



# Machine Learning (ML) at the Edge on Arm: A Practical Introduction

Michele Magno

Project-Based Learning Center | ETH Zurich

[michele.magno@pbl.ee.ethz.ch](mailto:michele.magno@pbl.ee.ethz.ch)

# Objectives of the Course

- **To Introduce ML especially for edge Resource-constrained devices.**
  - Algorithms (Support Vectors Machine, Neural Networks, Deep Learning)
  - Consideration to program a Microcontroller or low power Embedded systems
  - Learning how training machine on PC (Phyton/TensorFlow)
- **Hand-on on Microcontroller**
  - STM32 - Arm Cortex M4 Family and CubeAI
  - Energy efficient implementations (Including Quantization )
- Work with real data from Sensors (or Dataset)

# What is Artificial Intelligence?

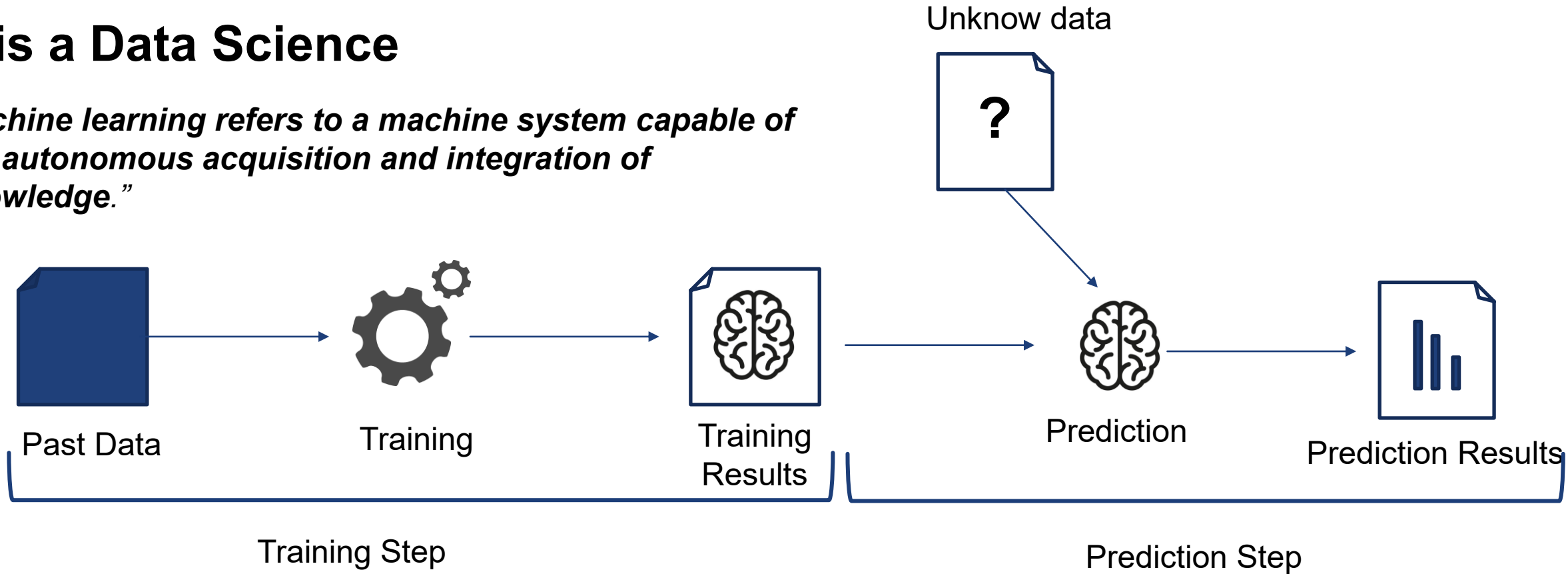
**AI is a superset of all the studies to replicate human reasoning with computer systems and is used everyday in our life**

- Face / voice recognition
- Autonomous driving
- Stock market trading strategy
- Disease symptom detection
- Predictive maintenance
- Hand writing recognition
- Content distribution on social media
- Fraudulent credit card transaction
- Translation engines
- Suggested shopping
- ...



