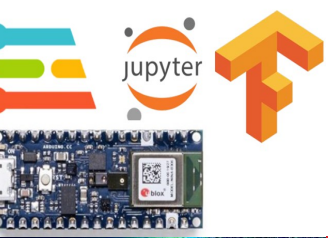
AI

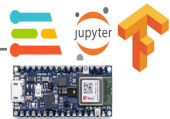
INTRODUCTION TO TINYML

Dennis A. N. Gookyi



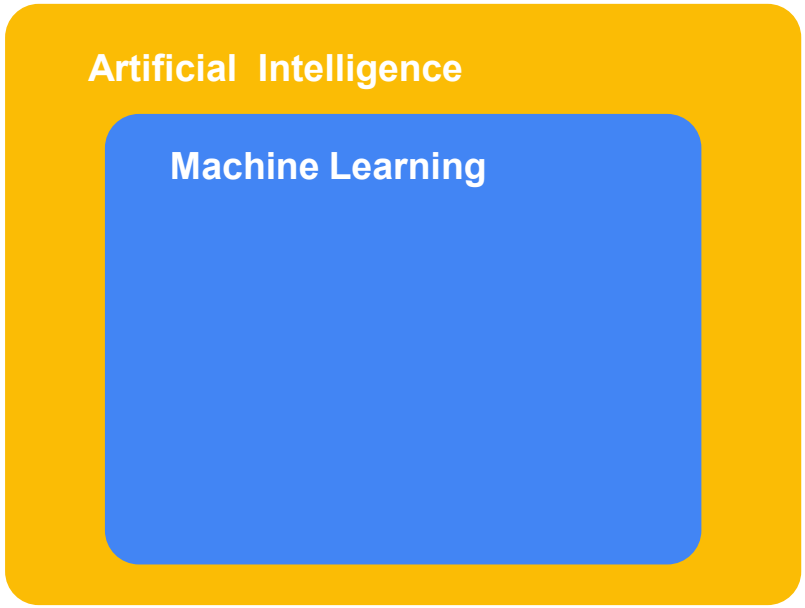
CONTENTS

- Introduction TO TinyML



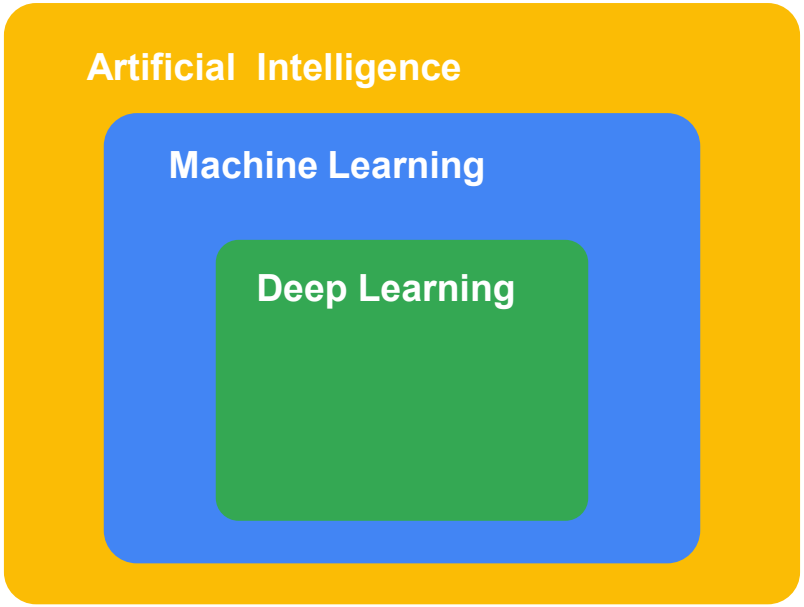
MACHINE LEARNING

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



MACHINE LEARNING

- Deep Learning
  - Deep Learning is a type of Machine Learning that leverages Neural Networks and Big Data



APPLICATIONS MACHINE LEARNING

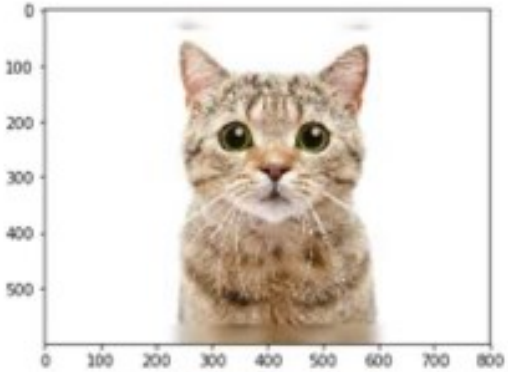
- Applications of machine learning

APPLICATIONS MACHINE LEARNING

- Image classification

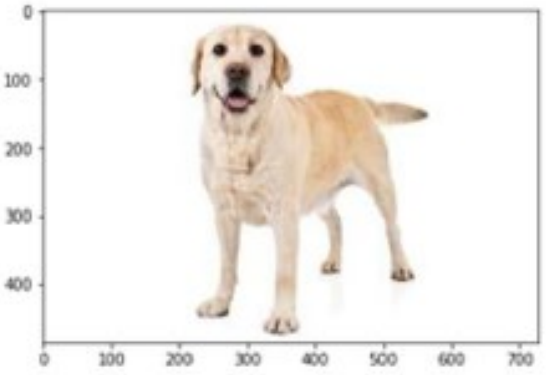
```
[PREDICTION]        [Prob]

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



```
[PREDICTION]        [Prob]

