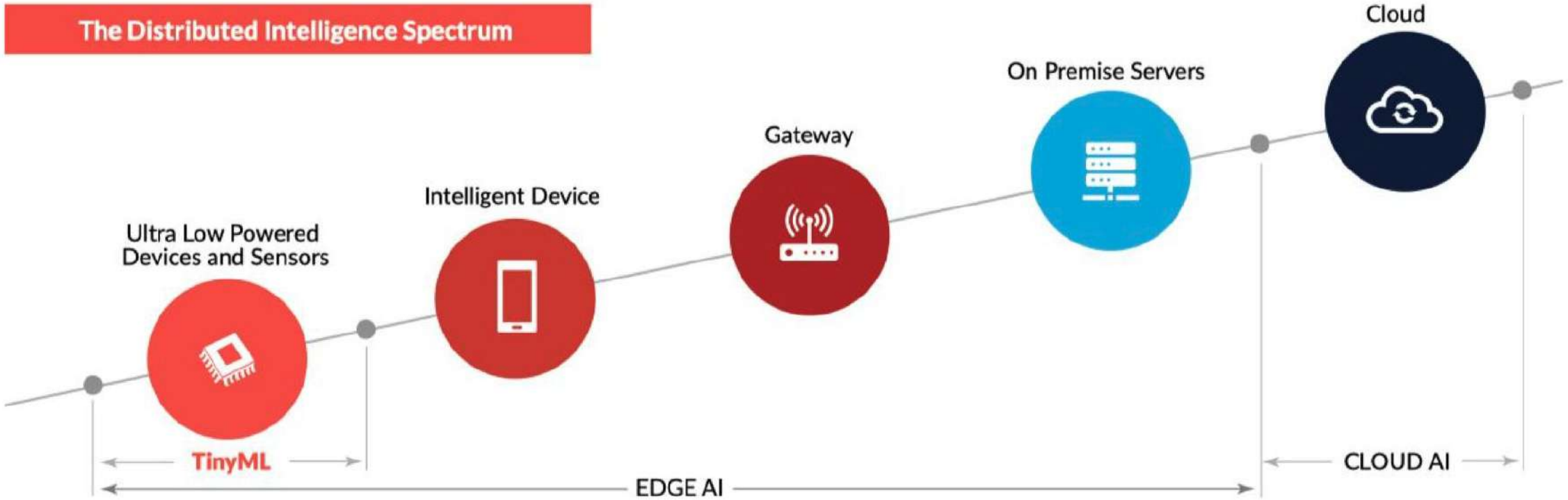
… Solution ?

IoT 2.0 * – Edge AI/ML

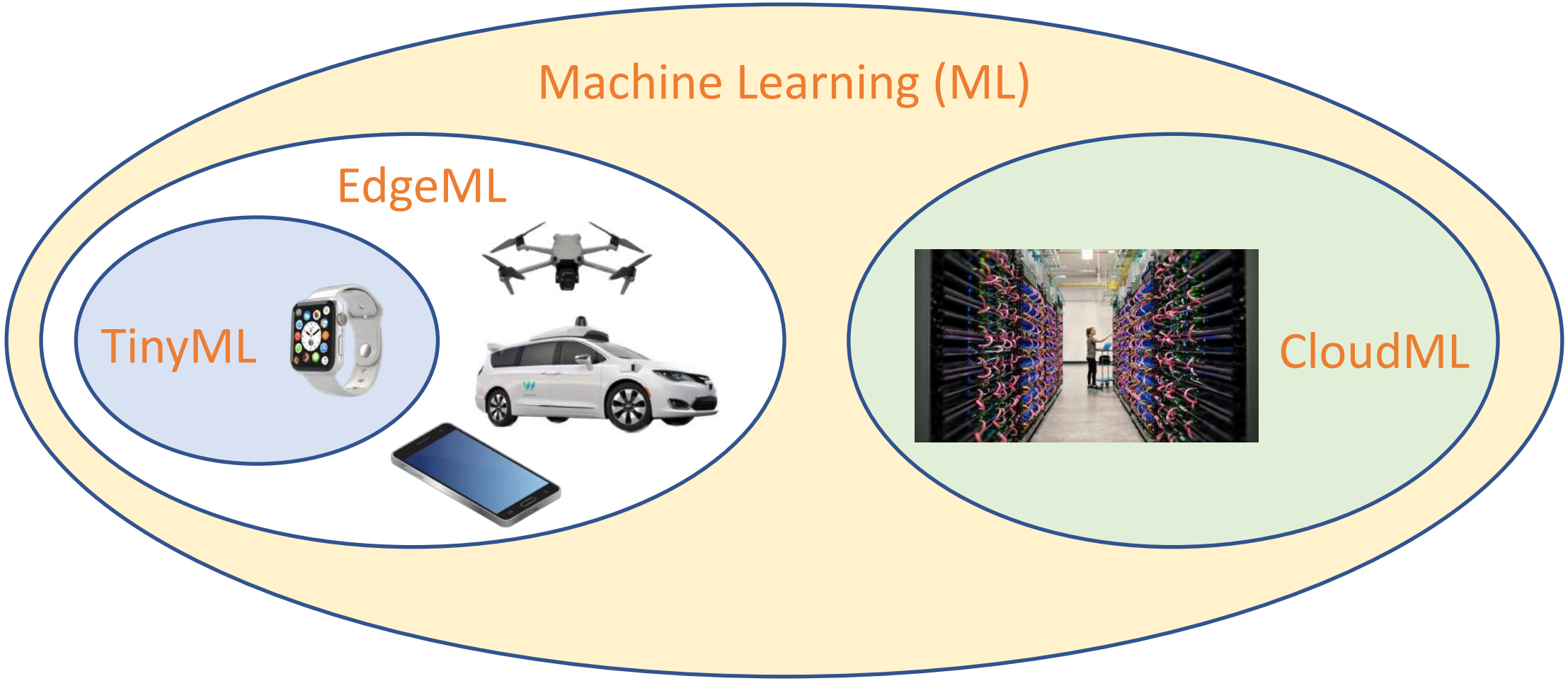
* Intelligence of Things



… **Solution ->** ML goes close to the data



Source: ABI Research: TinyML


Machine Learning (ML)
EdgeML
TinyML
CloudML

Embedded ML
EdgeAI & TinyML

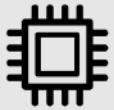
What is Tiny Machine Learning (TinyML)?

TinyML

Fastest-growing field of **ML**

Algorithms, hardware, software

On-device sensor analytics



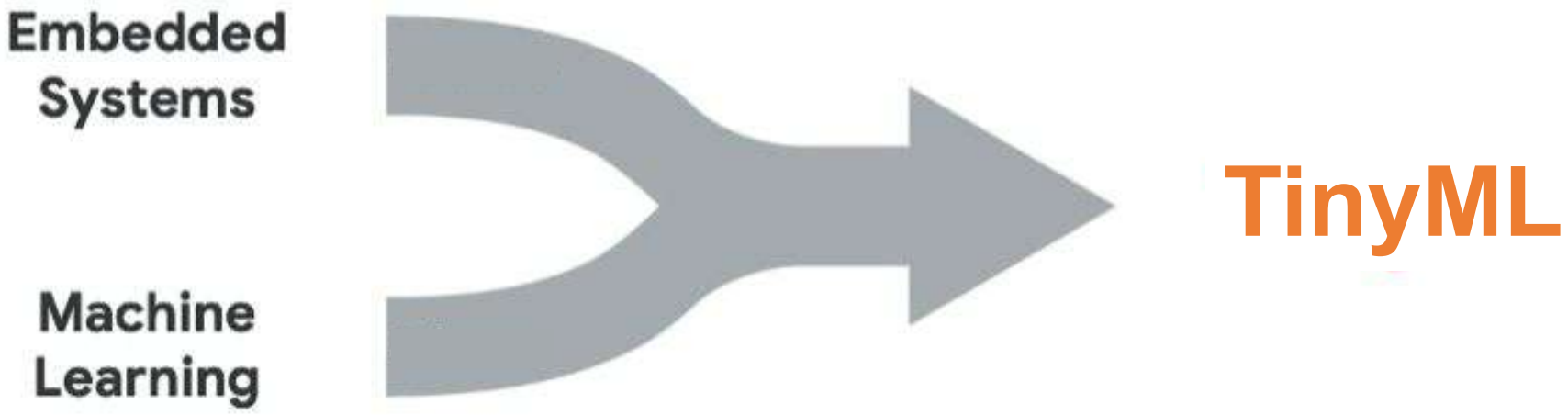
Low power consumption

Battery-operated

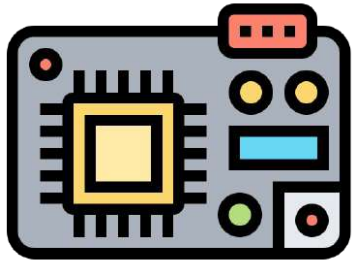
Always-on ML

15

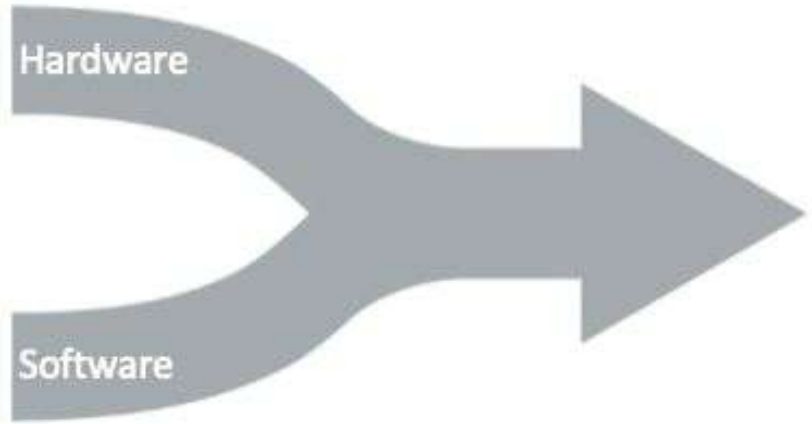
What Makes TinyML ?



What Makes TinyML ?

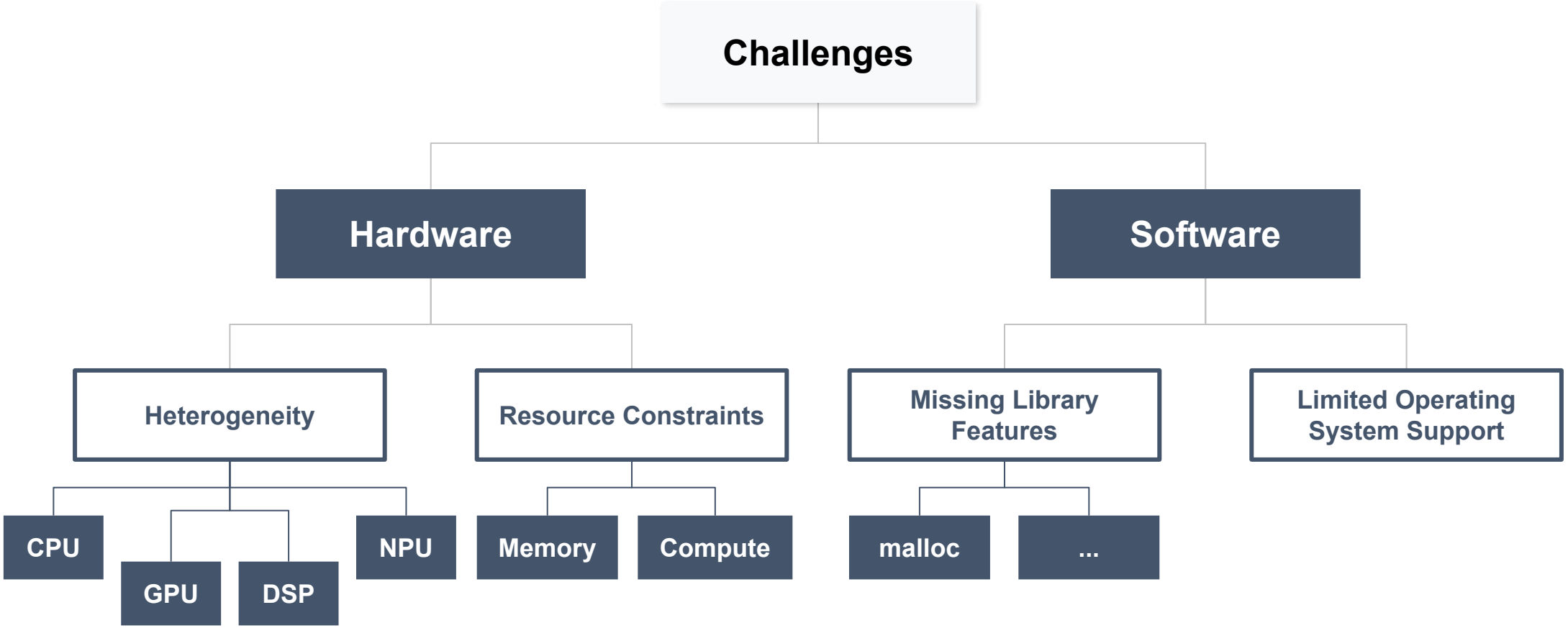


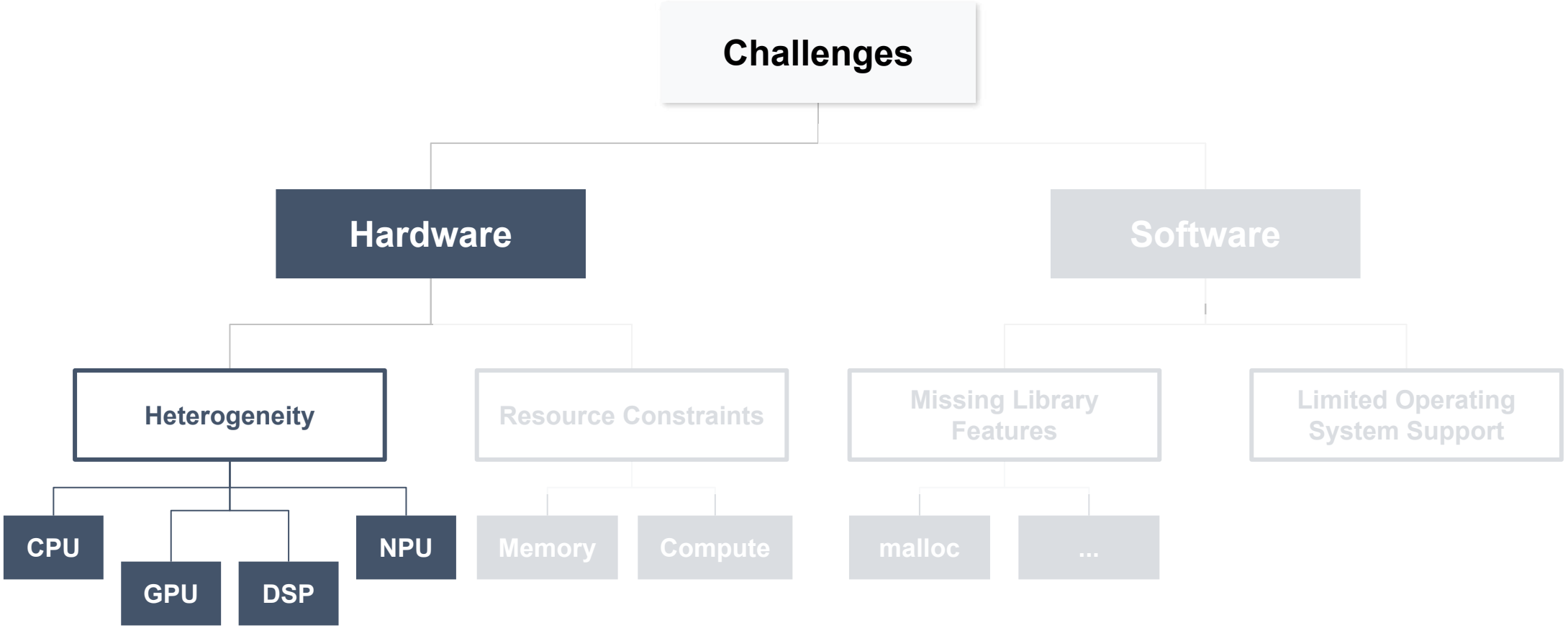
TensorFlow Lite



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
NPU
Memory
Compute
malloc
...

MOUSER ELECTRONICS

250 Billion

MCUs today

EMBARCADOS

21

Hardware

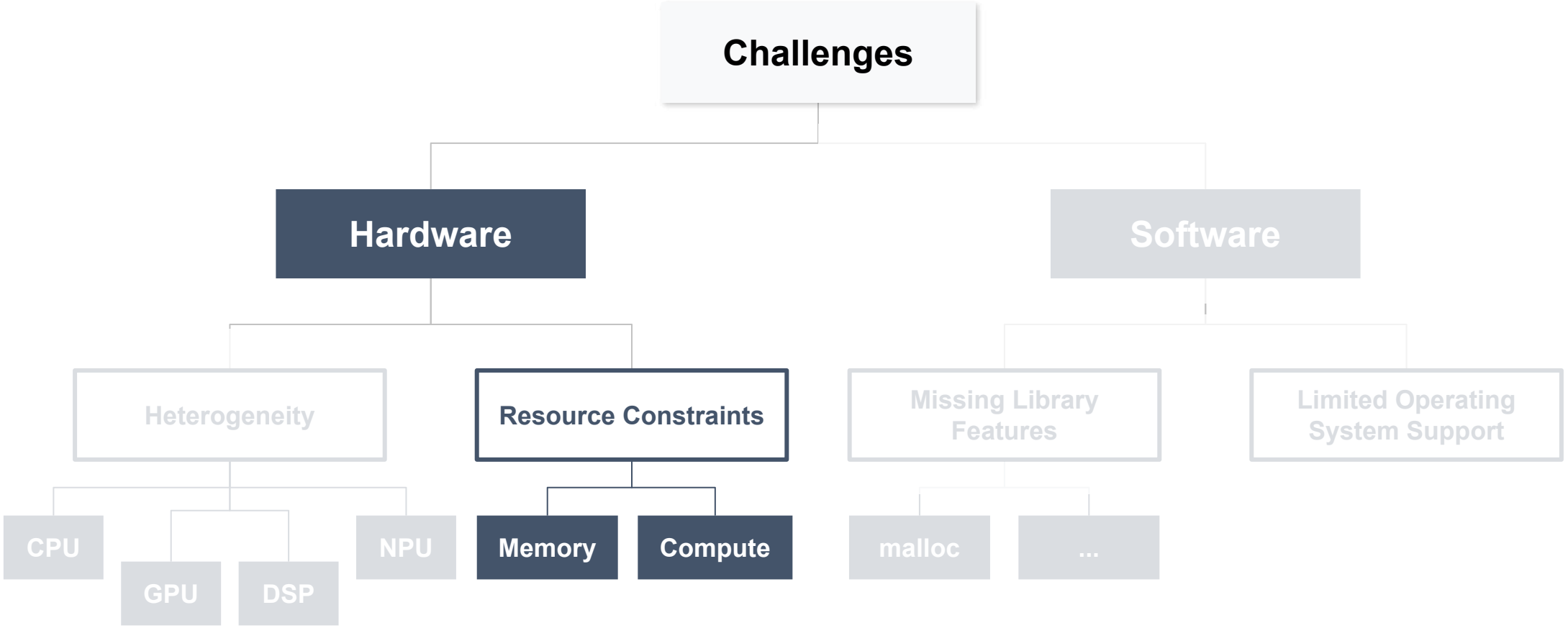
Hardware (Development Boards)





	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
NPU
Memory
Compute
malloc
...

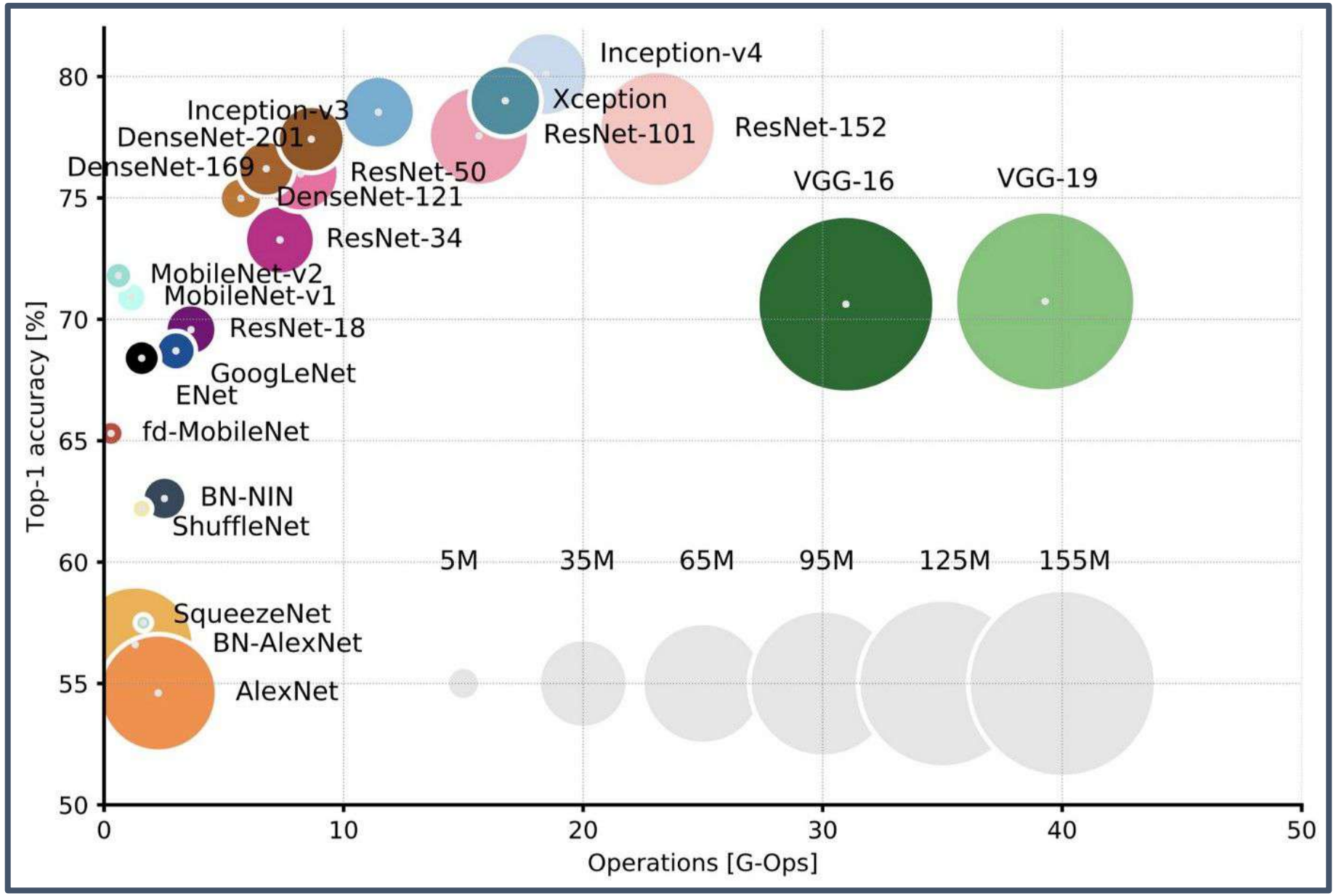
Hardware

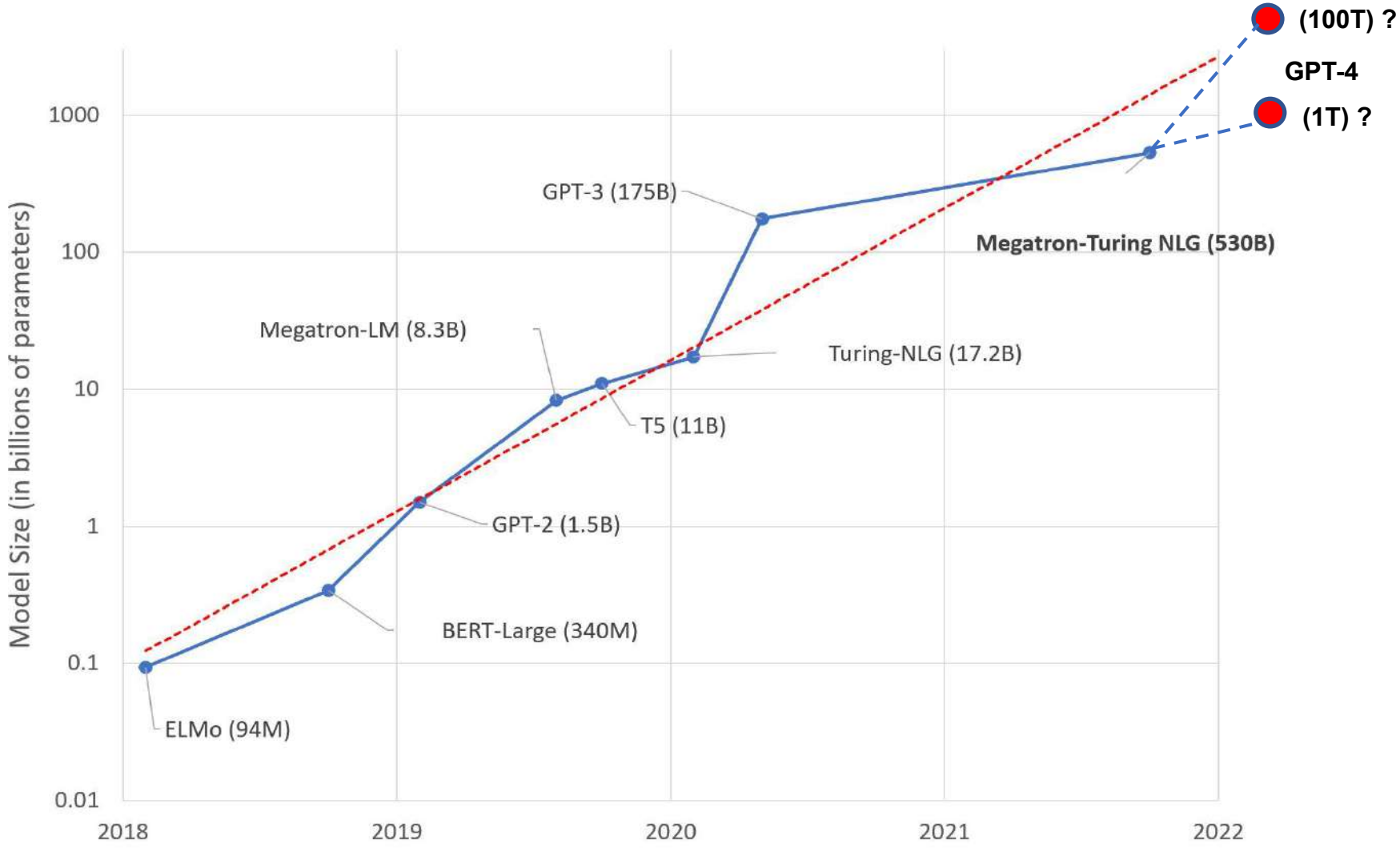



	Raspberry Pico (W)	Arduino Nano Sense	ESP 32	Seeed XIAO Sense / ESP32S3	Arduino Pro
32Bits CPU	Dual-core Arm Cortex-M0+	Arm Cortex-M4F	Xtensa LX6 Dual Core	Arm Cortex-M4F (BLE) Xtensa LX7 Dual Core	Dual Core Arm Cortex M7/M4
CLOCK	133MHz	64MHz	240MHz	64 / 240MHz	480/240MHz
RAM	264KB	256KB	520KB (part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