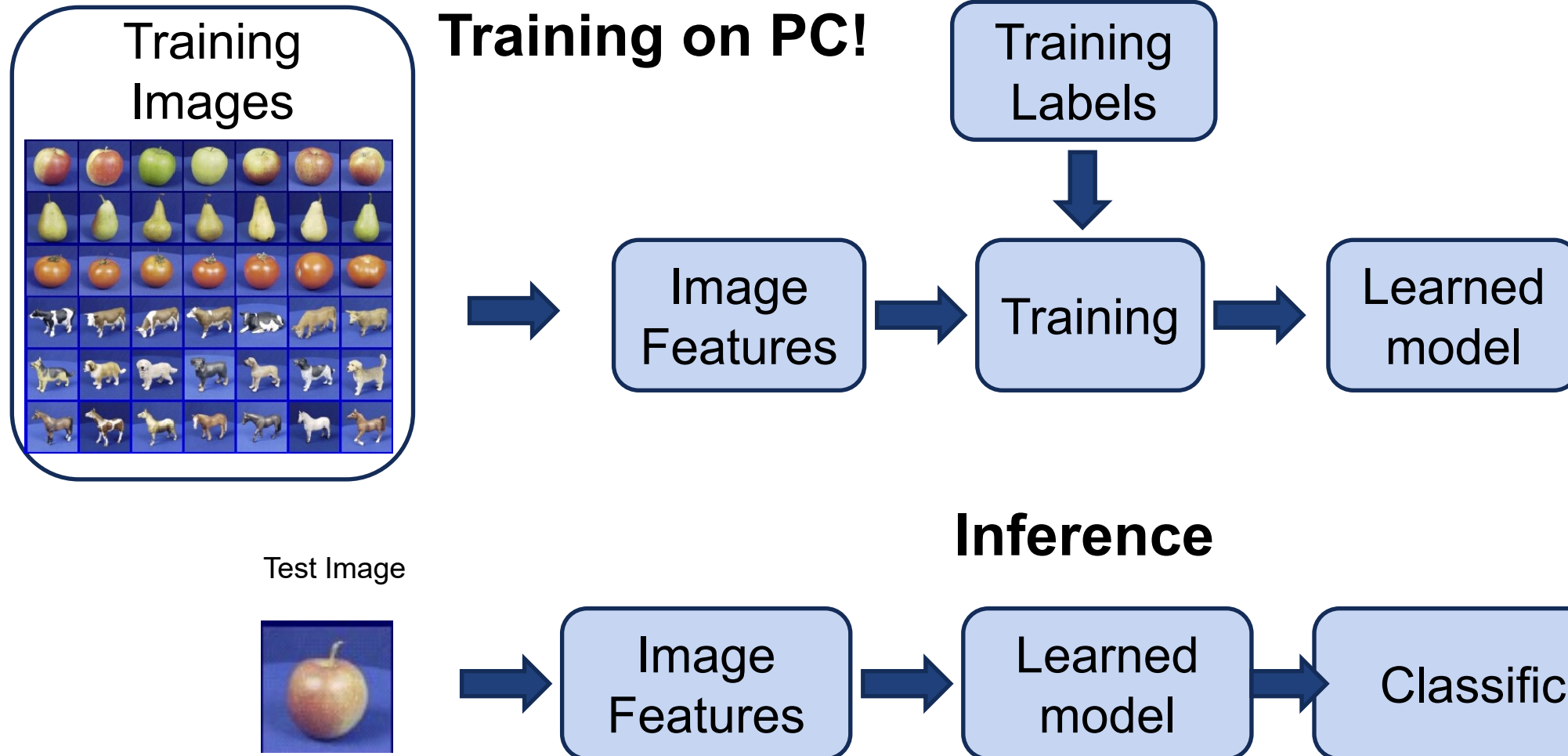
# ML is a Data Science

*Machine learning refers to a machine system capable of the autonomous acquisition and integration of knowledge."*








- During the training step, you use past or known data to train a model.
- During the prediction step, you use the trained model to make predictions from new data.

