

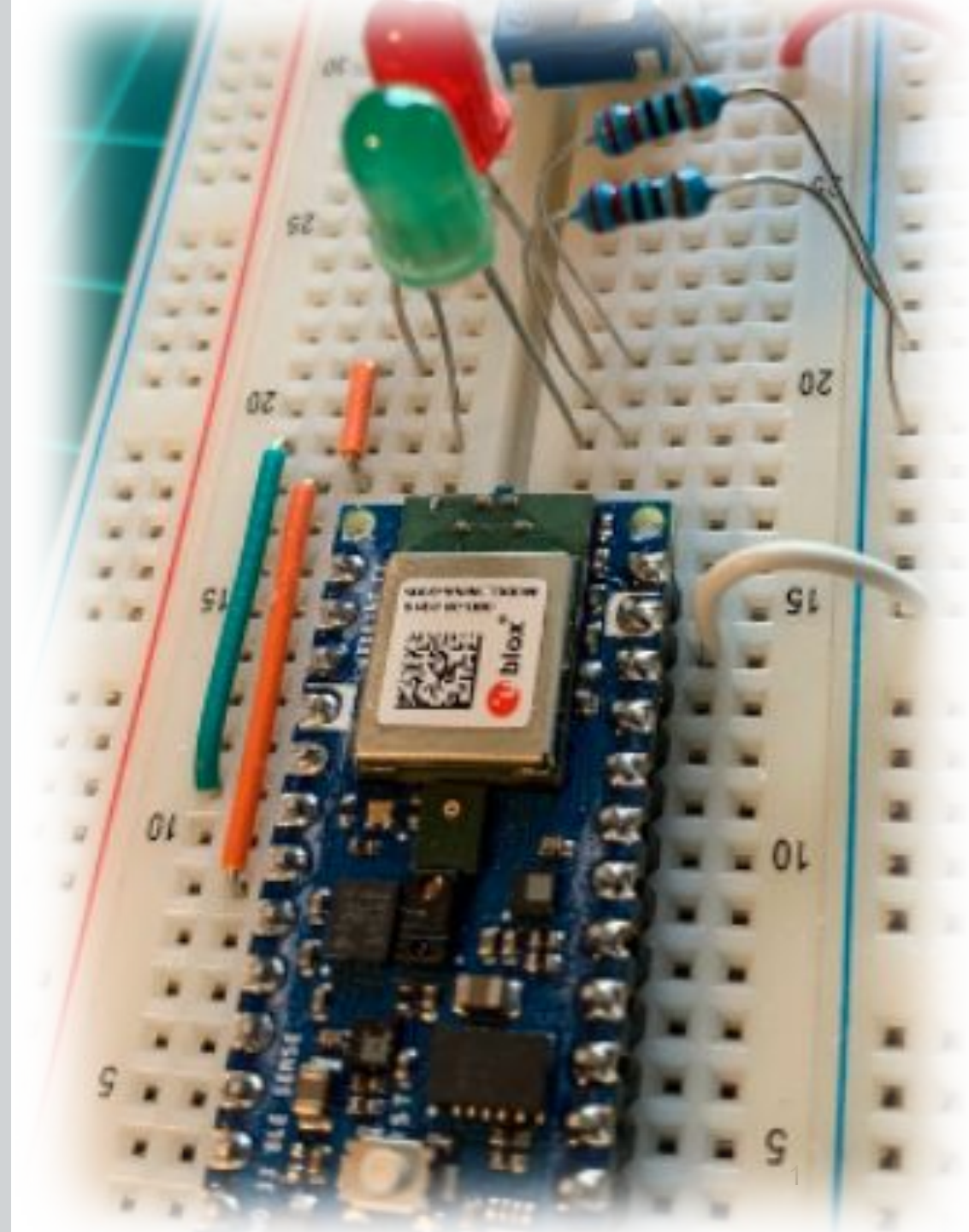
IESTI01 – TinyML

Embedded Machine Learning

2. Introduction to TinyML



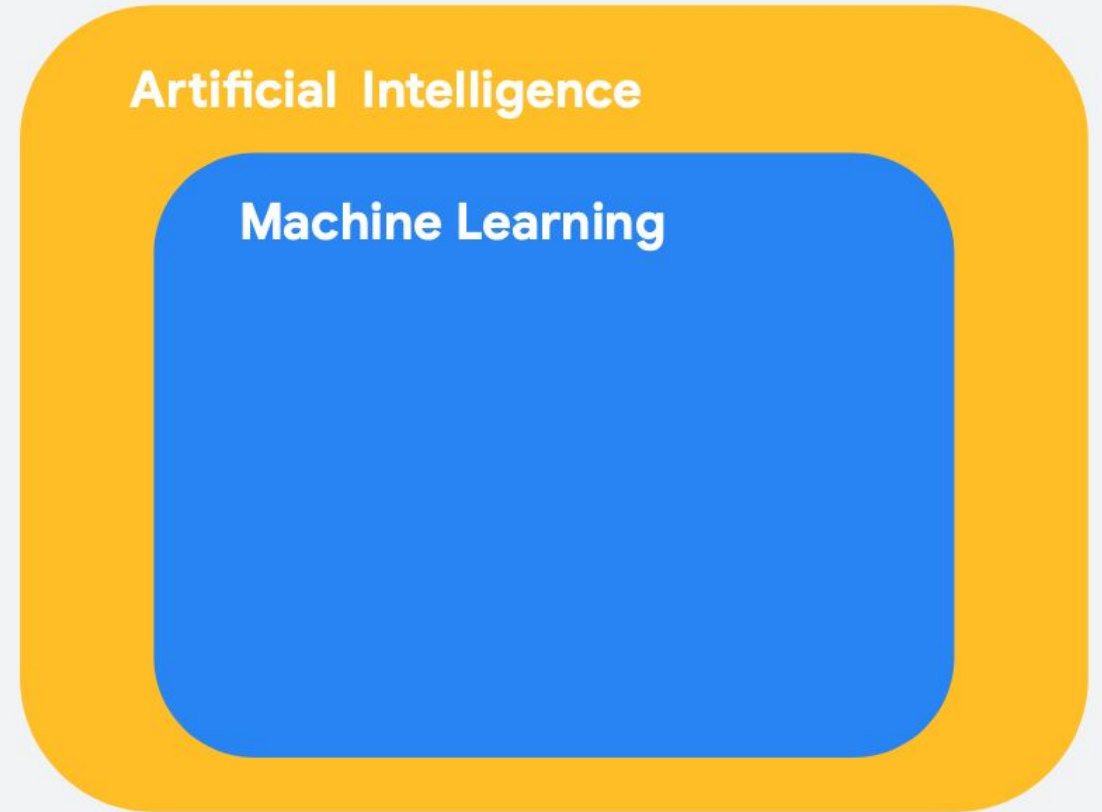
Prof. Marcelo Rovai
UNIFEI



What is Machine Learning?

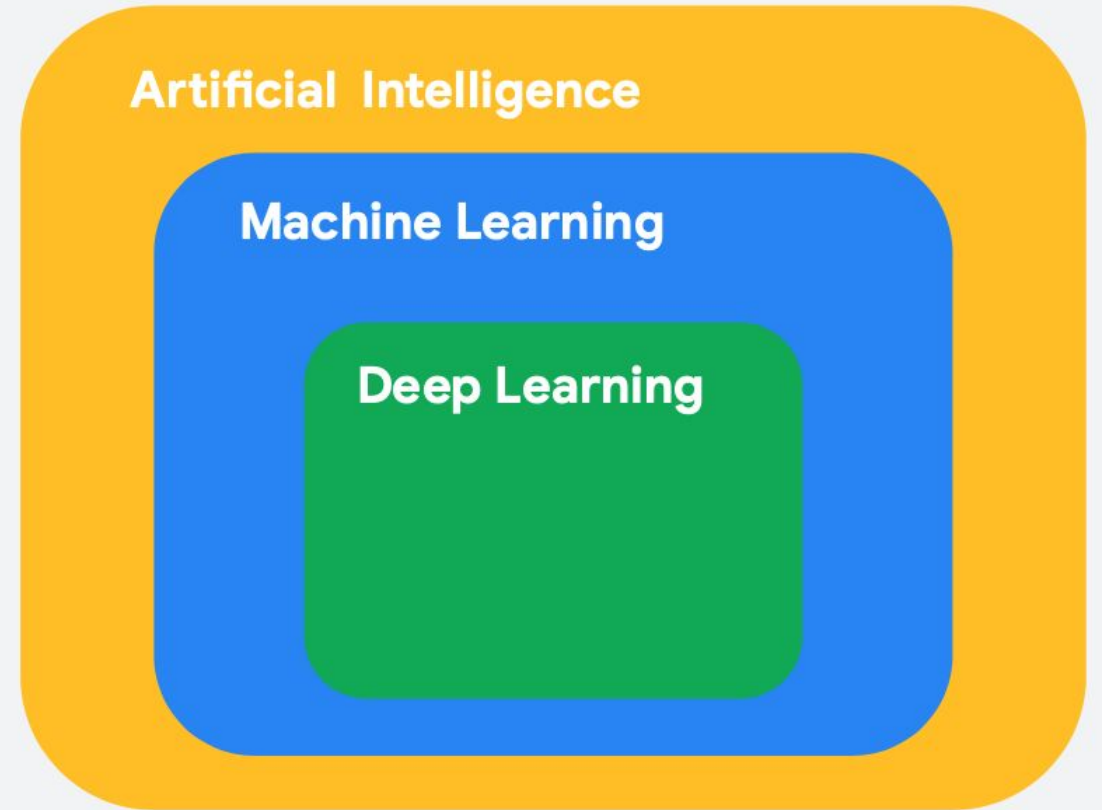
What is Machine Learning?

1. **Machine Learning** is a subfield of **Artificial Intelligence** focused on developing algorithms that learn to **solve problems by analyzing data for patterns**



What is (Deep) Machine Learning?