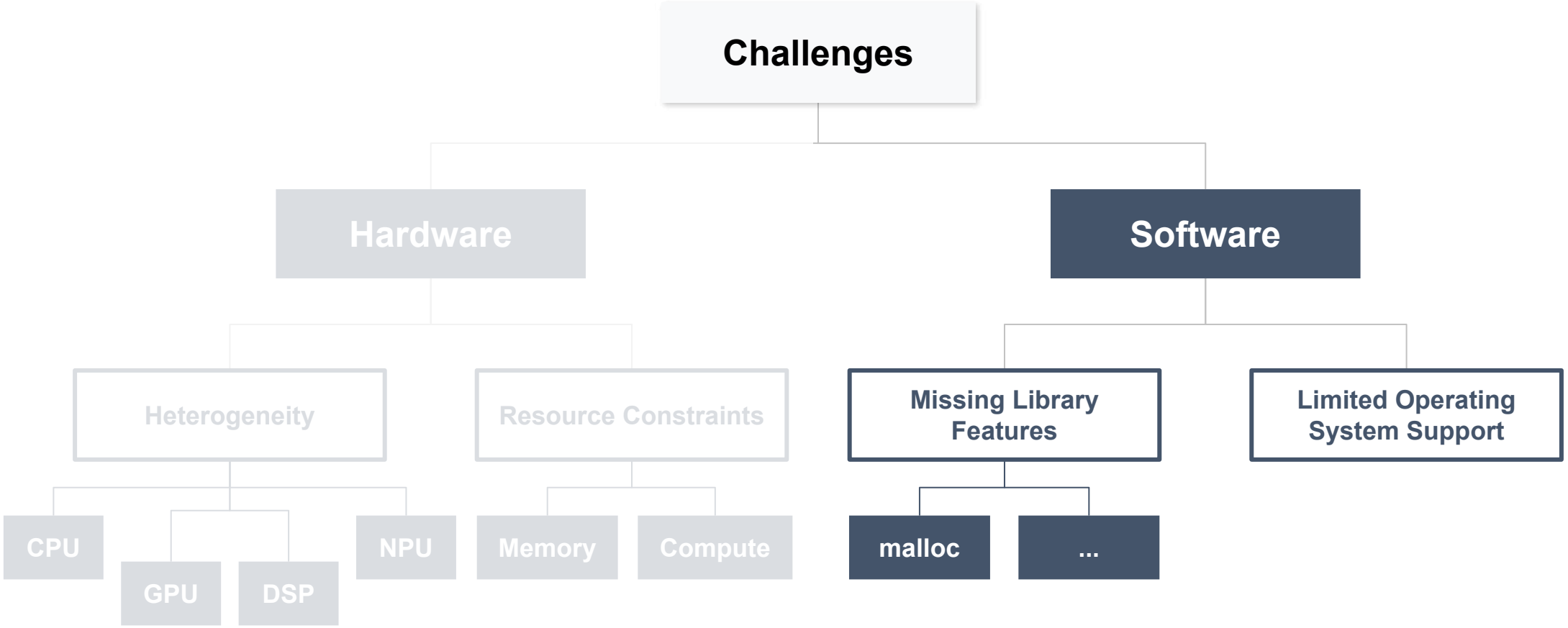
https://media.digikey.com/Resources/Maker/the-original-guide-to-boards-2022.pdf

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

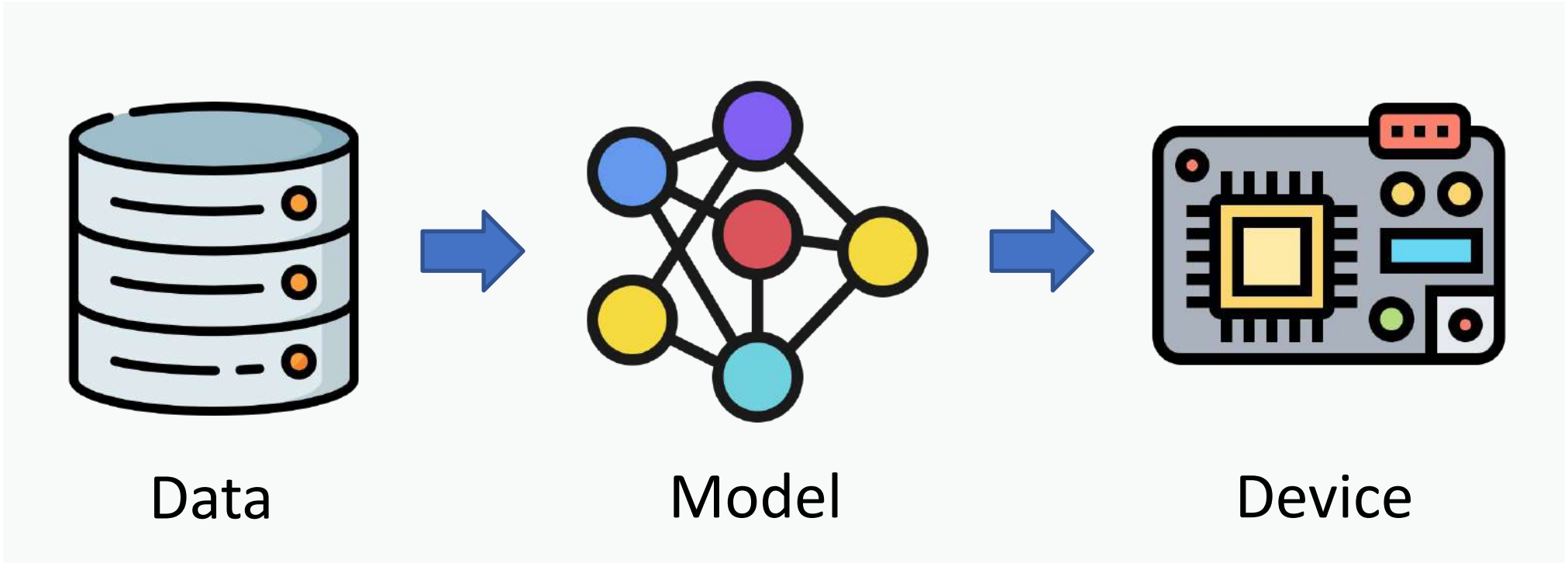




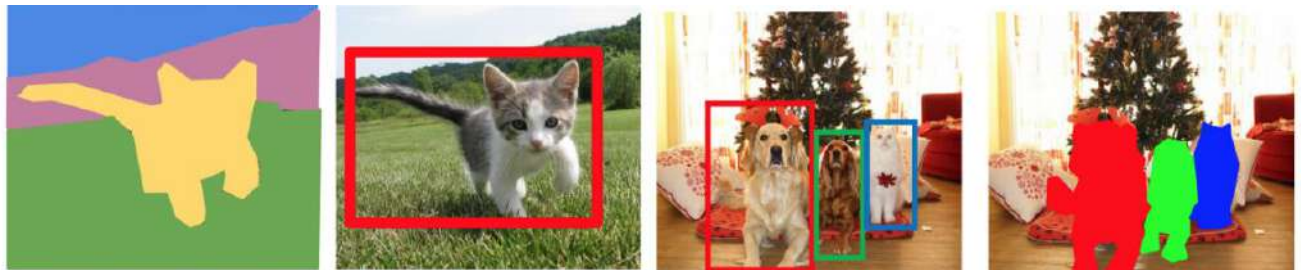
https://huggingface.co/blog/large-language-models https://arxiv.org/abs/2303.08774

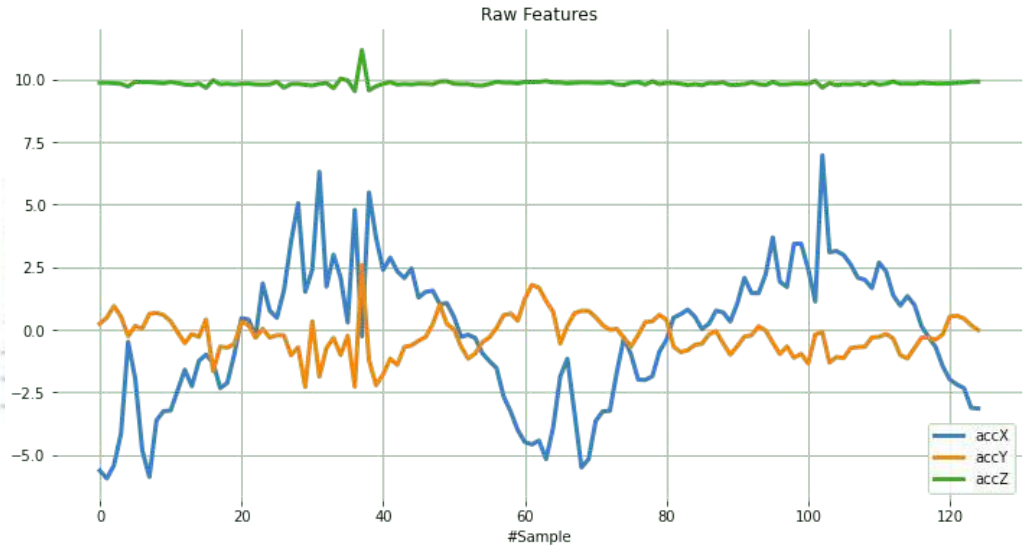

Challenges
Hardware
Software
Heterogeneity
Resource Constraints
Missing Library Features
Limited Operating System Support
CPU
GPU
DSP
NPU
Memory
Compute
malloc
...

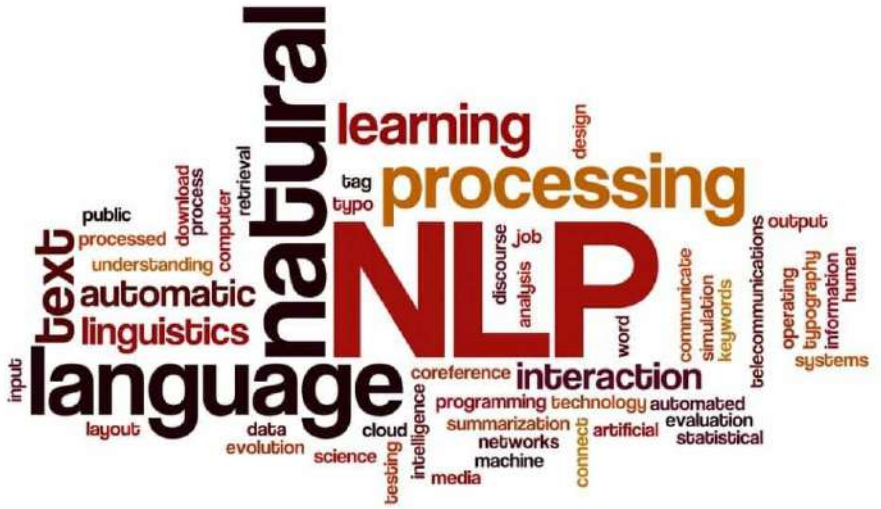
ML Deployment Pipeline

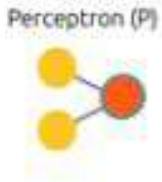


Unstructured Data









Neural Network Architectures

MOUSER ELECTRONICS

Vibration Analysis

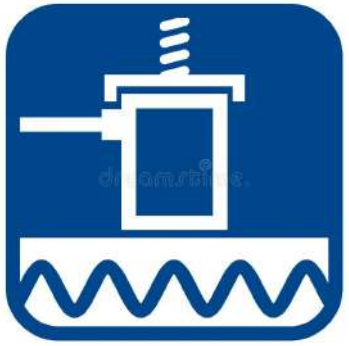
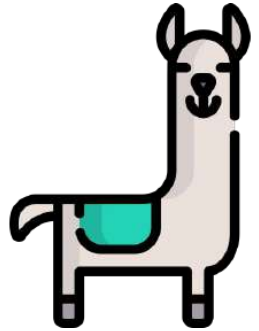


Image Classification

Text Generation

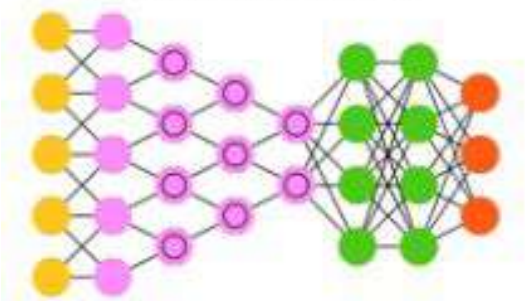
Image Generation

Large Language Models- LLMs

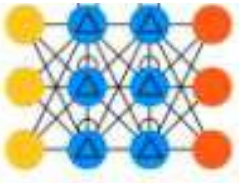


MLP - Deep Neural Network

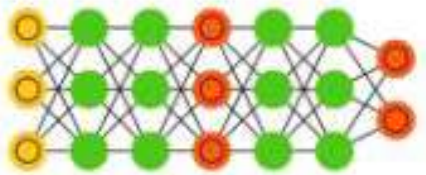
CNN - Convolutional NN



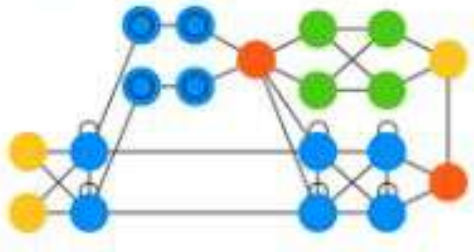
RNN - Recurrent NN (GRU/LSTM)



GAN - Generative Adversarial N.



AN - Attention (Transformers)



EMBARCADOS

Datasets Preprocessing

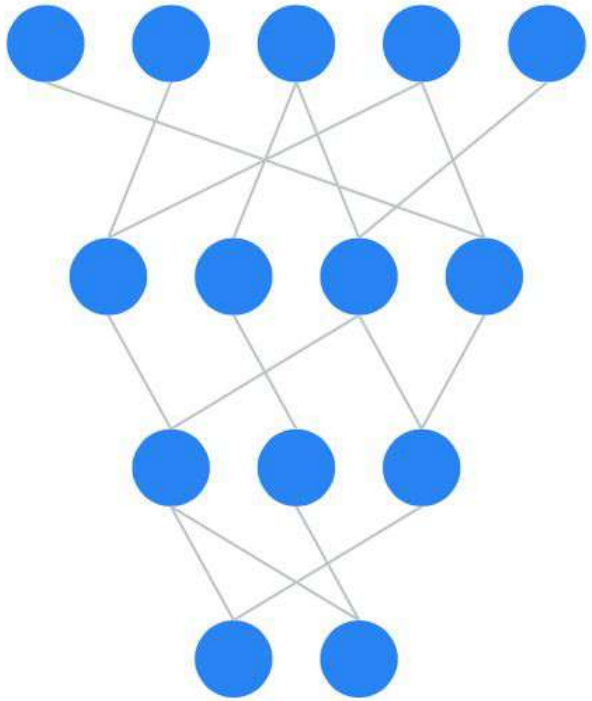
Sound

Vision

Vibration

Text

Quantization Pruning, Distillation



Resource constraints

End-to-end TinyML/EdgeAI Application

Application Complexity vs. HW

Power

EdgeML

TinyML

LLMs/VLMs
1 GB+

Object Detection
Complex Voice
Processing
1 MB+

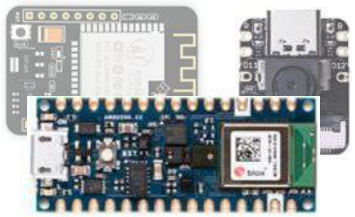
Image
Classification
250 KB+

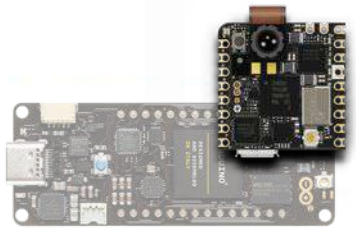
KeyWord Spotting
Audio Classification
50 KB

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 Orin
(Cortex-A + GPU)

Application Complexity

CPU Power / Memory

Application Complexity vs. HW

Power

EdgeML

TinyML

LLMs/VLMs
1 GB+

Object Detection
Complex Voice
Processing
1 MB+

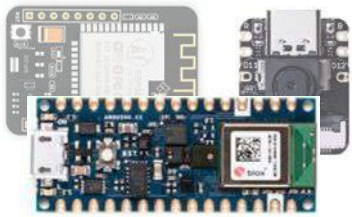
Image
Classification
250 KB+

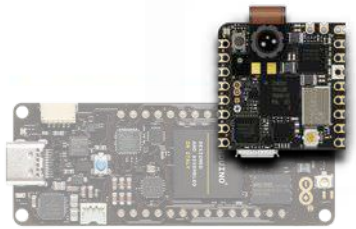
KeyWord Spotting
Audio Classification
50 KB

Anomaly Detection
Sensor Classification
20 KB

ESP32

XIAO





micro NPU

Rpi-Pico
(Cortex-M0+)

Arduino Nano
(Cortex-M4)

Arduino Pro
(Cortex-M7)

RaspberryPi SmartPhone
(Cortex-A)

Jetson Orin
(Cortex-A + GPU)

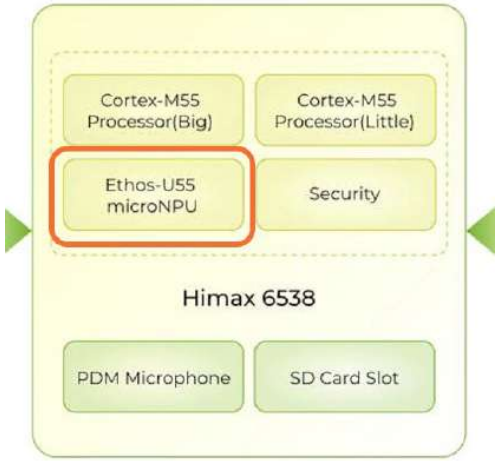
Application Complexity

CPU Power / Memory

microNPU

NN Inference Accelerator

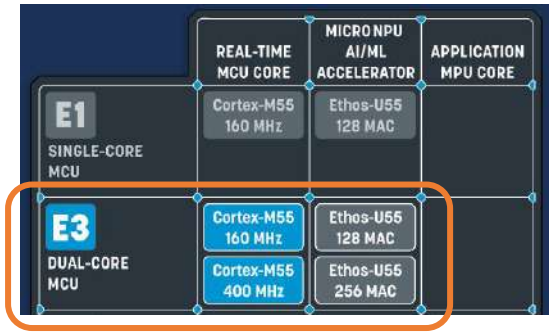
Grove Vision AI v2



350mW

0.5 TOPS

OpenMV AE3

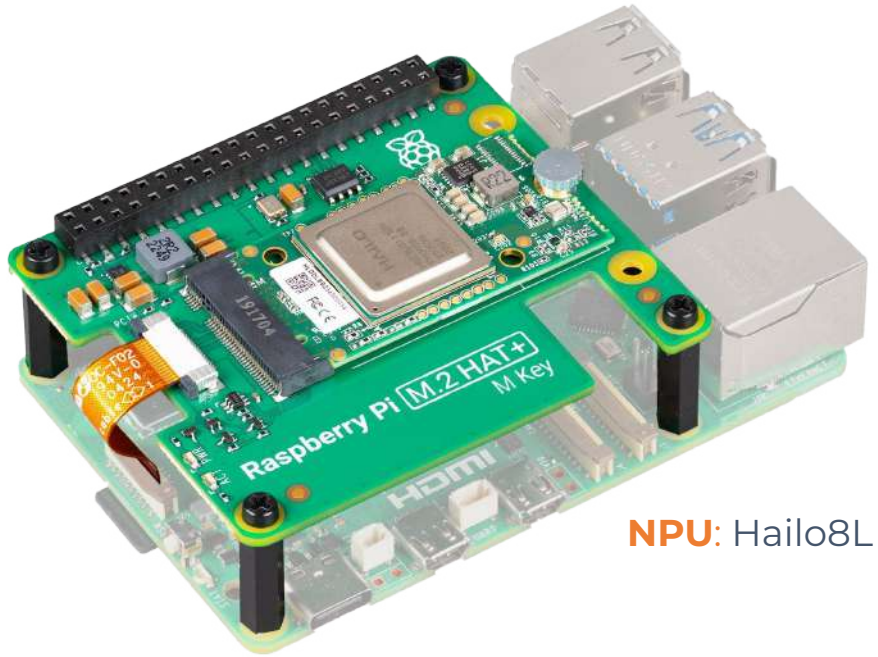


Alif E3

150mW

0.25 TOPS

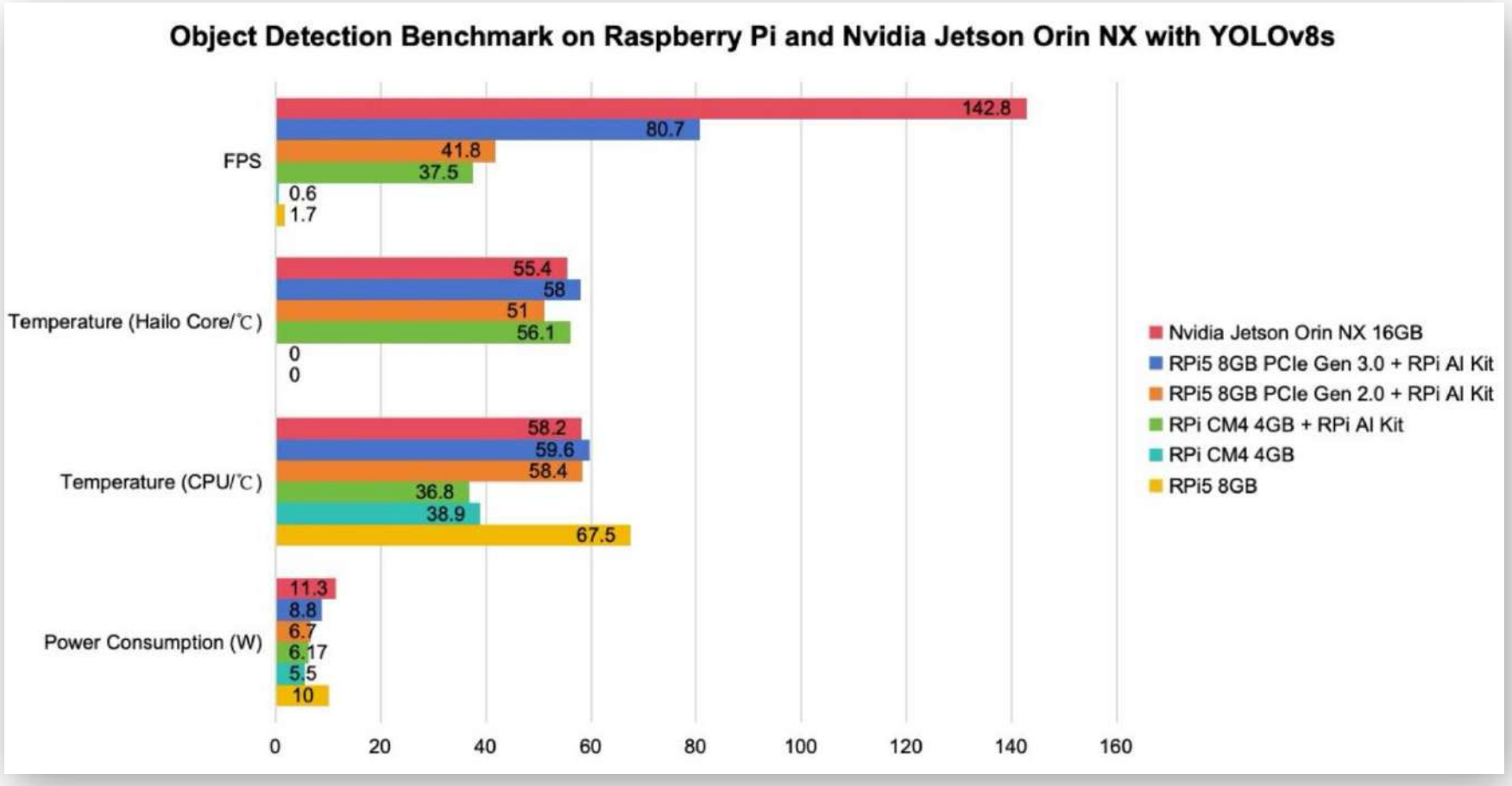
Raspberry Pi AI Kit



NPU: Hailo8L

30 W

13 - 25 TOPS

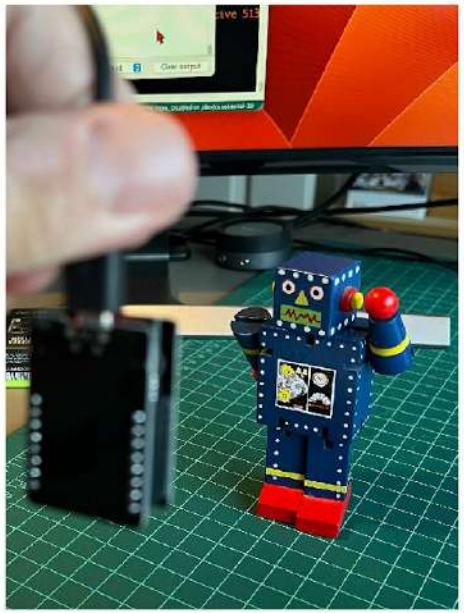


GPU: NVIDIA Ampere 100 TOPS

NPU: Hailo8L*

* **Note:** Hailo-10H for LLMs is planned for future release

Tiny Image Classification Benchmark (MobileNetV2 96x96 0.1)



1.5 FPS

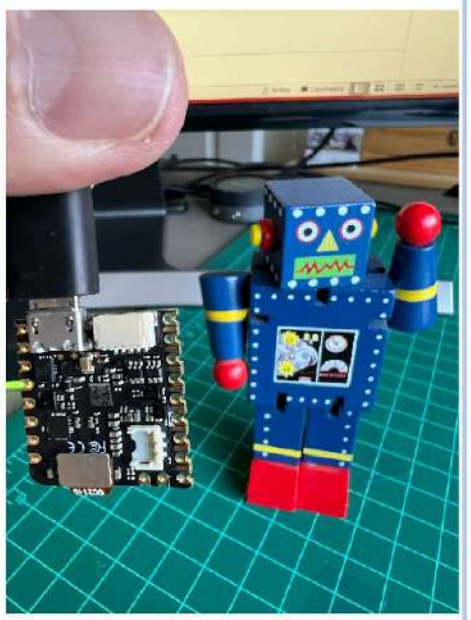
ESP - CAM
Xtensa LX6
240 MHz

300 mW

7.0 FPS

XIAO ESP32S3
Xtensa LX7
240 MHz

525 mW



11.6 FPS

Nicla-Vision
ARM M7
480 MHz

600 mW

91.0 FPS

Raspi Zero W2
ARM A53
1 GHz

1,500 mW

167 FPS

Grove Vision AI V2
ARM Ethus-U55
400 MHz

420 mW

Real-World Applications

Real-World Applications

- Agriculture
- Healthcare
- Industry
- Environment

Real-World Applications

Agriculture

Healthcare

Industry

Environment

Classifying mosquito wingbeat sound using TinyML

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

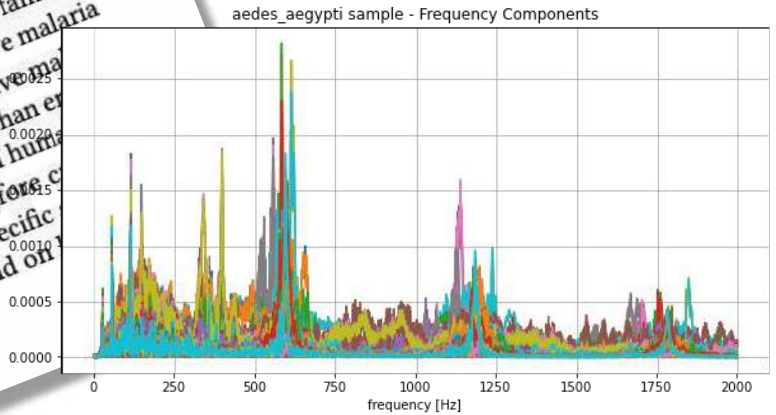
Marcelo Rovai
Universidade Federal de Itajubá
Itajubá, Brazil
rovai@unifei.edu.br

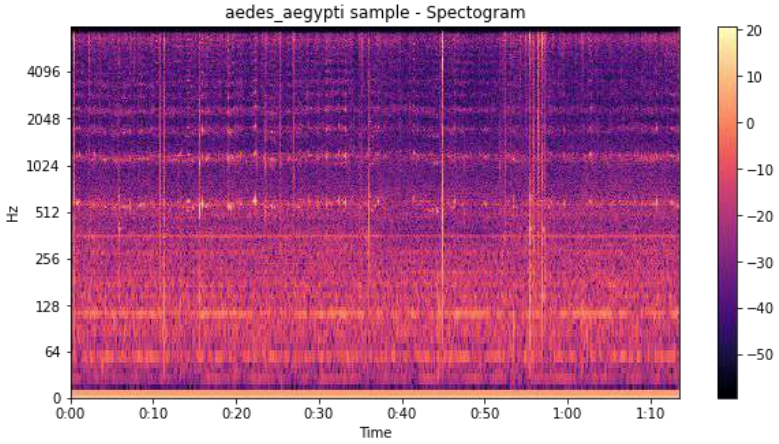
Marco Zennaro
ICTP
Trieste, Italy
mzennaro@ictp.it

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 malaria have much lower income levels than those that don't have ma... Countries affected by malaria turn to control rather than er... 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.