# In this course we focus on Supervised ML



# Where the inference is processed? - Edge Vs Cloud

- Latency/reliability 
- Data Protection 
- No Wireless Communication Needed – Lower Bandwidth requirements 
- **Lower Power Consumption** 
- Lower Cost 

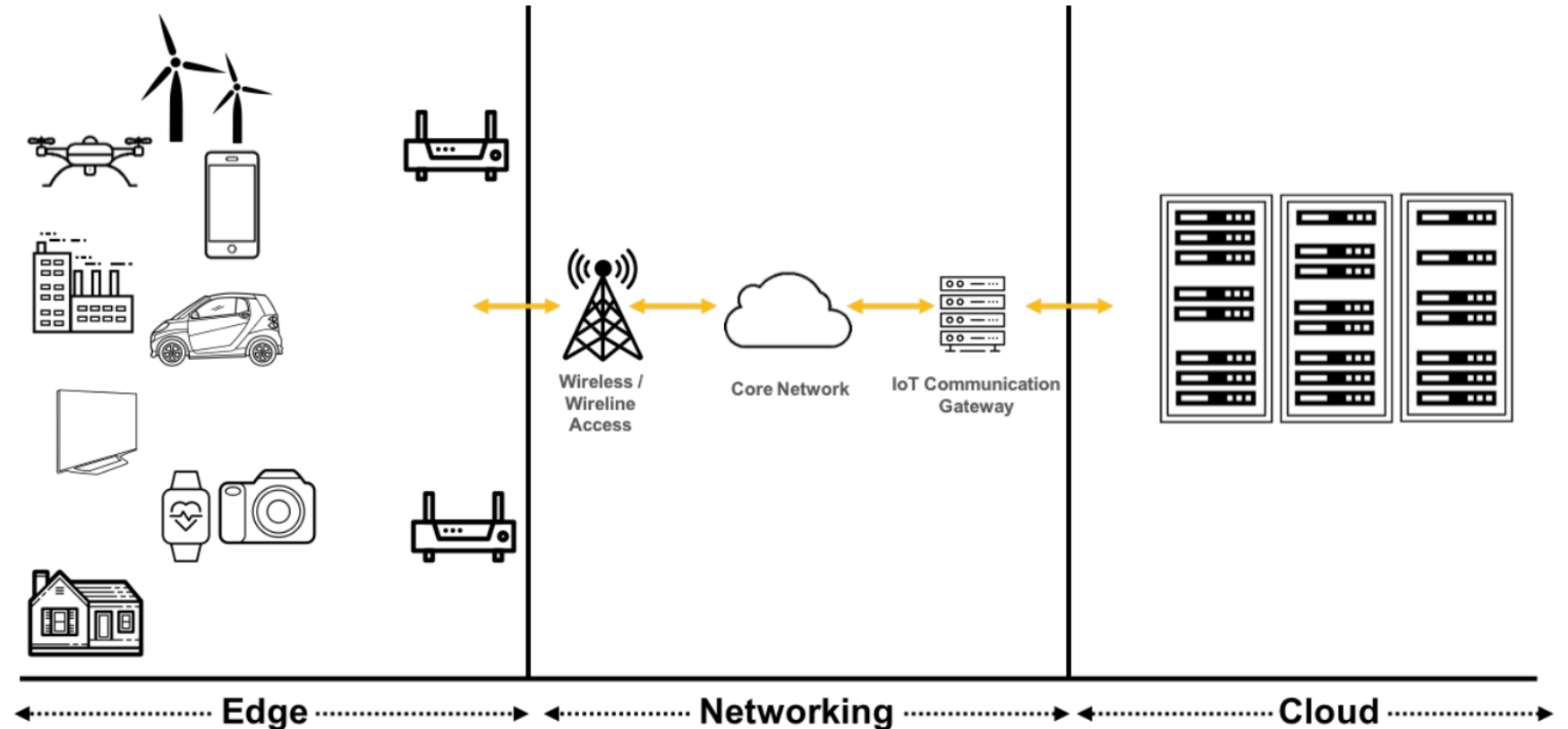


Figure reference: Accelerating Implementation of Low Power Artificial Intelligence at the Edge, A Lattice Semiconductor White Paper, November 2018

# Internet of Things pushes AI and ML at the edge

The world is producing excessive amounts of "unstructured data" that need to be reconstructed

(IBM's CTO Rob High)

"A PC will generate 90 megabytes of data a day, an autonomous car will generate 4 terabytes a day, a connected plane will generate 50 terabytes a day."

