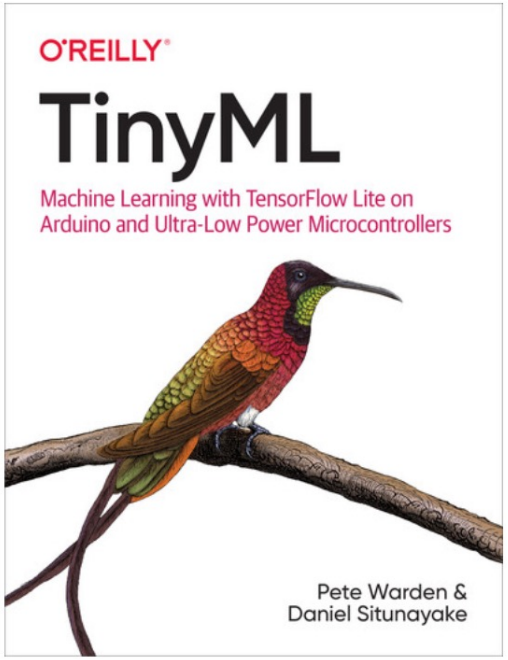
A brief introduction to TinyML



https://learning.oreilly.com/library/view/tinyml/9781492052036/

This presentation is strongly based on:

SciTinyML

Scientific Use of Machine Learning on Low-Power Devices

October 18-22 2021

https://tinyml.seas.harvard.edu/SciTinyML/schedule/

Thanks to:

- **Marcelo Rovai**
- **Vijay Janapa Reddi**
- Archana Vaidheeswaran
- ...and many more!

https://github.com/Mjrovai/UNIFEI-IESTI01-TinyML-2021.2

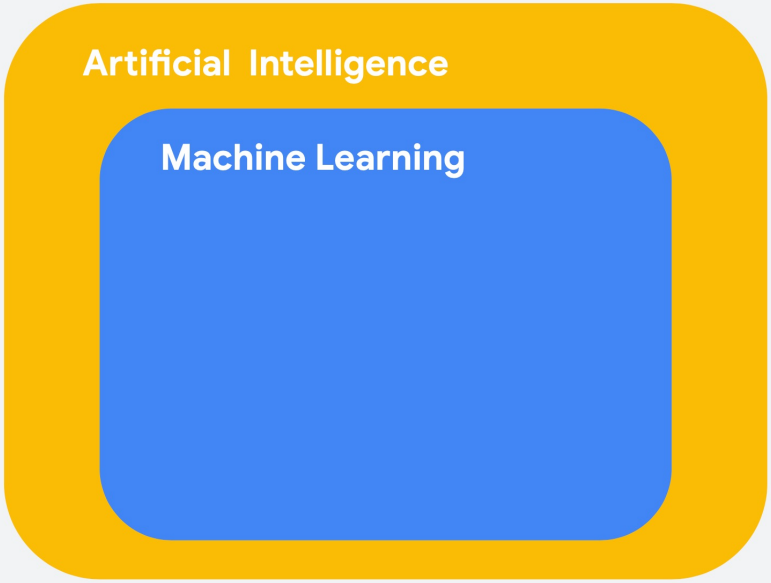
On-line references

- Tiny Machine Learning Open Education Initiative (TinyMLedu)
  - https://tinyml.seas.harvard.edu
  - Harvard School of Engineering and Applied Sciences - CS249r: Tiny Machine Learning
  - HarvardX Profession Certificate in Tiny Machine Learning (TinyML)
- Tiny Machine Learning Community
  - https://discuss.tinymlx.org
- TinyML study group
  - https://github.com/scaledown-team/study-group
- Google TensorFlow Lite for Microcontrollers
  - https://experiments.withgoogle.com/collection/tfliteformicrocontrollers
- Tensorflow lite:
  - https://www.tensorflow.org/lite