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

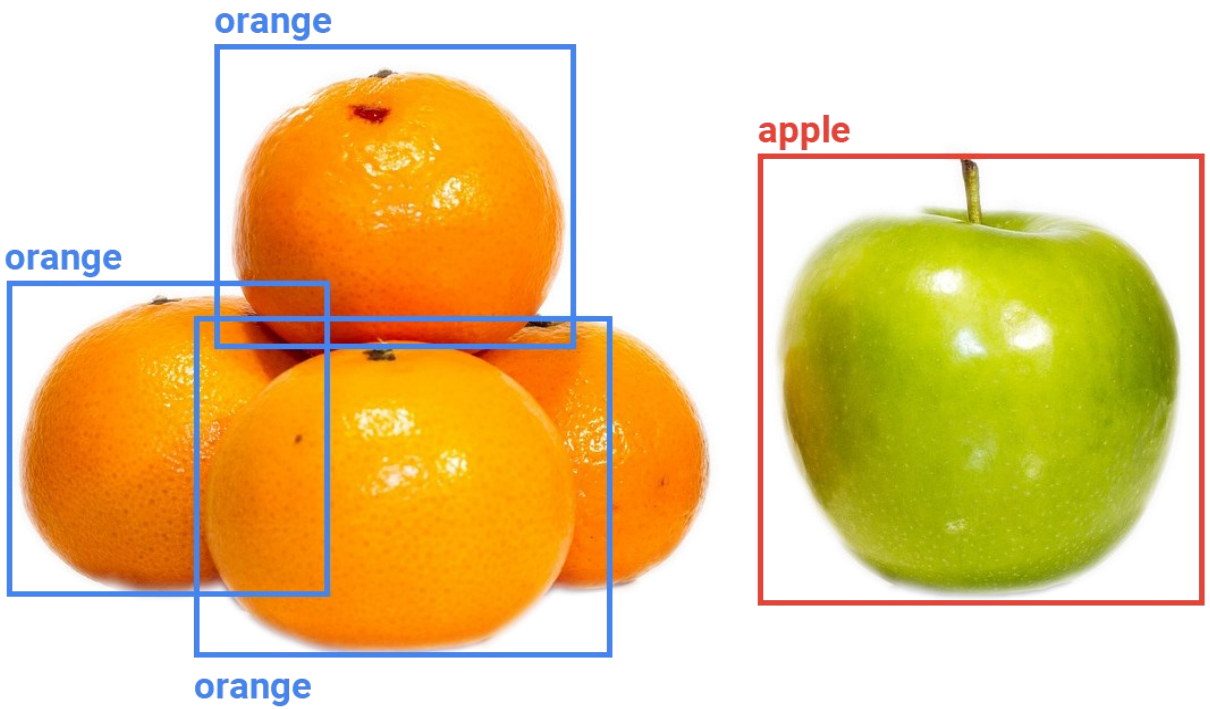


```
[PREDICTION]        [Prob]

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

APPLICATIONS MACHINE LEARNING

- Object detection

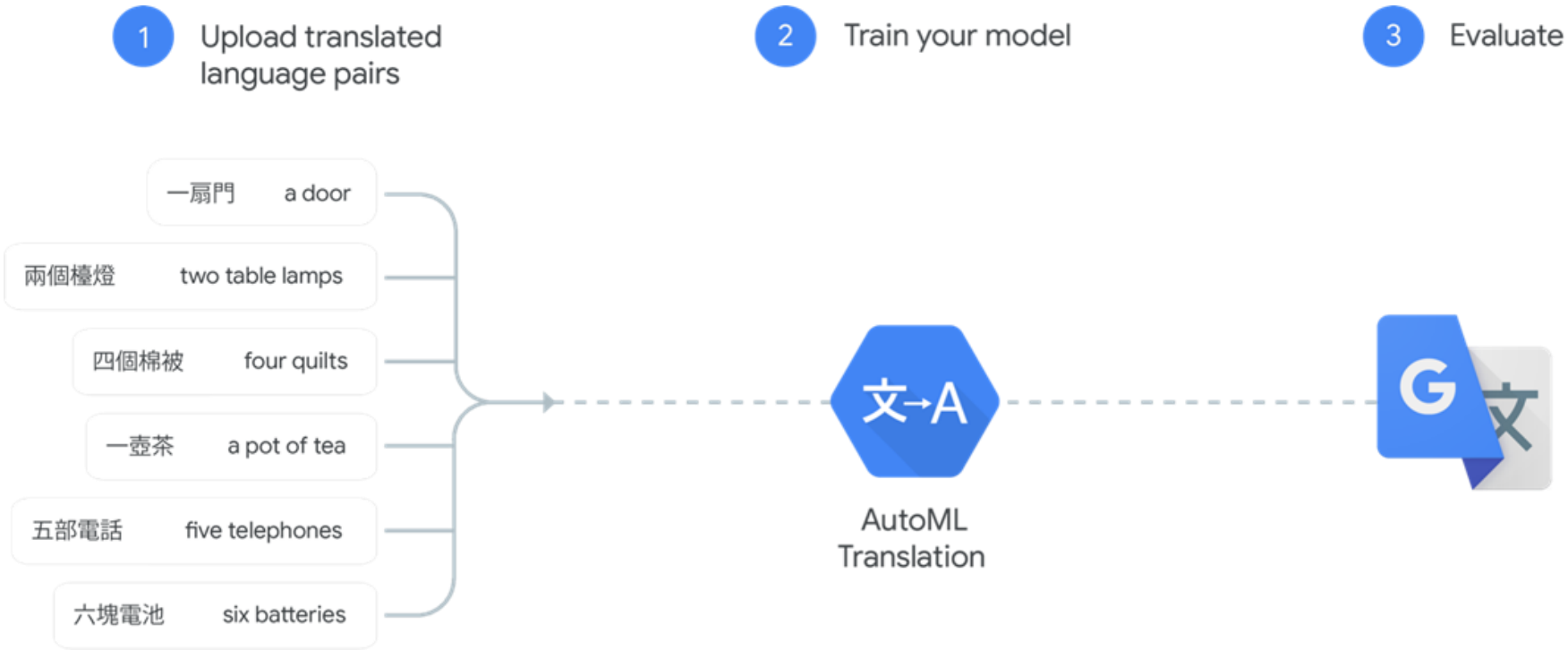


APPLICATIONS MACHINE LEARNING

- Segmentation

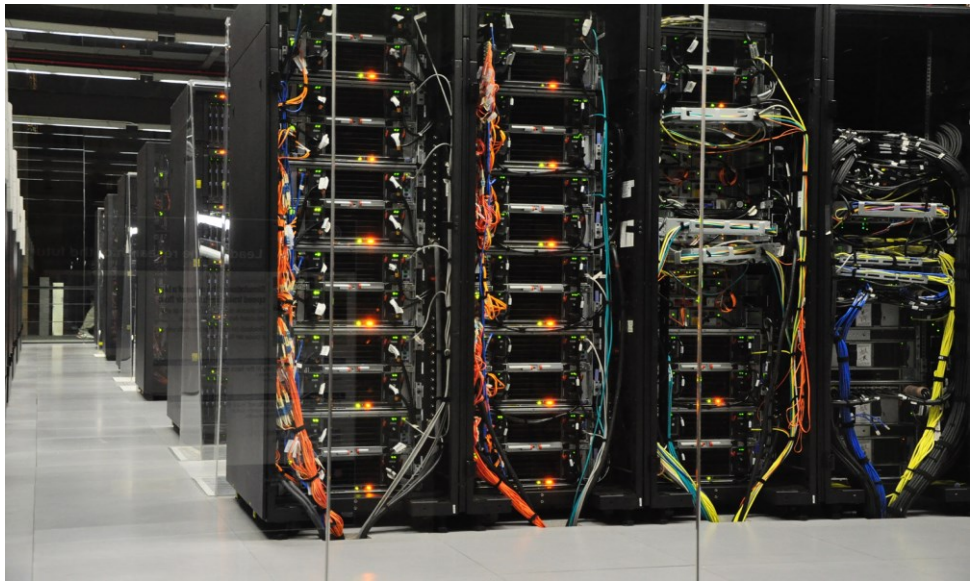
APPLICATIONS MACHINE LEARNING

- Machine translation



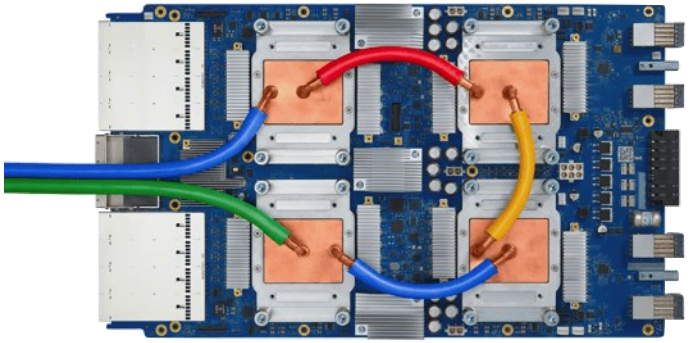
APPLICATIONS MACHINE LEARNING

- Data centers
  - 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



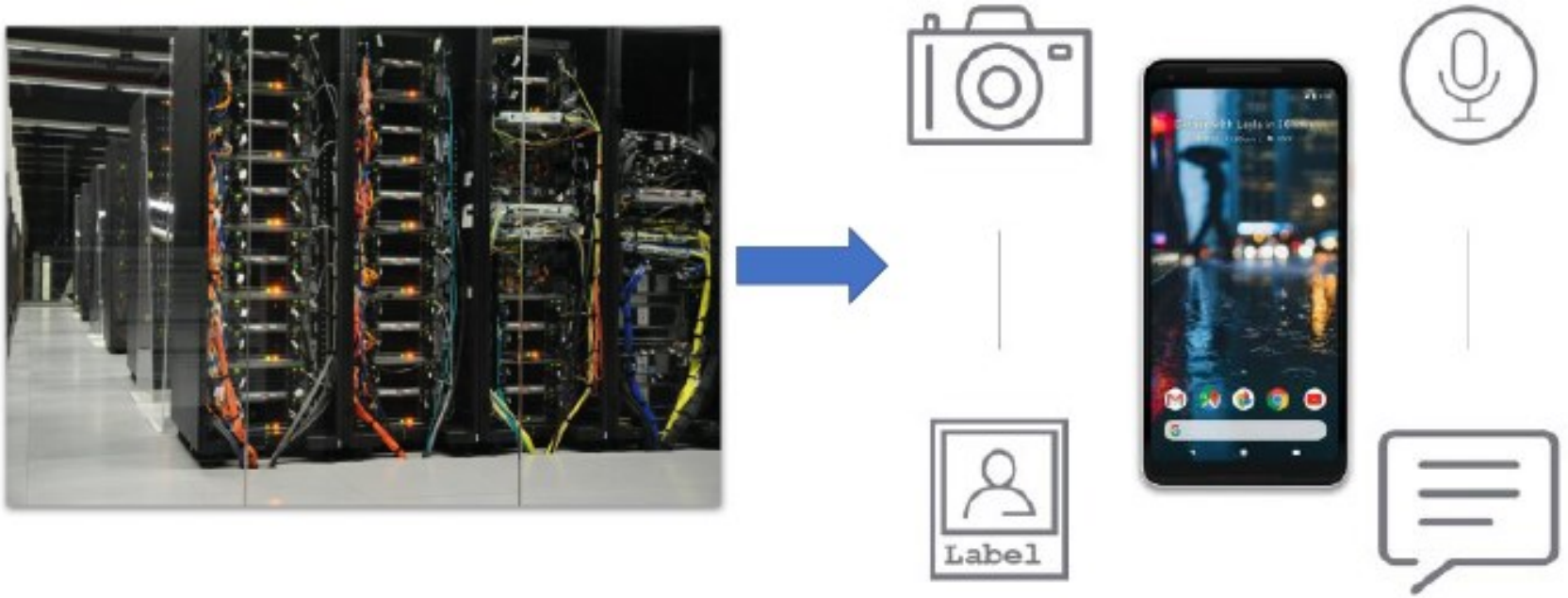
APPLICATIONS MACHINE LEARNING

- 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



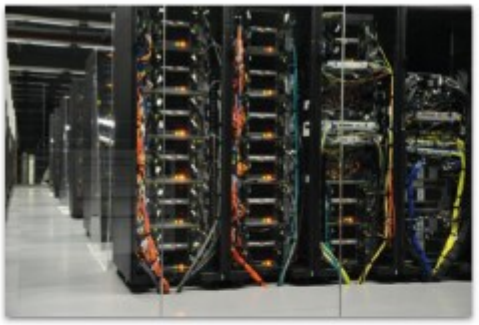
APPLICATIONS MACHINE LEARNING

- Bigger is not always better
  - Because we can not have a Datacenter to do ML inside our phone