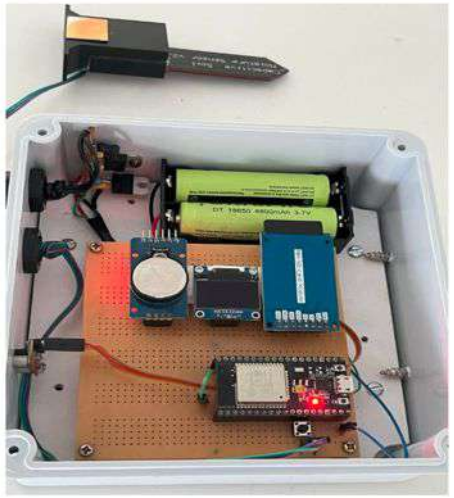
https://www.edgeimpulse.com/blog/the-world-is-all-abuzz-about-tinyml

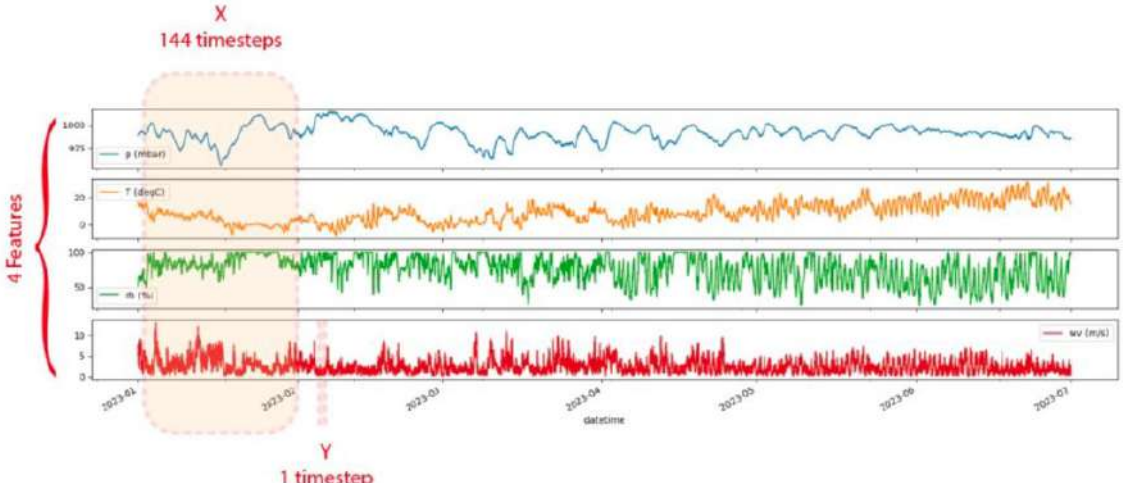
Coffee Disease Classification

João Vitor Yukio Bordin Yamashita
Mestre - UNIFEI

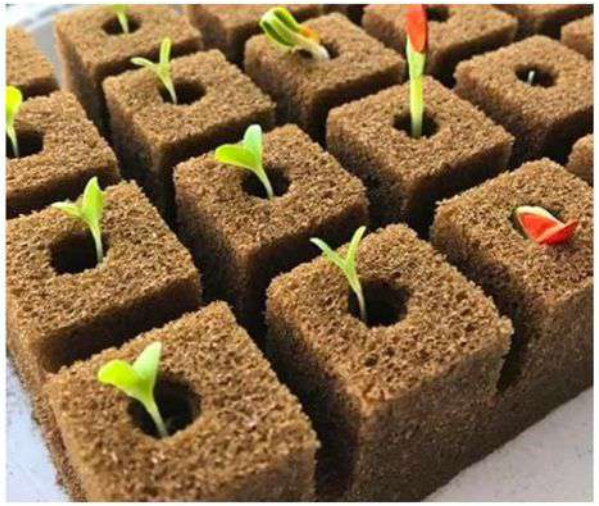
https://www.hackster.io/Yukio/coffee-disease-classification-with-ml-b0a3fc

LSTM Phenolic Sponge Moisture



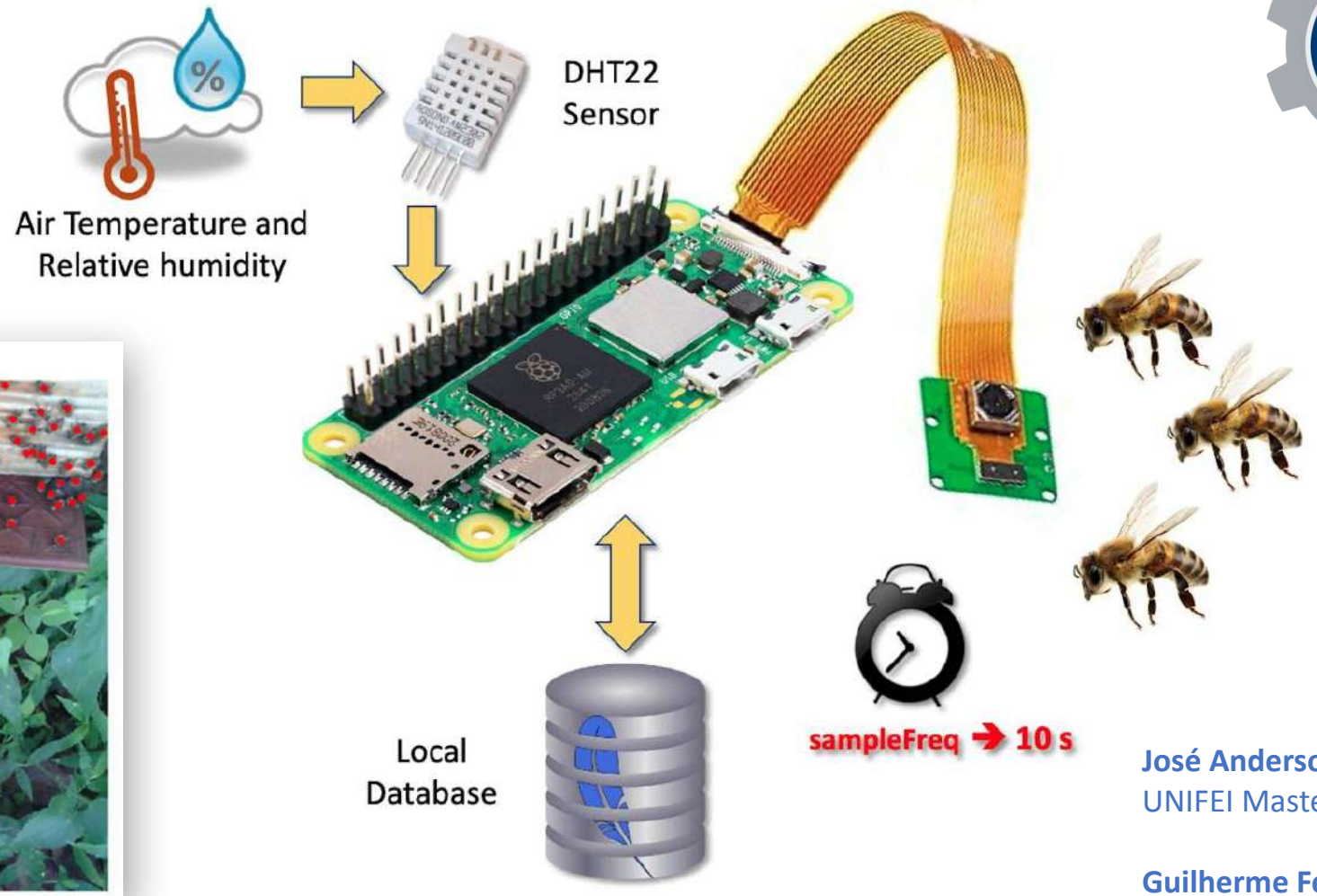


FURG UTEC UNRaf



Gustavo Pablo Castro Abdallah
PRIA

Bee Counting



José Anderson Reis
UNIFEI Master's Student

Guilherme Fernandes
UNIFEI Grad Student

Ant Detection

FURG UTEC UNRaf

Juan Camilo Abedala
PRIA

Real-World Applications

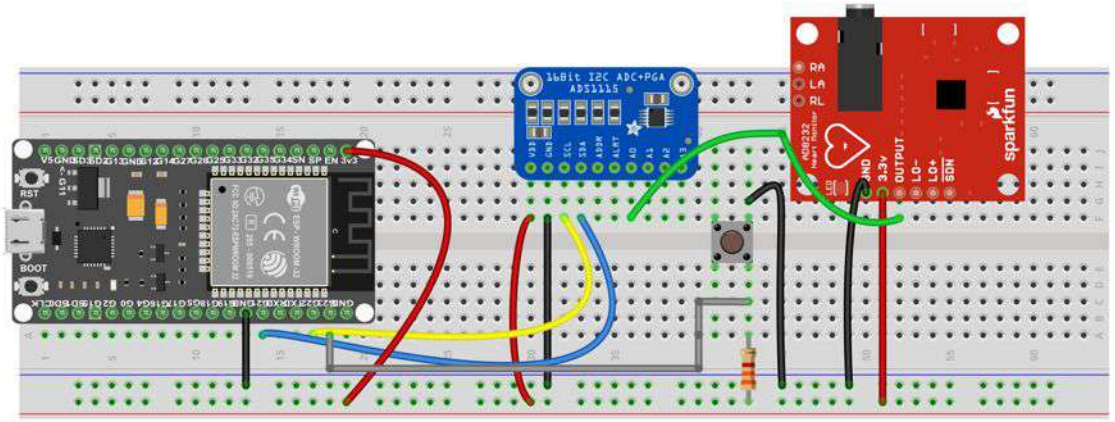
Agriculture

Healthcare

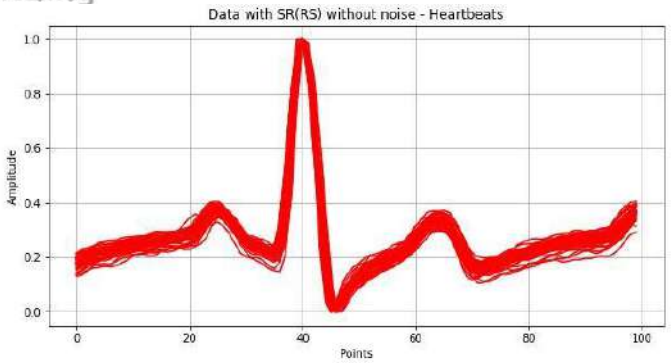
Industry

Environment

AD8232 - Single Lead Heart Rate Monitor



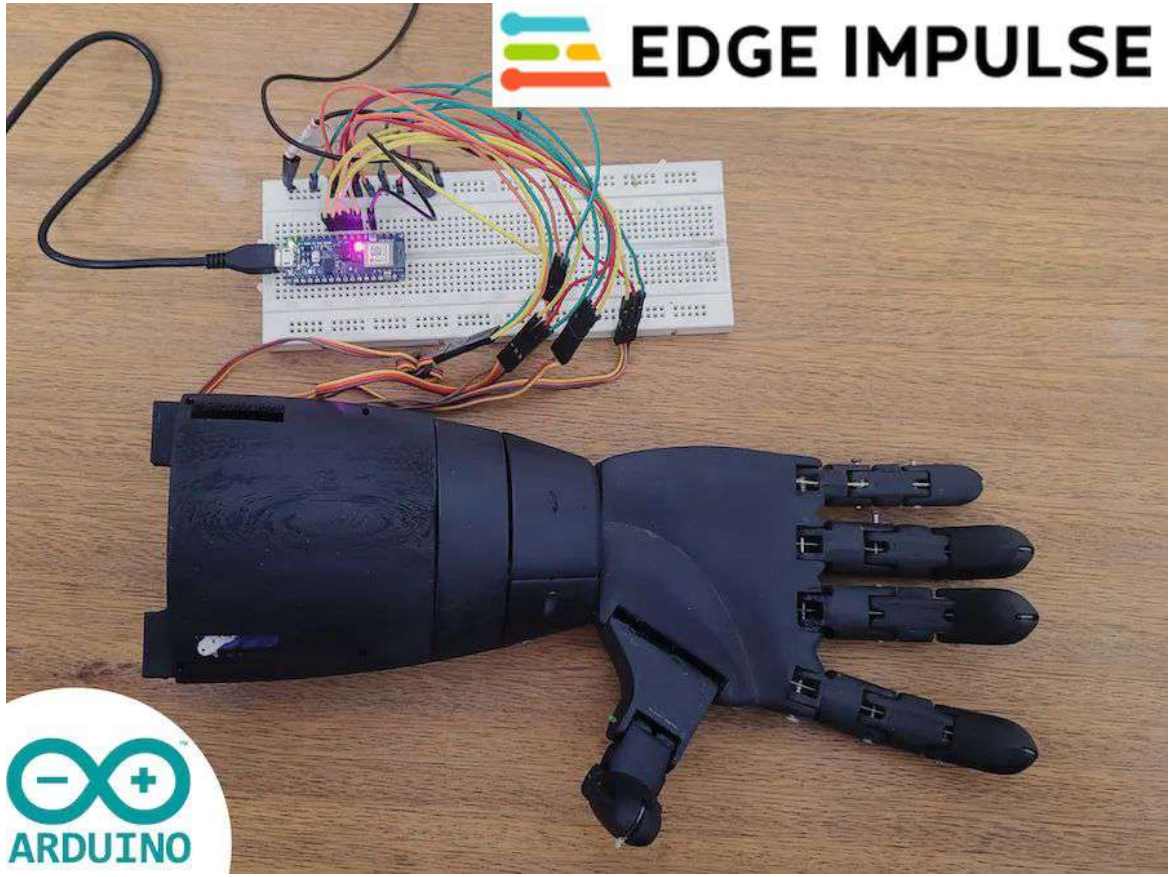
fritzing



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

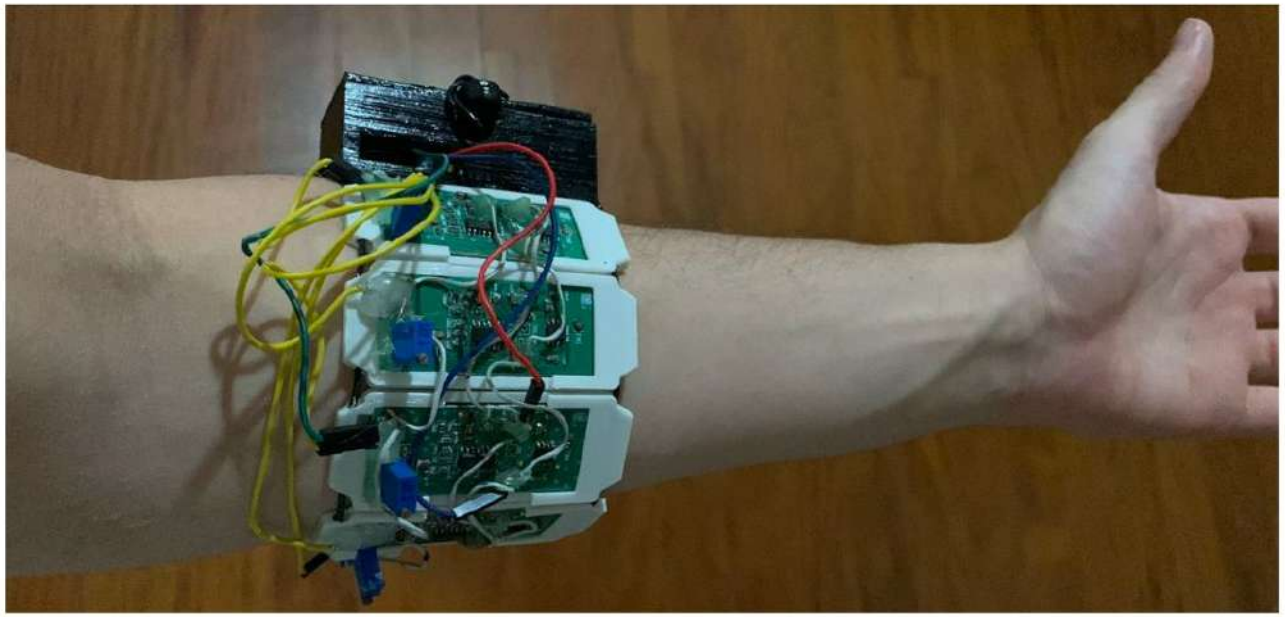
Guilherme Silva
Matheus Lima
Engenheiros - UNIFEI

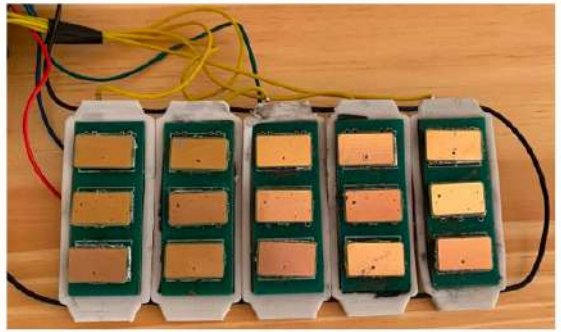
Bionic Hand Voice Commands

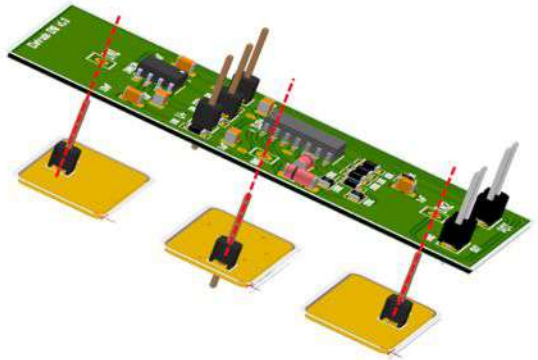


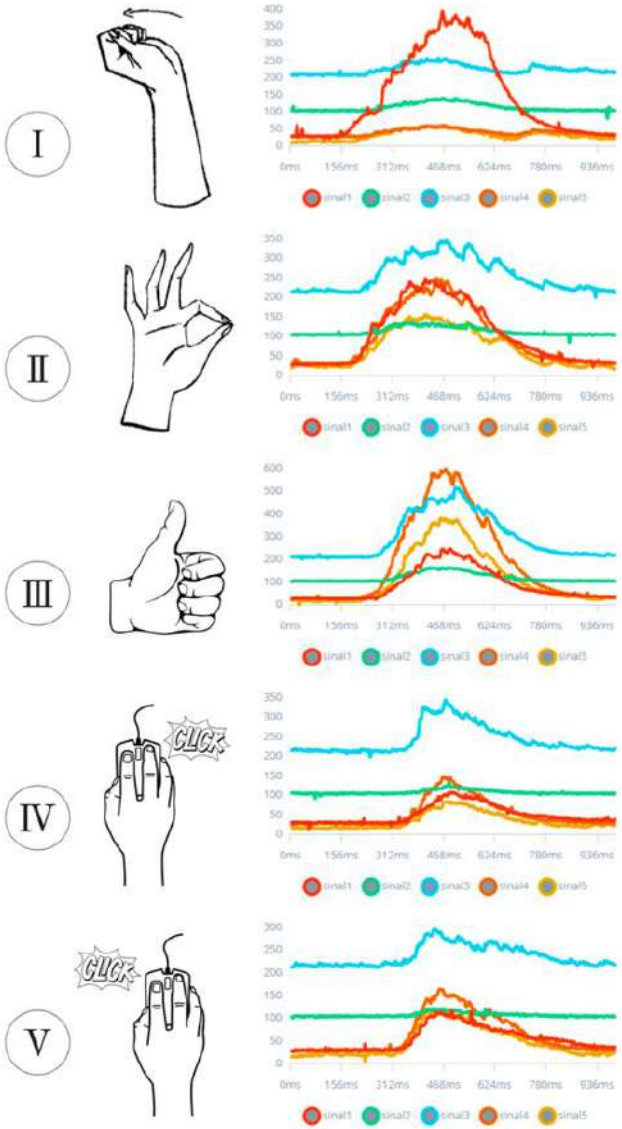
https://www.hackster.io/ex-machina/bionic-hand-voice-commands-module-w-edge-impulse-arduino-aa97e3

Surface electromyography



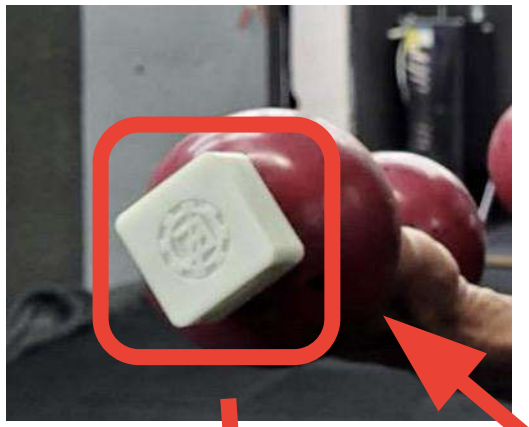


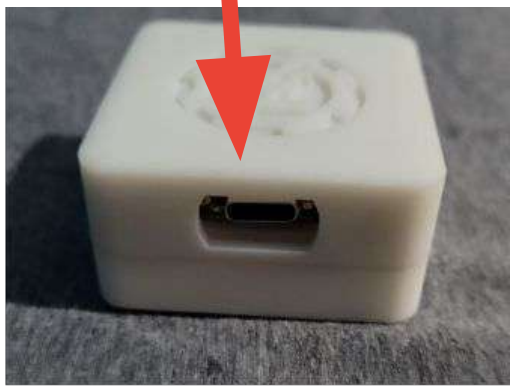


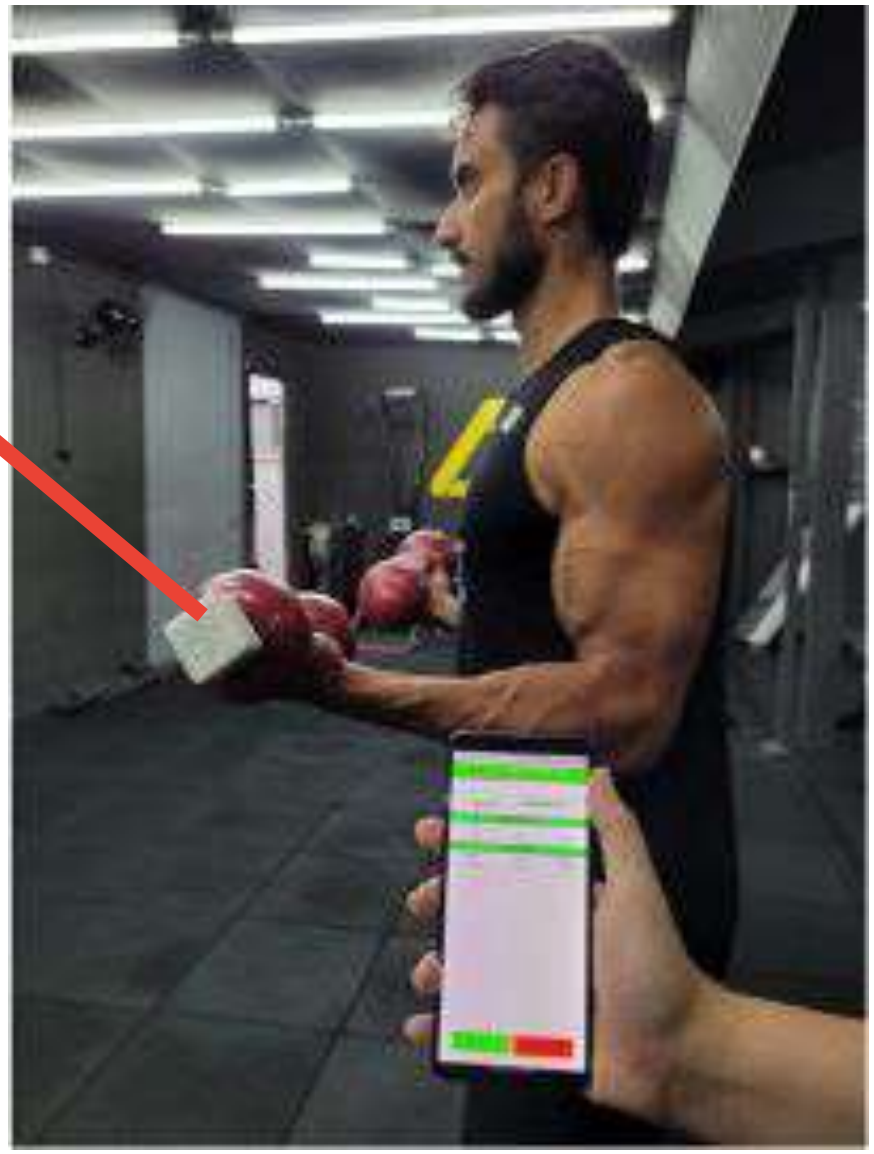


Mateus F. Delangélica e Renato M. Neto, UNIFEI

Personal Trainer

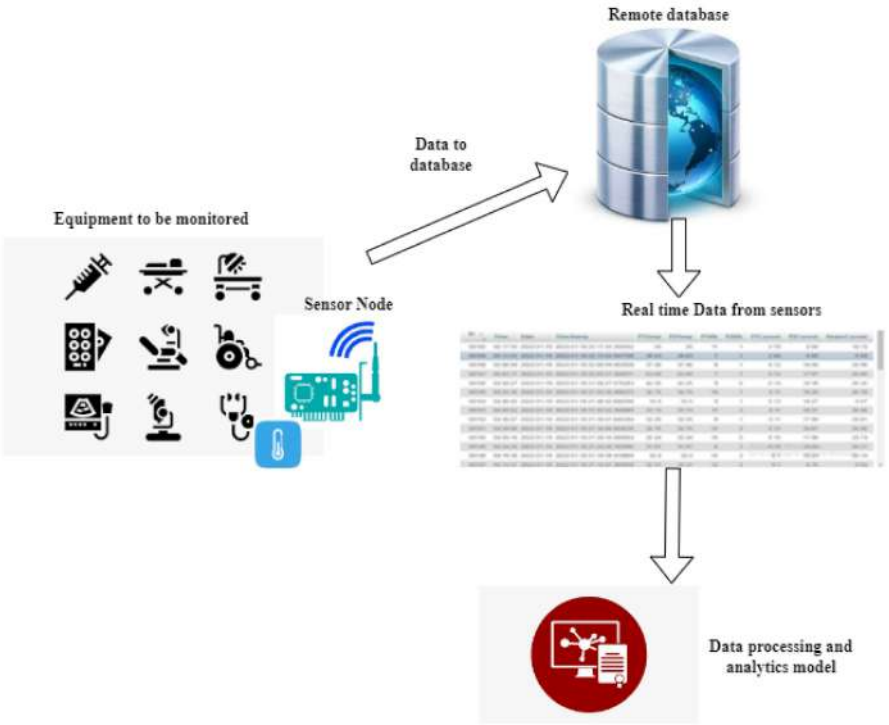






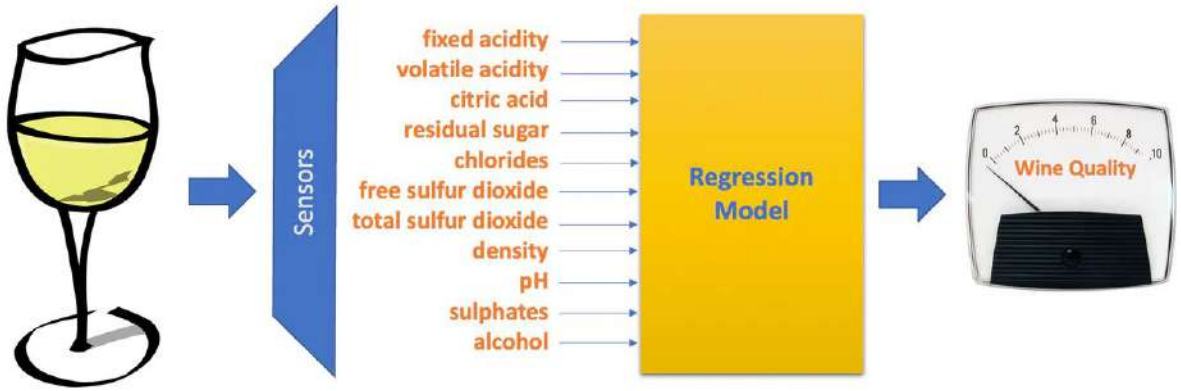
Ricardo Magno C. Junior
Luiz Fernando Kikuchi
UNIFEI