1. Machine Learning is a subfield of Artificial Intelligence focused on developing algorithms that learn to solve problems by analyzing data for patterns
2. **Deep Learning** is a type of Machine Learning that leverages **Neural Networks** and **Big Data**



Applications of Machine Learning

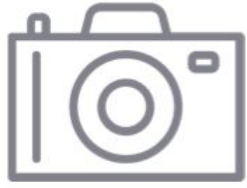
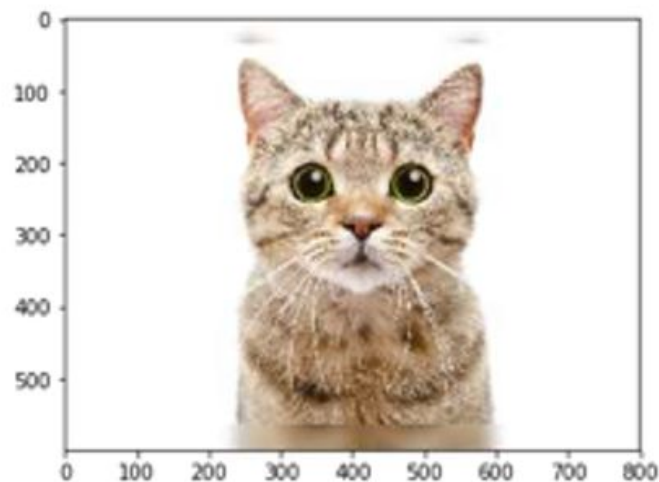


Image Classification

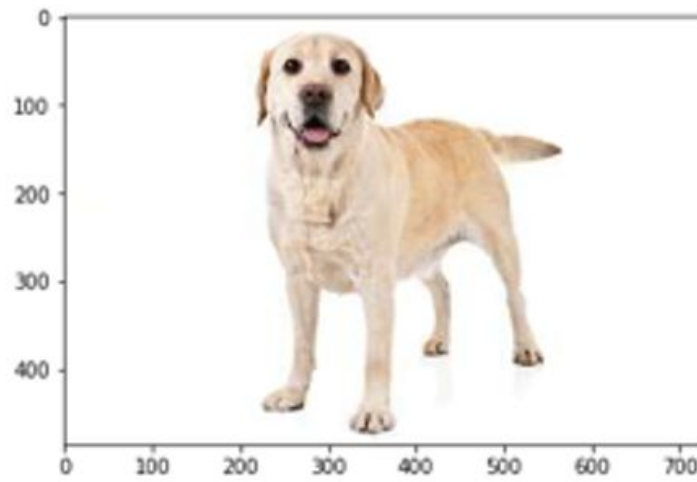
[PREDICTION] [Prob]

Egyptian cat : 64%
tabby : 14%
bucket : 3%



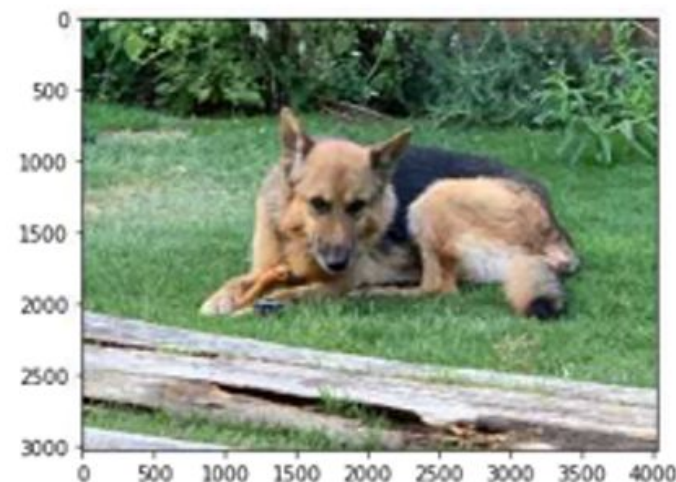
[PREDICTION] [Prob]

Labrador retriever : 83%
golden retriever : 13%
bloodhound : 0%



[PREDICTION] [Prob]

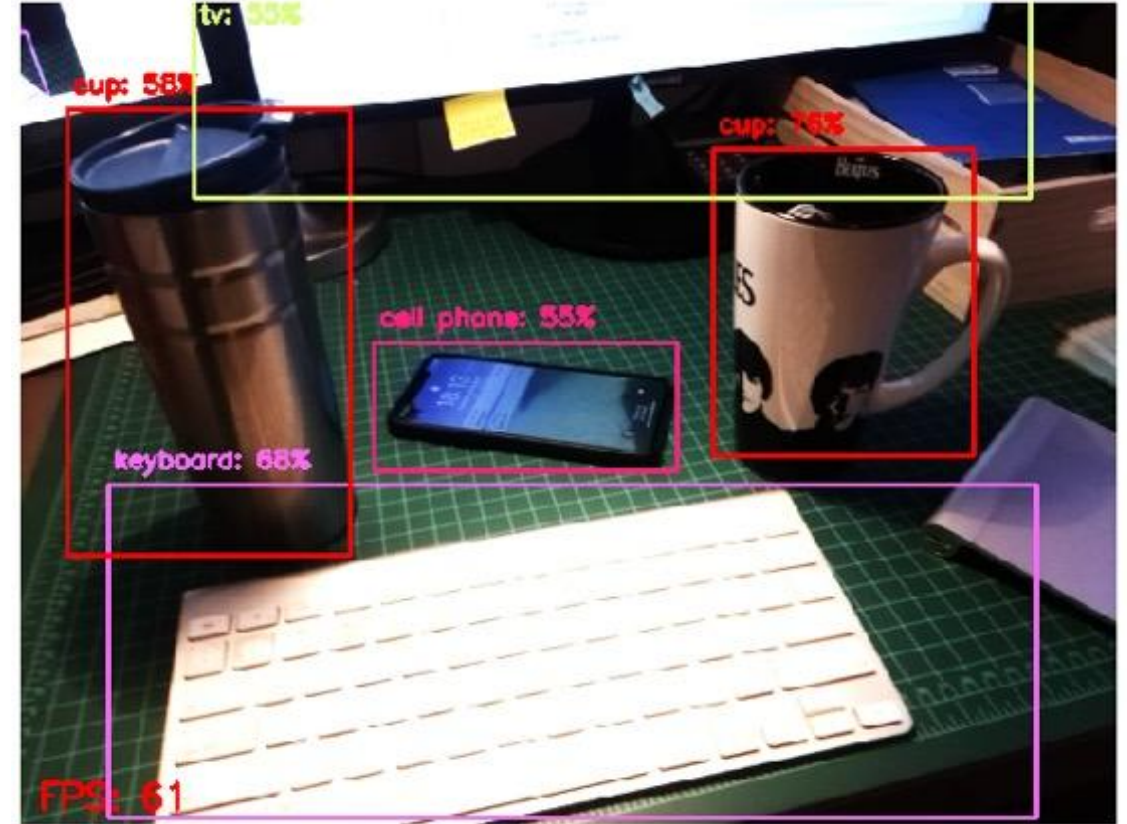
German shepherd : 60%
dhole : 16%
malinois : 7%



Object Detection

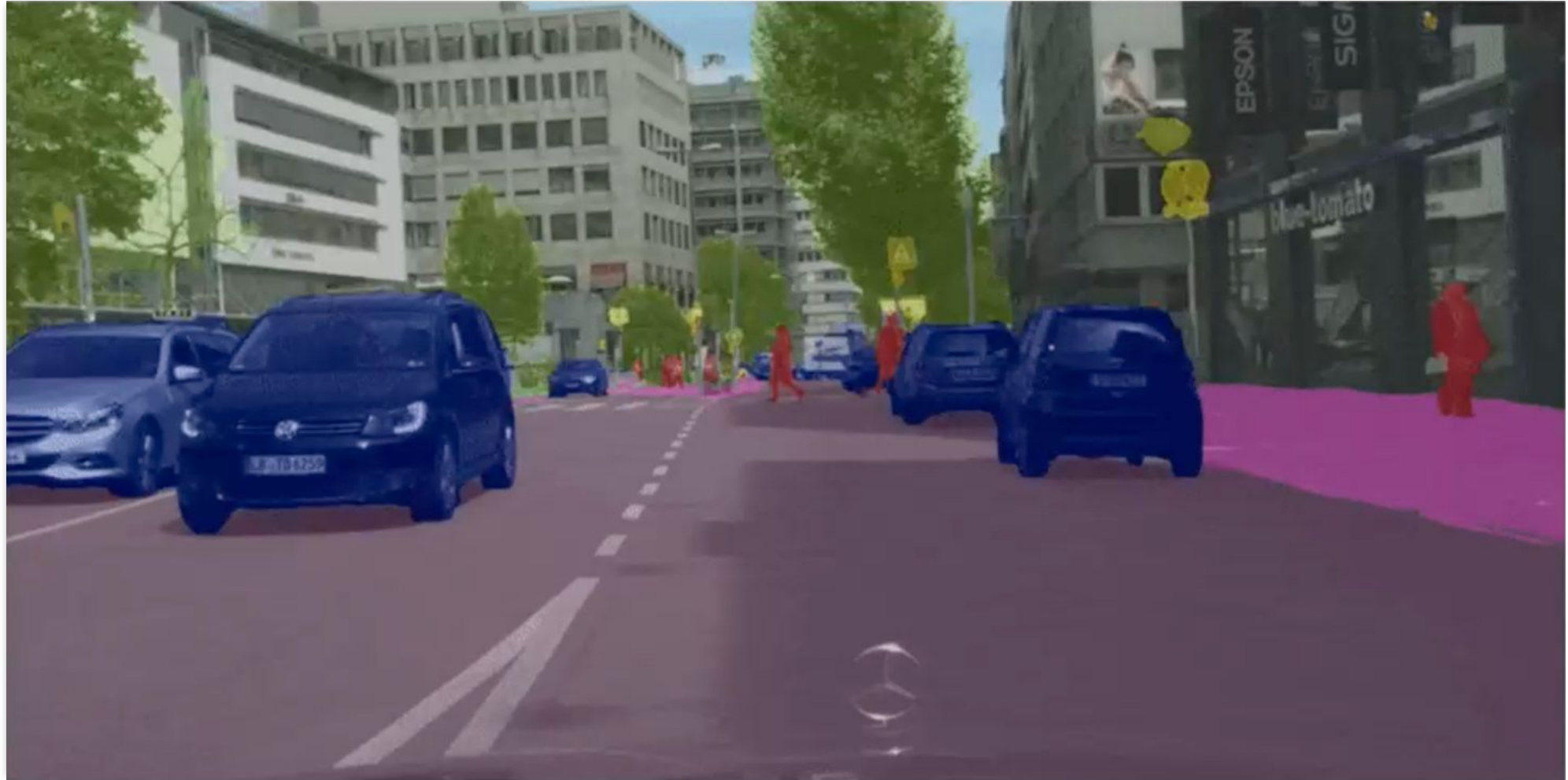


Photos



Live Video

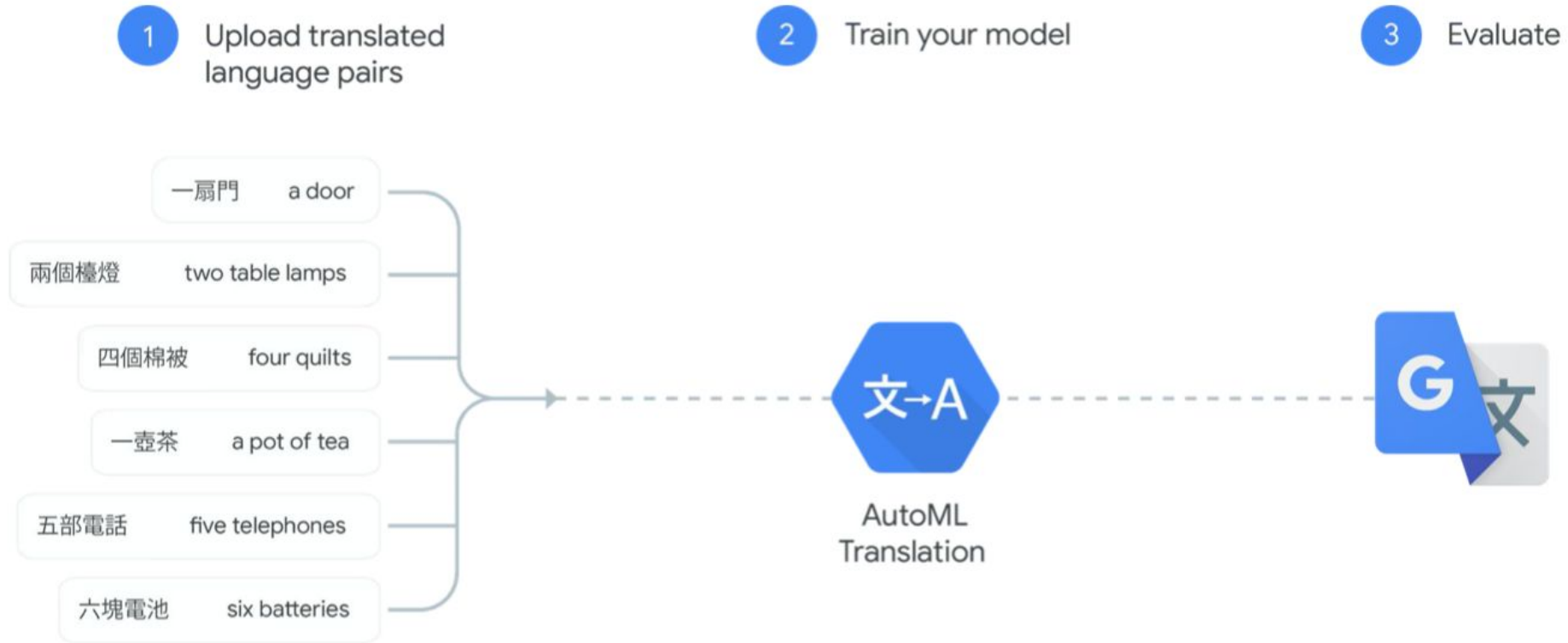
Segmentation



Pose Estimation



Machine Translation



Recommendations

				
		✓	✗	✓
	✓	✓	✗	
	✗	✓	✓	
	✓			✗
		✗	✓	

General AI does not exist (yet)