APPLICATIONS MACHINE LEARNING

- Bigger is not always better
  - Because we can not have a Datacenter to do ML inside our phone



High power
High bandwidth
High latency

Why?

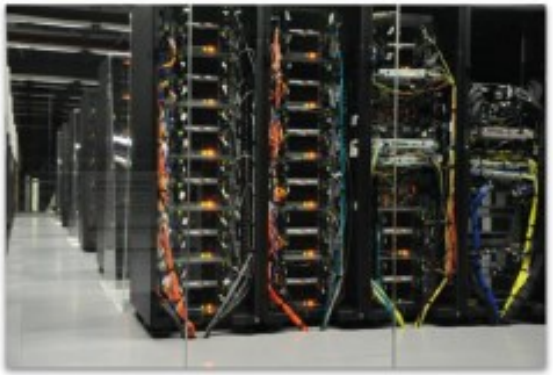
Low power
Low bandwidth
Low latency

APPLICATIONS MACHINE LEARNING

- Bigger is not always better
  - Because we can not have a Datacenter to do ML inside our phone

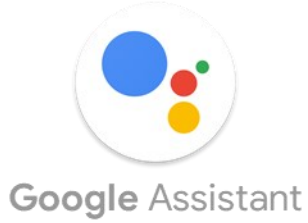
APPLICATIONS MACHINE LEARNING

- Bigger is not always better
  - Because we can not have a Datacenter to do ML inside our phone



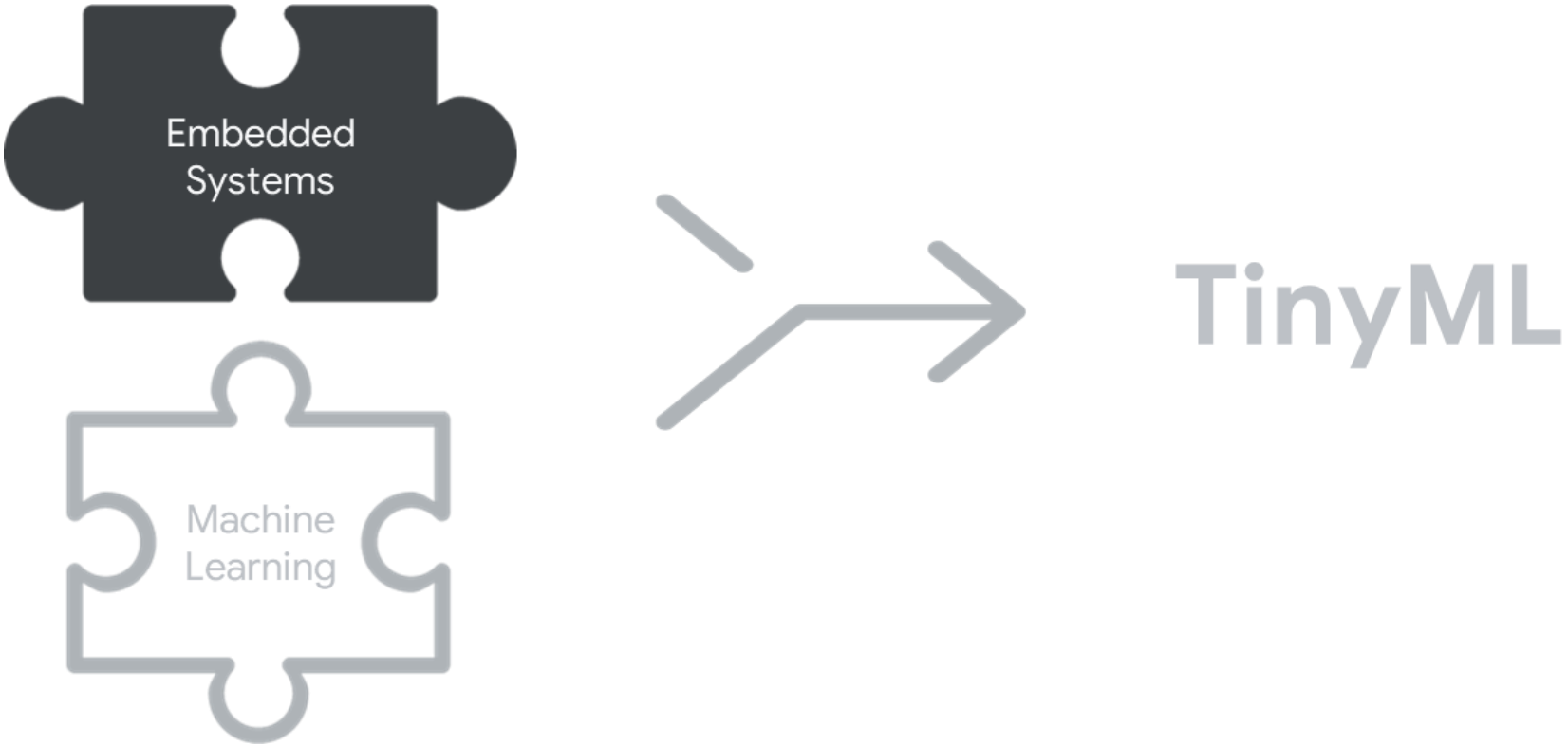
APPLICATIONS MACHINE LEARNING

- End devices



ENABLING TINYML

- What Makes TinyML?



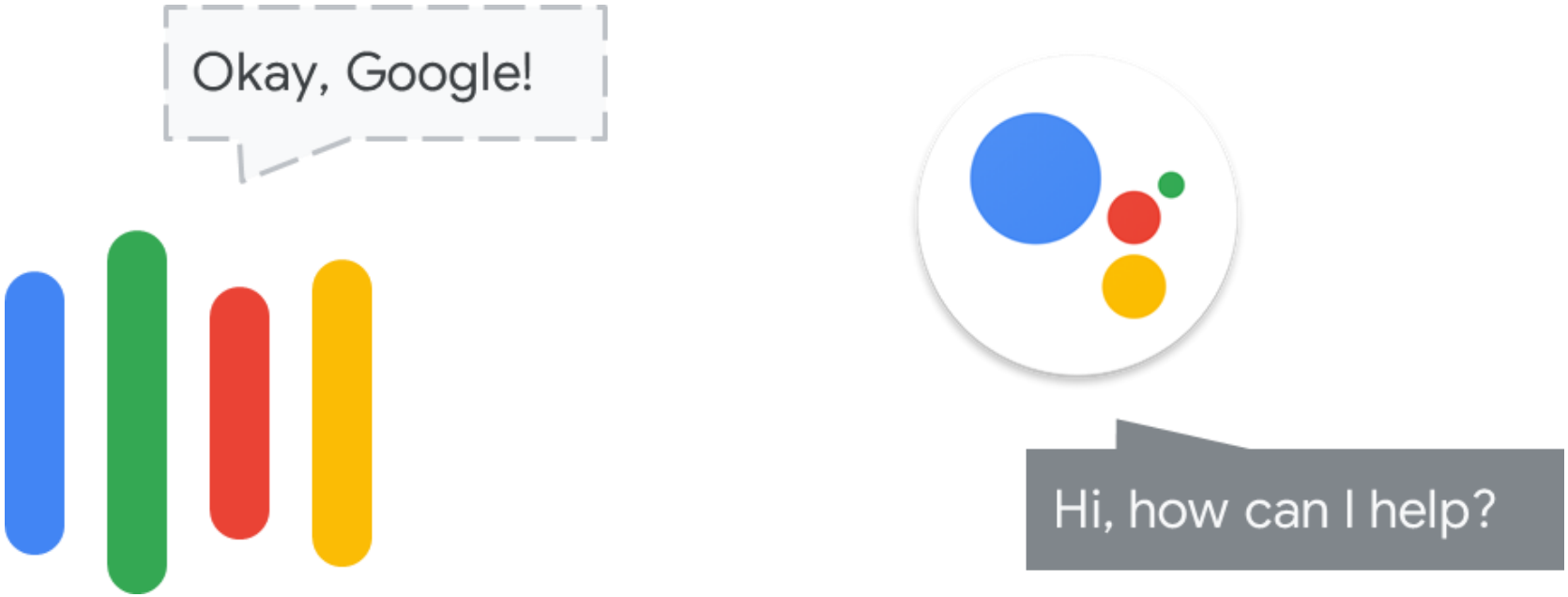
ENABLING TINYML

- Example

Google Assistant

ENABLING TINYML

- Example

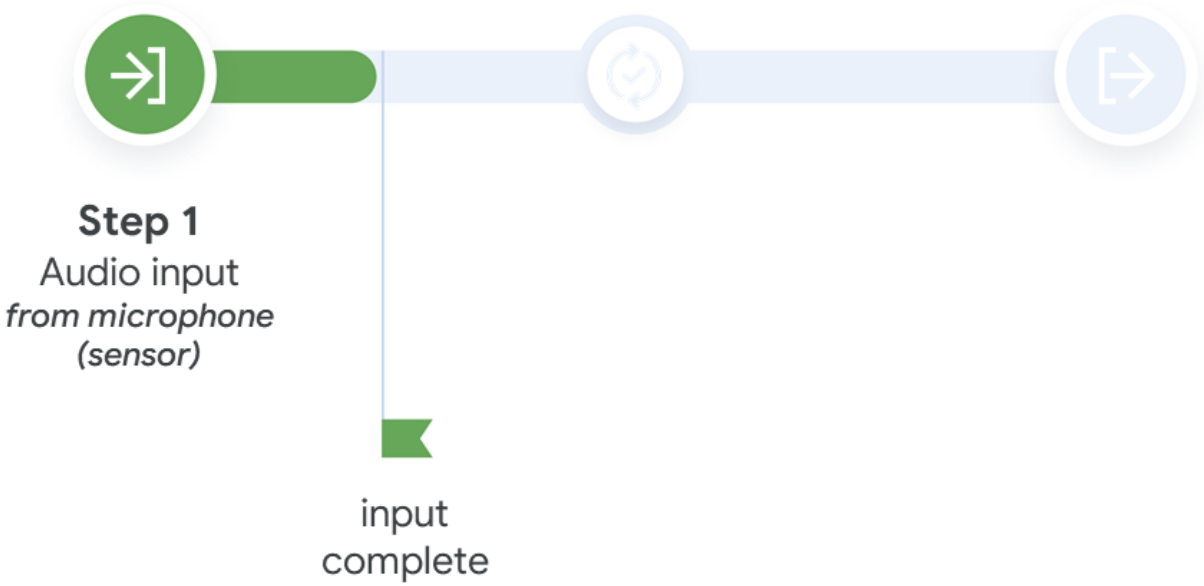