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]

Real-World Applications

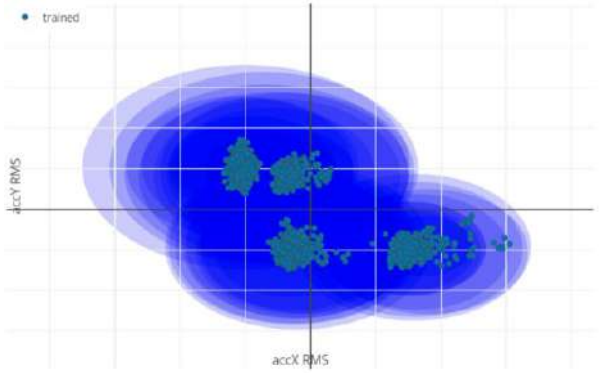
Agriculture

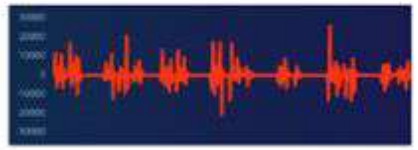
Healthcare

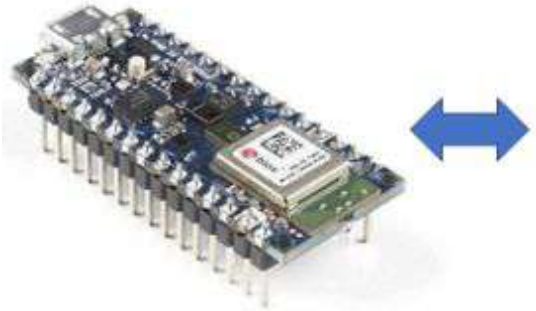
Industry

Environment

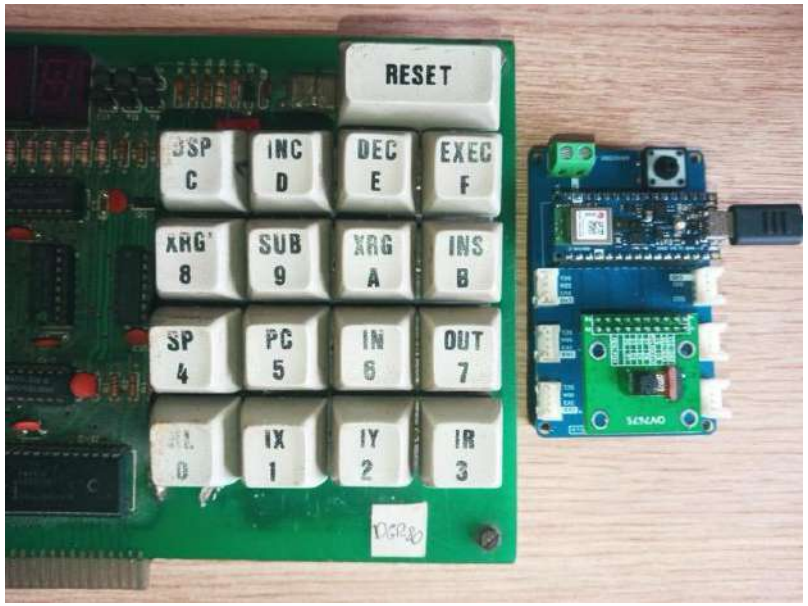
Vibration – Anomaly Detection



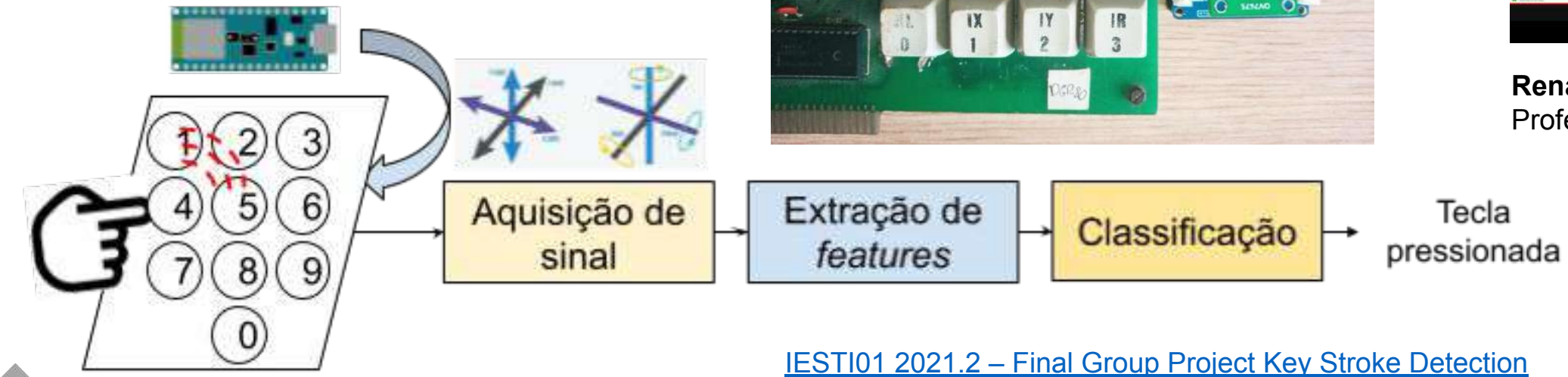




Keystroke **Sound** Detection



Renam Castro
Professor IFESP



[IESTI01 2021.2 – Final Group Project Key Stroke Detection]

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

Real-World Applications

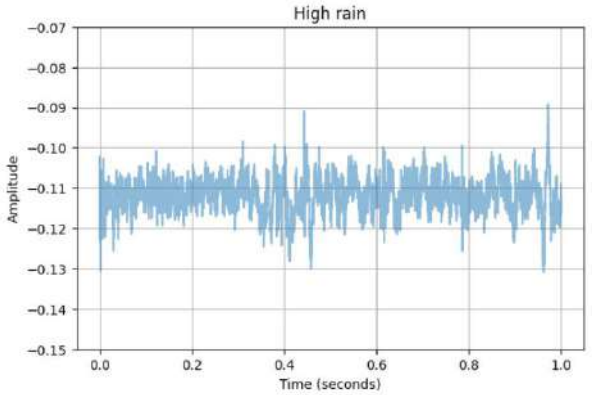
Agriculture

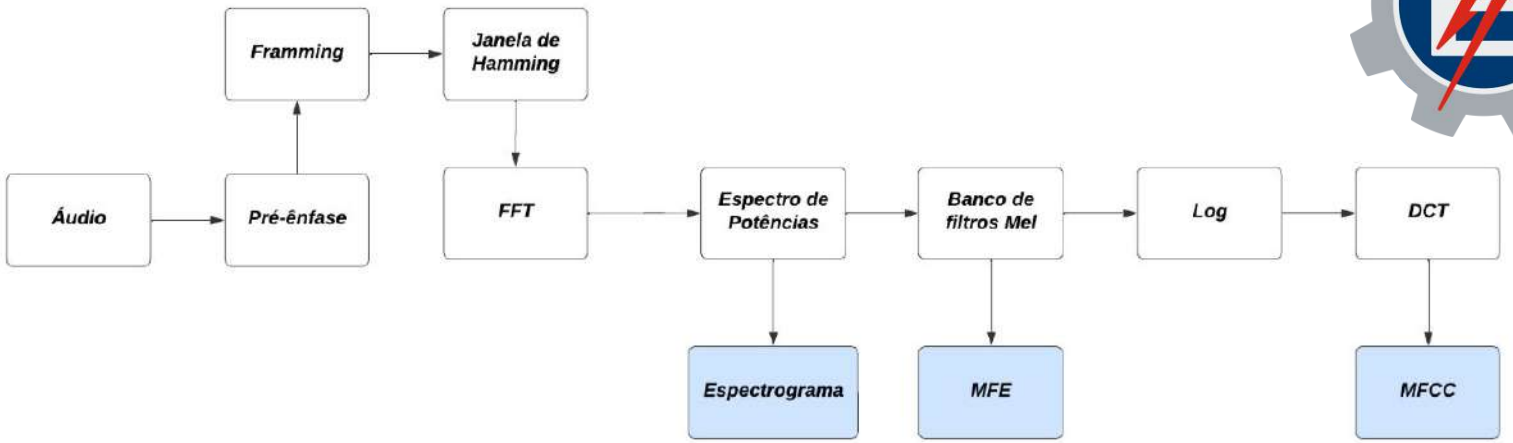
Healthcare

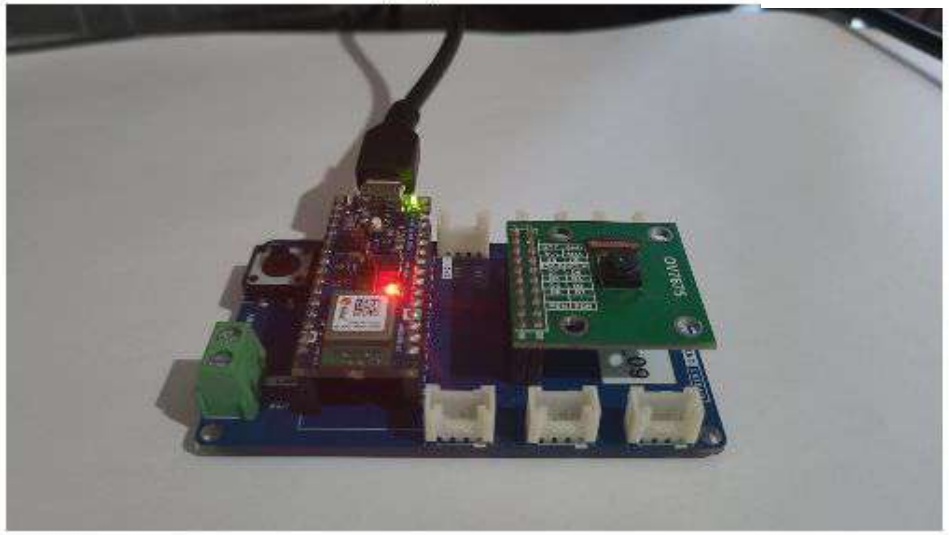
Industry

Environment

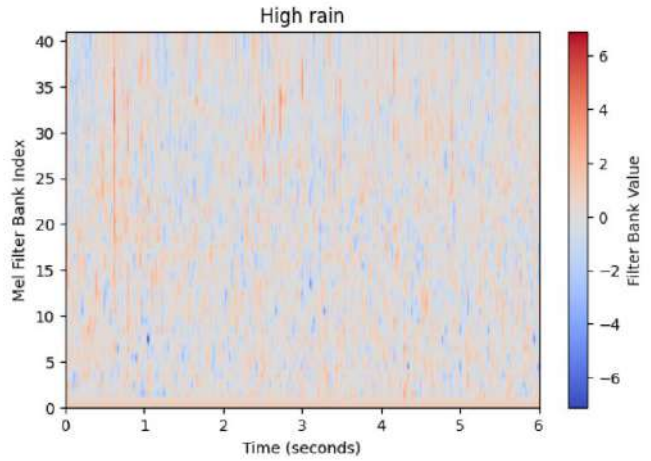
Measure rainfall using sound detection







```
Edge Impulse standalone inferencing
run_classifier returned: 0
Timing: DSP 630 ms, inference 47 ms,
Predictions:
  HighRain: 0.99609
  LowRain: 0.00000
  MediumRain: 0.00000
  NoRain: 0.00000
```



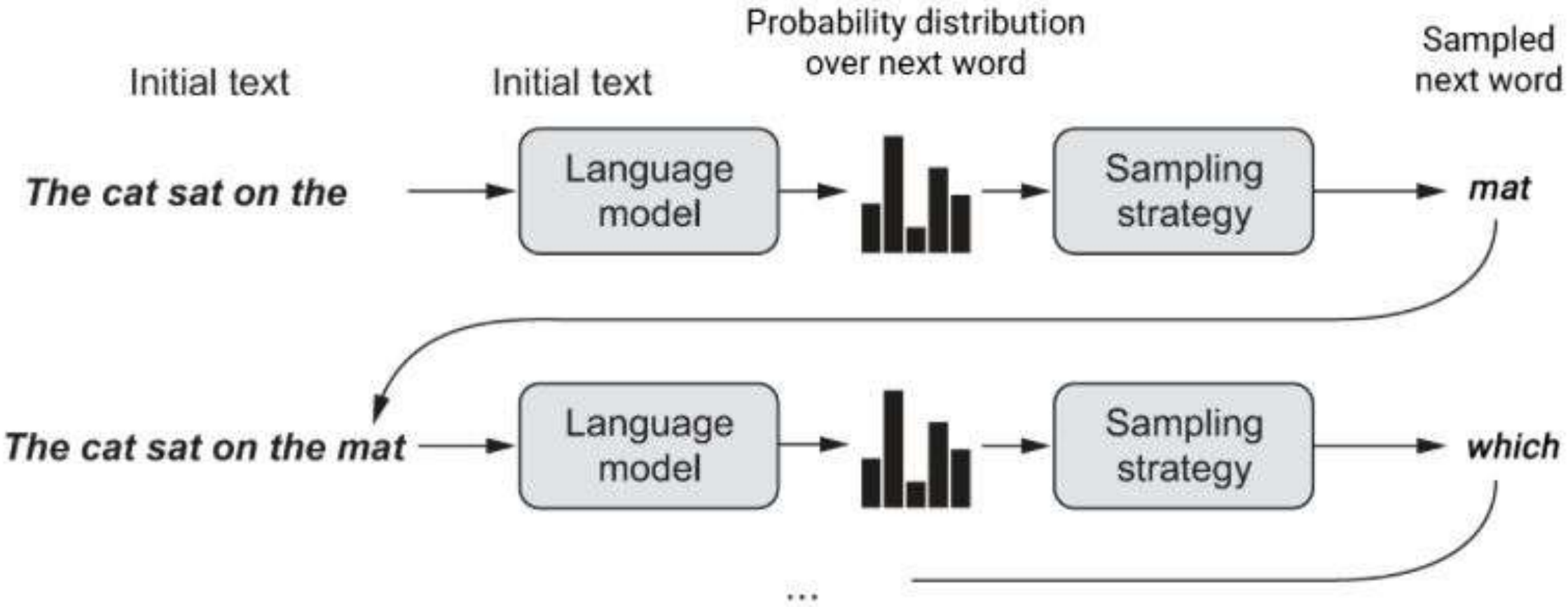
Generative AI at the Edge

Generative AI (GenAI)

Generative AI is an artificial intelligence system capable of creating new, original content across various mediums such as text, images, audio, and video. These systems learn patterns from existing data and use that knowledge to generate novel outputs that didn't previously exist.

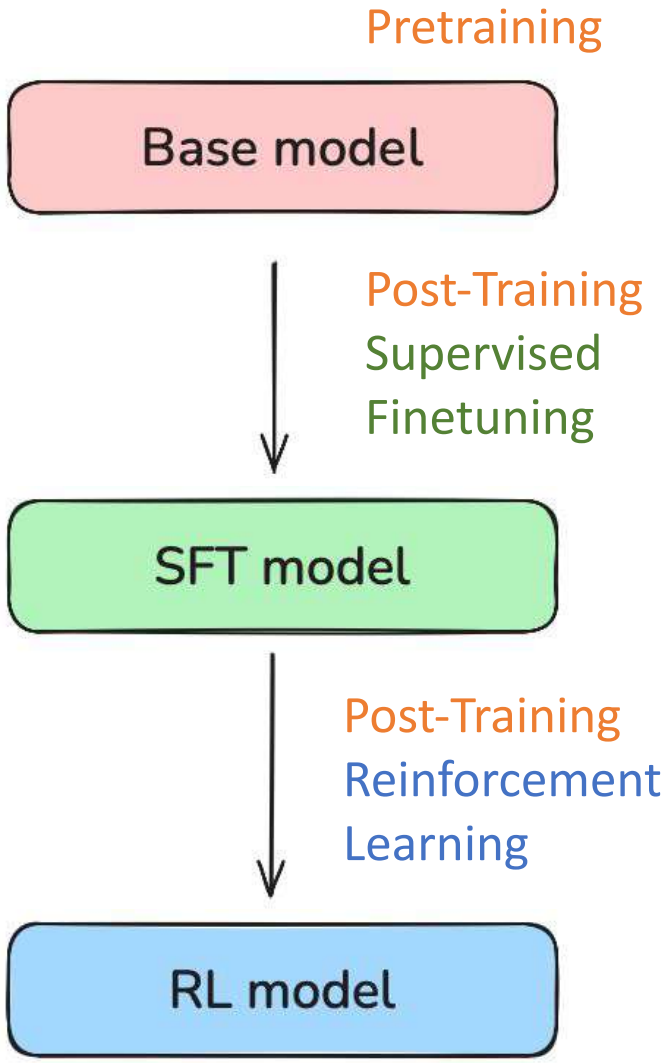
When used for generative tasks, Large Language Models (LLMs), Small Language Models (SLMs), and Visual-Language Models (VLMs) can all be considered types of GenAI.

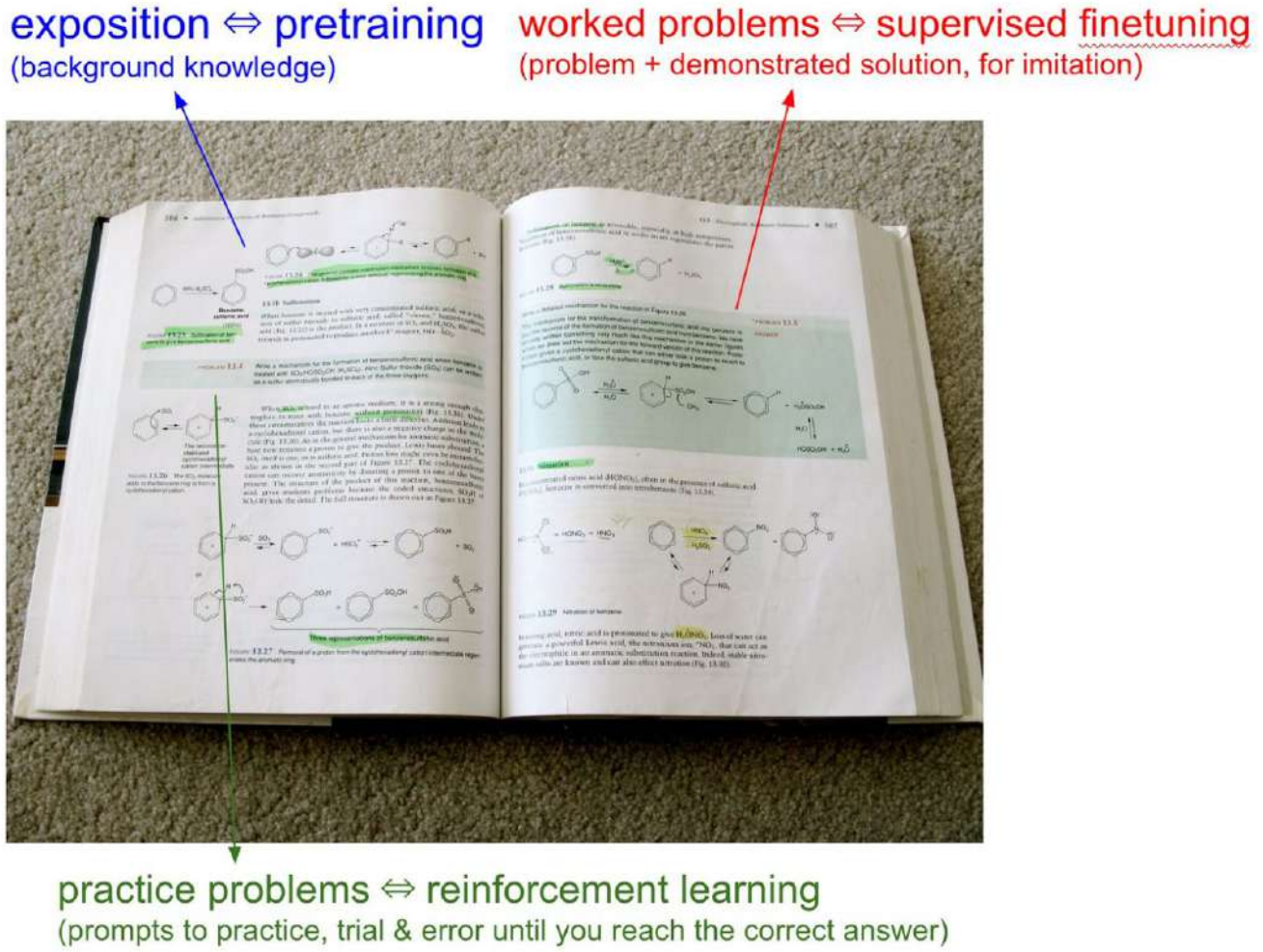
Language Models



https://platform.openai.com/tokenizer

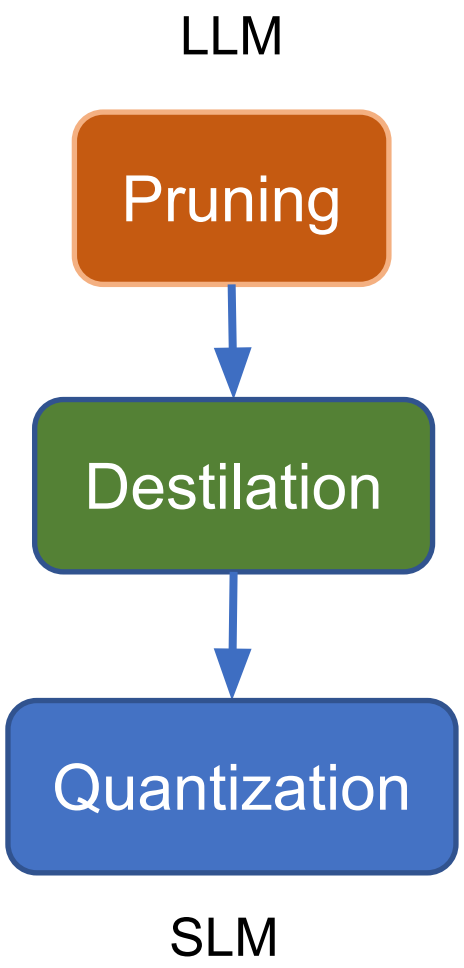
Large Language Model (LLM) Training

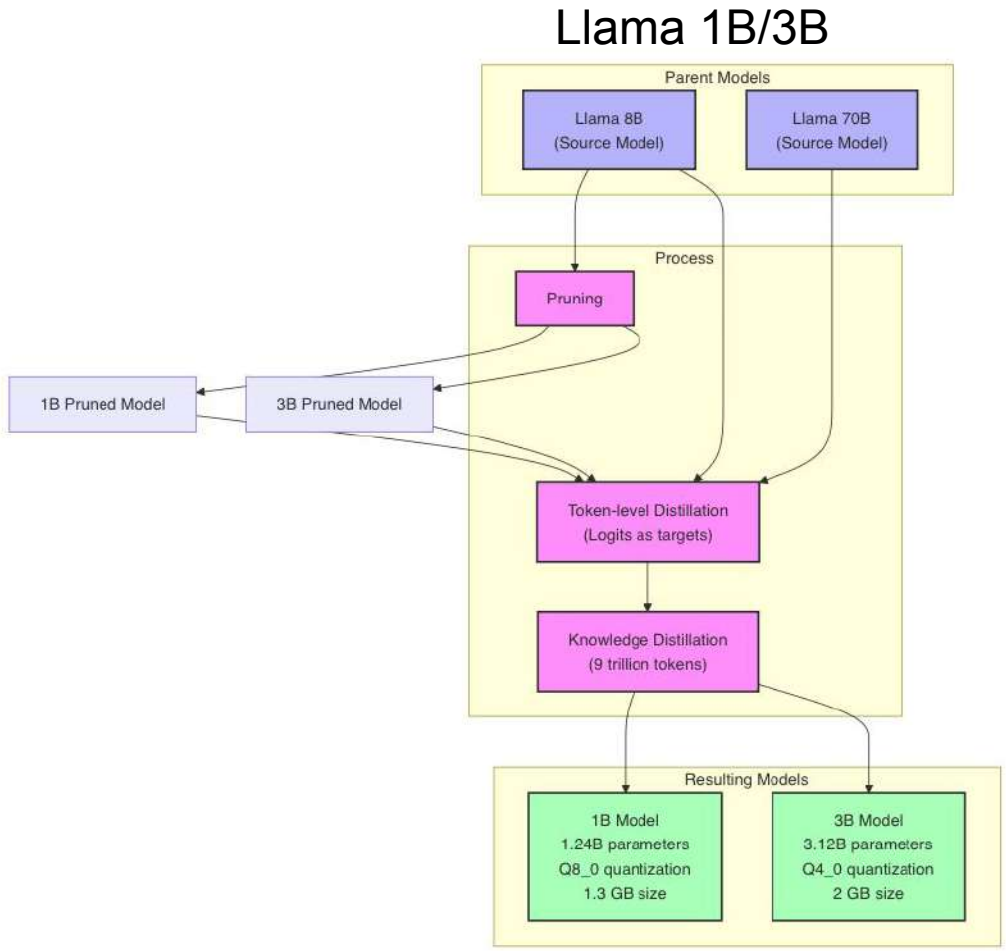




Deep Dive into LLMs like ChatGPT

Small Language Models (SLMs)



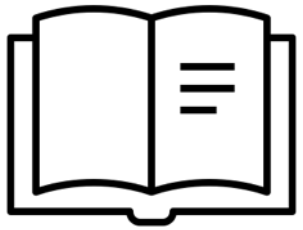


Microsoft Phi-3

1 word = ~ 1.4 token

~ 300 words/page

~ 350 pages



= 147K tokens

A 4-bit quantized 3.8 billion parameter* language model trained on 3.3 trillion tokens**, whose overall performance, as measured by both academic benchmarks and internal testing, rivals that of models such as Mixtral 8x7B and GPT-3.5