Source: Samsung HBM

## Bandwidth



1 Billion cameras WW  
(2020)  
30B Inference/sec

## Latency



Communication latency  
also with 5G or other  
networks is in the range  
of hundred of  
milliseconds

## Availability

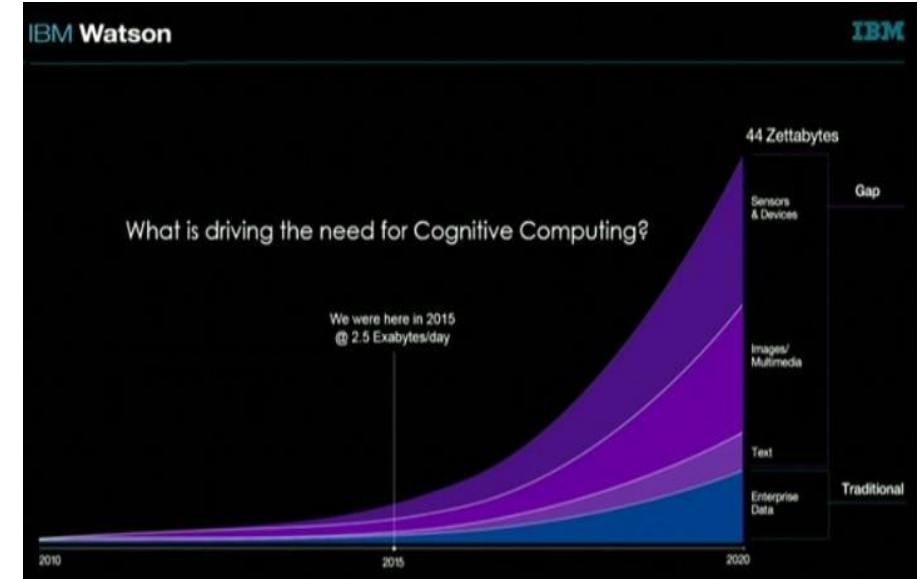


50% of world at less than  
8mbps Only 73% 3G/4G  
availability WW

## Security



Data traveling in the  
network are more  
vulnerable.  
Attacks to networks and  
communication towers

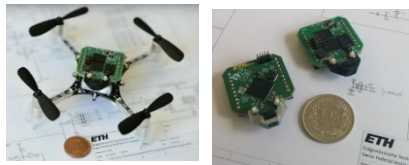
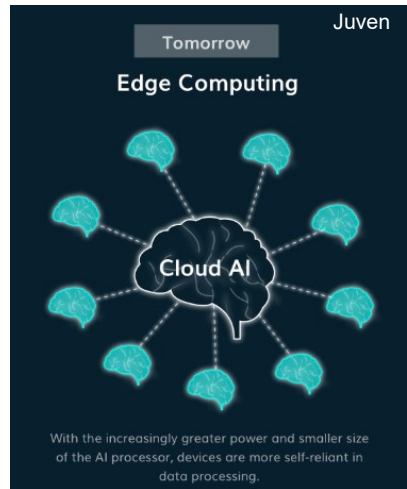


Source: IBM

Since 2015, roughly 2.5 Exabyte of data are being generated per day. Projection shows a 44 Zettabytes of data per day by 2020.