ENABLING TINYML

- The three basic steps

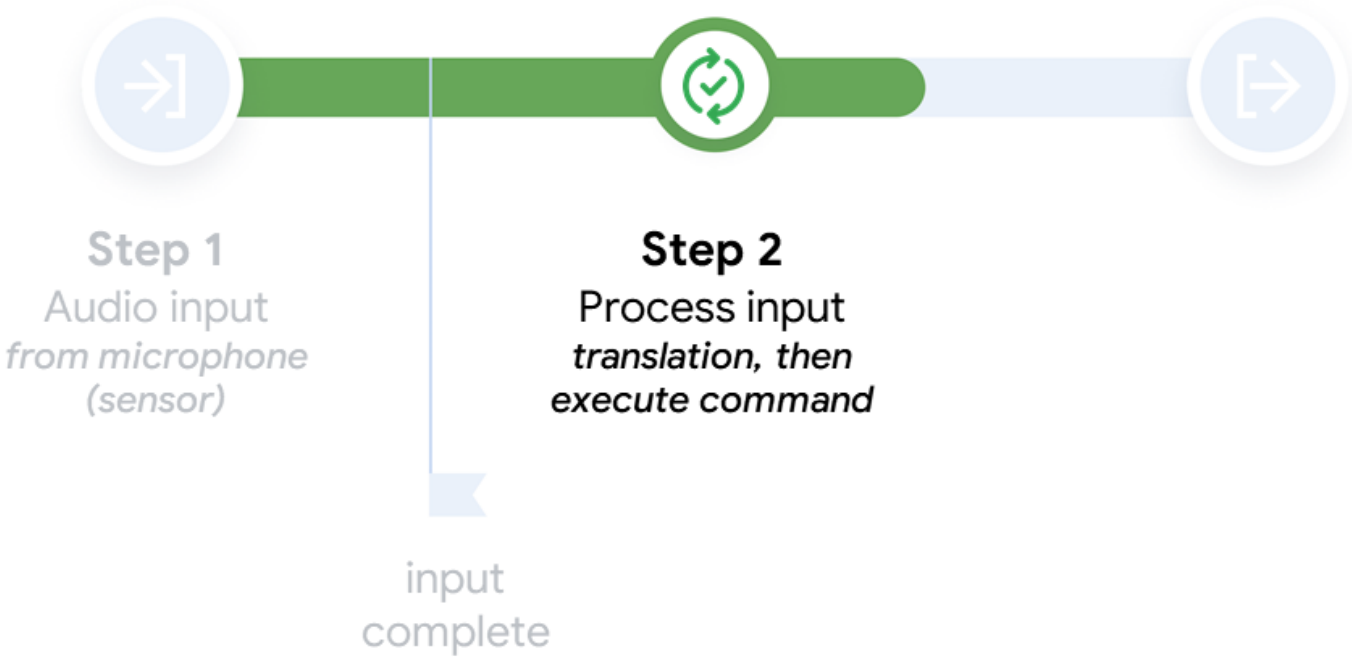
ENABLING TINYML

- The three basic steps



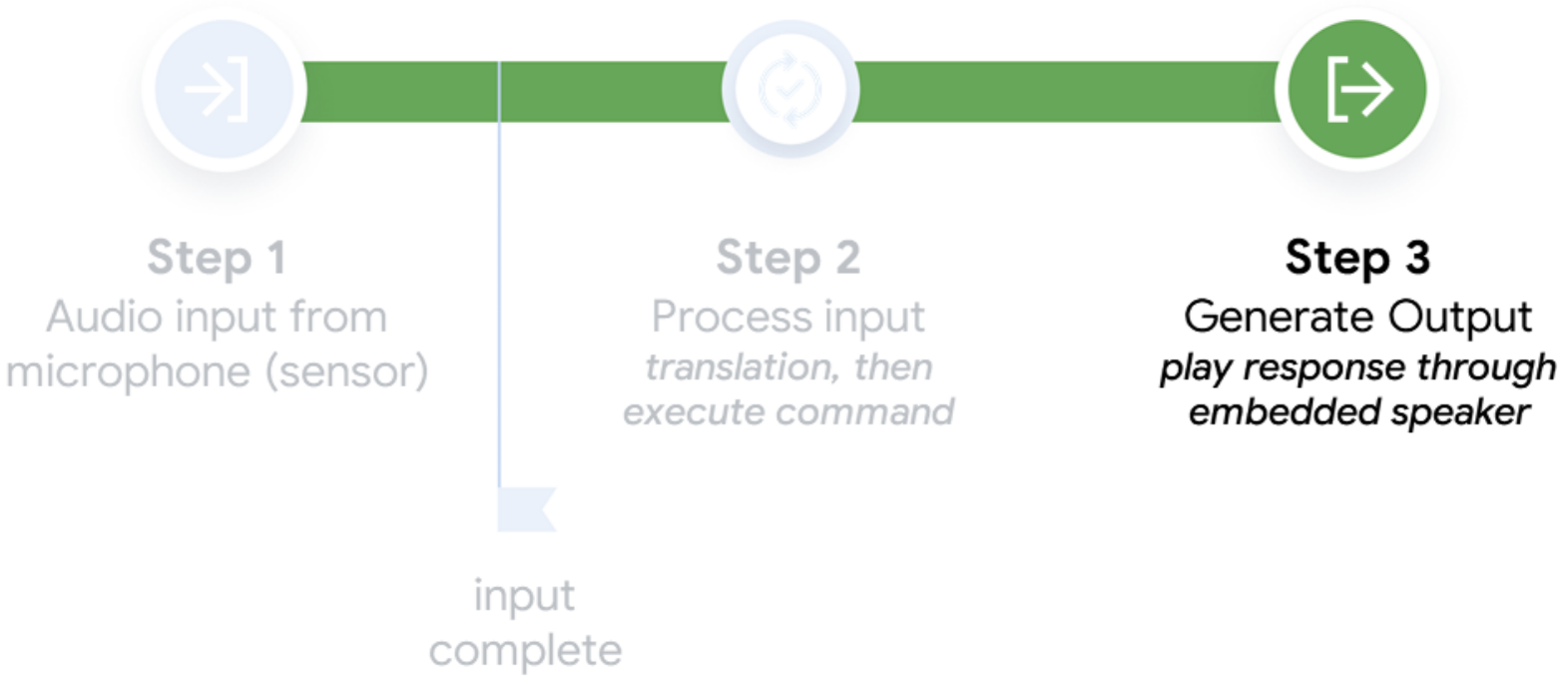
ENABLING TINYML

- The three basic steps



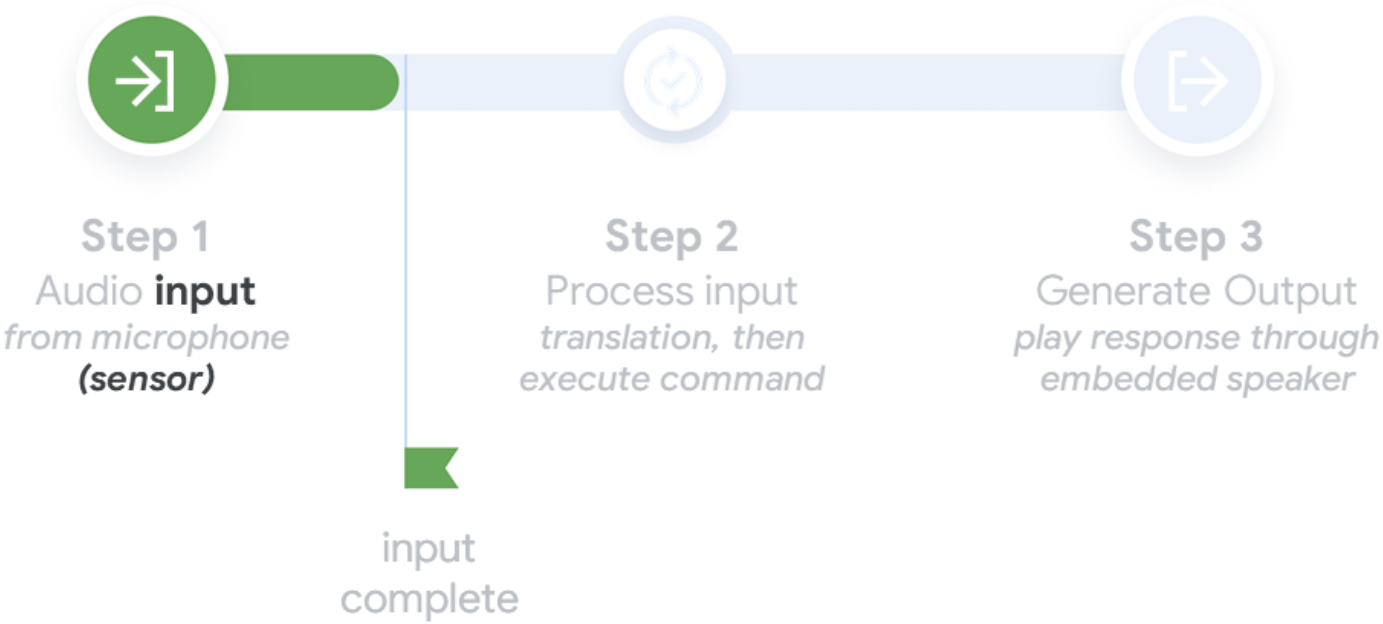
ENABLING TINYML

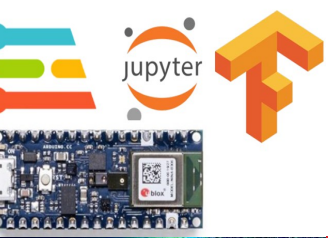
- The three basic steps



ENABLING TINYML

- Inputs





ENABLING TINYML

- 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.

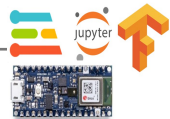
Touchscreen Sensors
Capacitive, IR

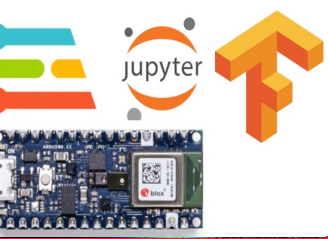
Image Sensors
Thermal, Image

Biometric Sensors
Fingerprint, Heart rate, etc.

Force Sensors
Pressure, Strain

Rotation Sensors
Encoders





ENABLING TINYML

- 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

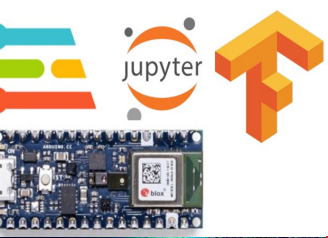
Rotation Sensors
Encoders