* Needs 2.4 GB of RAM

** Equivalent to 22.5 million books - 17% of all books written in the world

GenAi at the Edge: Llama3.2:1B

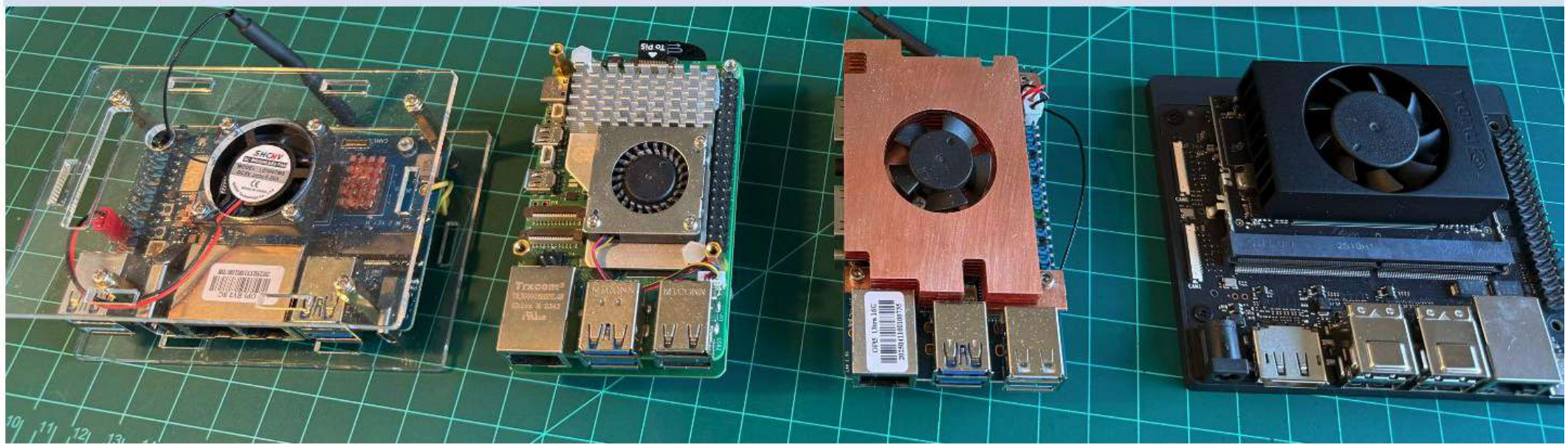
Desktop Reference

PC Linux (i7): 20 tokens/s

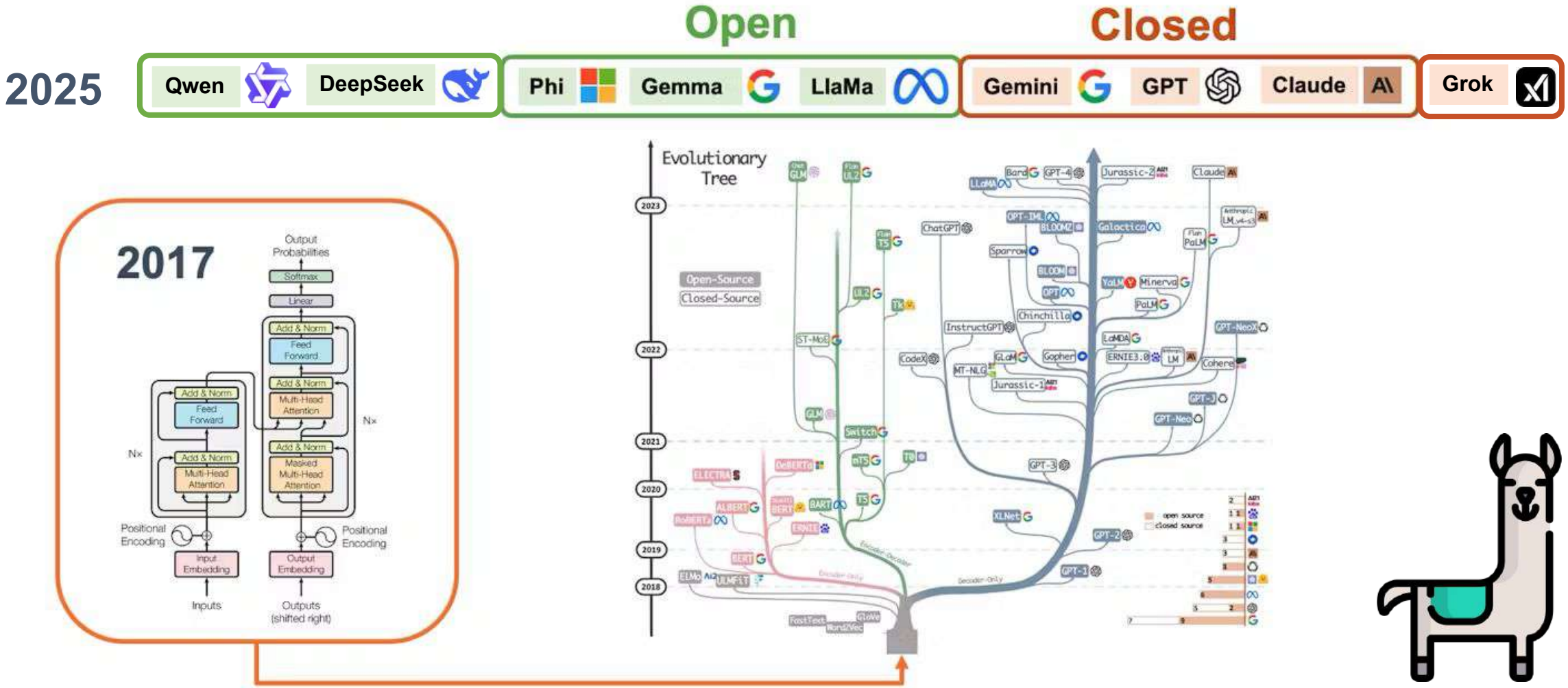
Mac (M1-Pro): 111 tokens/s

(*) Running on GPU

Orange Pi RV2	Raspi-5	Orange Pi 5 Ultra	Jetson Orin Nano
1 token/s	7.5 tokens/s	12 tokens/s	26 tokens/s (*)



Transformers to LLMs and SLMs



Small Models

LlaMa

Gemma

```
marcelo_rovai — mjrovai@raspi-5: ~ — ssh mjrovai@192.168.4.209 — 74×12
(ollama) mjrovai@raspi-5:~ $ ollama run llama3.2:3b --verbose
>>> What is the capital of France?
The capital of France is Paris.

total duration:       1.808927736s
load duration:        39.854862ms
prompt eval count:    32 token(s)
prompt eval duration: 221.506ms
prompt eval rate:     144.47 tokens/s
eval count:           8 token(s)
eval duration:        1.506376s
eval rate:            5.31 tokens/s
```

```
marcelo_rovai — mjrovai@raspi-5: ~ — ssh mjrovai@192.168.4.209 — 67×13
(ollama) mjrovai@raspi-5:~ $ ollama run gemma2:2b --verbose
>>> What is the capital of France?
The capital of France is **Paris**.


total duration:       4.373339337s
load duration:        48.129697ms
prompt eval count:    16 token(s)
prompt eval duration: 1.968114s
prompt eval rate:     8.13 tokens/s
eval count:           13 token(s)
eval duration:        2.313284s
eval rate:            5.62 tokens/s
```

llava-phi-3 (2.9 GB)

Ollama

```
mjrovai@rpi-5: ~/Documents/OLLAMA
Help
ute.

/Documents/OLLAMA $
/Documents/OLLAMA $ python calc_distance_image.py /
home/mjrovai/Documents/OLLAMA/image_test_1.jpg

 The image shows Paris, with lat:48.86 and long: 2.35, located in
France and about 11,630 kilometers away from Santiago, Chile.

 [INFO] ==> The code (running llava-phi3), took 232.60845186299412
 seconds to execute.

mjrovai@rpi-5:~/Documents/OLLAMA $
```

```
mjrovai@rpi-5: ~
File  Edit  Tabs  Help
>>> Awswer with one short sentence, what is the capital of France and its distanc
... e in Km from Santiago, Chile
 The capital of France is Paris and it is around 12,674 kilometers away
from Santiago, Chile.

total duration:       13.860074968s
load duration:        1.537039ms
prompt eval count:    27 token(s)
prompt eval duration: 5.925386s
prompt eval rate:     4.56 tokens/s
eval count:           26 token(s)
eval duration:        7.539223s
eval rate:            3.45 tokens/s
>>> Send a message (/? for help)
```

(13 seconds)

```
mjrovai@rpi-5: ~/Documents/OLLAMA
Help
/Documents/OLLAMA $
/Documents/OLLAMA $ python calc_distance_image.py
/home/mjrovai/Documents/OLLAMA/image_test_3.jpg

 The image shows Machu Picchu, with lat:-13.16 and long: -72.54,
 located in Peru and about 2,250 kilometers away from Santiago,
Chile.

 [INFO] ==> The code (running llava-phi3), took 267.579568572007
7 seconds to execute.

mjrovai@rpi-5:~/Documents/OLLAMA $
```

(4 minutes)

LLMs: Optimization Techniques

Techniques for Enhancing SLM at the Edge

Fundamentals: Optimizing Prompting Strategies

- Chain-of-Thought Prompting
- Few-Shot Learning
- Task Decomposition

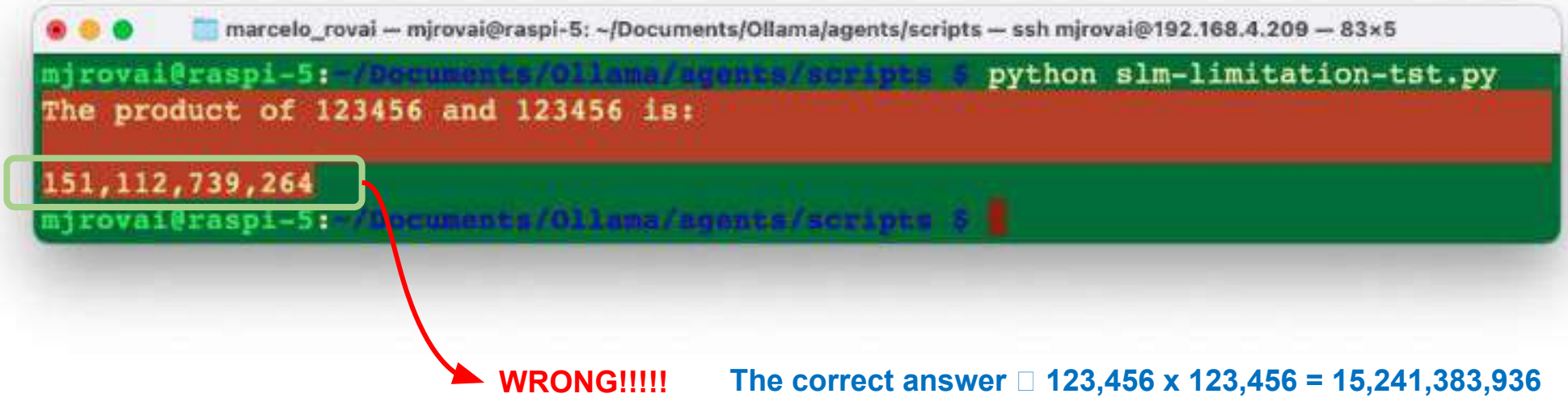
Intermediate: Building Intelligence Systems

- Building Agents with SLMs
- General Knowledge Router
- Function Calling
- Response Validation

Advanced: Extending Knowledge and Specialization

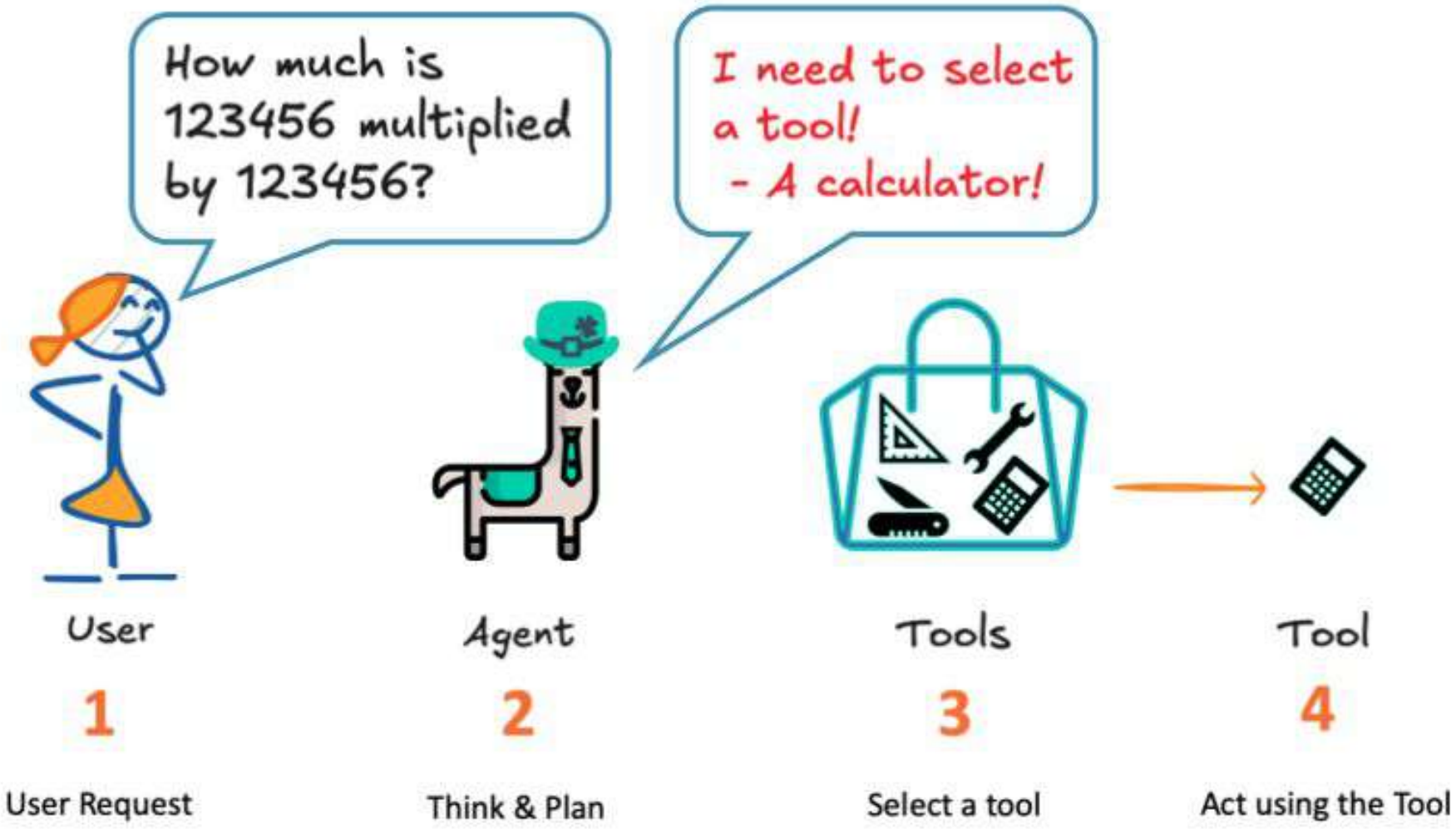
- Retrieval-Augmented Generation (RAG)
- Fine-Tuning for Domain Specialization

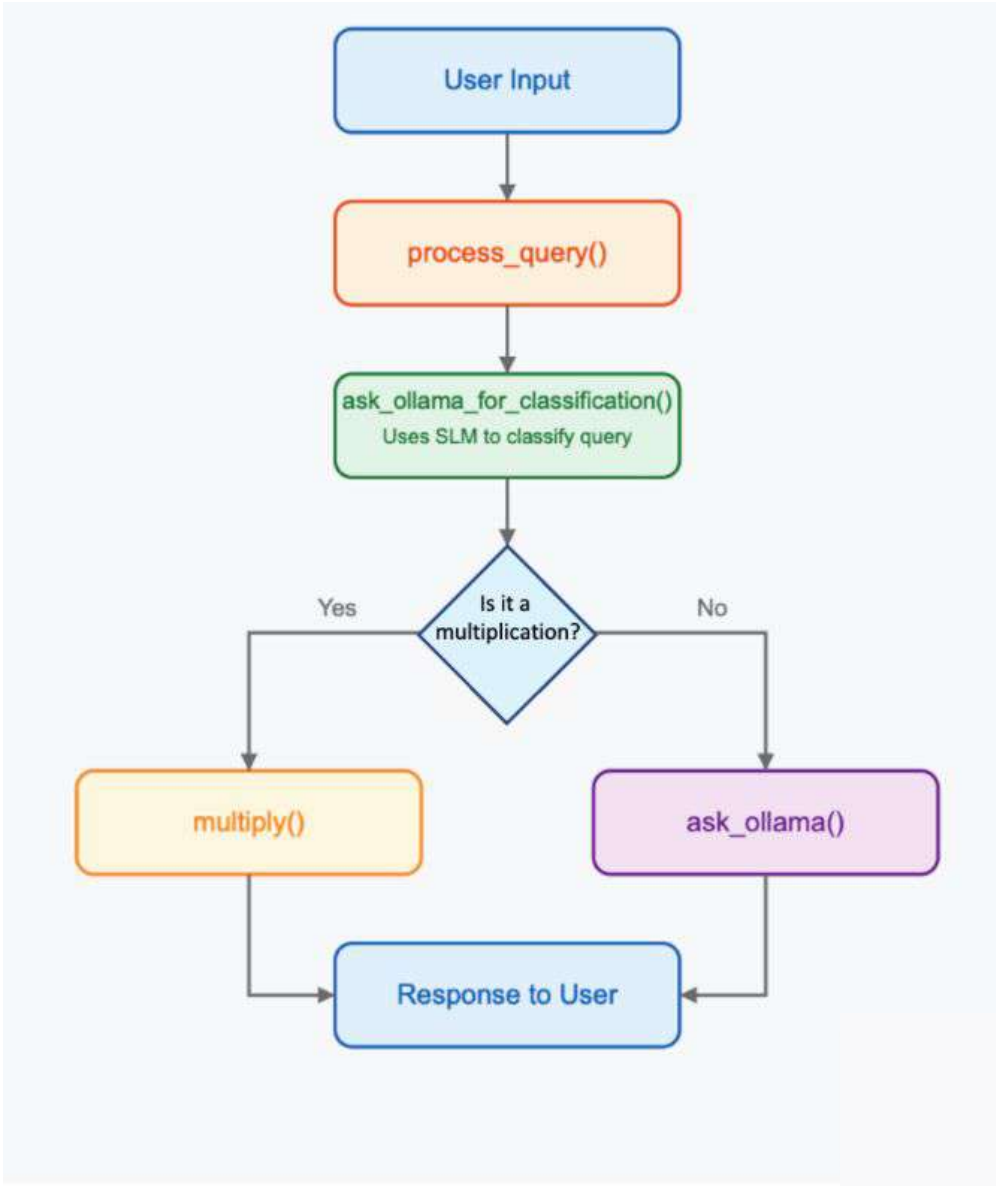
```python
import ollama

response = ollama.generate(
    model="llama3.2:3b",
    prompt="Multiply 123456 by 123456"
)
print(response['response'])
```



WRONG!!!!! The correct answer □ 123,456 x 123,456 = 15,241,383,936

Agents


How much is 123456 multiplied by 123456?
I need to select a tool! - A calculator!
User
Agent
Tools
Tool
1
2
3
4
User Request
Think & Plan
Select a tool
Act using the Tool

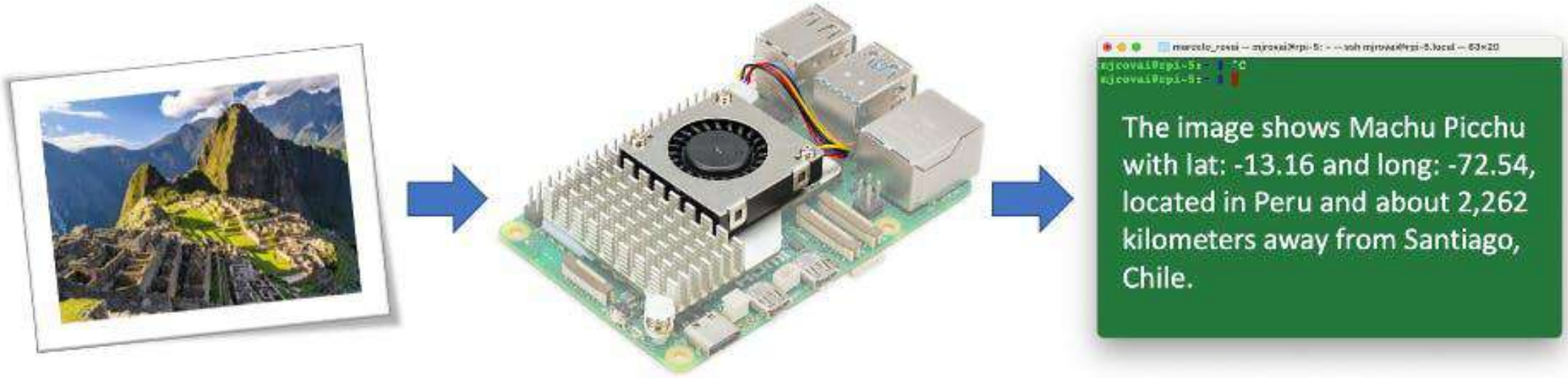


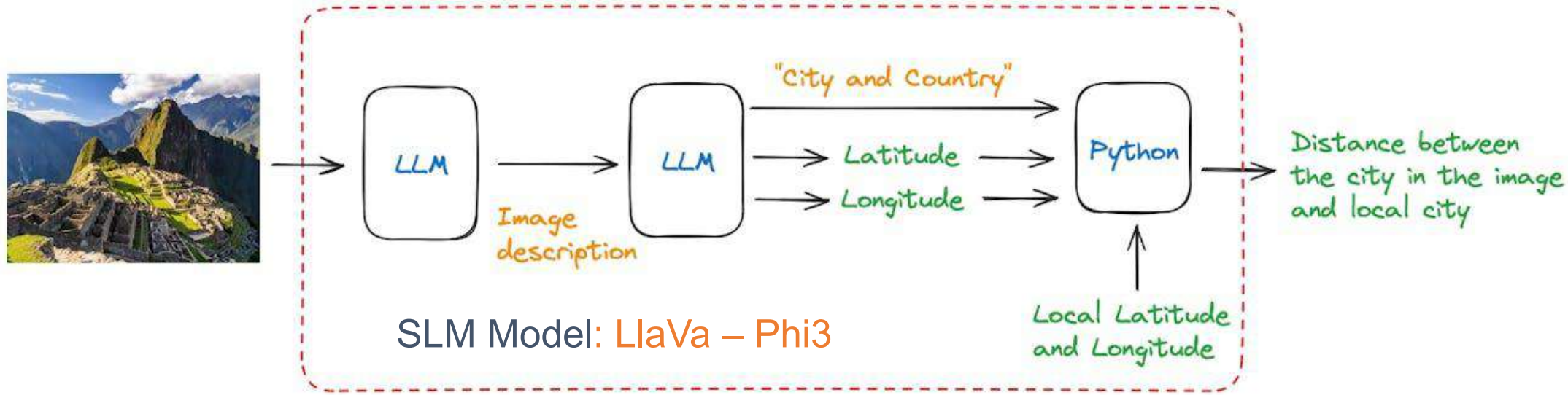

```
marcelo_rovai — mjrovai@raspi-5: ~/Documents/Ollama/agents/scripts — ssh mjrovai@192.168.4.209 — 83×14
mjrovai@raspi-5:~/Documents/Ollama/agents/scripts $ python 2-simple_agent.py
Ollama Agent (Type 'exit' to quit)
----------------------------------

You: Multiply 123456 by 123456
Sending classification request to Ollama
Classification response: {
  "type": "multiplication",
  "numbers": [123456, 123456]
}
Ollama classification: {'type': 'multiplication', 'numbers': [123456, 123456]}

Agent: The product of 123456 and 123456 is 15241383936.
```

It is correct □ 123,456 x 123,456 = 15,241,383,936

```
marcelo_rovai — mjrovai@raspi-5: ~/Documents/Ollama/agents/scripts — ssh mjrovai@192.168.4.209 — 83×12

You: What is the capital of Brazil?
Sending classification request to Ollama
Classification response: {
  "type": "general_question"
}
Ollama classification: {'type': 'general_question'}
Sending query to Ollama

Agent: The capital of Brazil is Brasília.

You:
```

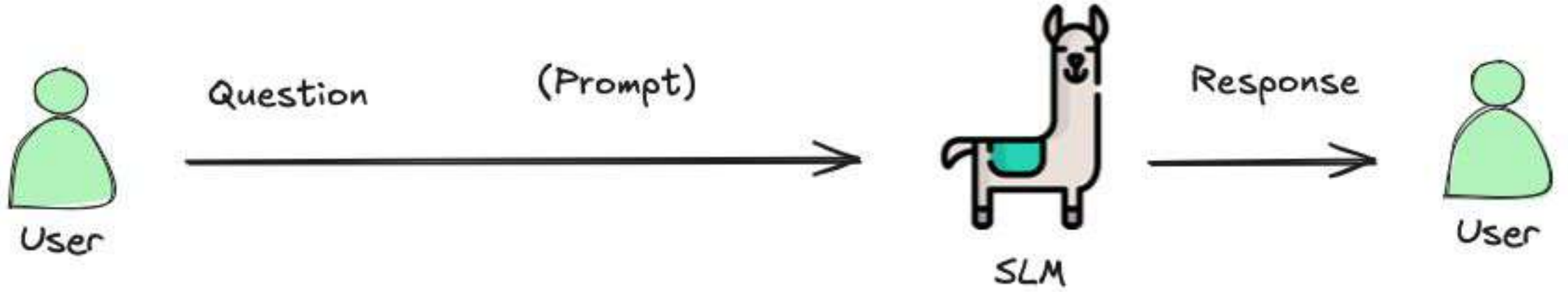
Function Calling

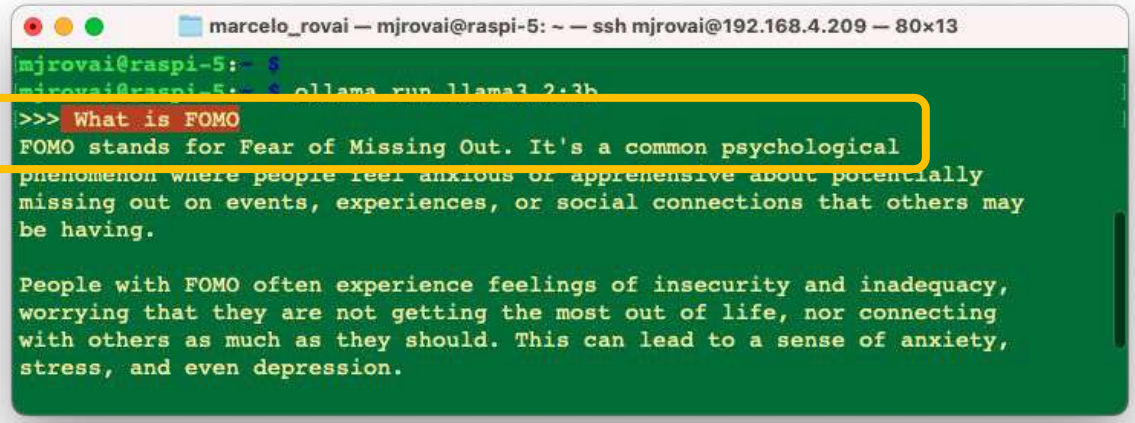




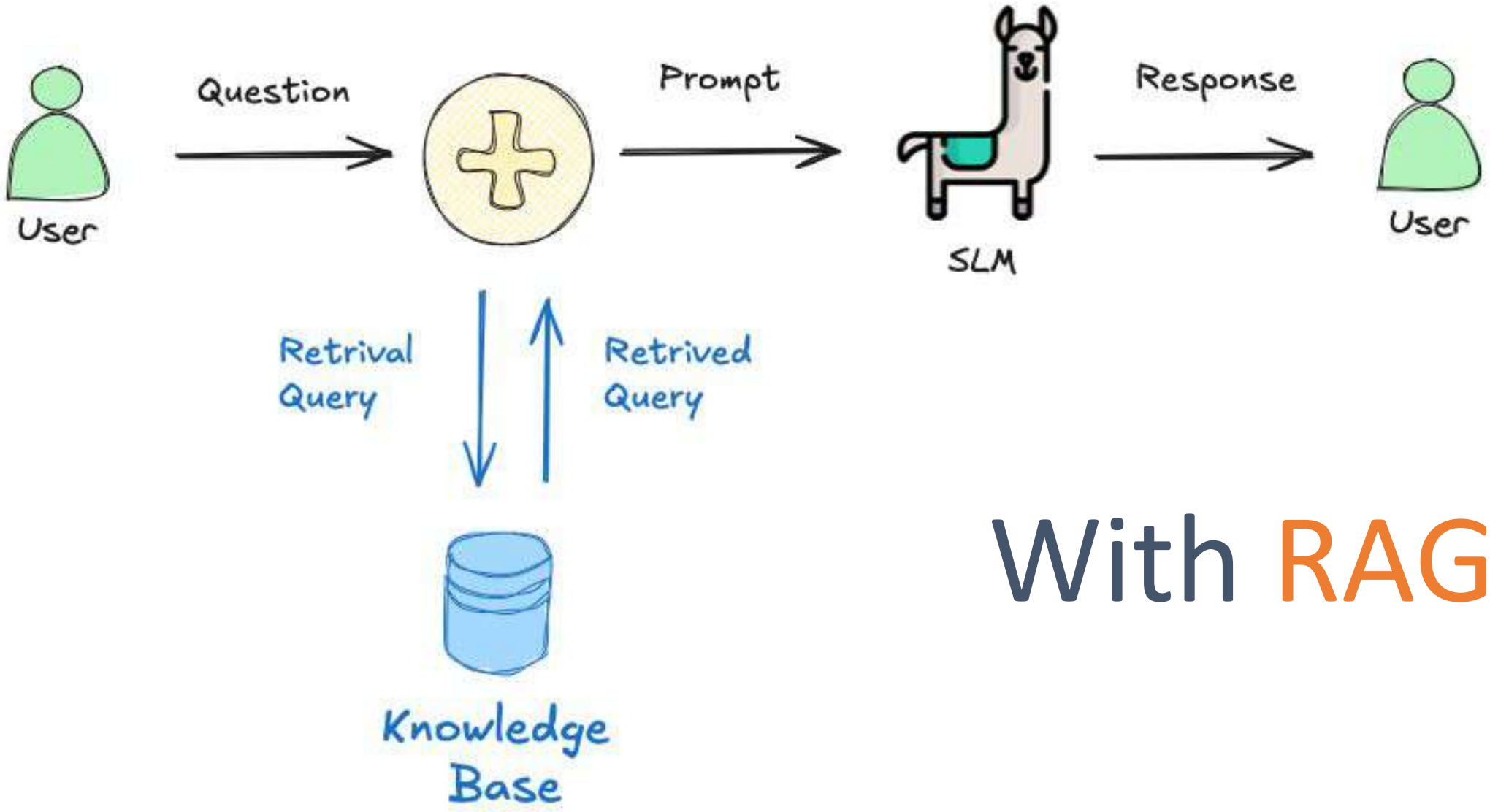
Retrieval-Augmented Generation (RAG)

“A method created by the FAIR team at Meta to enhance the accuracy of Large Language Models (LLMs) and reduce false information or “hallucinations.”