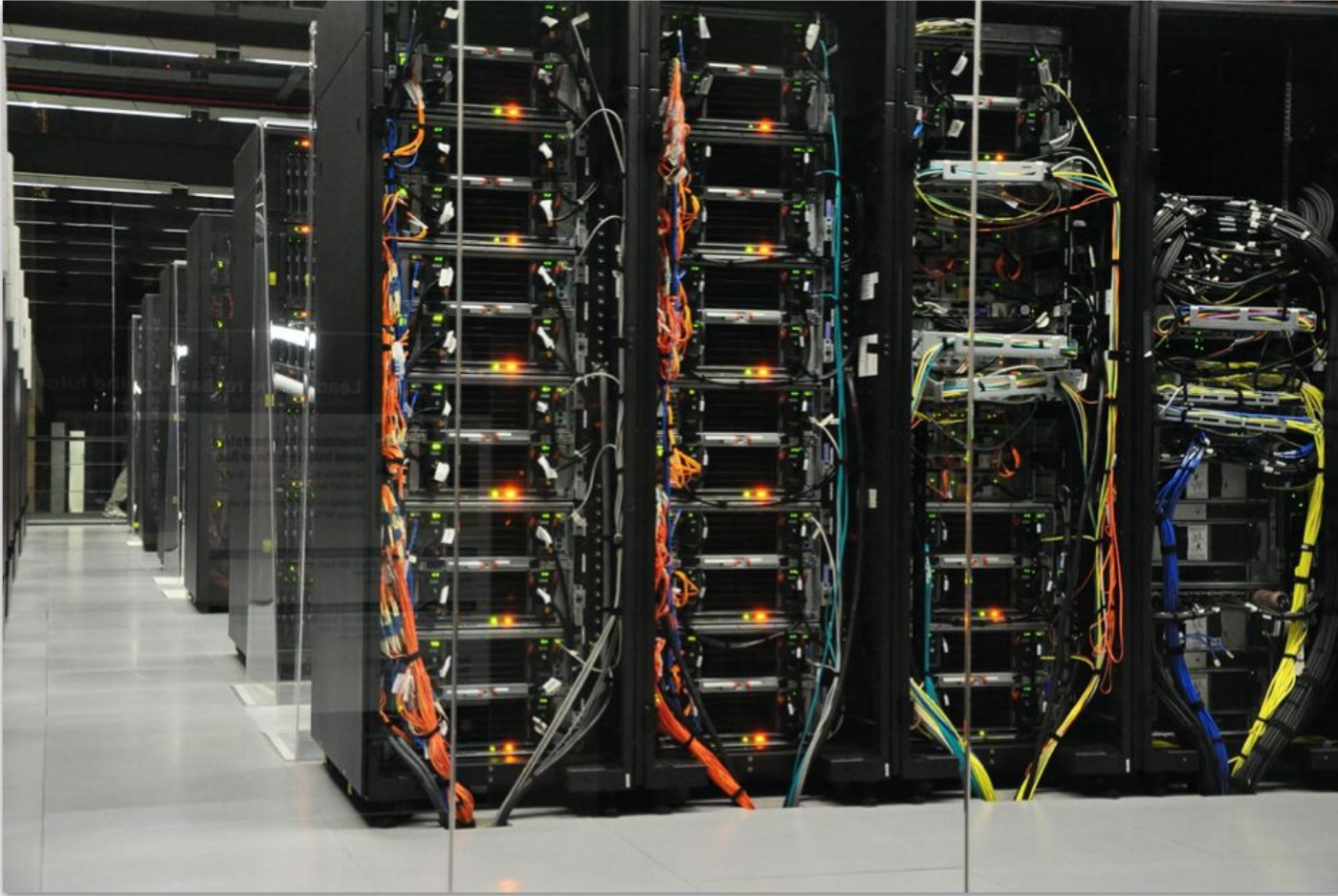
Dedicated ML Application examples

- Image Classification
- Object Detection
- Pose Estimation
- Voice Recognition
- Gesture Recognition
- Anomaly Detection
- Natural Language Processing (NLP)

Dedicated TinyML Application Examples

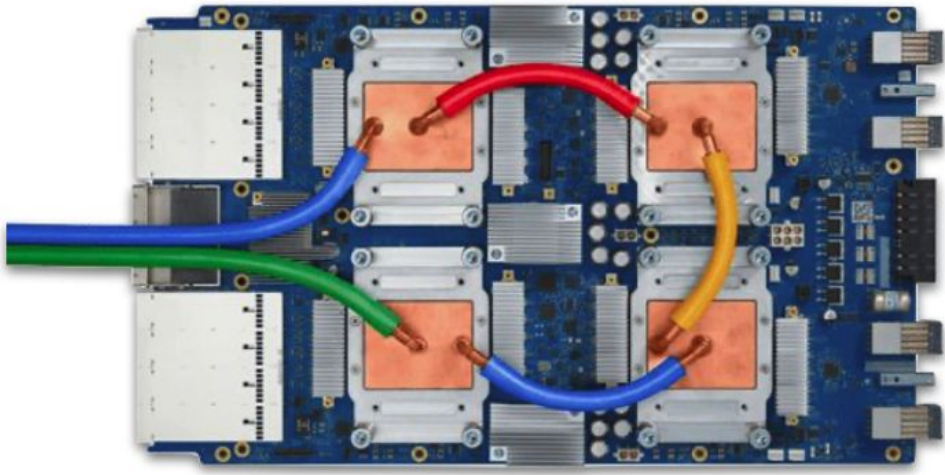
- ➡ Image Classification (Camera)
- ➡ Object Detection (||)
 - Pose Estimation
- ➡ Voice Recognition (Microphone)
- ➡ Gesture Recognition (Accelerometer)
- ➡ Anomaly Detection (||)
 - Natural Language Processing (NLP)

Datacenter



All the capabilities on previous examples, required a **remarkable amount of horsepower and computing capabilities**, so what companies are doing, they are taking all these computers and jam packing them into **data centers**, that are all just being dedicated in order to provide **machine learning** capabilities today.

TPUs/GPUs

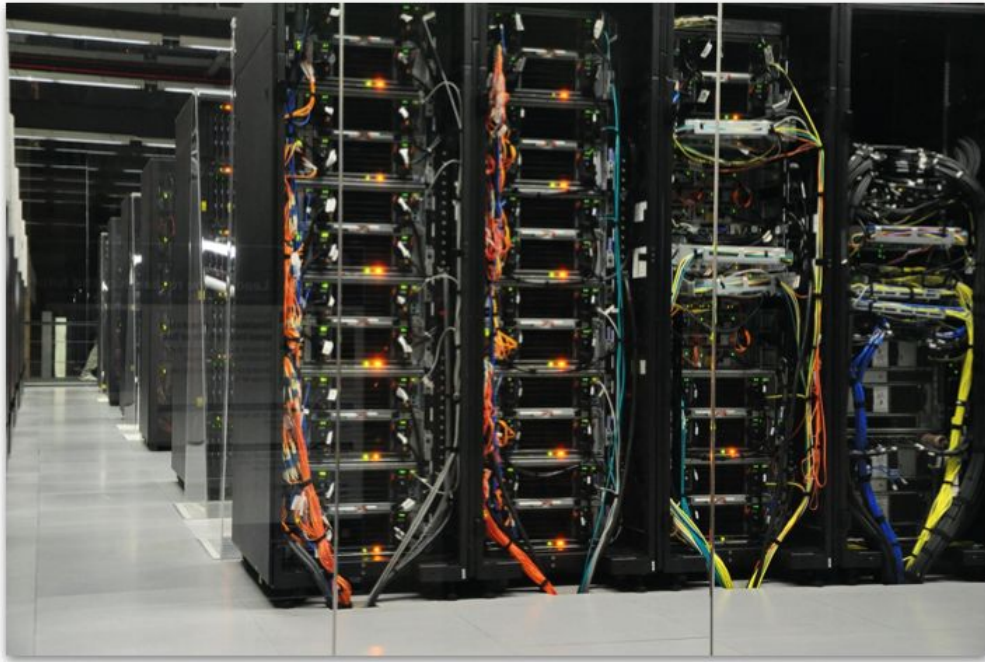


In order to be able to provide ML capability, companies like **Google** are building **TPUs** (Tensor Processing Units) and **NVIDIA GPUs** (Graphics Processing Units). Both of these computing systems are capable of **running machine learning extremely fast**.

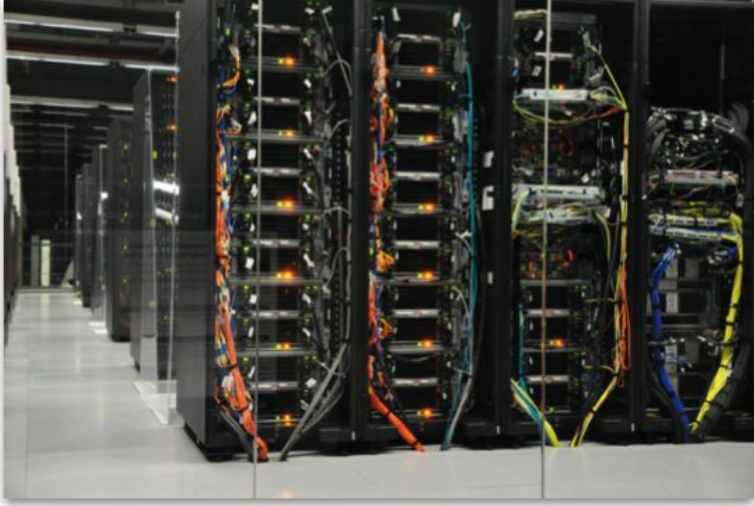


Bigger Is Not
Always Better.

Why?



Because we can not have a DataCenter to do ML inside our phone, for example!



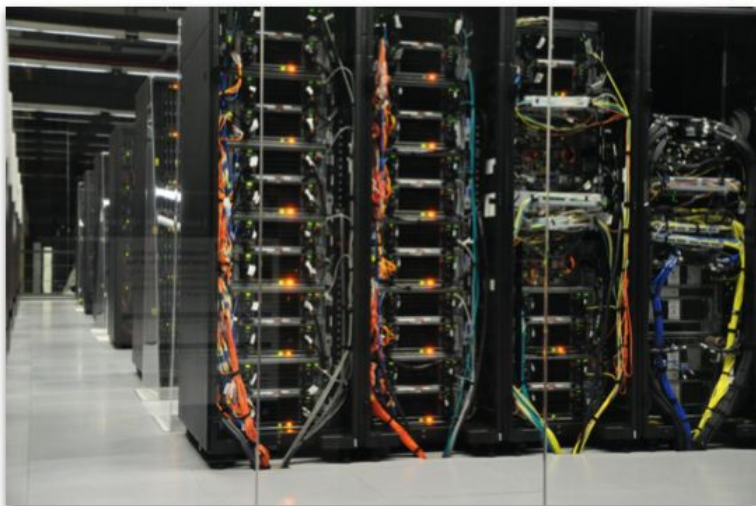
High power

Why?