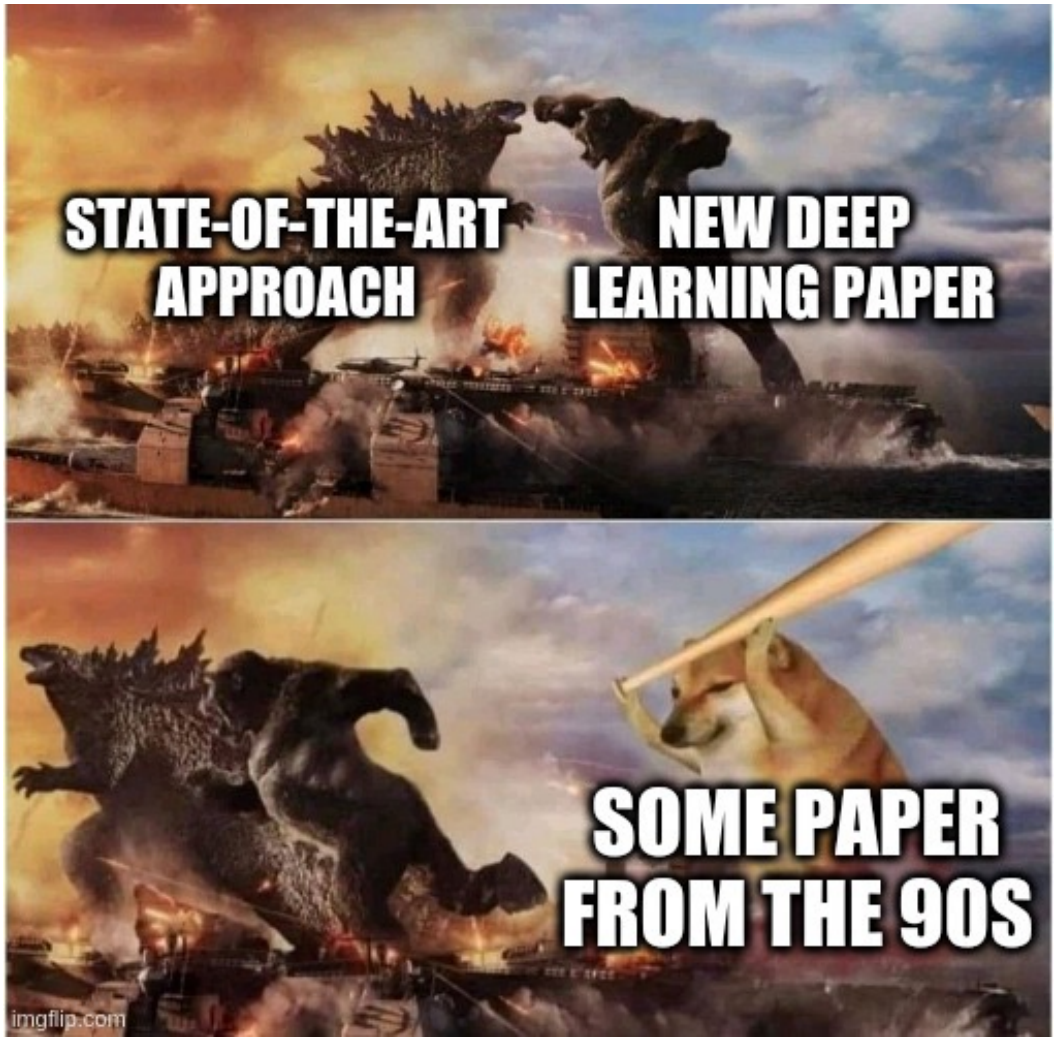
Some initial definitions

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**

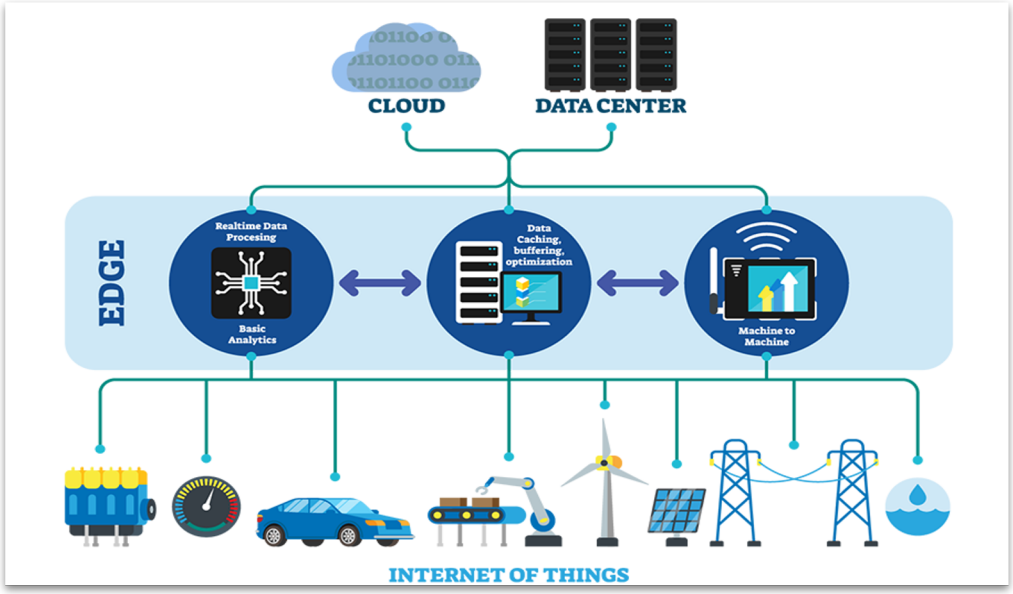


Warning


STATE-OF-THE-ART APPROACH
NEW DEEP LEARNING PAPER
SOME PAPER FROM THE 90S
imgflip.com

Internet of Things and Edge Computing

- Gartner defines edge computing as: "a part of a distributed computing topology in which information processing is located close to the edge — where things and people produce or consume that information."



Edge Computing Use Cases; innovationatwork.ieee.org

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

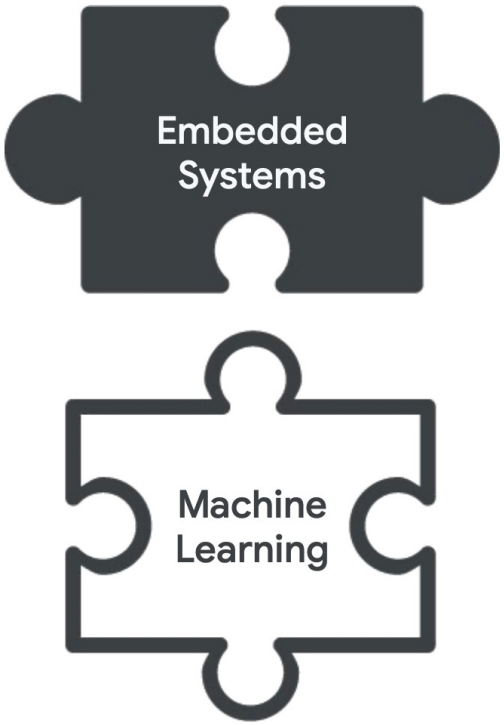
Image Sensors
Thermal, Image

Biometric Sensors
Fingerprint, Heart rate, etc.