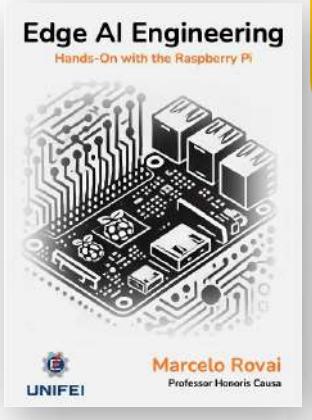




Usual Prompt


User
Question
Prompt
SLM
Response
User
Retrival Query
Retrived Query
Knowledge Base

With RAG



Knowledge Base

```
marcelo_rovai — mjrovai@raspi-5: ~/Documents/Ollama/Rag/edgeai — ssh mjrovai@192.168.4.209 — 100×36

Question: What is FOMO?
Retrieving documents...
Retrieved 2 document chunks
Generating answer...
Response latency: 107.41 seconds using model: llama3.2:3b

ANSWER:
==================================================
FOMO stands for Faster Objects, More Objects.
==================================================

Your question: How to setup a Raspi5?

Generating answer...

Question: How to setup a Raspi5?
Retrieving documents...
Retrieved 2 document chunks
Generating answer...
Response latency: 85.09 seconds using model: llama3.2:3b

ANSWER:
==================================================
To set up a Raspberry Pi 5, follow these steps:

1. Download and install the Raspberry Pi Imager on your computer.
2. Insert a 32GB microSD card into your computer.
3. Open Raspberry Pi Imager and select "Raspberry Pi OS (64-bit)" as the operating system.
4. Select the Raspberry Pi 5 model.
5. Set the hostname, username, password, configure WiFi, and enable SSH in the advanced options.
6. Write the image to the microSD card.

Note: The full 64-bit version of Raspberry Pi OS is recommended for the Raspberry Pi 5.
==================================================

Your question:
```

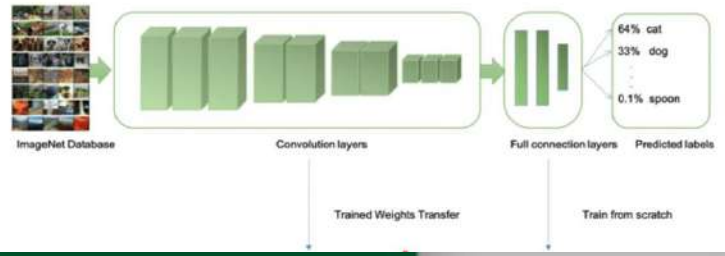
Training a FOMO Model at Edge Impulse Studio

The inference with the SSD MobileNet model worked well, but the latency was significantly high. The inference varied from 0.5 to 1.3 seconds on a Raspi-Zero, which means around or less than 1 FPS (1 frame per second). One alternative to speed up the process is to use FOMO (Faster Objects, More Objects).

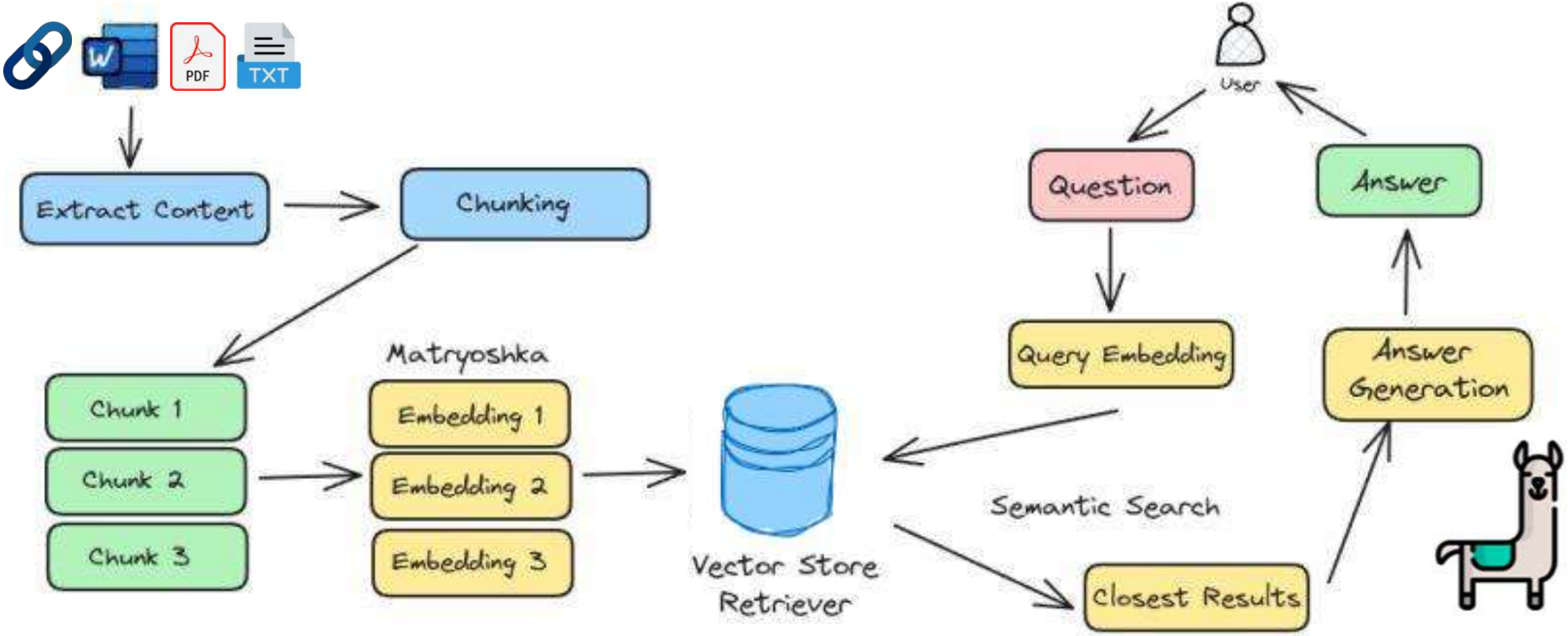
This novel machine learning algorithm lets us count multiple objects and find their location in an image in real-time using up to 30x less processing power and memory than MobileNet SSD or YOLO. The main reason this is possible is that while other models calculate the object's size by drawing a square around it (bounding box), FOMO ignores the size of the image, providing only the information about where the object is located in the image through its centroid coordinates.

How FOMO works?

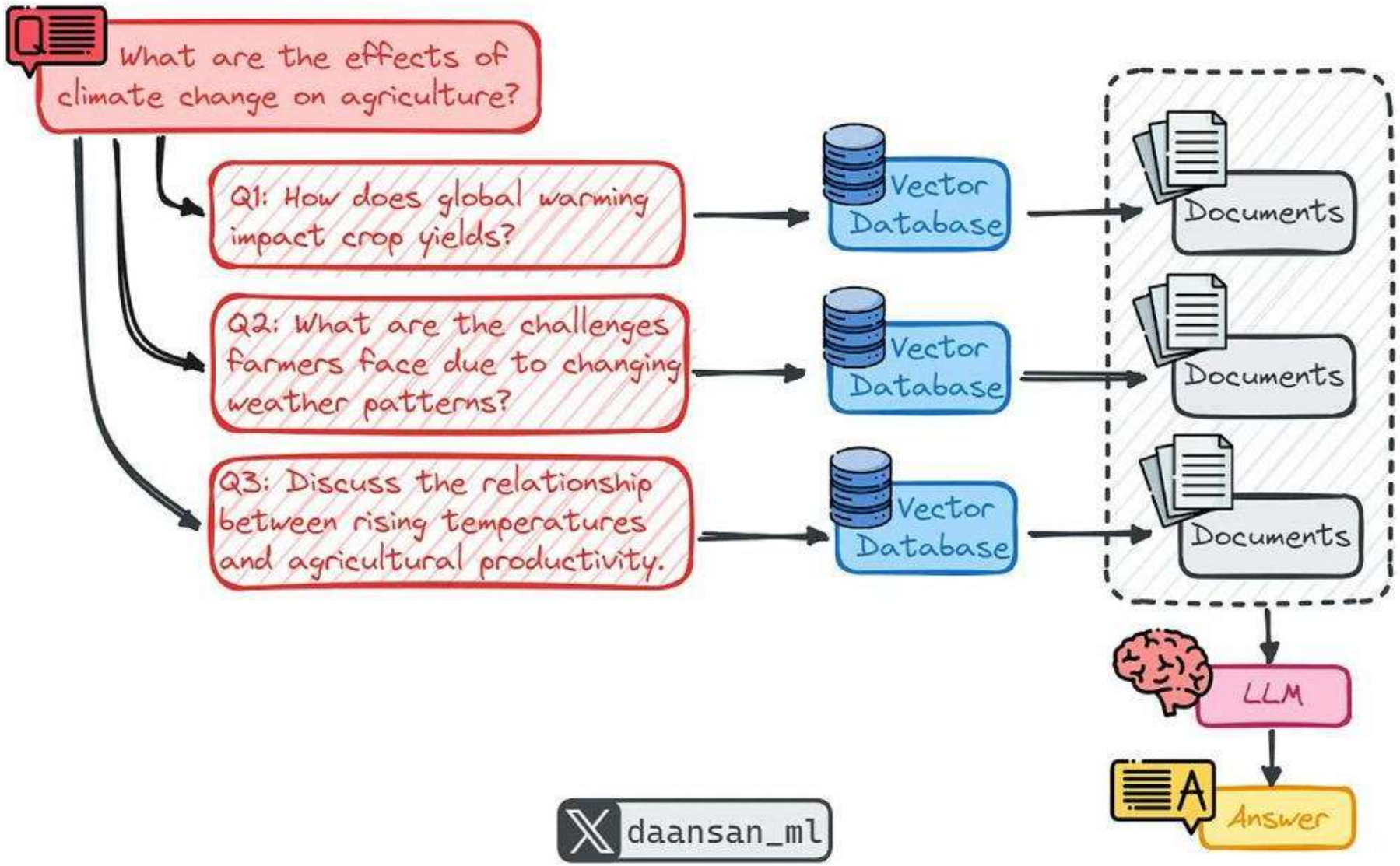
In a typical object detection pipeline, the first stage is extracting features from the input image. **FOMO leverages MobileNetV2 to perform this task**. MobileNetV2 processes the input image to produce a feature map that captures essential characteristics, such as textures, shapes, and object edges, in a computationally efficient way.



RAG: Simple Query



Advanced RAG: Multi Query



VLM

Vision Language Models

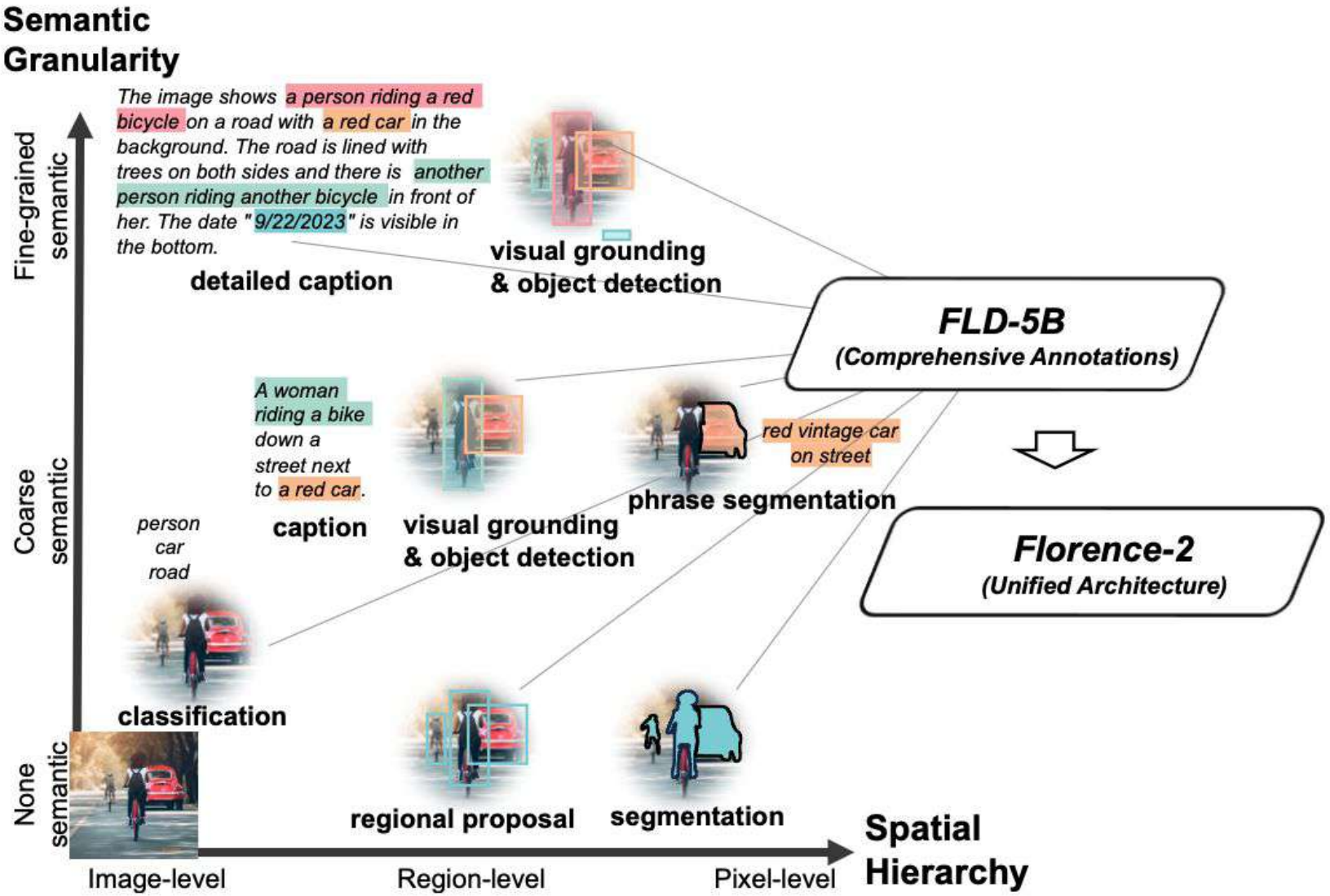
Vision Language Models (VLMs) are artificial intelligence systems integrating computer vision and natural language processing capabilities. This fusion enables them to process, understand, and generate both visual (images, videos) and textual data, allowing for a wide range of multimodal tasks that require joint reasoning across these domains.

Florence-2 stands out for its zero-shot performance, compactness, and ability to handle multiple vision-language tasks without extensive fine-tuning. It is particularly well-suited for scenarios where rapid deployment and efficiency are crucial, such as edge devices or when a unified model is preferred.

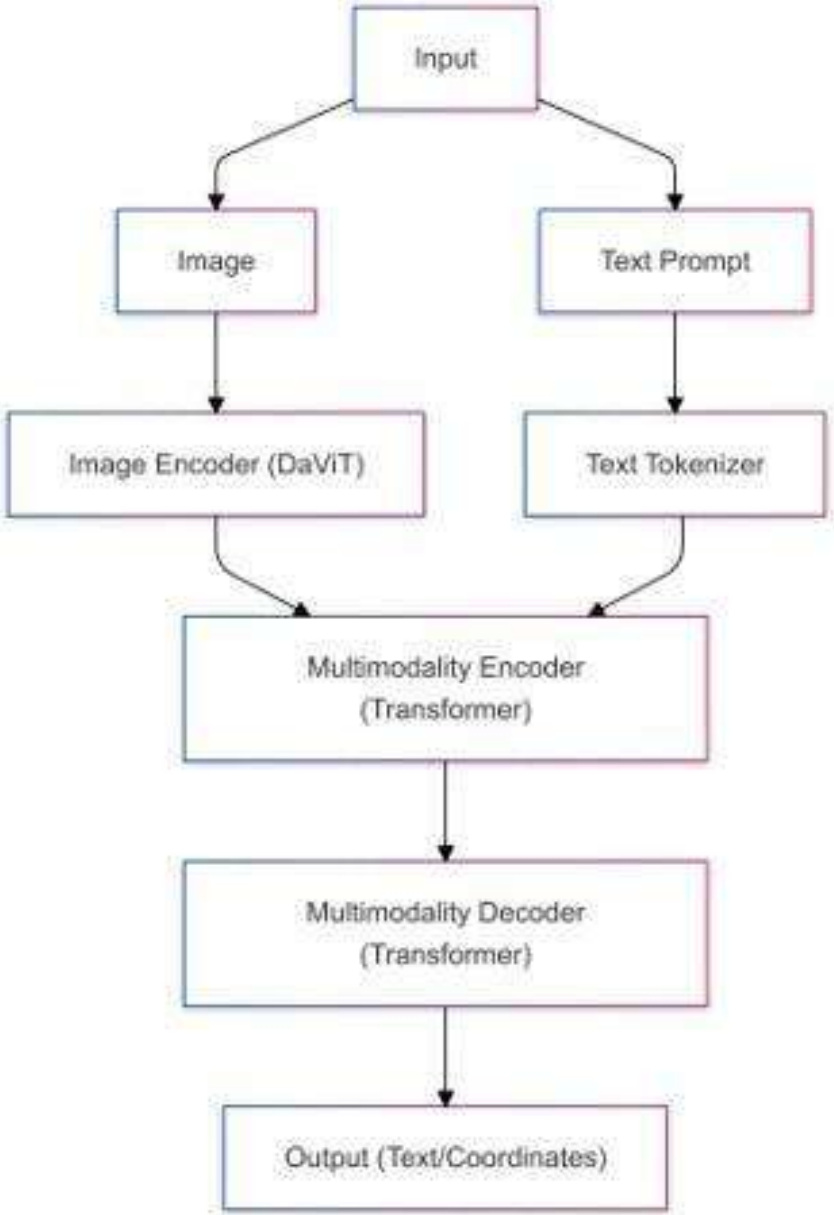
PaliGemma (especially PaliGemma 2) is designed for flexibility and scalability, excelling when fine-tuned for specific tasks or domains. It supports a broader range of languages and higher-resolution images, making it a strong choice for complex, custom, or multilingual applications.

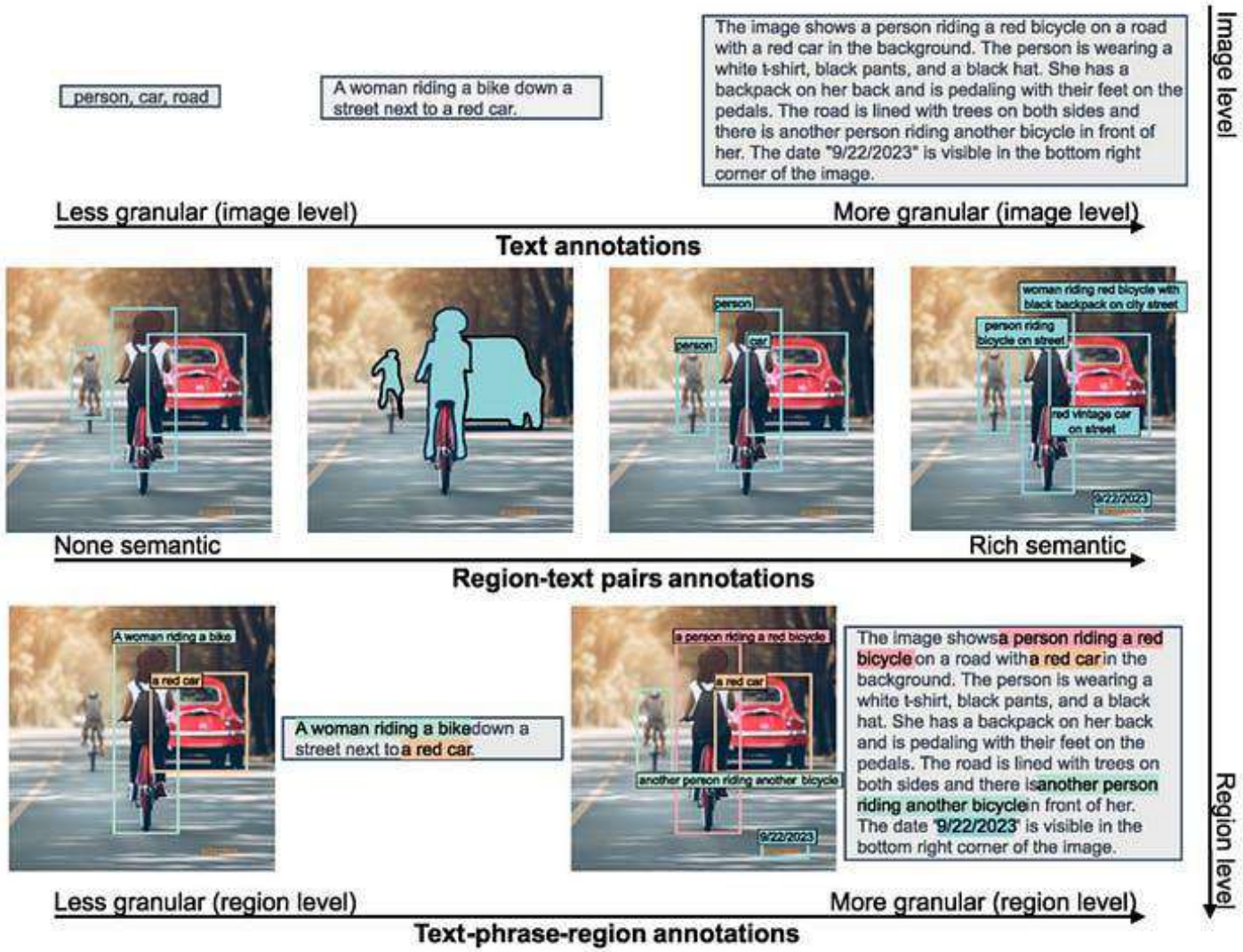
Florence-2

Advancing a Unified Representation for a Variety of Vision Tasks



Paper: https://arxiv.org/abs/2311.06242


Input
Image
Text Prompt
Image Encoder (DaViT)
Text Tokenizer
Multimodality Encoder
(Transformer)
Multimodality Decoder
(Transformer)
Output (Text/Coordinates)


person, car, road
A woman riding a bike down a street next to a red car.
The image shows a person riding a red bicycle on a road with a red car in the background. The person is wearing a white t-shirt, black pants, and a black hat. She has a backpack on her back and is pedaling with their feet on the pedals. The road is lined with trees on both sides and there is another person riding another bicycle in front of her. The date "9/22/2023" is visible in the bottom right corner of the image.
Image level
Less granular (image level)
More granular (image level)
Text annotations
person
car
woman riding red bicycle with black backpack on city street
person riding bicycle on street
red vintage car on street
9/22/2023
None semantic
Rich semantic
Region-text pairs annotations
A woman riding a bike
a red car
A woman riding a bike down a street next to a red car.
a person riding a red bicycle
another person riding another bicycle
The image shows a person riding a red bicycle on a road with a red car in the background. The person is wearing a white t-shirt, black pants, and a black hat. She has a backpack on her back and is pedaling with their feet on the pedals. The road is lined with trees on both sides and there is another person riding another bicycle in front of her. The date '9/22/2023' is visible in the bottom right corner of the image.
Region level
Less granular (region level)
More granular (region level)
Text-phrase-region annotations

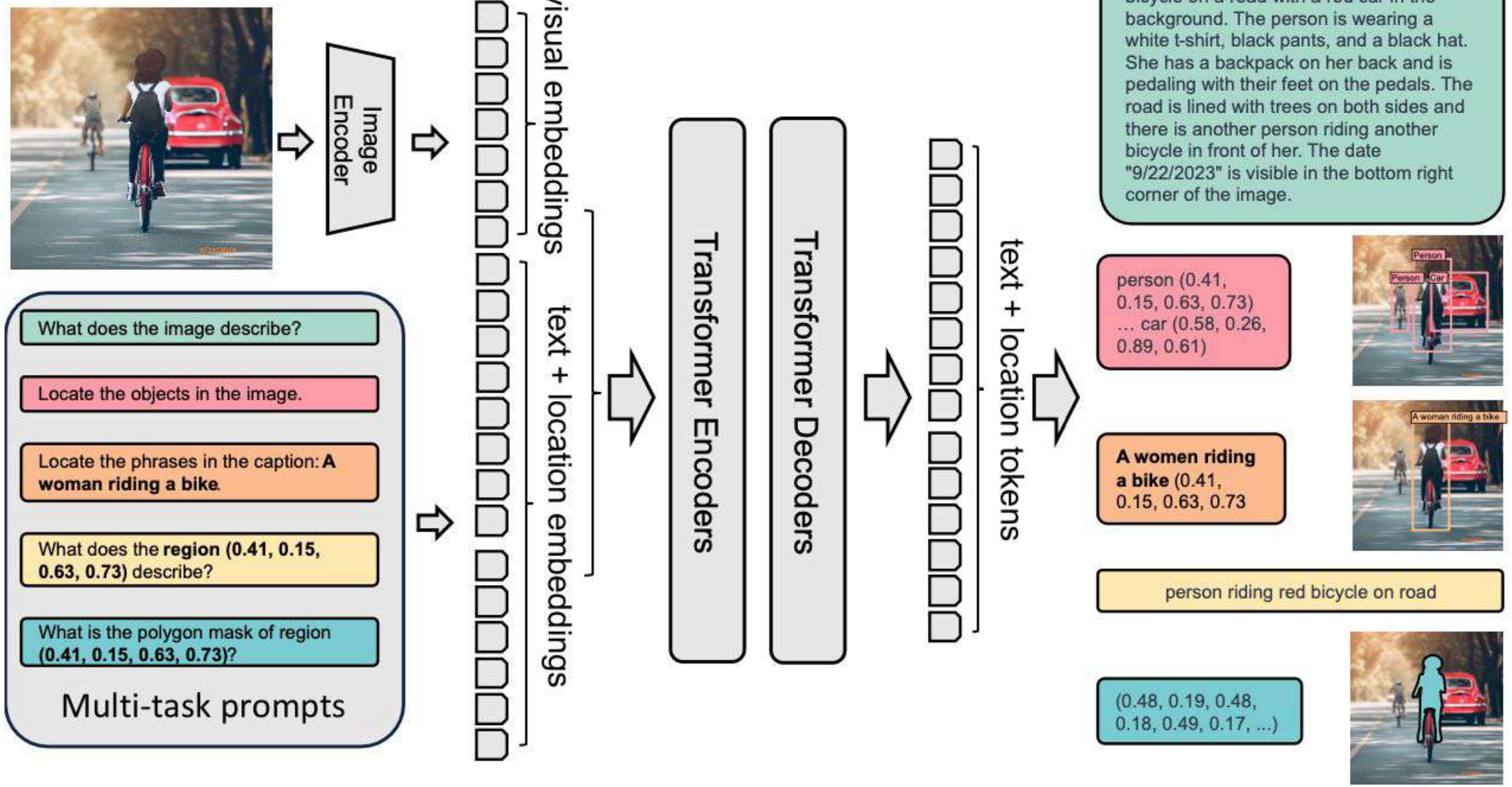


Image Encoder
visual embeddings
text + location embeddings
What does the image describe?
Locate the objects in the image.
Locate the phrases in the caption: A woman riding a bike.
What does the region (0.41, 0.15, 0.63, 0.73) describe?
What is the polygon mask of region (0.41, 0.15, 0.63, 0.73)?
Multi-task prompts
Transformer Encoders
Transformer Decoders
text + location tokens
The image shows a person riding a red bicycle on a road with a red car in the background. The person is wearing a white t-shirt, black pants, and a black hat. She has a backpack on her back and is pedaling with their feet on the pedals. The road is lined with trees on both sides and there is another person riding another bicycle in front of her. The date "9/22/2023" is visible in the bottom right corner of the image.
person (0.41, 0.15, 0.63, 0.73) … car (0.58, 0.26, 0.89, 0.61)
Person
Person
Car
A women riding a bike (0.41, 0.15, 0.63, 0.73
A woman riding a bike
person riding red bicycle on road
(0.48, 0.19, 0.48, 0.18, 0.49, 0.17, ...)