Low power

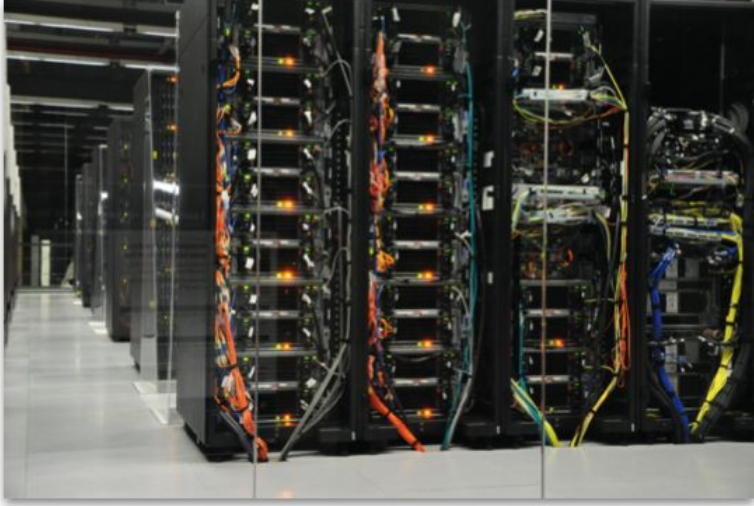
Why?



High power
High bandwidth



Low power
Low bandwidth



High power
High bandwidth
High latency

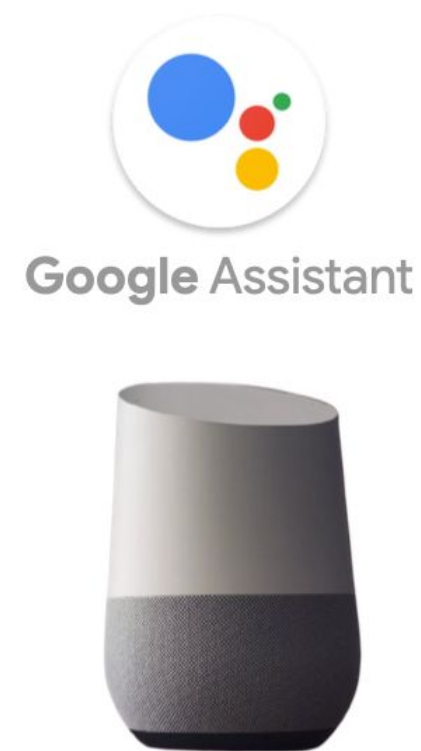
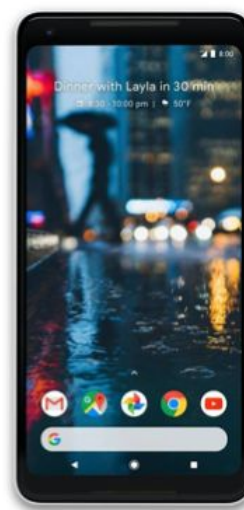
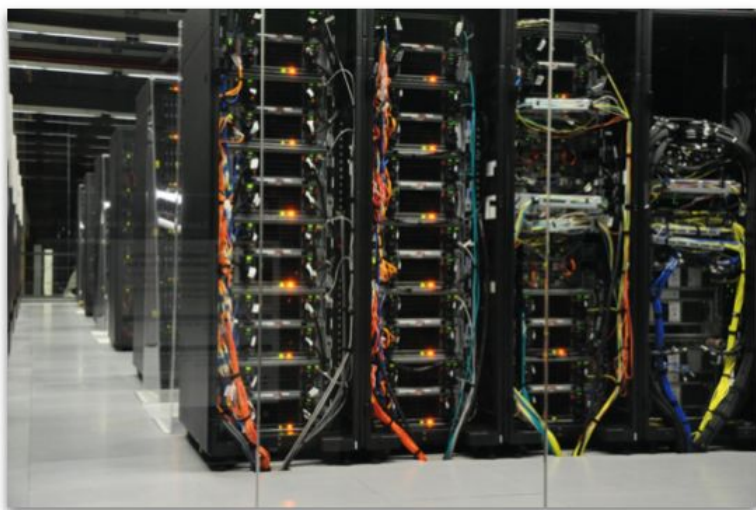


Low power
Low bandwidth
Low latency

Why?







Endpoint Devices



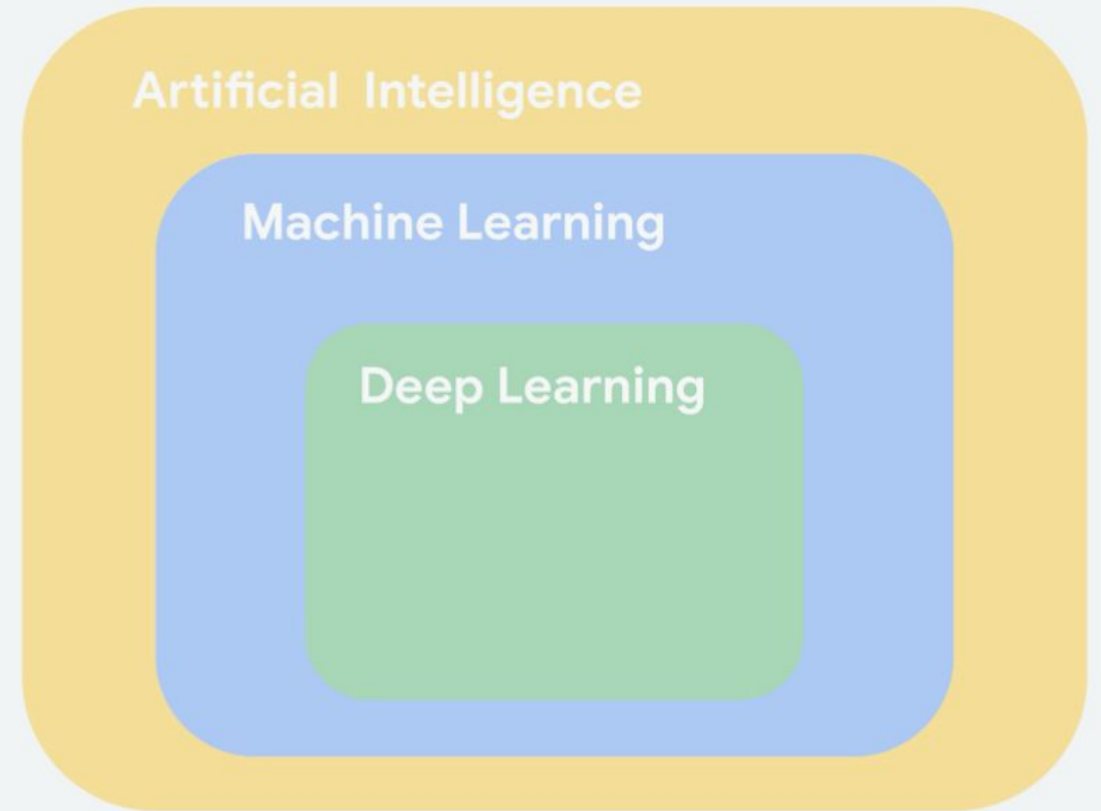
Google Assistant



What is (Deep) Machine Learning?

1. Machine Learning is a subfield of Artificial Intelligence focused on developing algorithms that learn to solve problems by analyzing data for patterns
2. **Deep Learning** is a type of Machine Learning that leverages **Neural Networks** and

Big Data



No Good Data Left Behind

5 Quintillion

bytes of data produced
every day by IoT

<1%

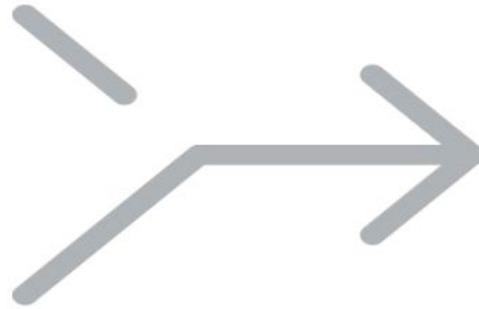
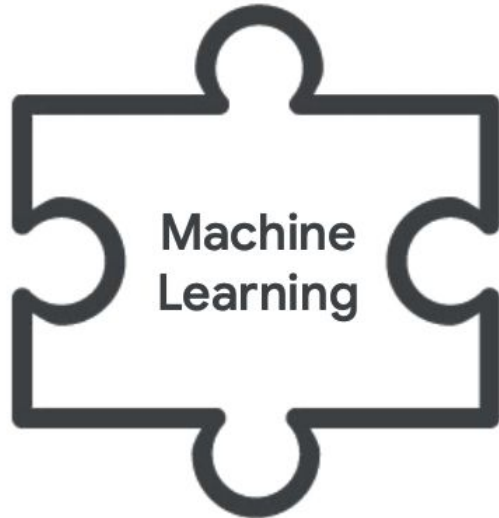
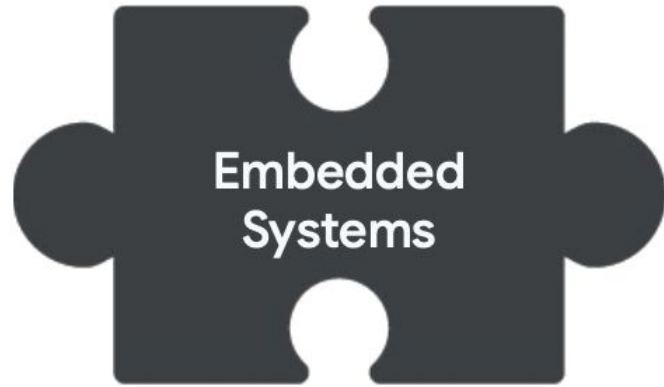
of unstructured data is
analyzed or used at all

Summary

- ML has several diverse applications in the real-world
- ML is increasingly moving from the cloud to endpoint devices
- Endpoint devices are everywhere around us

How do we enable TinyML?

What Makes **TinyML**?