ENABLING TINYML

- Biometric sensors

Non-invasive Glucose Monitoring

Fingerprint + Photoplethysmography (**PPG**)



ENABLING TINYML

- 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

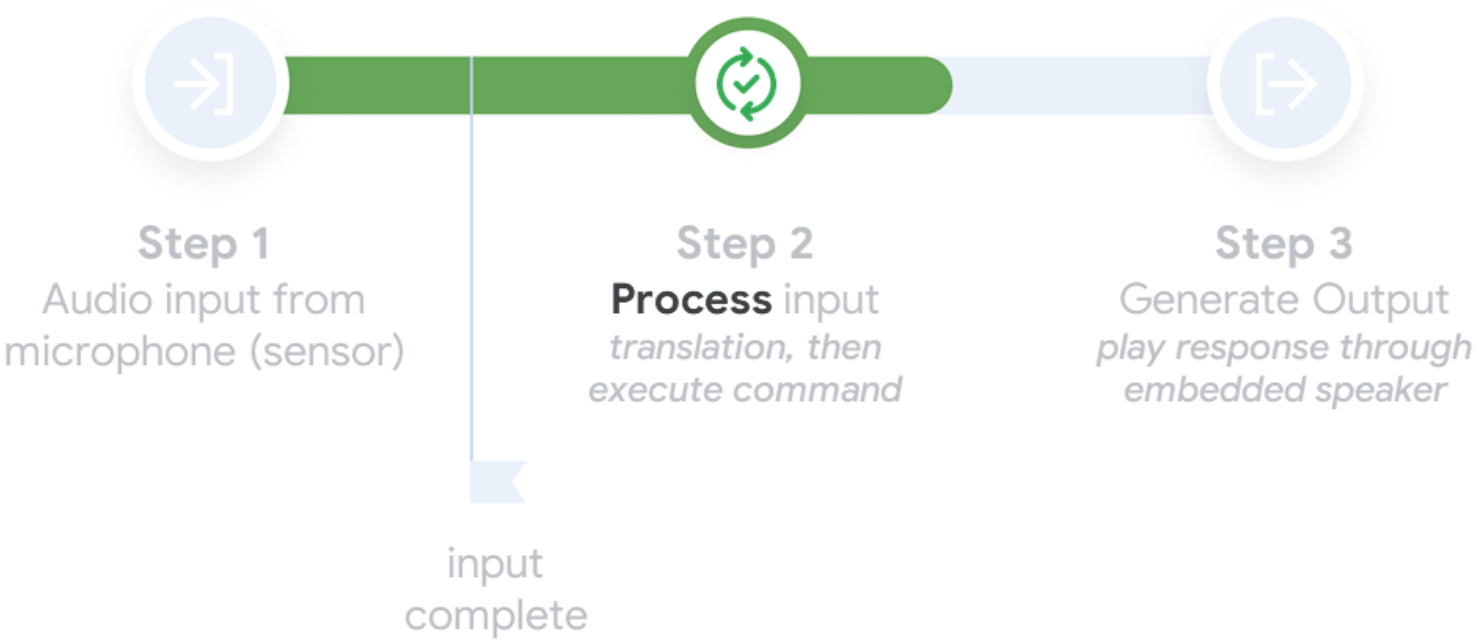
Rotation Sensors
Encoders

ENABLING TINYML

- Endpoints have sensors, tons of sensors

ENABLING TINYML

- Processing

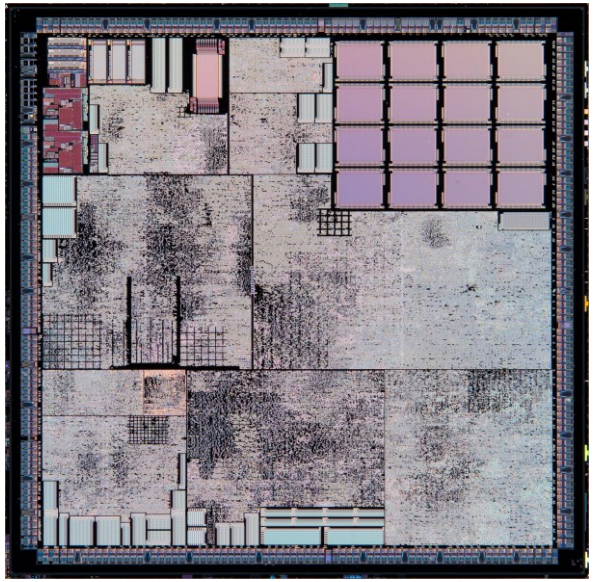


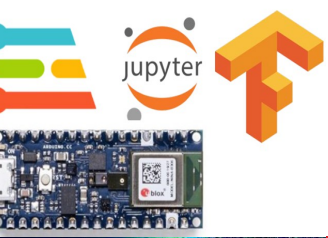
ENABLING TINYML

- Thinking big

ENABLING TINYML

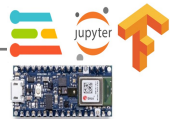
- Thinking big





ENABLING TINYML

- Thinking big



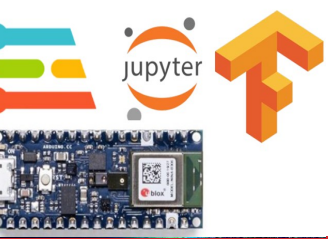
ENABLING TINYML

- Thinking small

BIG

GPU / CPU

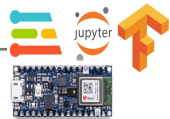
561mm²

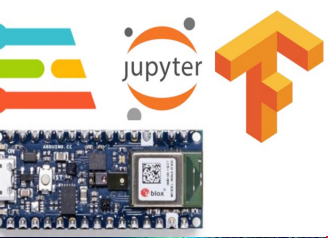


ENABLING TINYML

- Thinking small

BIG
GPU / CPU
561mm²





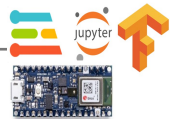