Force Sensors
Pressure, Strain

Rotation Sensors
Encoders

What is TinyML?

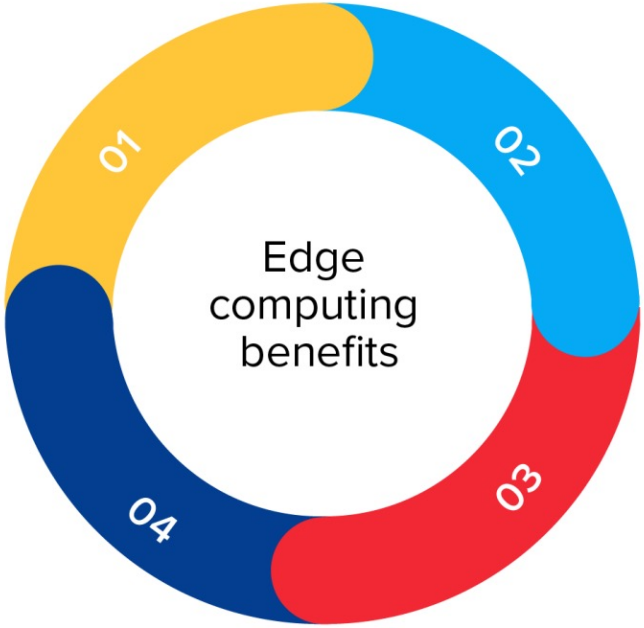


Tentative definition: "…a neural network model that runs at an energy cost of below 1 mW."
© "TinyML" by Pete Warden, Daniel Situnayake

TinyML

"Tiny machine learning (TinyML) is a fast-growing field of machine learning technologies and applications including algorithms, hardware, and software capable of performing on-device sensor data analytics at extremely low power consumption, **typically in the mW range and below**, enabling a variety of always-on ML use-cases on battery-operated devices."

Advantages of Edge Computing



Latency
Reduction of latency by processing the data closer to the customer

Security ➔ Privacy
Computing at the edge provides more security than computing at the cloud because it is less vulnerable to numerous variety of threats due to its scope

Bandwidth
Sending data from edge to the cloud takes up spectral resources; there's just not enough bandwidth for data transportation

Reliability
By processing data at the edge, you eliminate network reliability problems

https://www.wwt.com/article/show-me-the-money-drive-new-revenue-streams-with-edge-computing

… and power usage

BIG
GPU / CPU

300W
NVIDIA Tesla K80

SMALL

3.64W
Apple A12

Neural Decision Processor

Always-on deep learning speech/audio recognition

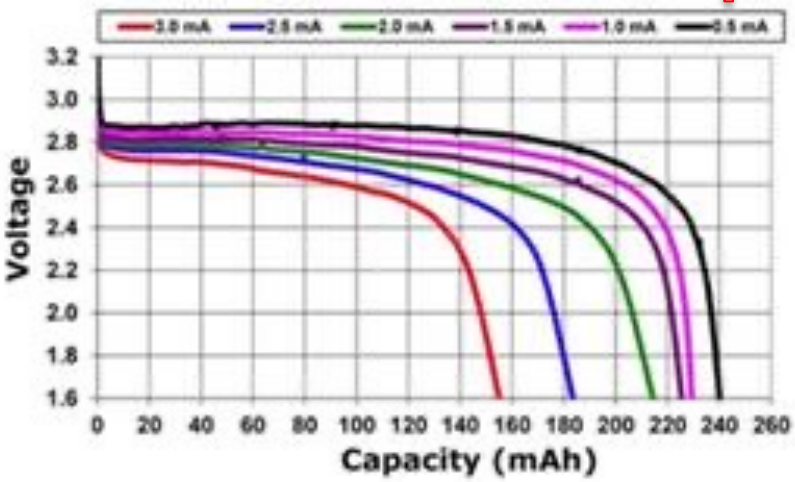
Ultra low power, 128KB SRAM, 12-pin, 2.52mm^2