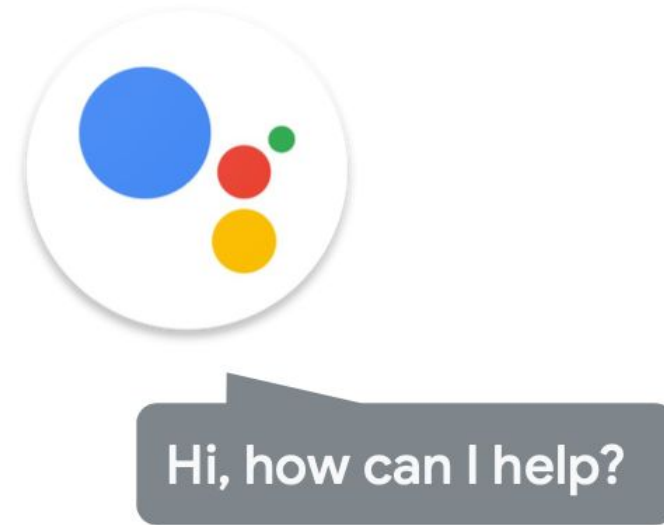
TinyML

Let's Take an Example

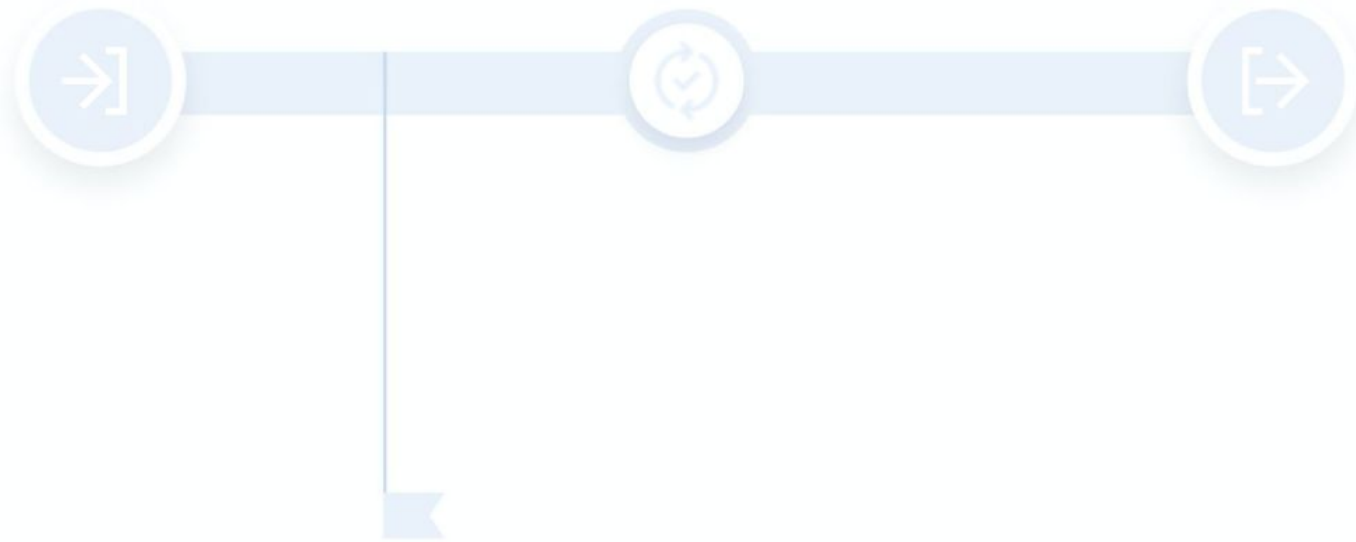


Google Assistant

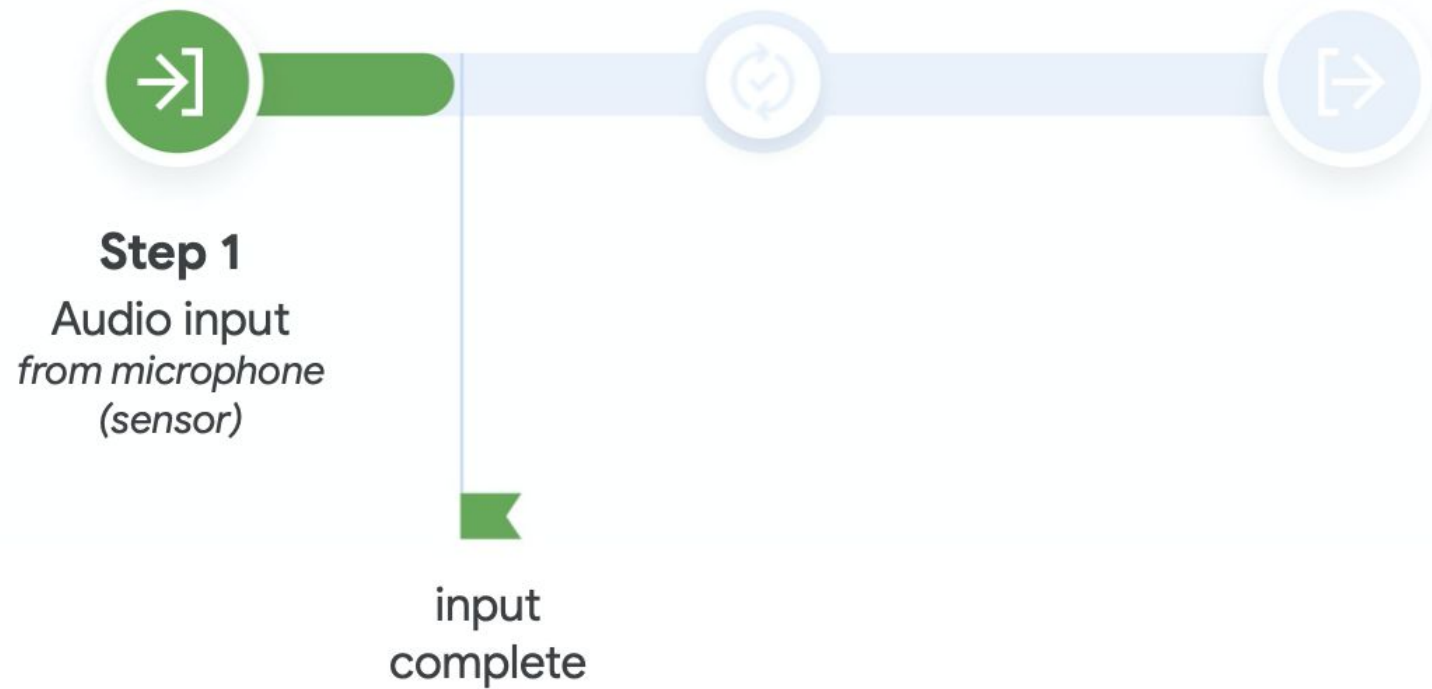
Let's Take an Example



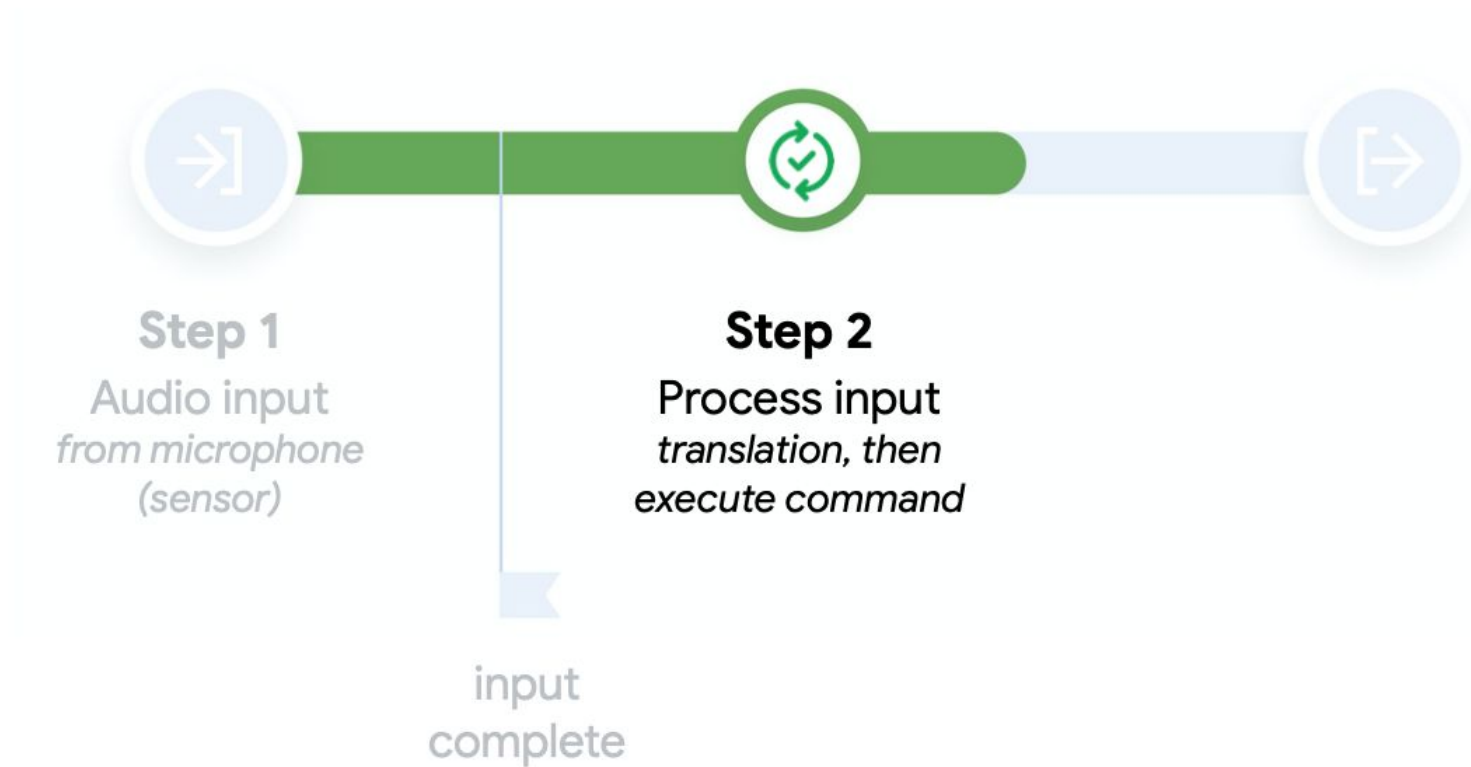
The Three Basic Steps



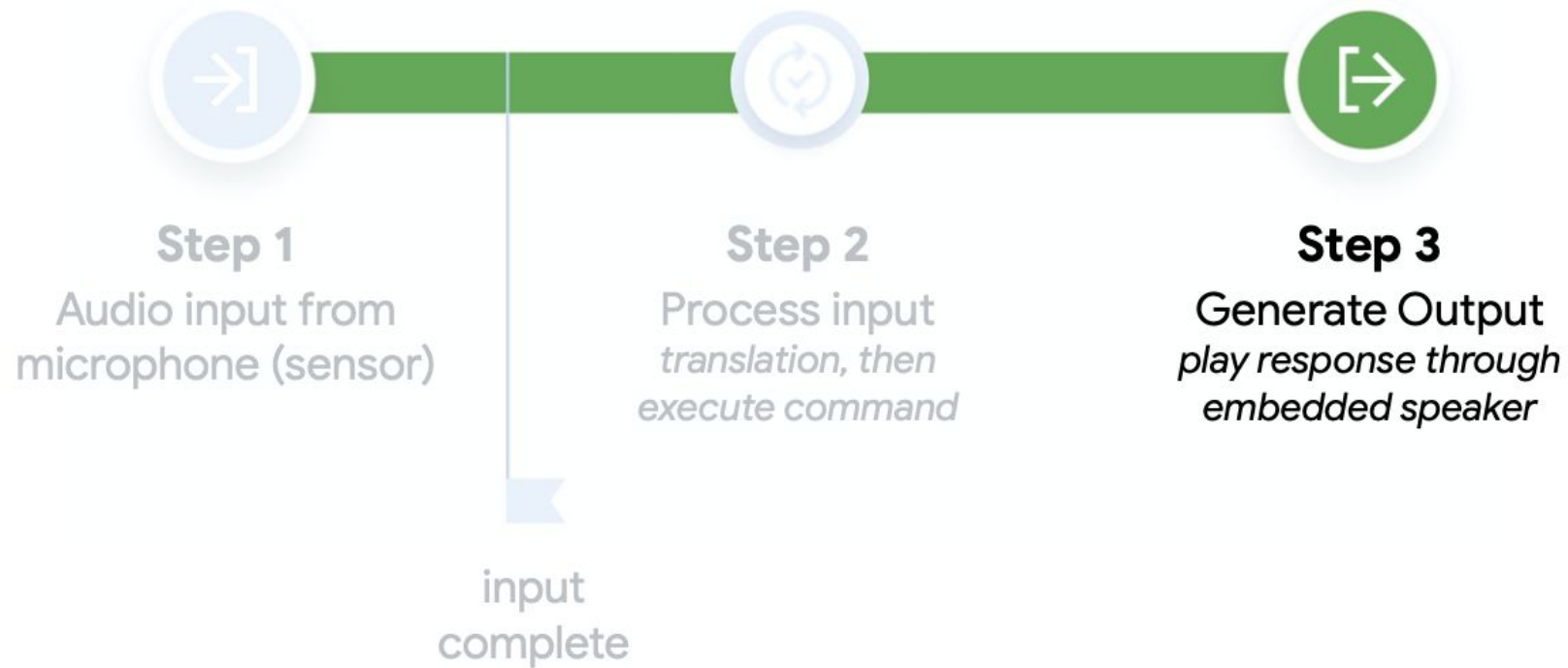
The Three Basic Steps



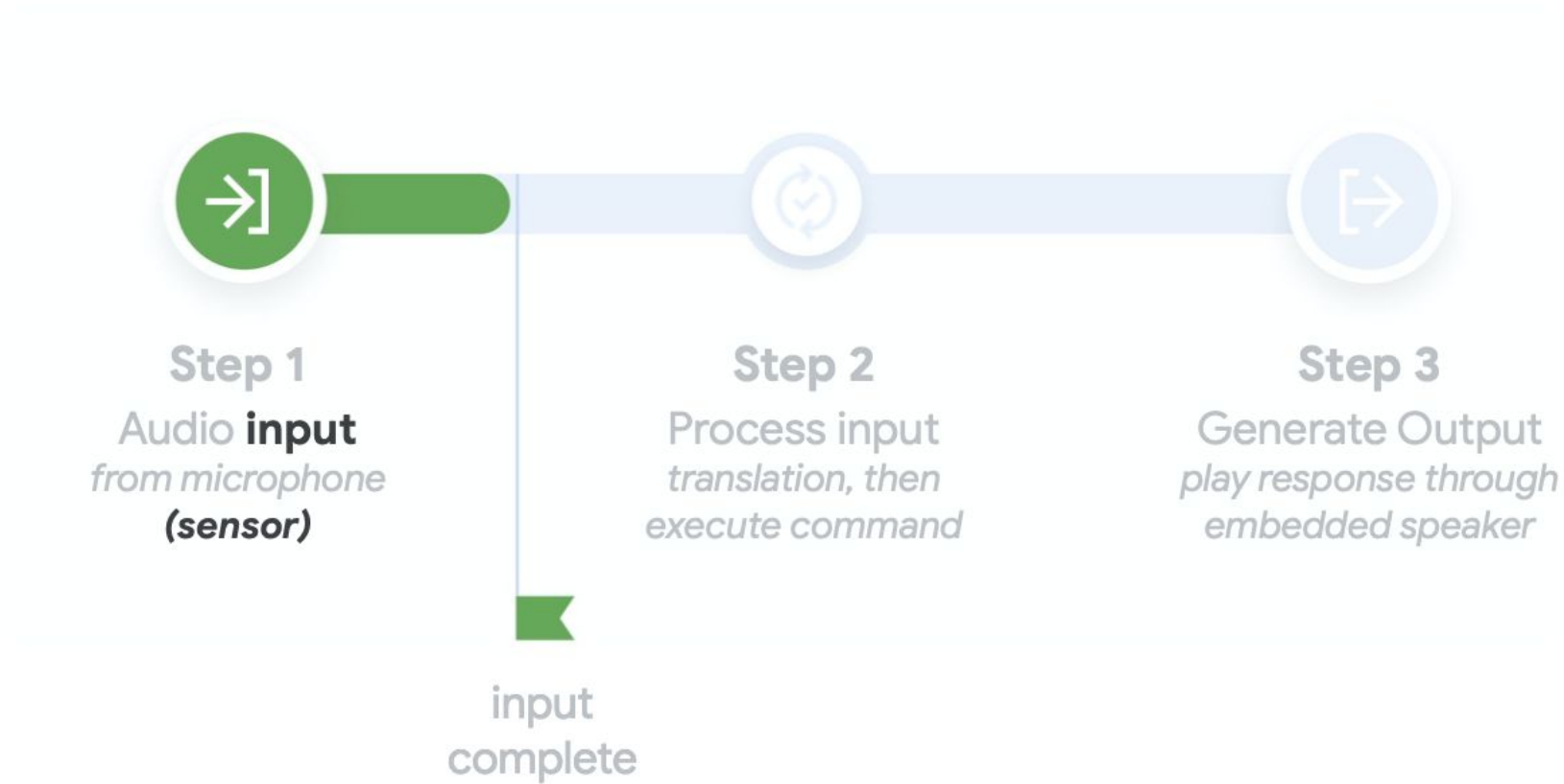
The Three Basic Steps



The Three Basic Steps



Input



Endpoints Have **Sensors**, Tons of Sensors

Motion Sensors

Gyroscope, radar,
magnetometer, accelerator

Acoustic Sensors

Ultrasonic, Microphones,
Geophones, Vibrometers

Environmental Sensors

Temperature, Humidity,
Pressure, IR, etc.

Touchscreen Sensors

Capacitive, IR

Image Sensors

Thermal, Image

Biometric Sensors

Fingerprint, Heart rate, etc.

Force Sensors

Pressure, Strain

Rotation Sensors

Encoders

Endpoints Have **Sensors**, Tons of Sensors

Motion Sensors

Gyroscope, radar,
magnetometer, accelerator

Acoustic Sensors

Ultrasonic, Microphones,
Geophones, Vibrometers

Environmental Sensors

Temperature, Humidity,
Pressure, IR, etc.

Touchscreen Sensors

Capacitive, IR

Image Sensors

Thermal, Image

Biometric Sensors

Fingerprint, Heart rate, etc.

Force Sensors

Pressure, Strain

Rotation Sensors

Encoders

Biometric Sensors



Non-invasive Glucose Monitoring



Fingerprint + Photoplethysmography (PPG)



ECG Sensor

Endpoints Have **Sensors**, Tons of Sensors

Motion Sensors

Gyroscope, radar,
magnetometer, accelerator

Acoustic Sensors

Ultrasonic, Microphones,
Geophones, Vibrometers

Environmental Sensors

Temperature, Humidity,
Pressure, IR, etc.

Touchscreen Sensors

Capacitive, IR

Image Sensors

Thermal, Image

Biometric Sensors

Fingerprint, Heart rate, etc.