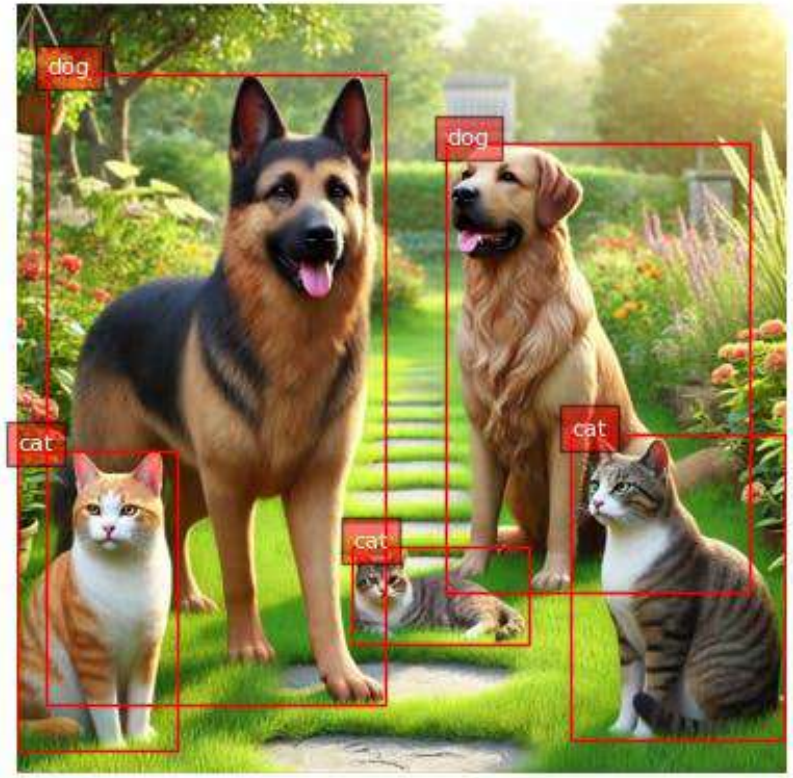
Caption

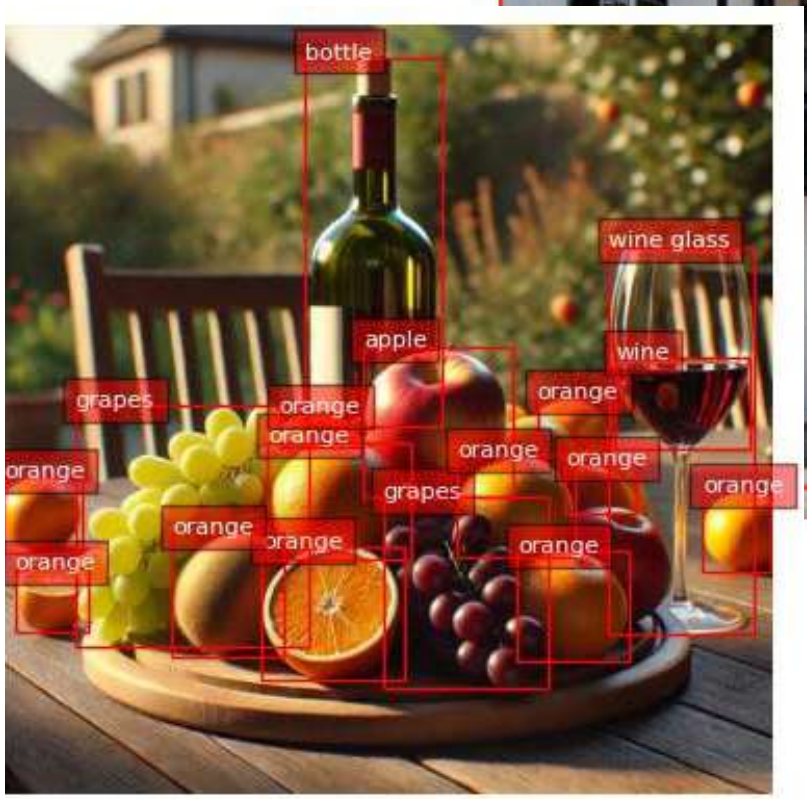
{'<MORE_DETAILED_CAPTION>': 'The image shows a wooden table with a wooden tray on it. On the tray, there are various fruits such as grapes, oranges, apples, and grapes. There is also a bottle of red wine on the table. The background shows a garden with trees and a house. The overall mood of the image is peaceful and serene.'}

{'<CAPTION>': 'A group of dogs and cats sitting in a garden.'}

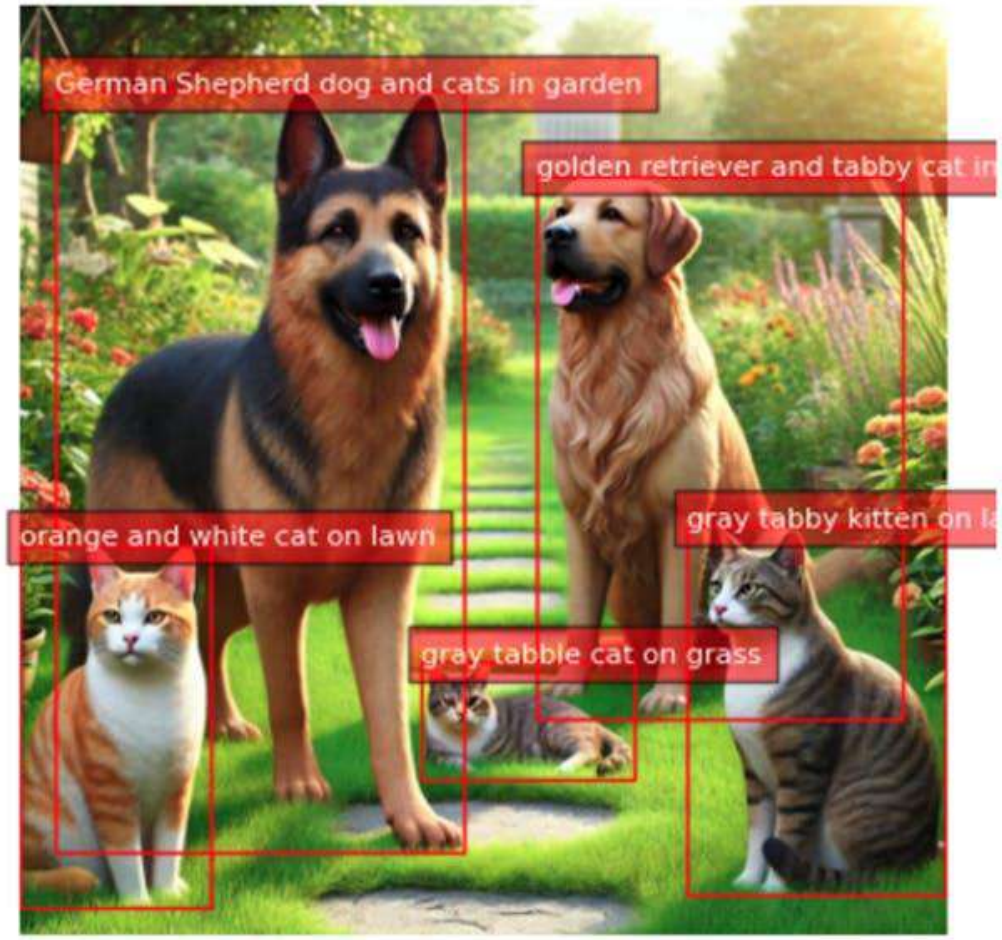
{'<DETAILED_CAPTION>': 'The image shows a street with cars and people walking down it, surrounded by buildings with windows, railings, and balconies. There is a tree in the foreground and a clock tower in the background. The sky is filled with clouds and there is a watermark on the image.'}

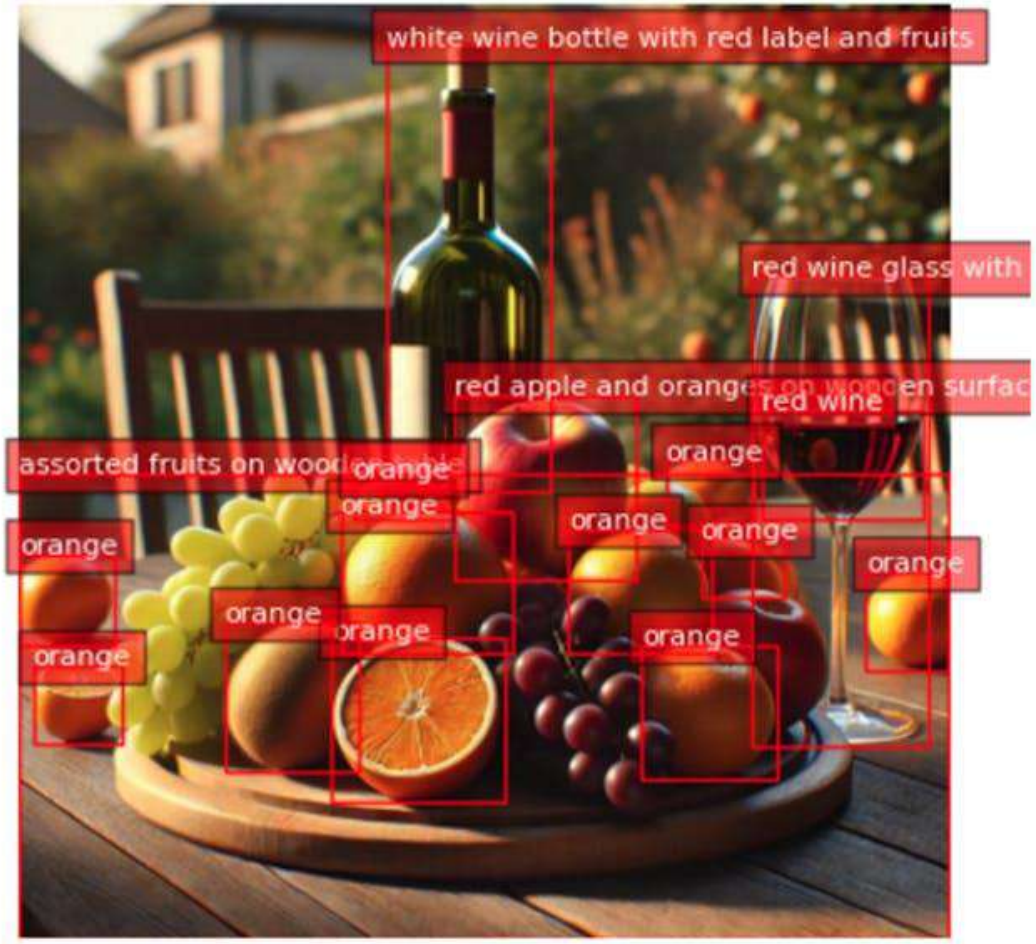
Object Detection





Dense Region Caption





```
task_prompt = '<CAPTION_TO_PHRASE_GROUNDING>'
results = run_example(task_prompt, text_input="a church clock tower",image=city)
plot_bbox(table, results['<CAPTION_TO_PHRASE_GROUNDING>'])
```

```
[INFO] ==> Florence-2-base (<CAPTION_TO_PHRASE_GROUNDING>), took 12.7 seconds to execute.
```


```
task_prompt = '<CAPTION_TO_PHRASE_GROUNDING>'
results = run_example(task_prompt, text_input="a person dressed in white",image=city)
plot_bbox(table, results['<CAPTION_TO_PHRASE_GROUNDING>'])
```

```
[INFO] ==> Florence-2-base (<CAPTION_TO_PHRASE_GROUNDING>), took 12.3 seconds to execute.
```

Segmentation

OCR


```
results['<OCR_WITH_REGION>']['labels']
```

```
['</s>Machine Learning',
 'Café',
 'com',
 'Embarcado',
 'Embarcados',
 'Democratizando a Inteligência',
 'Artificial para Paises em',
 '25 de Setembro ás 17h',
 'Desenvolvimento',
 'Toda quarta-feira',
 'Marcelo Roval',
 'Professor na UNIFIEI e',
 'Transmissão via',
 'in',
 'Co-Director do TinyML4D']
```

Fine-Tunning

{"<OD>": {"bboxes": [[0.1599999964237213, 133.59999084472656, 78.23999786376953, 232.1599884033203], [117.27999877929688, 139.0399932861328, 196.63999938964844, 243.67999267578125], [190.239990234375, 193.1199951171875, 270.239990234375, 319.5199890136719], [248.1599884033203, 91.04000091552734, 319.5199890136719, 189.27999877929688], [160.8000030517578, 27.680000305175781, 221.27999877929688, 118.23999786376953], [0.1599999964237213, 0.1599999964237213, 86.23999786376953, 57.119998931884766], [35.360000610351562, 36.31999969482422, 104.15999603271484, 112.15999603271484], [0.1599999964237213, 0.47999998927116394, 319.5199890136719, 319.5199890136719]], "labels": ["wheel", "wheel", "box", "box", "box", "box", "wheel", "box"]}}

The Future...

The Future of Generative AI at the Edge

Generative AI is rapidly transitioning from centralized cloud environments to edge computing, transforming how data is processed and utilized across industries. This shift is gaining significant momentum, with Gartner projecting that GenAI will be featured in 60% of edge computing deployments by 2029, up from just 5% in 2023. The Edge AI market is expected to reach $269.82 billion by 2032, growing at a compound annual rate of 33.3%.

Edge-based GenAI is enabling transformative applications across multiple sectors:

- Autonomous Vehicles: Real-time sensor analysis and decision-making
- Healthcare: On-device diagnostics and privacy-preserving patient monitoring
- Retail: Voice-assisted shopping and interactive customer service
- Industrial IoT: Predictive maintenance and anomaly detection
- Smart Devices: Real-time translation and augmented reality experiences

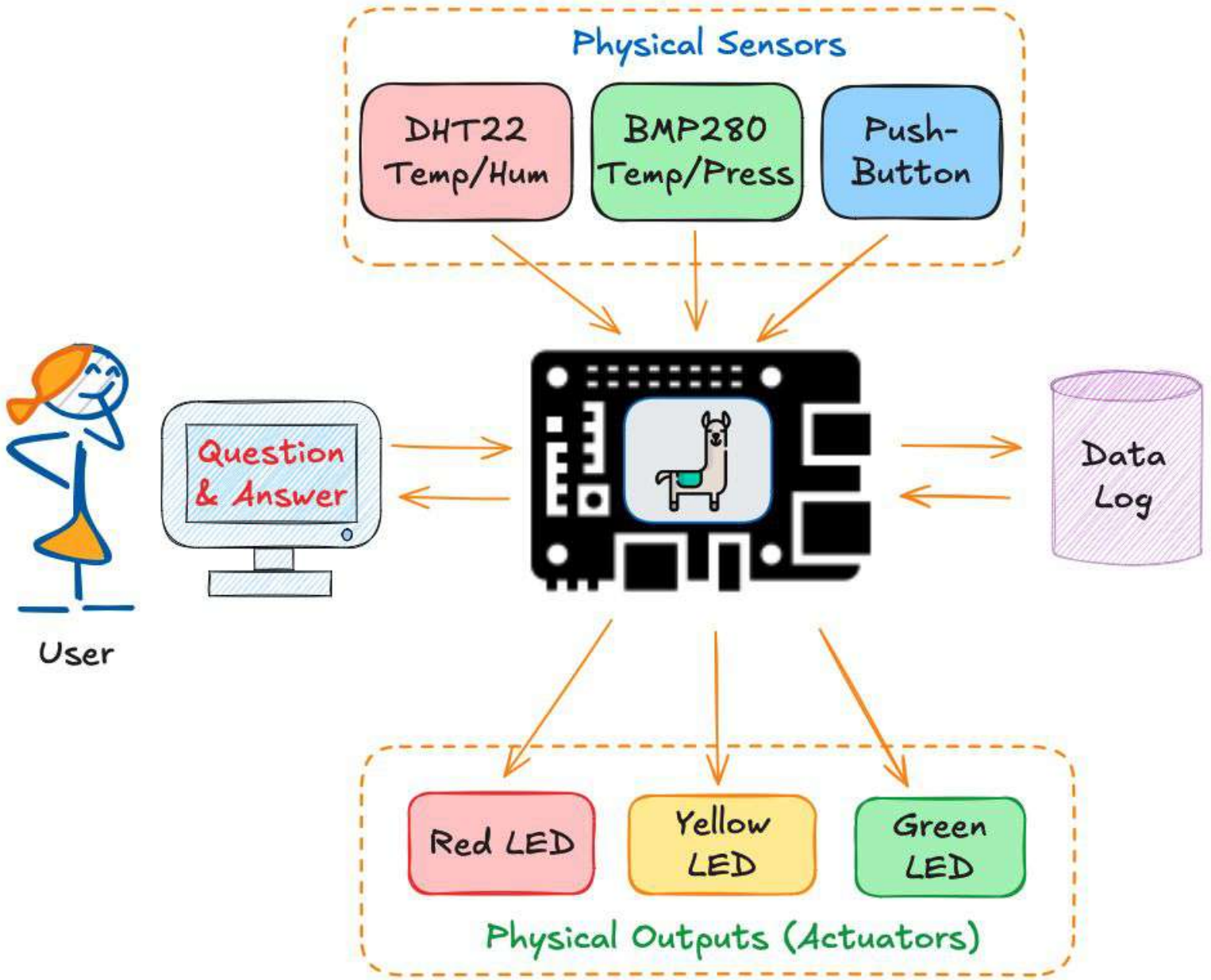
microsoft/BitNet

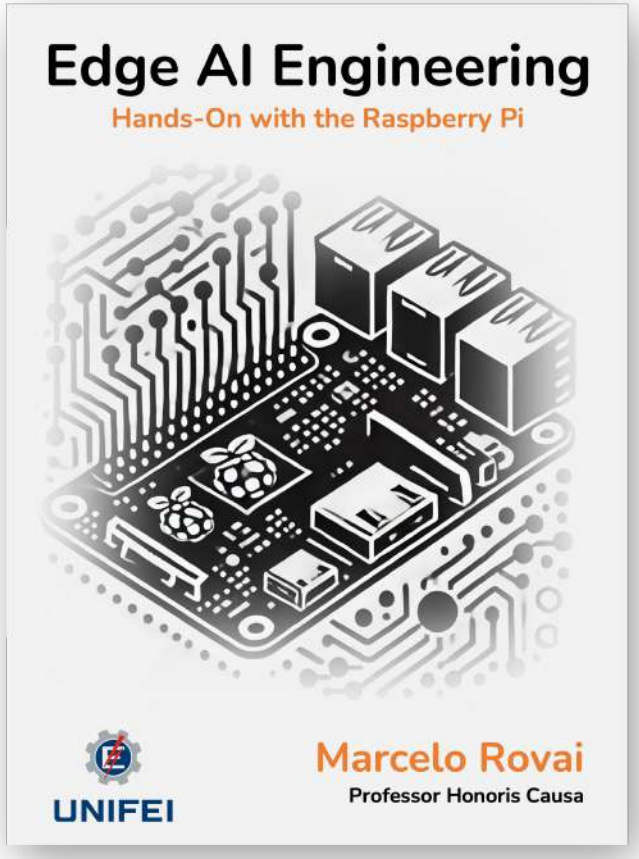
Official inference framework for 1-bit LLMs

Bitnet.cpp employs one-bit quantization, representing values with a ternary system (+1, -1, 0). This approach simplifies calculations by replacing complex multiplications with additions and subtractions, eliminating the need for GPUs.

- Speedups range from 1.37x to 6.1x on various CPUs.
- Power consumption reductions between 55.4% and 82.2% compared to traditional GPU-based inference.

bitnet.cpp



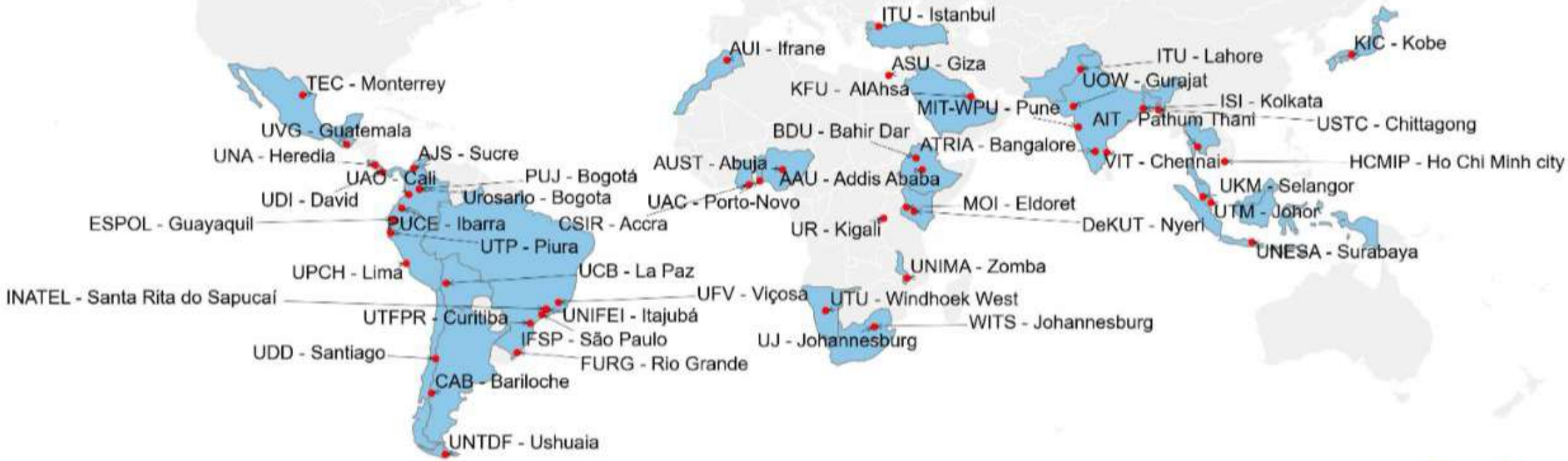


EdgeAI Engineering

TinyML4D Academic Network

Edge AIP Academic Industry Partnership

TinyML4D Academic Network



TinyML4D, an initiative to make TinyML education available to everyone globally.

More than 2,000 people trained!

MOUSER ELECTRONICS

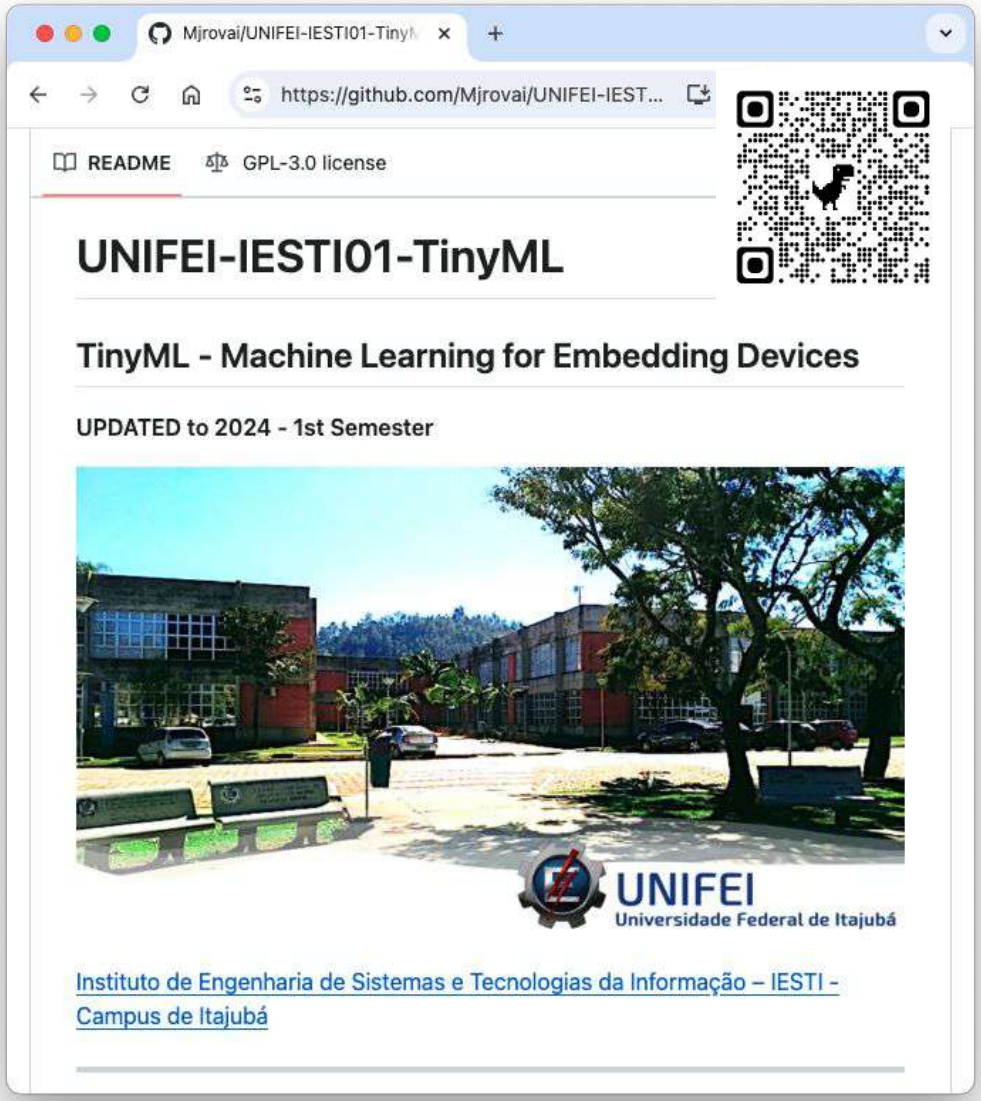
Brazil 2024

EMBARCADOS

Malawi 2025

IESTI01 - Course Structure

- Weekly video-recorded lectures (15 weeks)
  - Slides
  - Hands-on coding (by teacher & students)
- Weekly Additional Readings
- Assignments
  - Quizzes (Weekly)
  - Notebooks with codes (4)
  - Hands-on lab reports (4)
- Final Project (Groups of 3 students)
  - Report
  - Presentation



https://github.com/Mjrovai/UNIFEI-IESTI01-TinyML

TinyML Arduino Kit

(Principal)

Thanks to: ICTP The Abdus Salam International Centre for Theoretical Physics, IAEA, unesco

Wio Terminal Kit

(Optional)