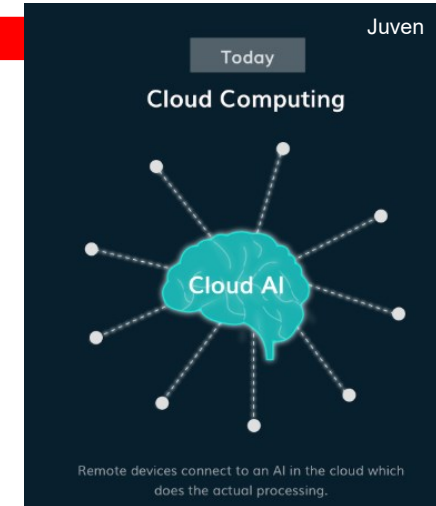
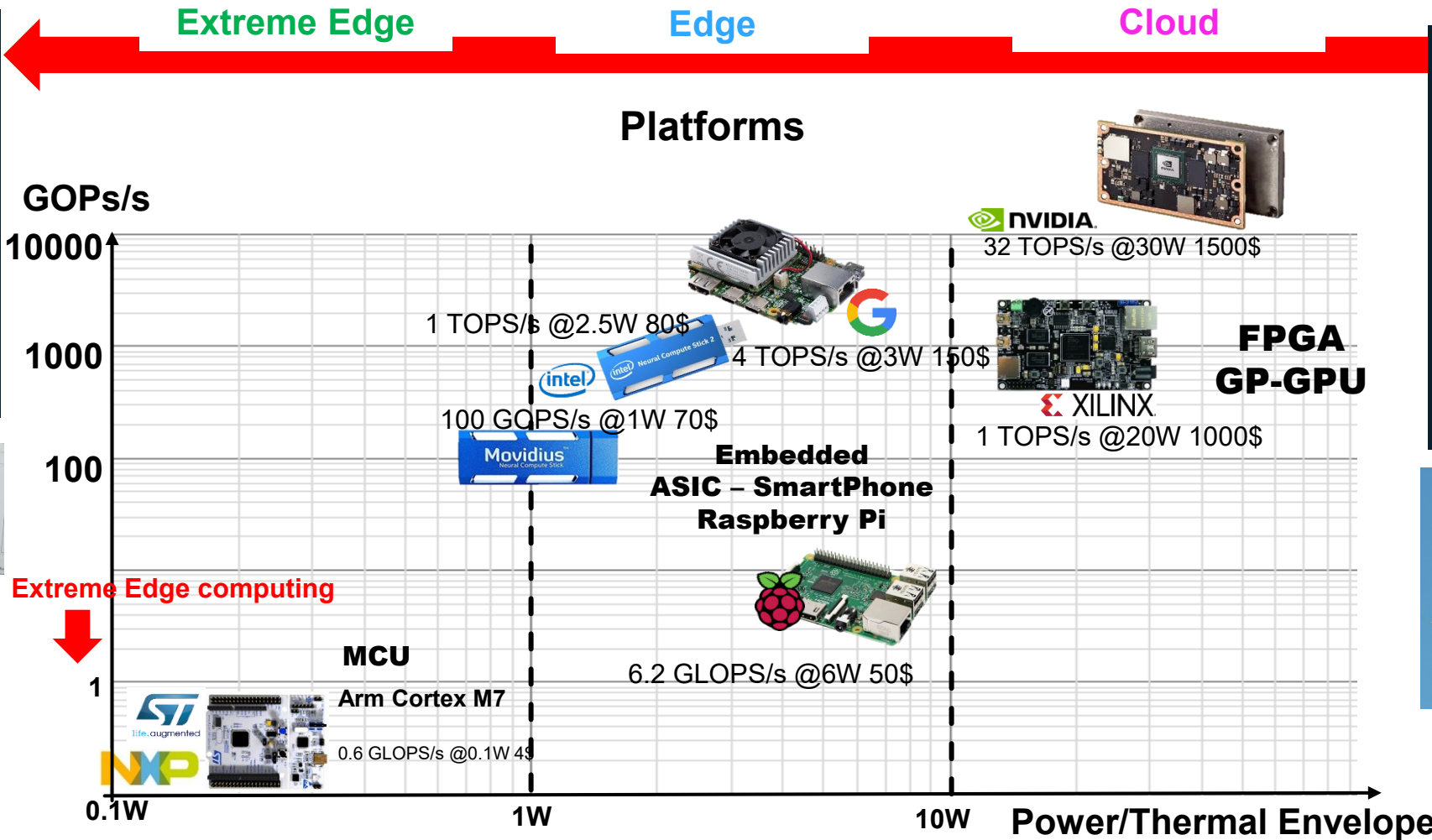