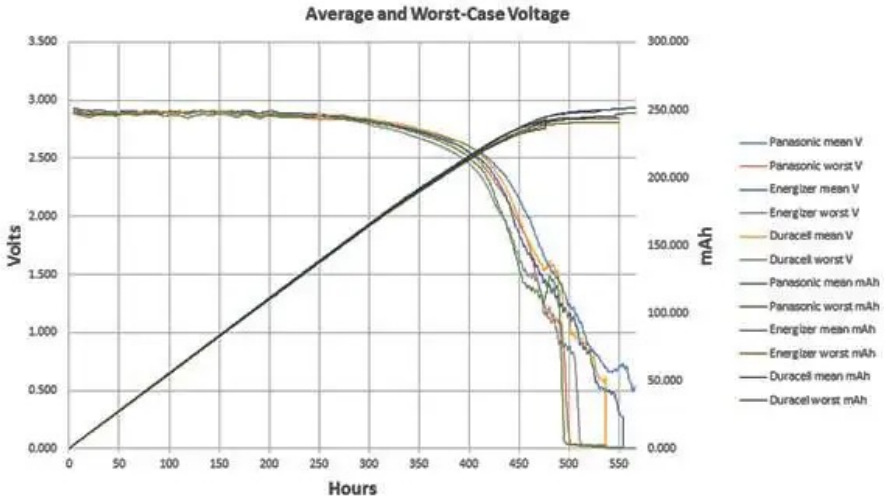
140 µW
Syntiant NDP100

https://www.syntiant.com/ndp100

Example: a CR2032 battery

- Lithium coin cell battery
- 3 V; 225/240 mAh





So, the syntiant ndp100 requires approximately 46µA ➔ 7 months

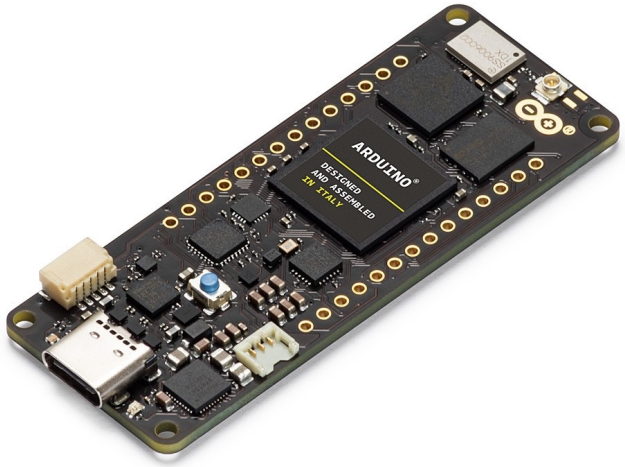
Orders of magnitude

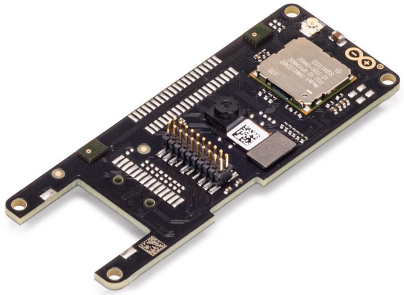
	Microprocessor	>	Microcontroller
Platform			
Compute	1GHz–4GHz	~10X	1MHz–400MHz
Memory	512MB–64GB	~10000X	2KB–512KB
Storage	64GB–4TB	~100000X	32KB–2MB
Power	30W–100W	~1000X	150µW–23.5mW

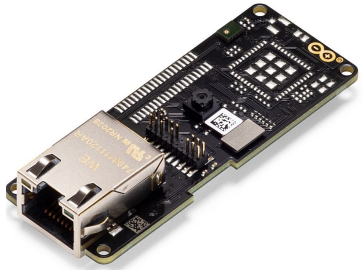
Some Hardware used for prototypes

	Board	MCU / ASIC	Clock	Memory	Sensors	Radio
	Himax WE-I Plus EVB	HX6537-A 32-bit EM9D DSP	400 MHz	2MB flash 2MB RAM	Accelerometer, Mic, Camera	None
	Arduino Nano 33 BLE Sense	32-bit nRF52840	64 MHz	1MB flash 256kB RAM	Mic, IMU, Temp, Humidity, Gesture, Pressure, Proximity, Brightness, Color	BLE
	SparkFun Edge 2	32-bit ArtemisV1	48 MHz	1MB flash 384kB RAM	Accelerometer, Mic, Camera	BLE
	Espressif EYE	32-bit ESP32-D0WD	240 MHz	4MB flash 520kB RAM	Mic, Camera	WiFi, BLE

Some Hardware used for prototypes: Portenta H7







Portenta H7 is probably one of the currently most powerful MCUs.

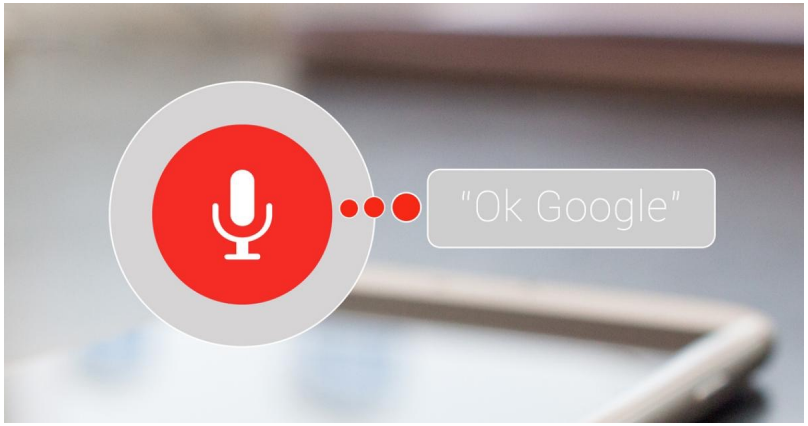
H7's main processor is the dual core STM32H747 including a Cortex® M7 running at 480 MHz and a Cortex® M4 running at 240 MHz.

The two cores communicate via a *Remote Procedure Call* mechanism.

Both processors share all the in-chip peripherals and can run:

- Arduino sketches on top of the Arm® Mbed™ OS
- Native Mbed™ applications
- MicroPython / JavaScript via an interpreter
- TensorFlow™ Lite