ENABLING TINYML

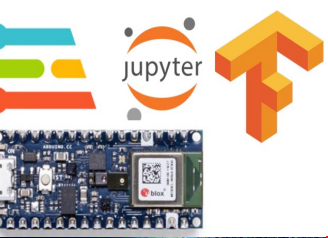
- Thinking small

BIG
GPU / CPU
561mm²

SMALL

Mobile SoC
83mm²





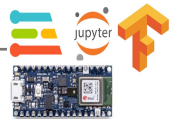
ENABLING TINYML

- Thinking tiny

BIG
GPU / CPU
561mm²

SMALL

Mobile SoC
83mm²

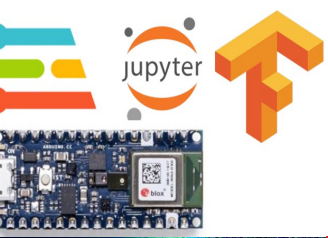


ENABLING TINYML

- Thinking tiny

BIG
GPU / CPU
561mm²

SMALL
Mobile SoC
83mm²



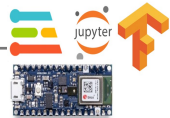
ENABLING TINYML

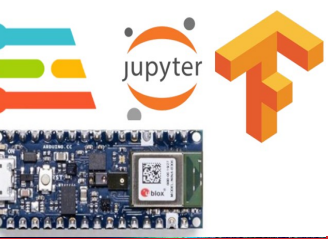
- Thinking tiny

BIG
GPU / CPU
561mm²

SMALL

Mobile SoC
83mm²





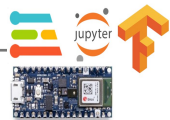
ENABLING TINYML

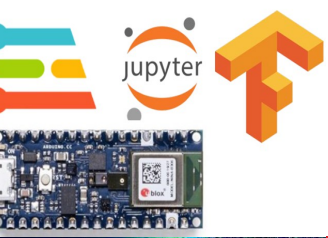
- Thinking tiny

BIG
GPU / CPU
561mm²

SMALL
Mobile SoC
83mm²

TINY
Apple 0778
30mm²





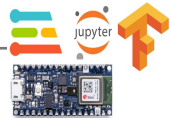
ENABLING TINYML

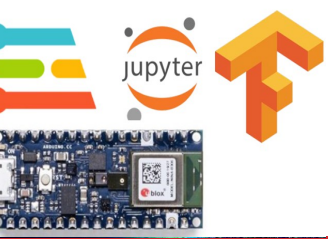