# Cloud → Edge → **Extreme Edge**



Wearable-Battery operated

Always-on - Self-sustaining





# Course Organization

---

Module 1: An Overview of Machine Learning at the Edge

Module 2: Introduction to Machine Learning on Constrained Devices

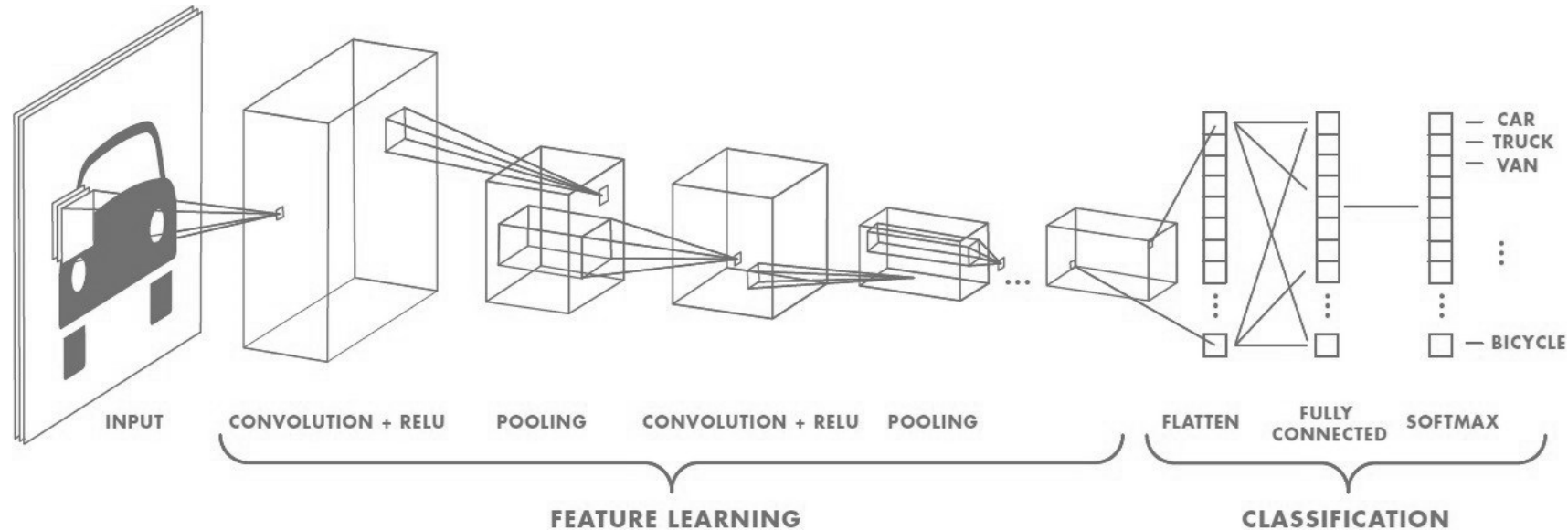
Module 3: Artificial Neural Networks

Module 4: Convolutional Neural Networks.

Module 5: Computer Vision and Models

Module 6: Optimizing Machine Learning on Constrained Devices

# Consideration to deploy ML on a Microcontroller

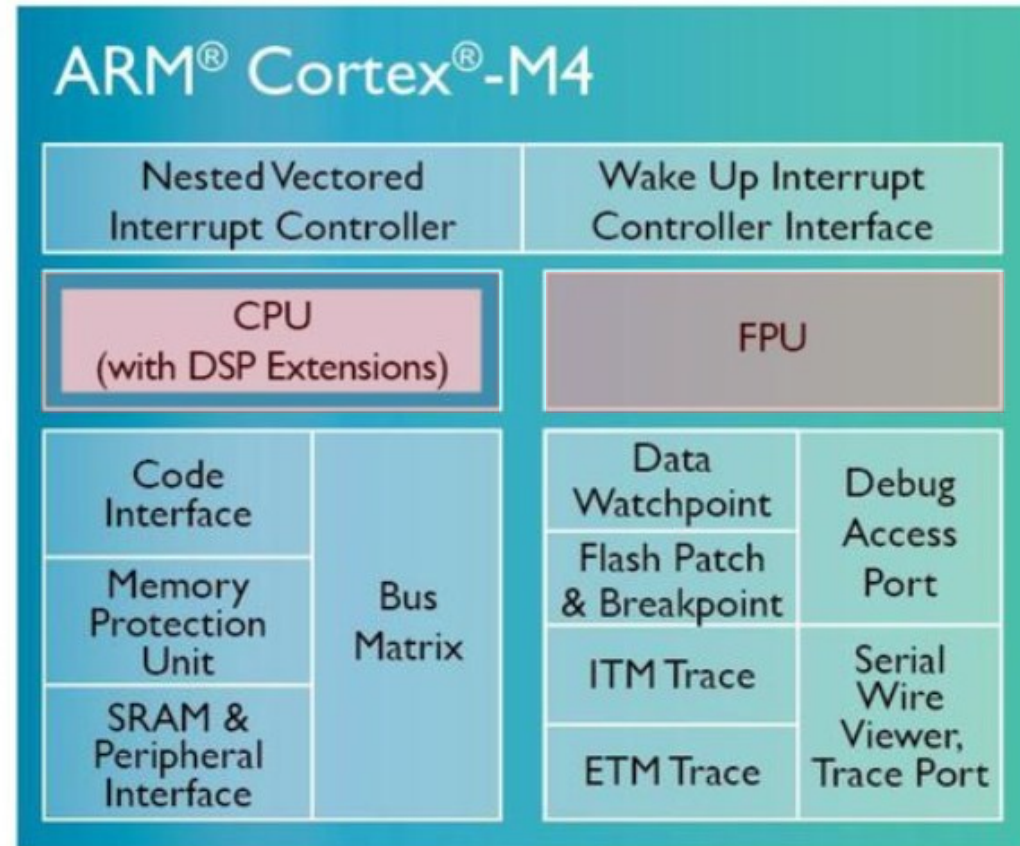


Requirements : Million of operations, mainly multiplication and add , and memory to store the parameters of Learning and run-time calculations result and data.