Applications of TinyML: the beginning



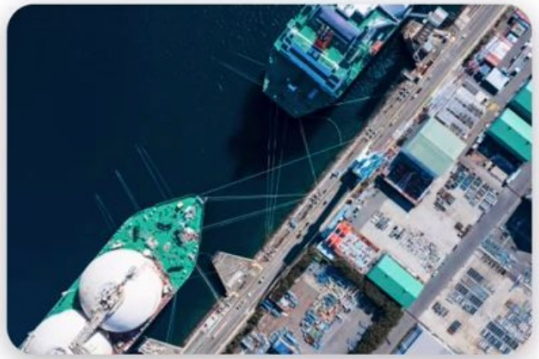
Applications examples

Predictive Maintenance

Motion, current, audio and camera

- Industrial
- White goods
- Infrastructure
- Automotive

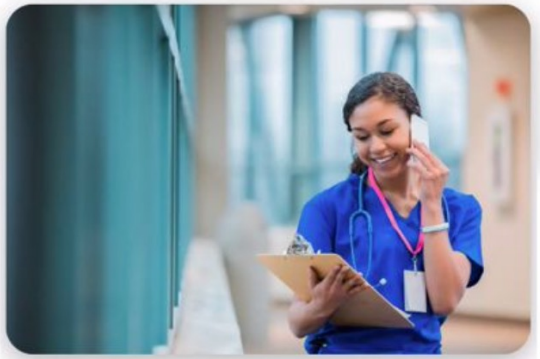
Asset Tracking & Monitoring



Motion, temp, humidity, position, audio and camera

- Logistics
- Infrastructure
- Buildings

Human & Animal Sensing



Motion, radar, audio, PPG, ECG

- Health
- Consumer
- Industrial

Applications examples

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

KeyWord Spotting

Environmental Control

Motion & biometric

Image Spot

RAM-1: An Advanced Grid Monitoring Solution

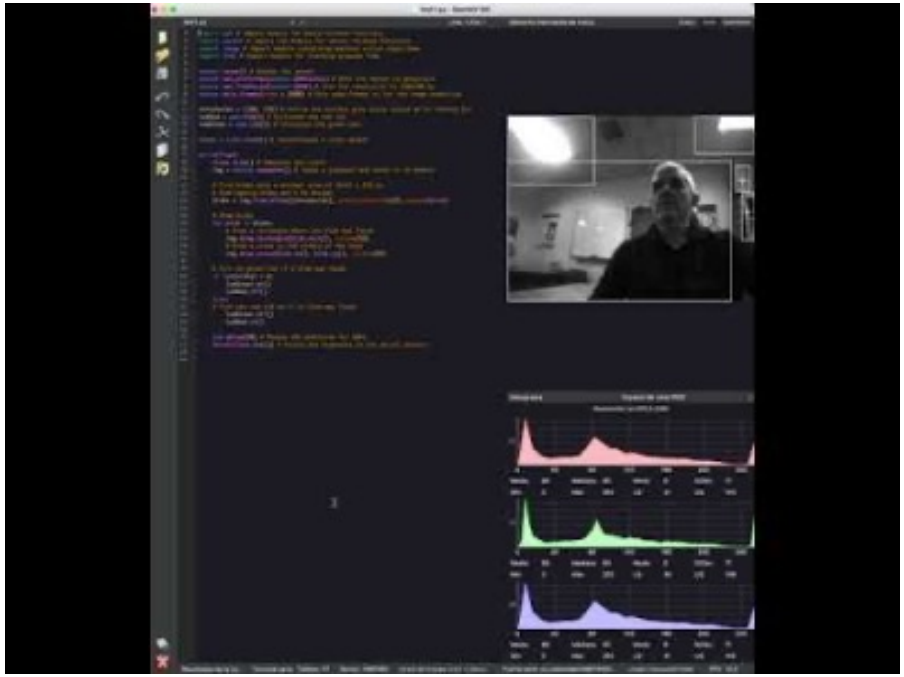
https://www.ram-center.com

Powered by Edge Impulse

TensorFlow Lite on the Raspberry Pi

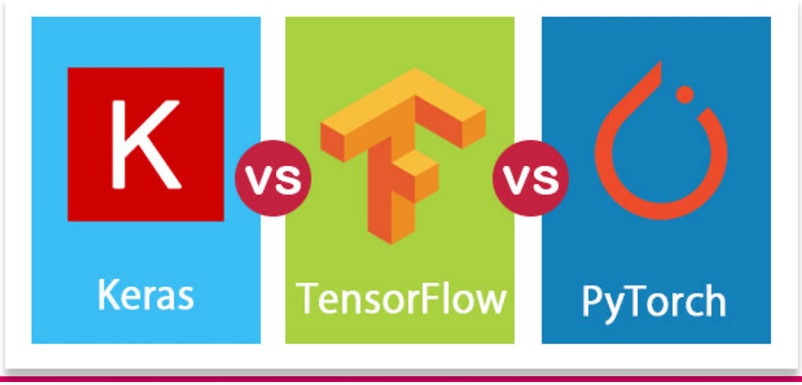
- The example below shows how to use TensorFlow Lite on the Raspberry Pi to run object detection models.

- This other example performs "Blob Detection" with a Portenta detecting the presence and the position of objects in a camera image.