- Thinking record breaking

Mobile SoC
83mm²

Apple 0778
30mm²

Kinetis KL03
3.2mm²





ENABLING TINYML

- Thinking record breaking

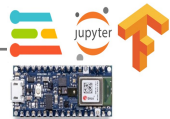
Mobile SoC
83mm²

Apple 0778
30mm²

world's smallest
ARM-Powered MCU

48MHz, 32KB flash, 20-pin

Kinetis KL03
3.2mm²

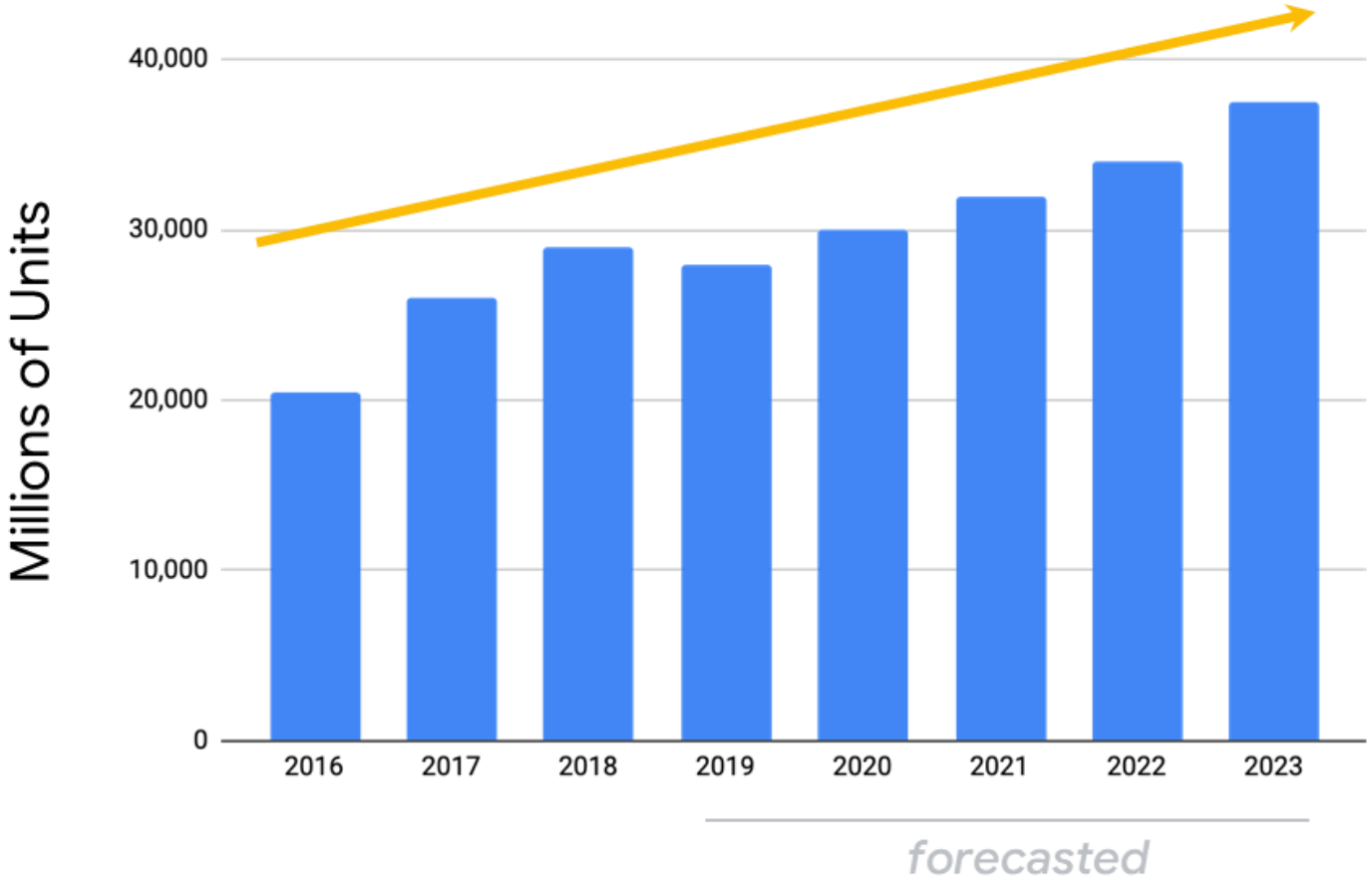


ENABLING TINYML

- Thinking record breaking

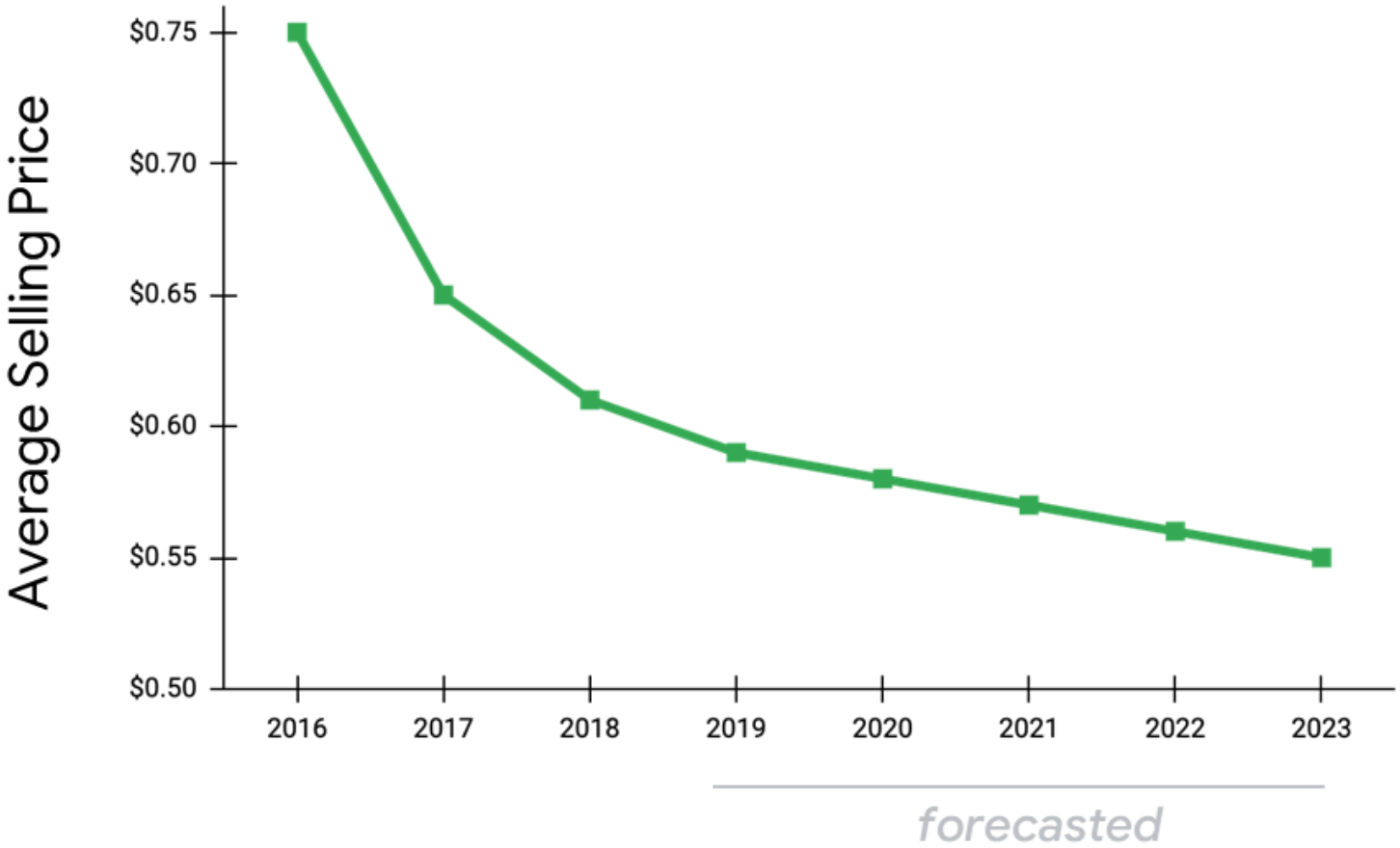
ENABLING TINYML

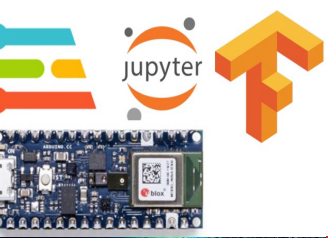
- MCU demand forecast



ENABLING TINYML

- MCU pricing forecast





ENABLING TINYML

- Comparing power

BIG
GPU / CPU

300W
NVIDIA Tesla K80

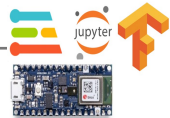
SMALL

3.64W
Apple A12

Neural Decision Processor
Always-on deep learning
speech/audio recognition

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

140 µW
Syntiant NDP100



ENABLING TINYML