Force Sensors

Pressure, Strain

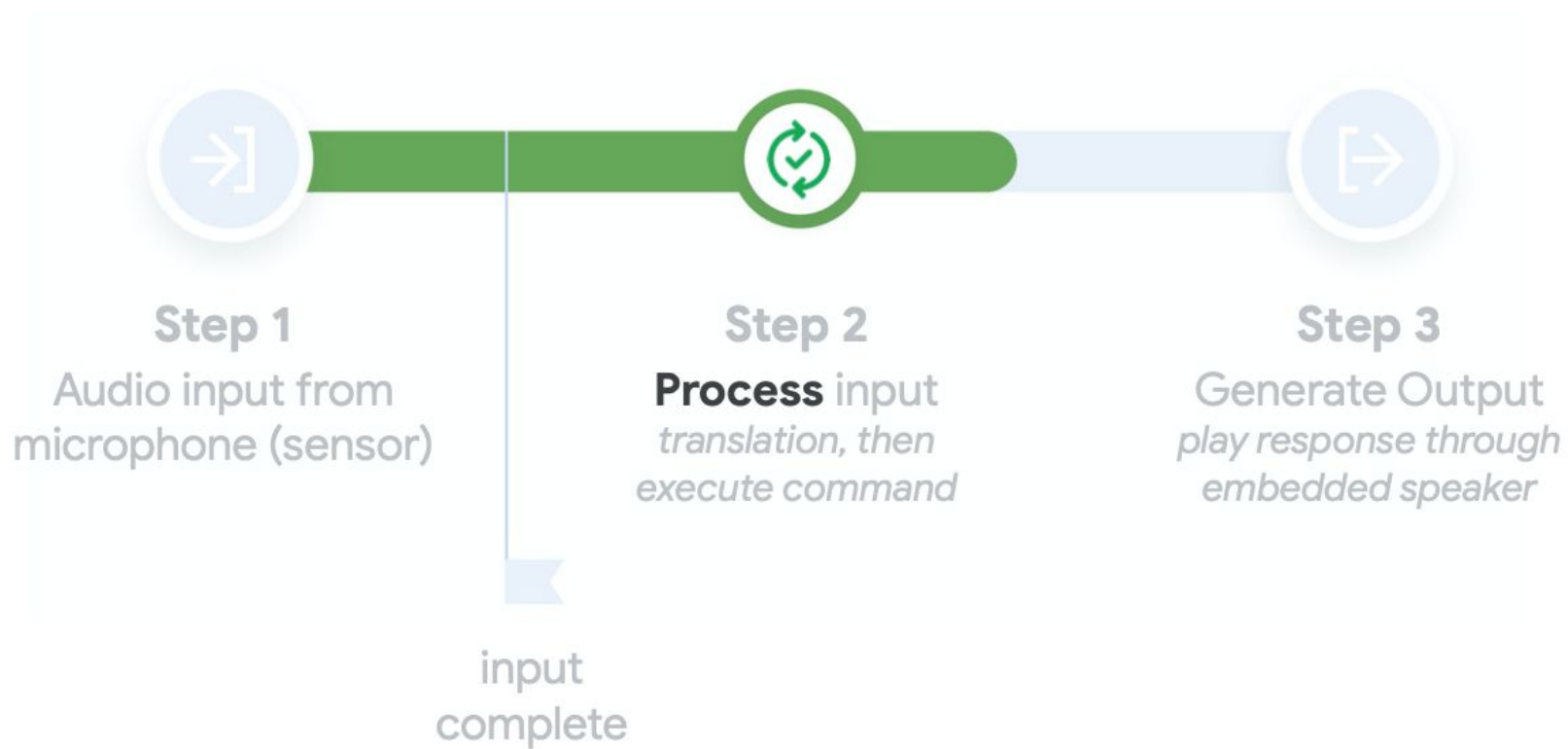
Rotation Sensors

Encoders

Endpoints Have **Sensors**, Tons of Sensors



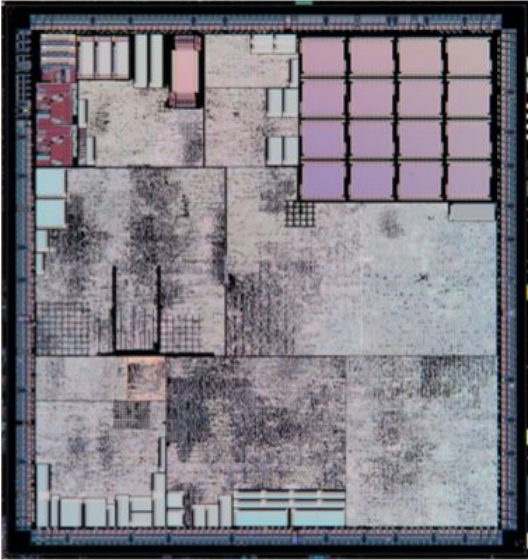
Processing



Thinking Big



Thinking **Big**



Thinking **Big**



BIG
GPU / CPU
561mm²

Thinking Small

BIG
GPU / CPU
561mm²



Thinking Small

BIG
GPU / CPU
561mm²



Thinking Small



Mobile SoC
83mm²

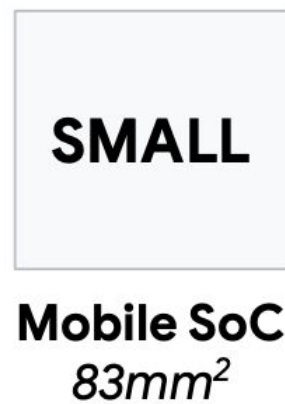
Thinking Tiny



Mobile SoC
83mm²



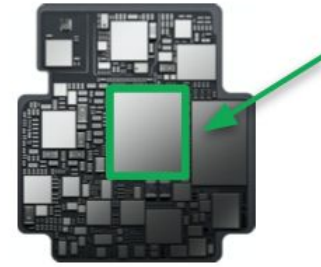
Thinking Tiny



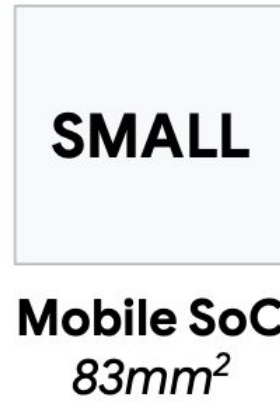
Thinking Tiny



Mobile SoC
83mm²



Thinking Tiny



Thinking Record-breaking



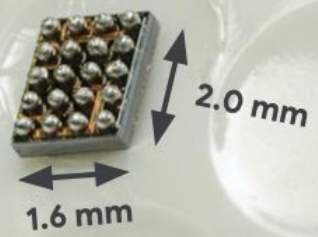
Mobile SoC
83mm²



Apple 0778
30mm²

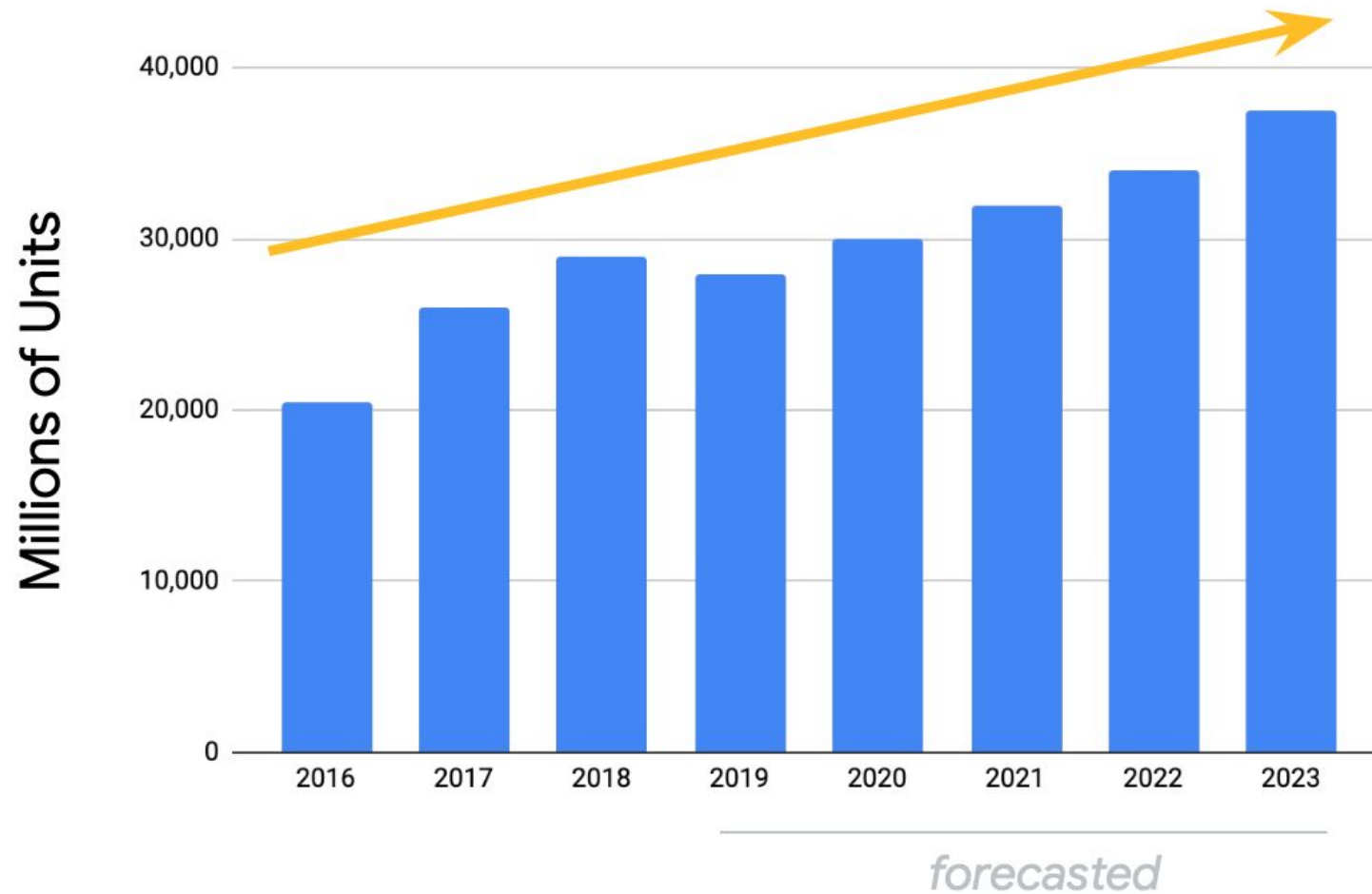


Kinetis KL03
3.2mm²

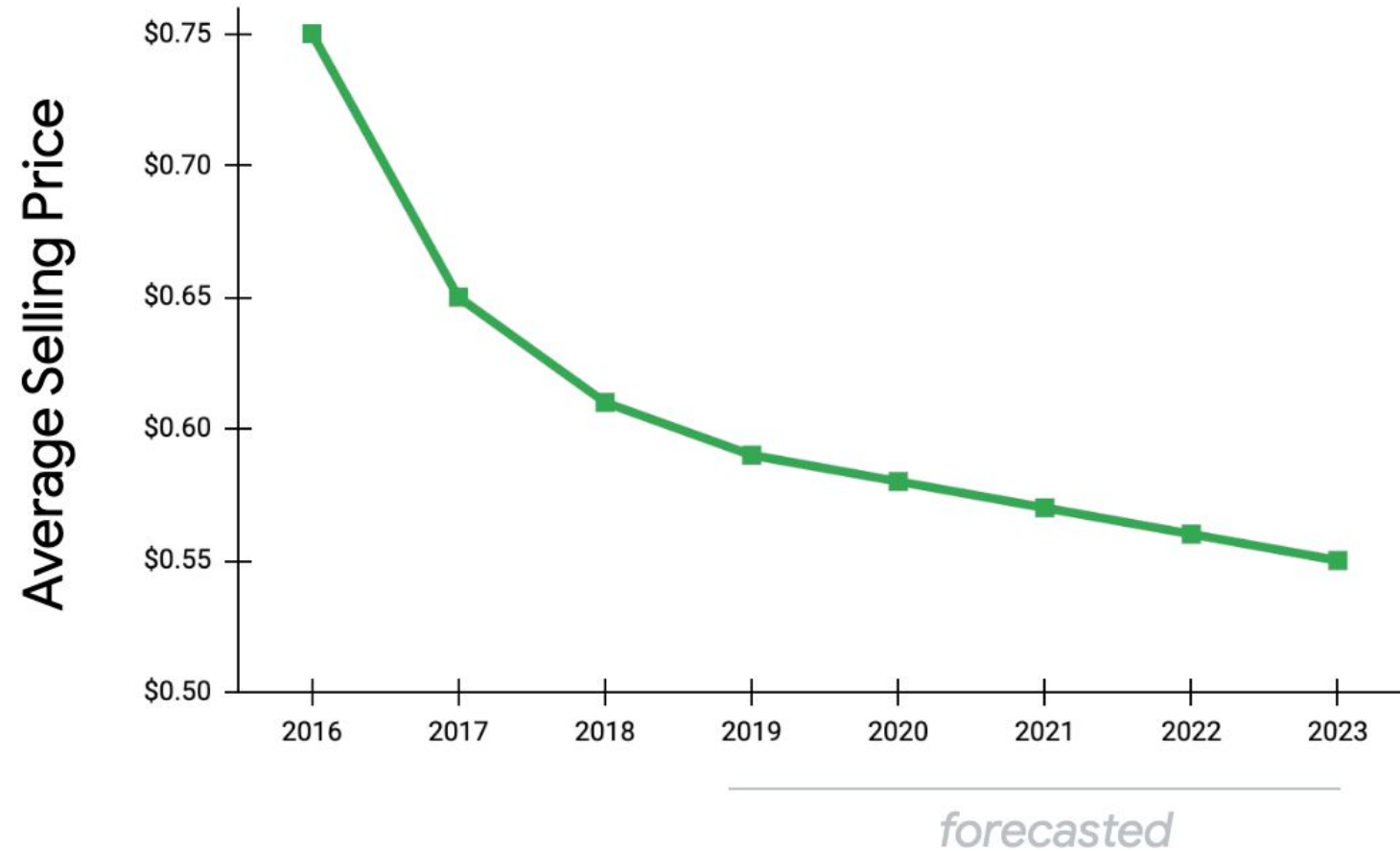


250 Billion
MCUs today

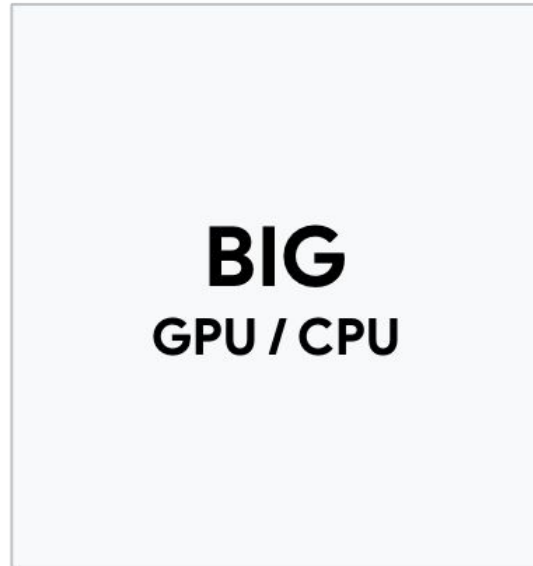
MCU Demand Forecast



MCU Pricing Forecast



Comparing Power



300W
NVIDIA Tesla K80



3.64W
Apple A12

Neural Decision Processor

*Always-on deep learning
speech/audio recognition*

Ultra low power, 128KB SRAM,
12-pin, 2.52mm²



140 μ W
Syntiant NDP100

Comparing Power



Use case: button cell battery

Neural Decision Processor

*Always-on deep learning
speech/audio recognition*