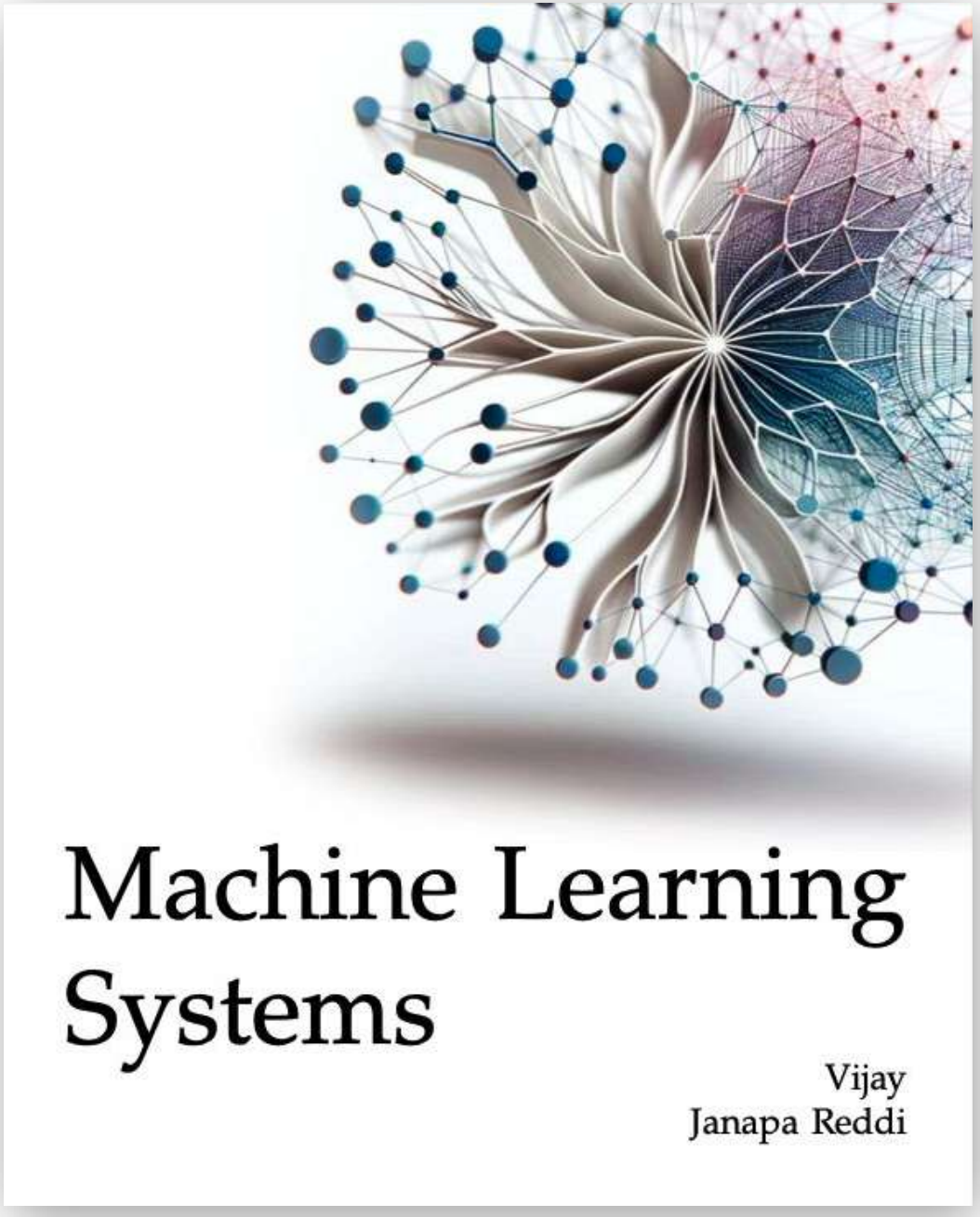
Thanks to: seeed studio

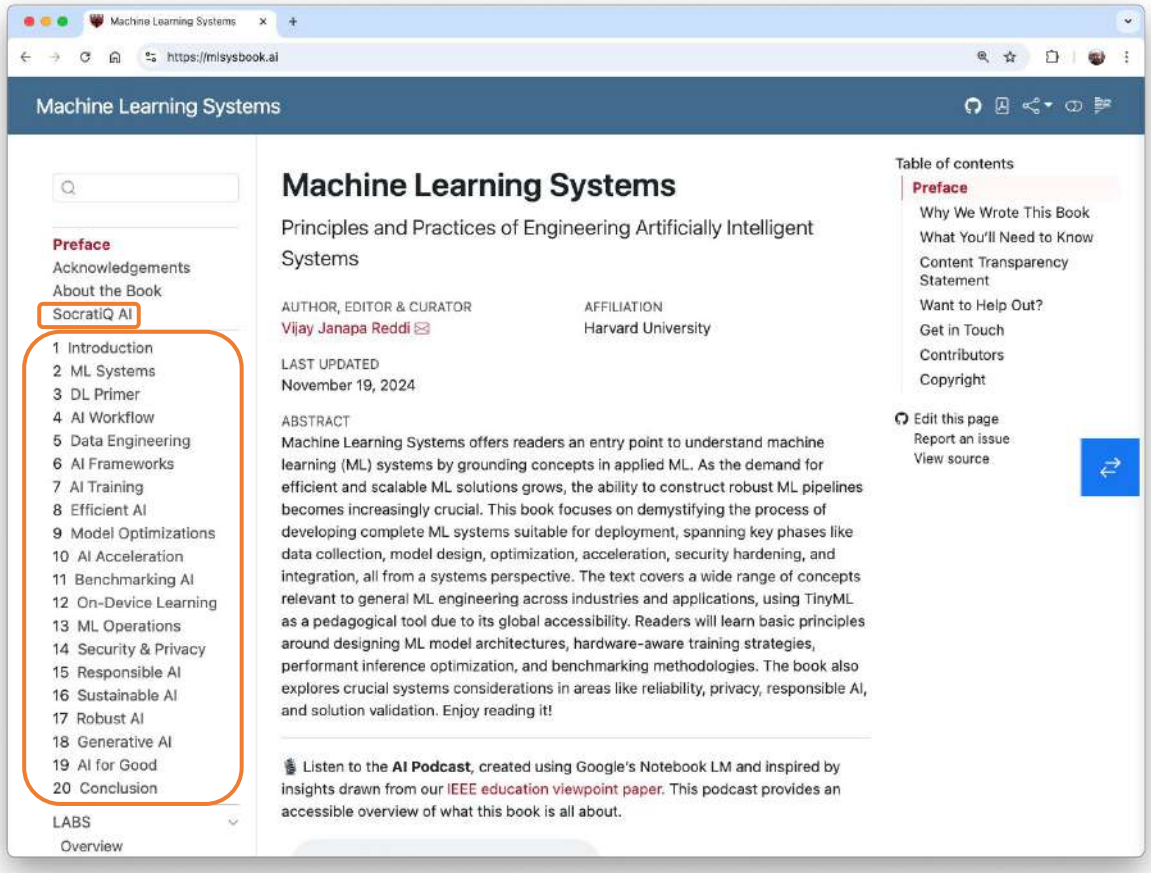
XIAO ESP32S3 Kit

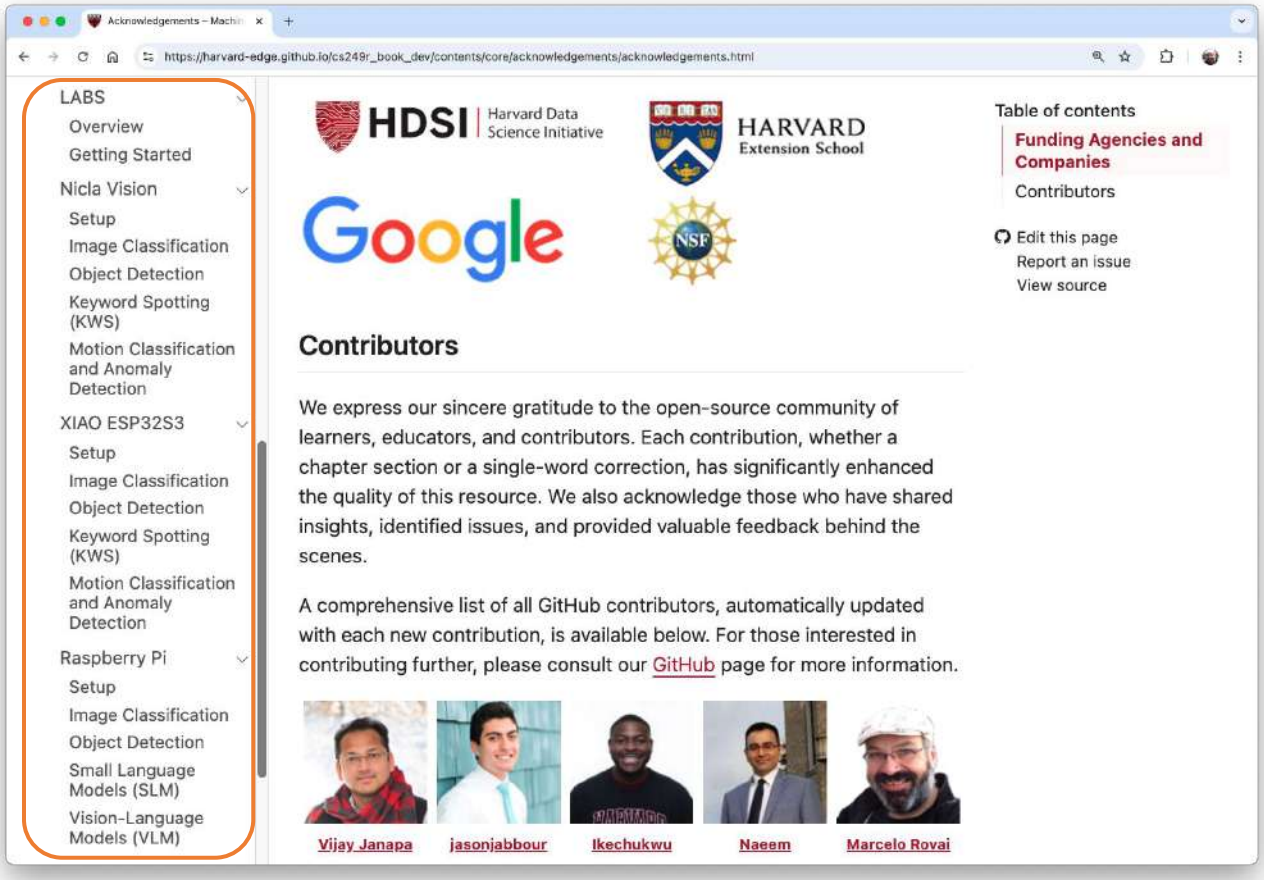
(Optional)

EMBARCADOS

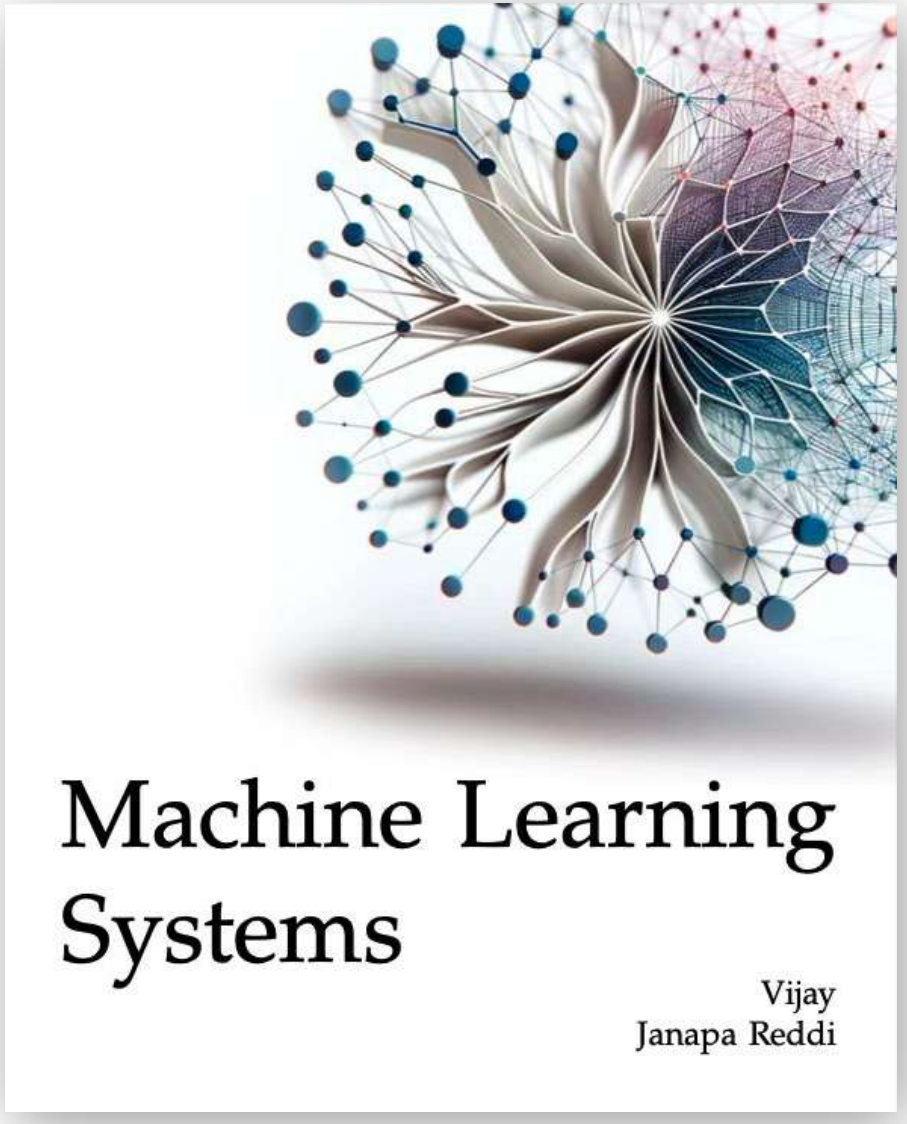
https://mlsysbook.ai/




Machine Learning Systems
https://mlsysbook.ai
Machine Learning Systems
Preface
Acknowledgements
About the Book
SocratiQ AI
1 Introduction
2 ML Systems
3 DL Primer
4 AI Workflow
5 Data Engineering
6 AI Frameworks
7 AI Training
8 Efficient AI
9 Model Optimizations
10 AI Acceleration
11 Benchmarking AI
12 On-Device Learning
13 ML Operations
14 Security & Privacy
15 Responsible AI
16 Sustainable AI
17 Robust AI
18 Generative AI
19 AI for Good
20 Conclusion
LABS
Overview
Machine Learning Systems
Principles and Practices of Engineering Artificially Intelligent Systems
AUTHOR, EDITOR & CURATOR
Vijay Janapa Reddi
AFFILIATION
Harvard University
LAST UPDATED
November 19, 2024
ABSTRACT
Machine Learning Systems offers readers an entry point to understand machine learning (ML) systems by grounding concepts in applied ML. As the demand for efficient and scalable ML solutions grows, the ability to construct robust ML pipelines becomes increasingly crucial. This book focuses on demystifying the process of developing complete ML systems suitable for deployment, spanning key phases like data collection, model design, optimization, acceleration, security hardening, and integration, all from a systems perspective. The text covers a wide range of concepts relevant to general ML engineering across industries and applications, using TinyML as a pedagogical tool due to its global accessibility. Readers will learn basic principles around designing ML model architectures, hardware-aware training strategies, performant inference optimization, and benchmarking methodologies. The book also explores crucial systems considerations in areas like reliability, privacy, responsible AI, and solution validation. Enjoy reading it!
Listen to the AI Podcast, created using Google's Notebook LM and inspired by insights drawn from our IEEE education viewpoint paper. This podcast provides an accessible overview of what this book is all about.
Table of contents
Preface
Why We Wrote This Book
What You'll Need to Know
Content Transparency Statement
Want to Help Out?
Get in Touch
Contributors
Copyright
Edit this page
Report an issue
View source


https://harvard-edge.github.io/cs249r_book_dev/contents/core/acknowledgements/acknowledgements.html
LABS
Overview
Getting Started
Nicla Vision
Setup
Image Classification
Object Detection
Keyword Spotting (KWS)
Motion Classification and Anomaly Detection
XIAO ESP32S3
Setup
Image Classification
Object Detection
Keyword Spotting (KWS)
Motion Classification and Anomaly Detection
Raspberry Pi
Setup
Image Classification
Object Detection
Small Language Models (SLM)
Vision-Language Models (VLM)
HDSI Harvard Data Science Initiative
HARVARD Extension School
Google
NSF
Contributors
We express our sincere gratitude to the open-source community of learners, educators, and contributors. Each contribution, whether a chapter section or a single-word correction, has significantly enhanced the quality of this resource. We also acknowledge those who have shared insights, identified issues, and provided valuable feedback behind the scenes.
A comprehensive list of all GitHub contributors, automatically updated with each new contribution, is available below. For those interested in contributing further, please consult our GitHub page for more information.
Vijay Janapa
jasonjabbour
Ikechukwu
Naeem
Marcelo Rovai
Table of contents
Funding Agencies and Companies
Contributors
Edit this page
Report an issue
View source

https://mlsysbook.ai/contents/labs/overview.html



Nicla Vision
Cortex-M4/M7

XIAO ML Kit
Xtensa LX7

Grove Vision AI V2
Cortex-M55/U55

Raspberry Pi
Cortex-A53/A76

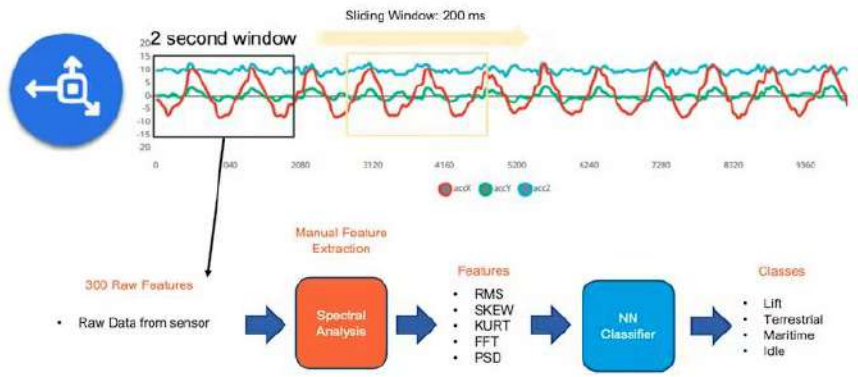
Motion Classification and An

https://harvard-edge.github.io/cs249r_book_dev/contents/labs/seeed/xiao_esp32s3/motion_classification/motion_classification.html

Section Quiz

Data Pre-Processing

The raw data type captured by the accelerometer is a "time series" and should be converted to "tabular data". We can do this conversion using a sliding window over the sample data. For example, in the below figure,



We can see 10 seconds of accelerometer data captured with a sample rate (SR) of 50Hz. A 2-second window will capture 300 data points (3 axis x 2 seconds x 50 samples). We will slide this window each 200ms, creating a larger dataset where each instance has 300 raw features.

> You should use the best SR for your case, considering Nyquist's theorem, which states that a periodic signal must be sampled at more than twice the signal's highest frequency component.

Data preprocessing is a challenging area for embedded machine learning. Still, Edge Impulse helps overcome this with its digital signal processing (DSP) preprocessing step and, more specifically, the Spectral Features.

On the Studio, this dataset will be the input of a Spectral Analysis block, which is excellent for analyzing repetitive motion, such as data from accelerometers. This block will perform a DSP (Digital Signal Processing), extracting features such as "FFT" or "Wavelets". In the most common case, FFT, the **Time Domain Statistical features** per axis/channel are:

SocratiQ

Q2: Which theorem helps determine an appropriate sample rate for periodic signals?

- (●) A1: Nyquist's theorem — Correct
  Nyquist's theorem dictates that a periodic signal should be sampled at more than twice the signal's highest frequency component, which aids in selecting a suitable sample rate.
- ( ) A2: Shannon's theorem
  Shannon's theorem is related to sampling, but does not offer a concrete numerical relationship between sample rate and the maximum frequency.
- ( ) A3: Whittaker–Kotel'nikov–Shannon theorem
  This theorem uses a similar assertion to Nyquist but often concerns itself additionally with optimal signal reconstruction.

Q3: Considering the Spectral Analysis block in Edge Impulse, why is it proper to analyze repetitive motion, such as accelerometer data?

- (●) A1: Because repetitive motion generally contains obvious spectral patterns in the frequency domain — Correct

+ Add Context

type '@' to reference a section...

Information provided here may not always be accurate. Provide feedback

The XIAOML Kit

A hands-on introduction to machine learning systems using TinyML®

Designed by Professor **Vijay Janapa Reddi (Harvard University)**, author of the **Machine Learning Systems** textbook.

What's inside

XIAO ESP32-S3 Sense CAM • IMU • Heatsinks • Labs • SD Toolkit

Build

Build keyword detection, image classification, motion detection, object detection, and more

For

For learners, educators, and real-world builders

Learners **mlsysbook.ai**

Builders **mlsysbook.ai/kits**

Developers **github.com/mlsysbook**

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

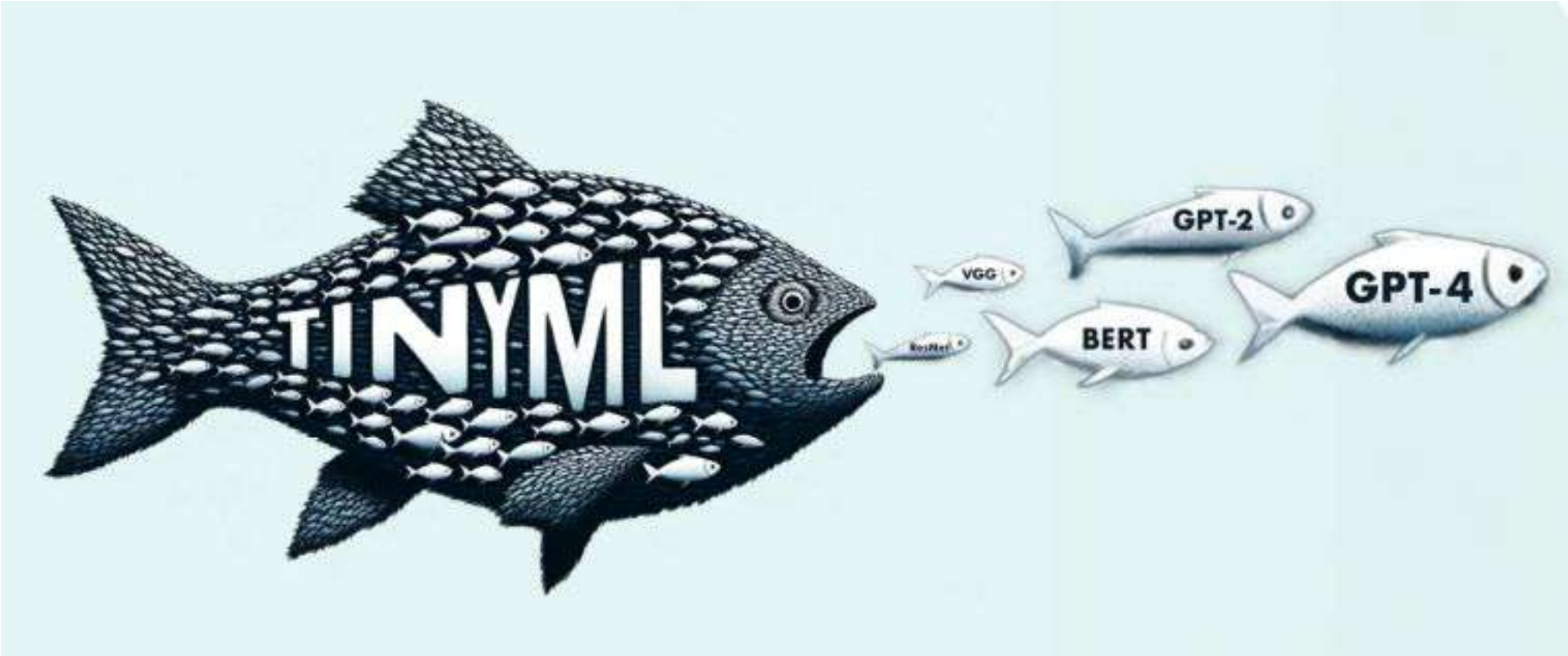
“Python for Data Analysis” by Wes McKinney
“Deep Learning with Python” by François Chollet - GitHub Notebooks
“TinyML” by Pete Warden and Daniel Situnayake
“TinyML Cookbook 2nd Edition” by Gian Marco Iodice
“Technical Strategy for AI Engineers, In the Era of Deep Learning” by Andrew Ng
“AI at the Edge” book by Daniel Situnayake and Jenny Plunkett
“XIAO: Big Power, Small Board” by Lei Feng and Marcelo Rovai
“Machine Learning Systems” by Vijay Janapa Reddi
"Tiny Machine Learning Quickstart" by Simone Salerno

Projects Repository

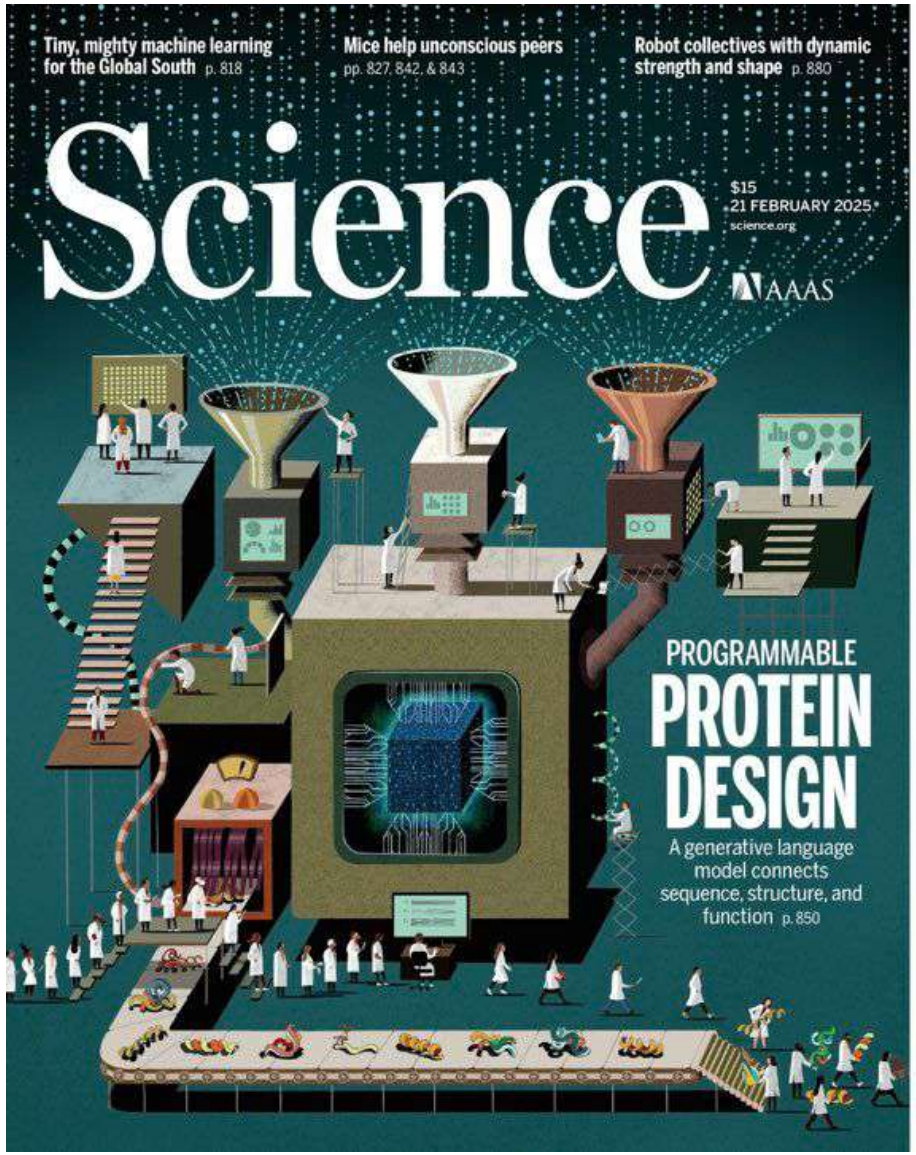
Edge Impulse Expert Network

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.

TinyML: Why the Future of Machine Learning is Tiny and Bright



Shvetank Prakash, Emil Njor, Colby Banbury, Matthew Stewart, Vijay Janapa Reddi



FEATURES

CUTTING AI DOWN TO SIZE

A $14 chip incorporating tinyML AI models, actual size shown.

https://www.science.org/content/article/what-s-tinyml-global-south-s-alternative-power-hungry-pricey-ai

Conclusion

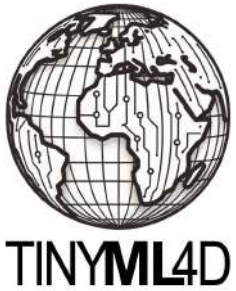
The Future of ML is Tiny and Bright

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

Questions?

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

UNIFEI - Federal University of Itajuba, Brazil
TinyML4D - Academic Network Co-Chair
EdgeAIP - Academia-Industry Partnership Co-Chair



OBRIGADO!

www.embarcados.com.br

linkedin.com/embarcados

@portalembarcados

youtube/Embarcados TV

EMBARCADOS

Patrocinado por

MOUSER ELECTRONICS