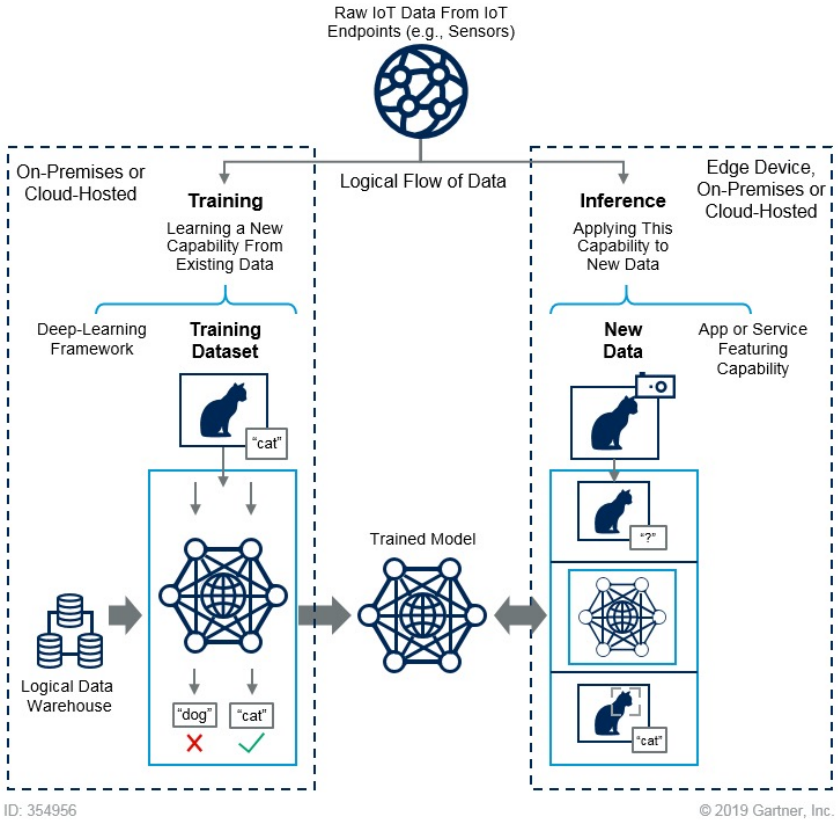
About the available software frameworks

- The three main frameworks which are available as an open-source library are opted by data scientist in deep learning are PyTorch, TensorFlow, and Keras.
  - **Keras** is a neural network library scripted in Python and can execute on the top layer of TensorFlow.
  - **TensorFlow** is used to perform multiple tasks in data flow programming and machine learning applications.
  - **PyTorch** is a machine learning library that is mainly used in natural language processing.



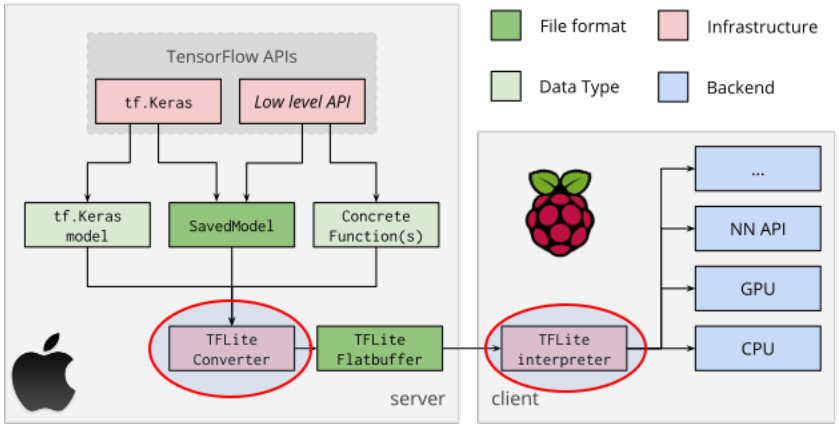
Training and Inference

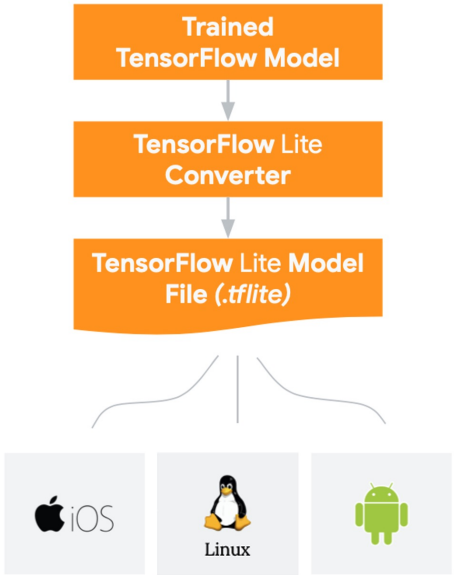
- **Training** refers to the process of creating a machine learning algorithm. Training involves using a machine learning framework and a training dataset.
  - **IoT data** provides a source of training data that data scientists and engineers can use to train machine learning models.
- **Inference** refers to the process of using a trained machine-learning algorithm to make a prediction.
  - **IoT data** can be used as the input to a trained machine learning model, enabling predictions that can guide decision logic on the device, at the edge.



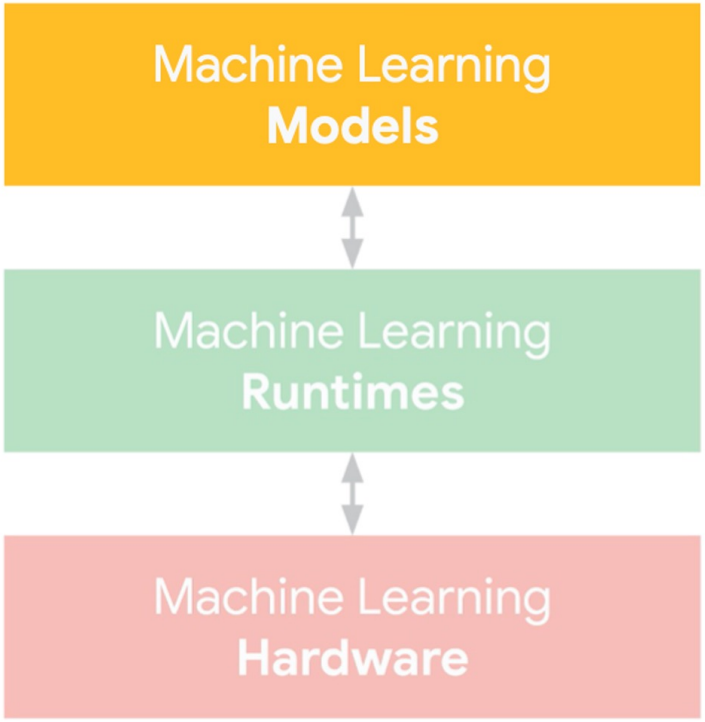
TensorFlow Lite (TFLite)