- Comparing power

Use case: button cell battery

Neural Decision Processor

Always-on deep learning speech/audio recognition

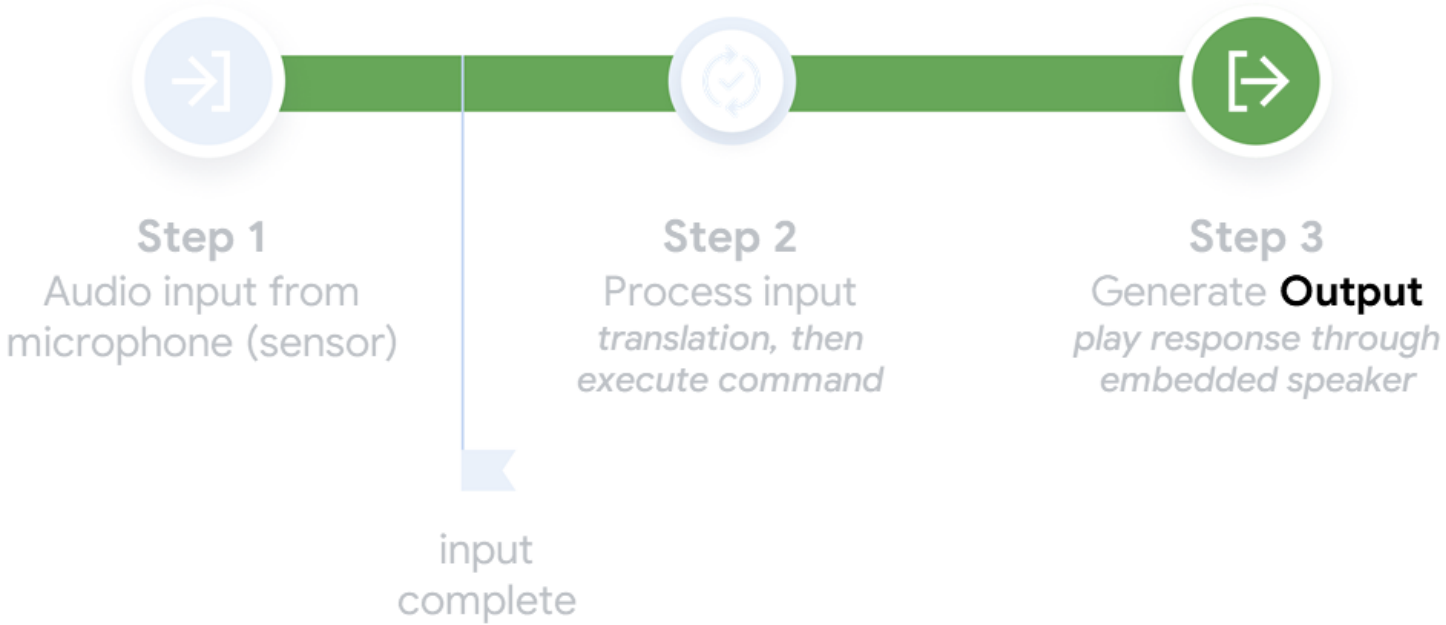
Ultra low power, 128KB SRAM, 12-pin, 2.52mm²

140 µW

Syntiant NDP100

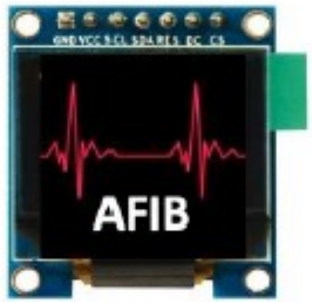
ENABLING TINYML

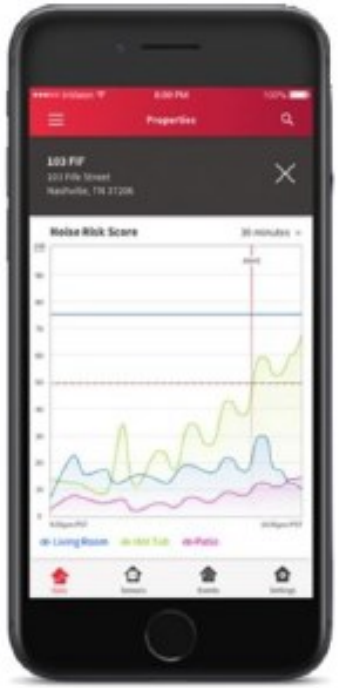
- Output



ENABLING TINYML

- Output





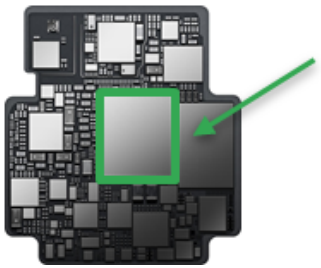
ENABLING TINYML

- MCUs enable TinyML

SIZE

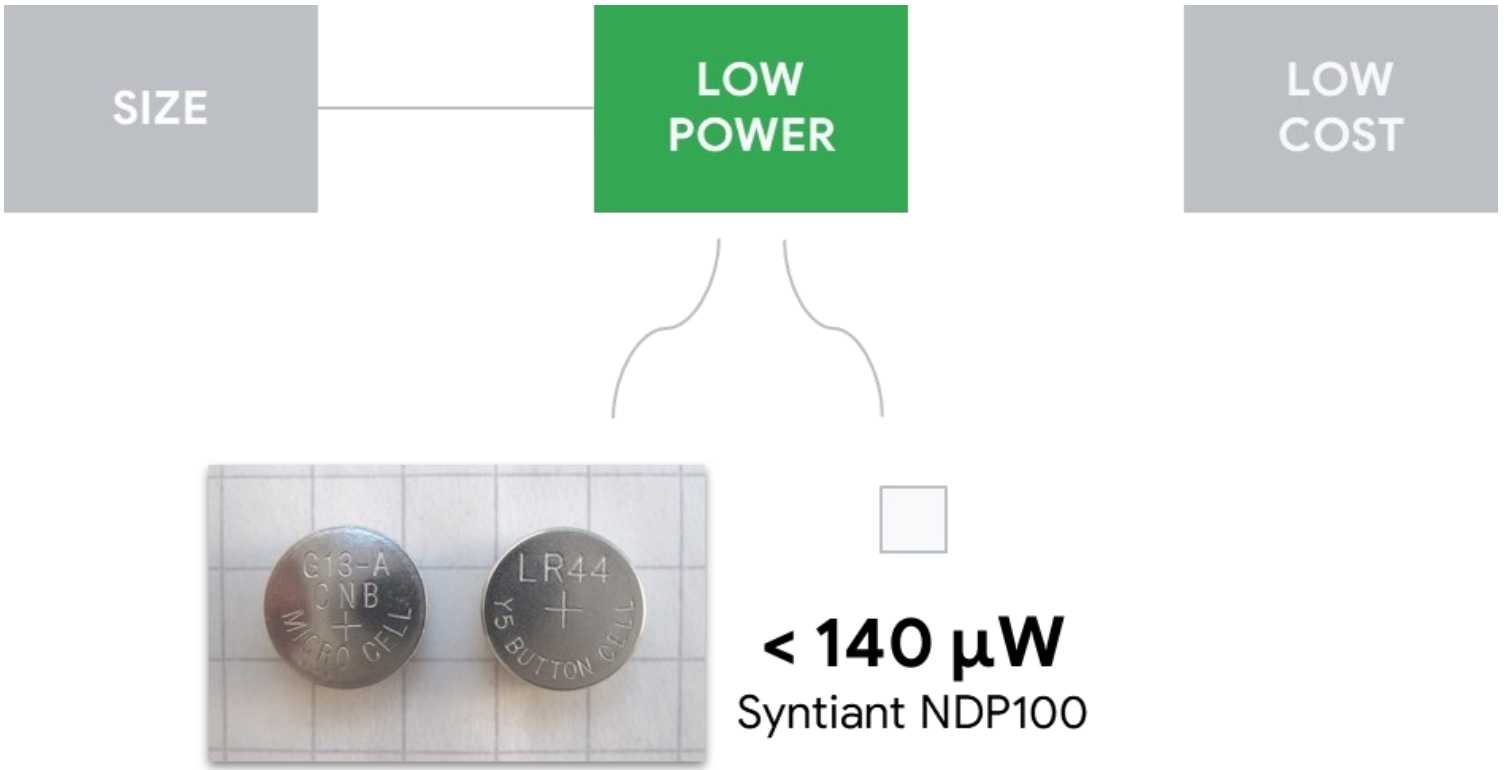
LOW POWER

LOW COST



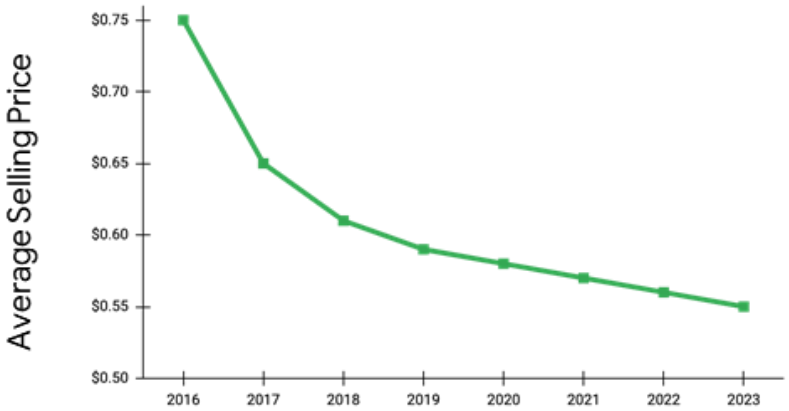
ENABLING TINYML

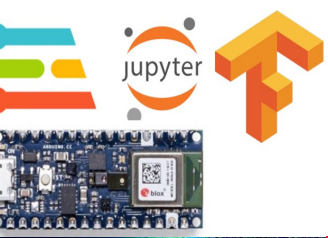
- MCUs enable TinyML



ENABLING TINYML

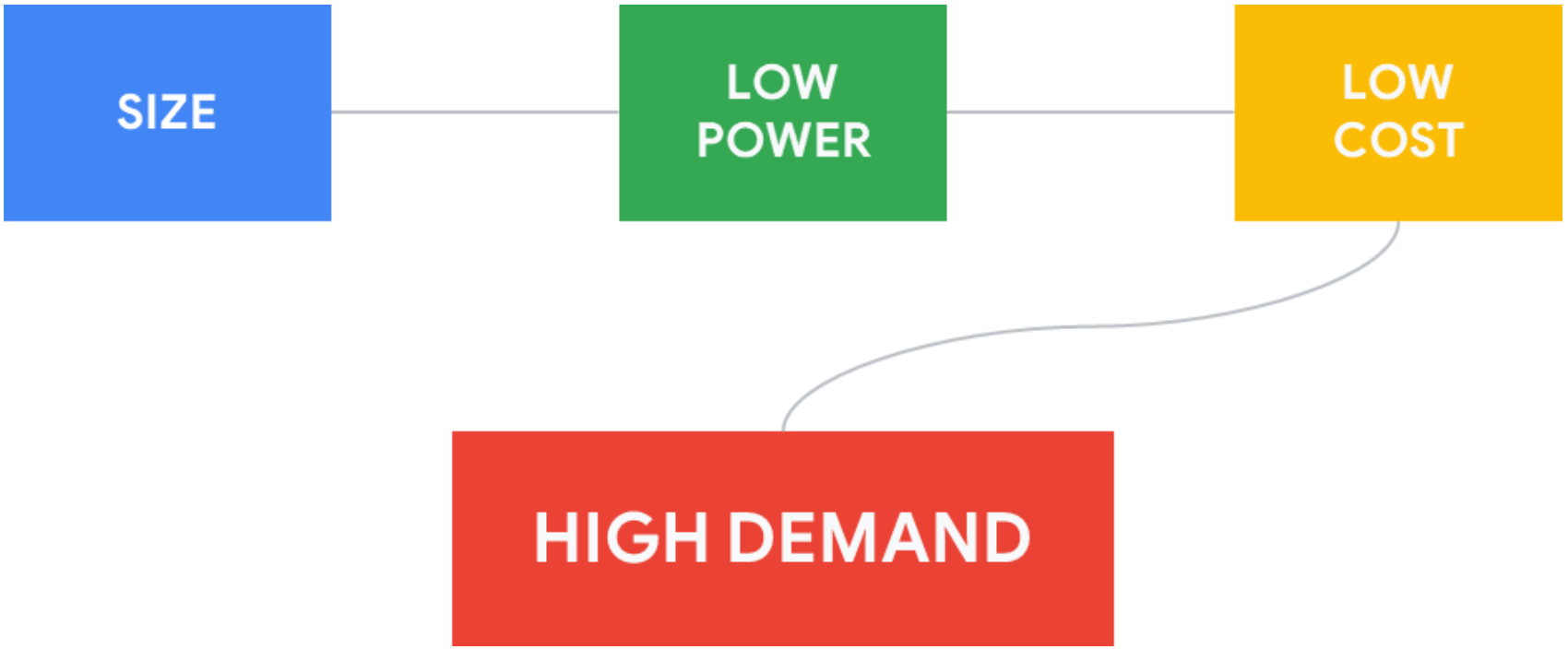
- MCUs enable TinyML

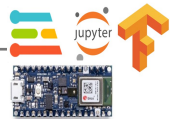




ENABLING TINYML

- MCUs enable TinyML





ENABLING TINYML

- What makes TinyML