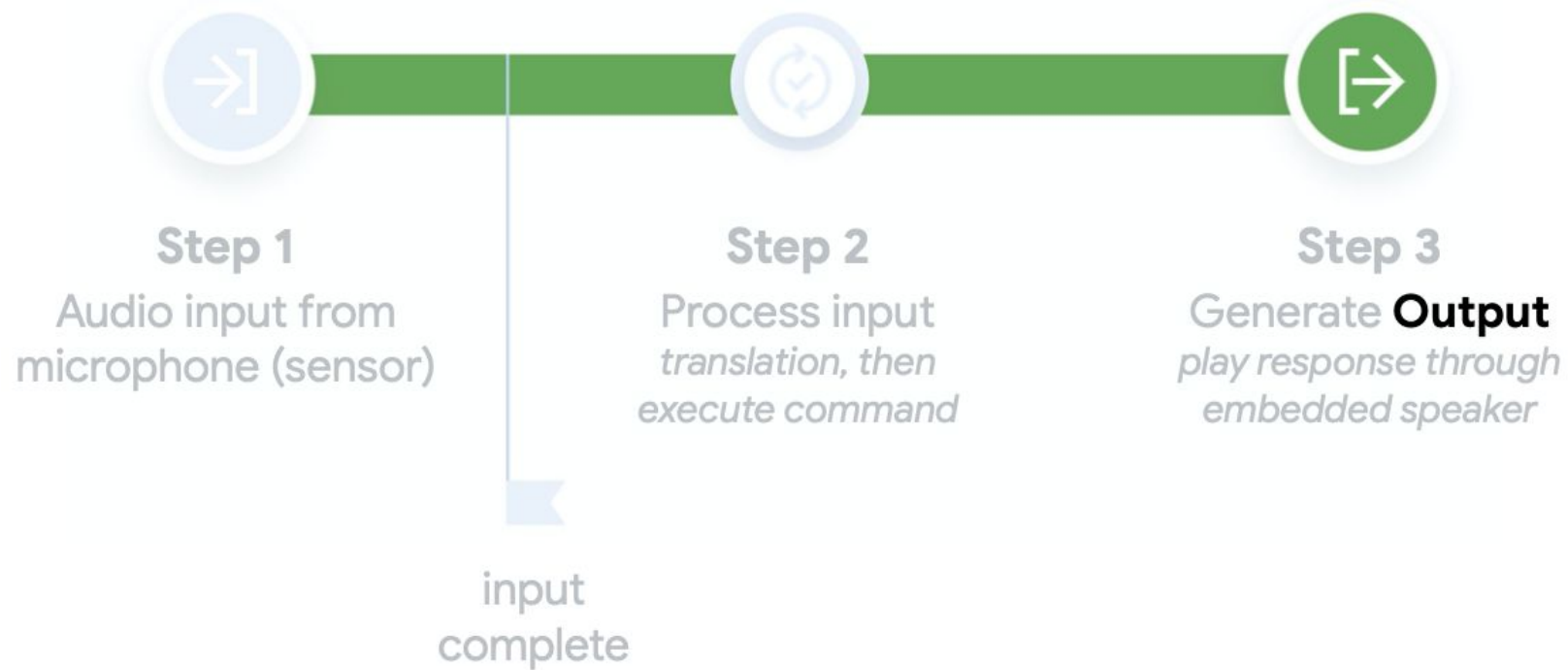
Ultra low power, 128KB SRAM,
12-pin, 2.52mm²



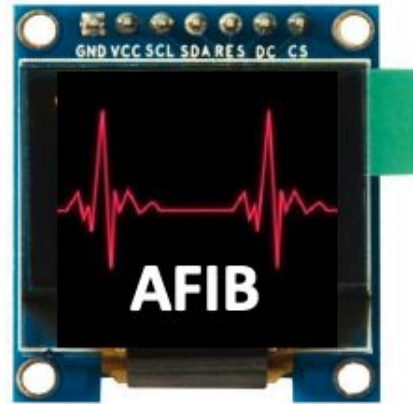
140 μ W

Syntiant NDP100

Output



Output

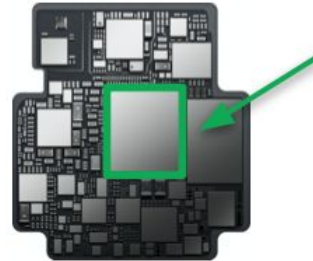


MCUs enable **TinyML**

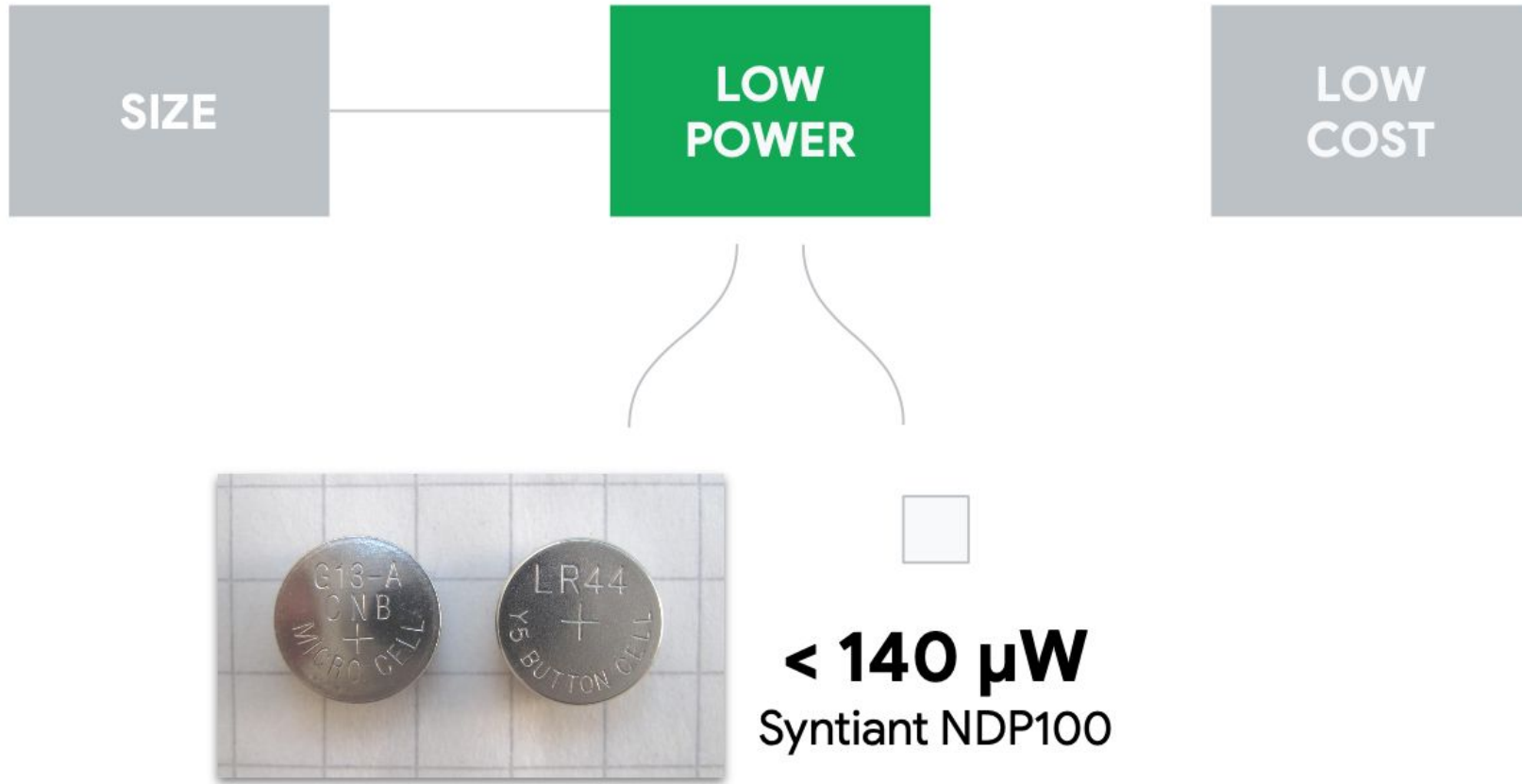
SIZE

LOW
POWER

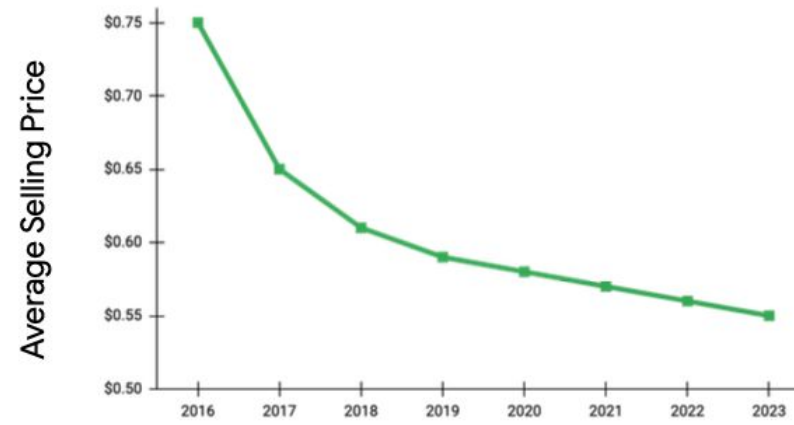
LOW
COST



MCUs enable **TinyML**



MCUs enable **TinyML**



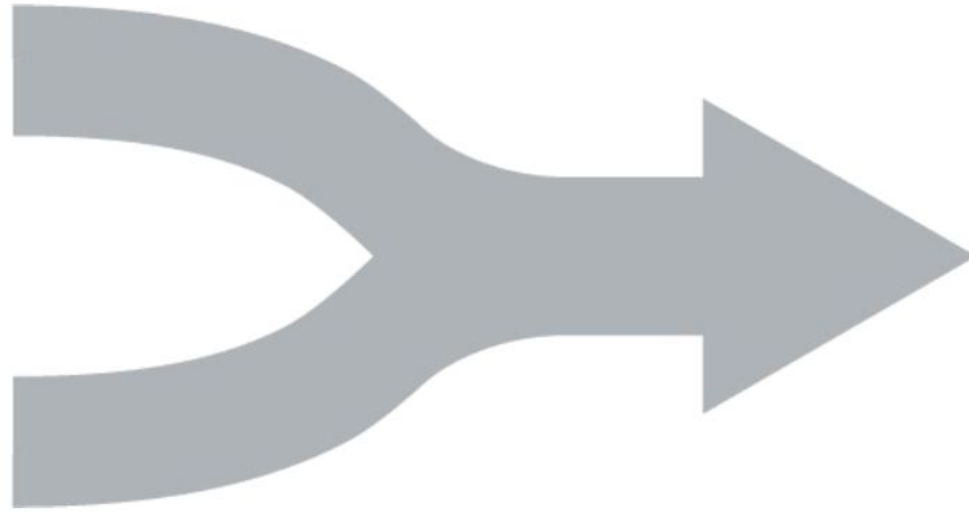
MCUs enable **TinyML**



What Makes **TinyML**?

**Embedded
Systems**

**Machine
Learning**



TinyML

Reading Material

Main references

- [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](#)
- Fundamentals textbook: [“Deep Learning with Python” by François Chollet](#)
- Applications & Deploy textbook: [“TinyML” by Pete Warden, Daniel Situnayake](#)
- Deploy textbook [“TinyML Cookbook” by Gian Marco Iodice](#)

I want to thank **Shawn Hymel** and Edge Impulse, **Pete Warden** and **Laurence Moroney** from Google, Professor **Vijay Janapa Reddi** and **Brian Plancher** from Harvard, and the rest of the **TinyMLedu** team for preparing the excellent material on TinyML that is the basis of this course at UNIFEI.

The IESTI01 course is part of the **TinyML4D**, an initiative to make TinyML education available to everyone globally.

Thanks



UNIFEI