- **TensorFlow Lite (TFLite)** is an open-source deep learning framework that enables on-device machine learning inference.
- Consists of two main components:
  - The **TFLite converter**, which converts TensorFlow models into an efficient form for use by the interpreter and can introduce optimizations to improve binary size and performance.
  - The **TFLite interpreter** runs with specially optimized models on many different hardware types, including mobile phones, embedded Linux devices, and microcontrollers.



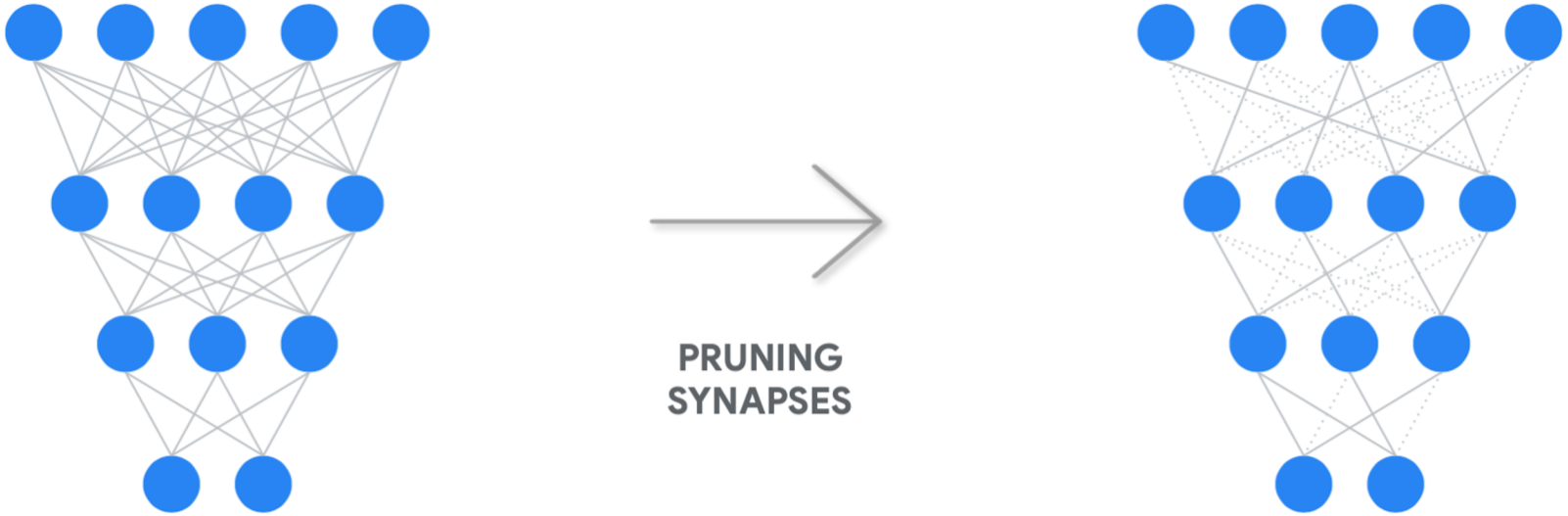


Model Compression Techniques

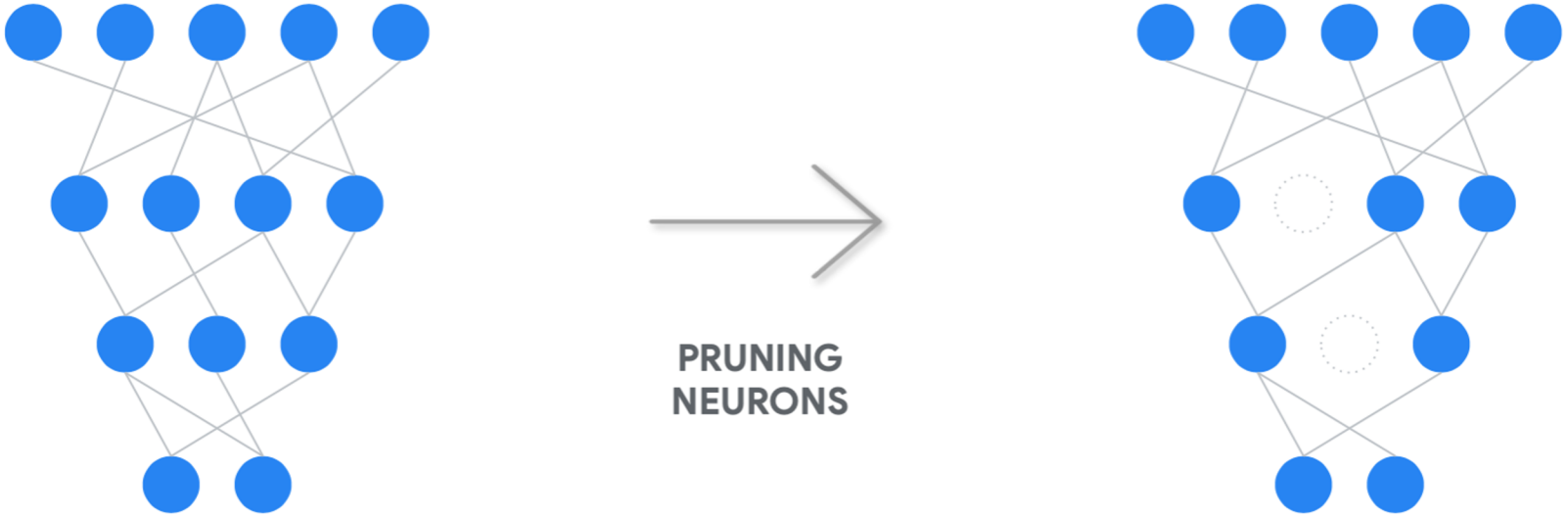


- Pruning
- Quantization
- Knowledge Distillation
- …

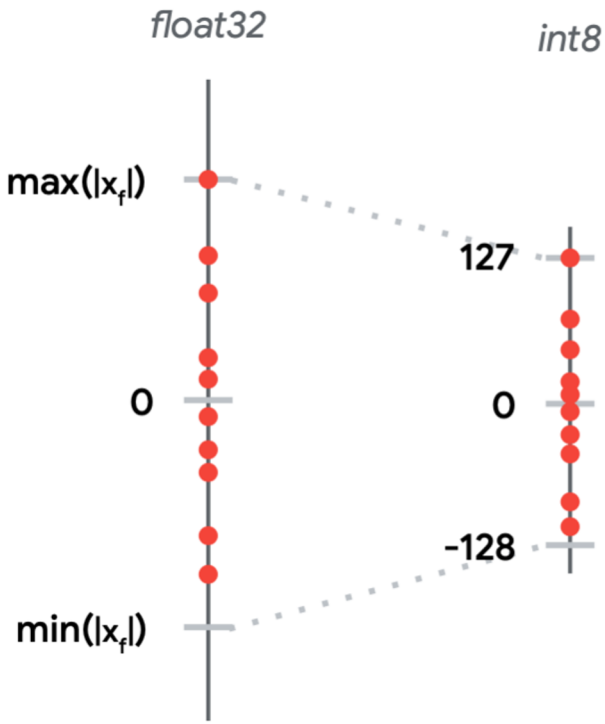
Pruning



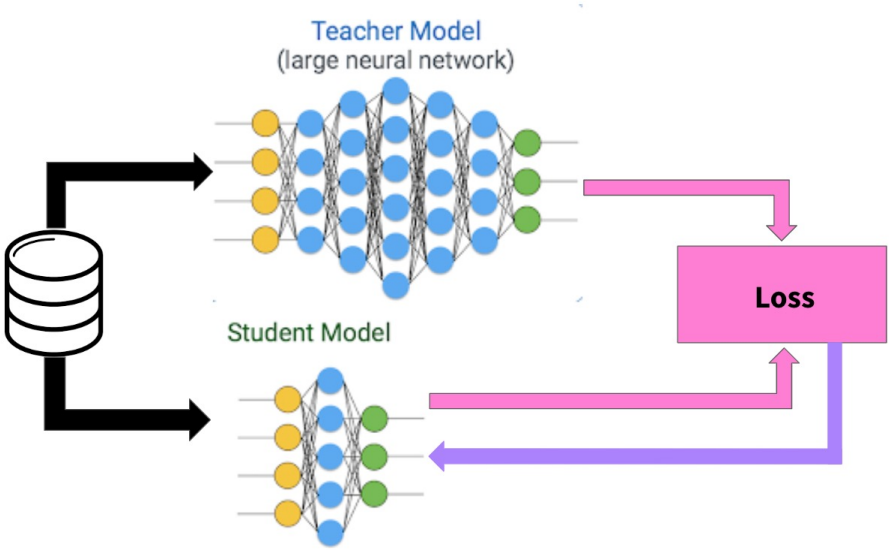
Pruning

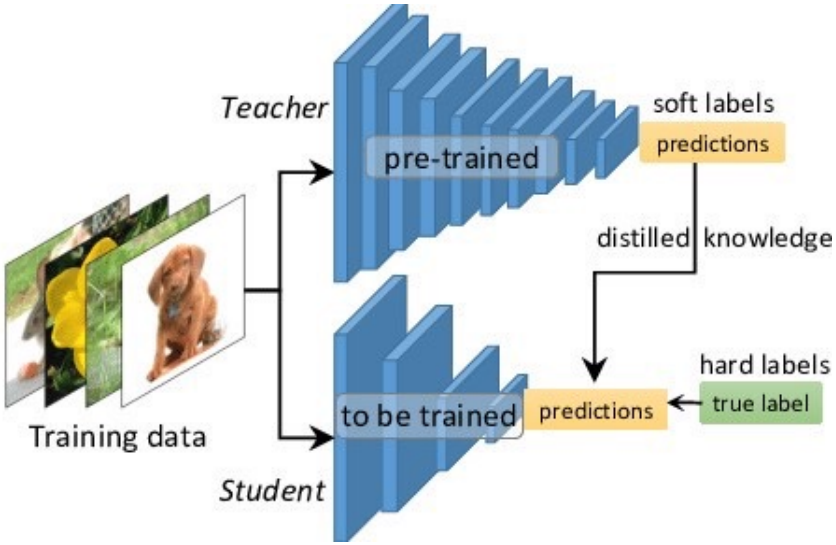


Quantization



Knowledge Distillation





A brief introduction to TinyML