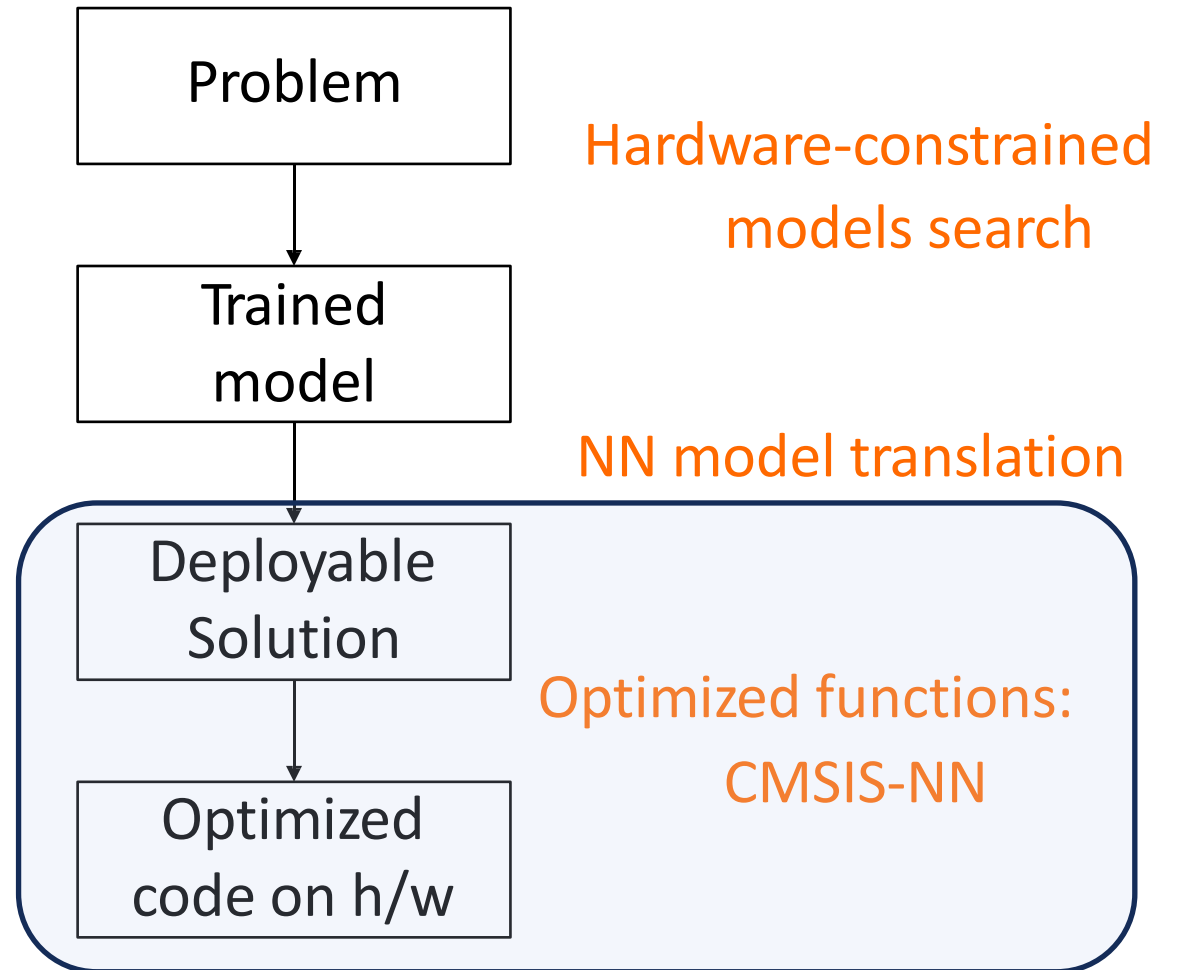
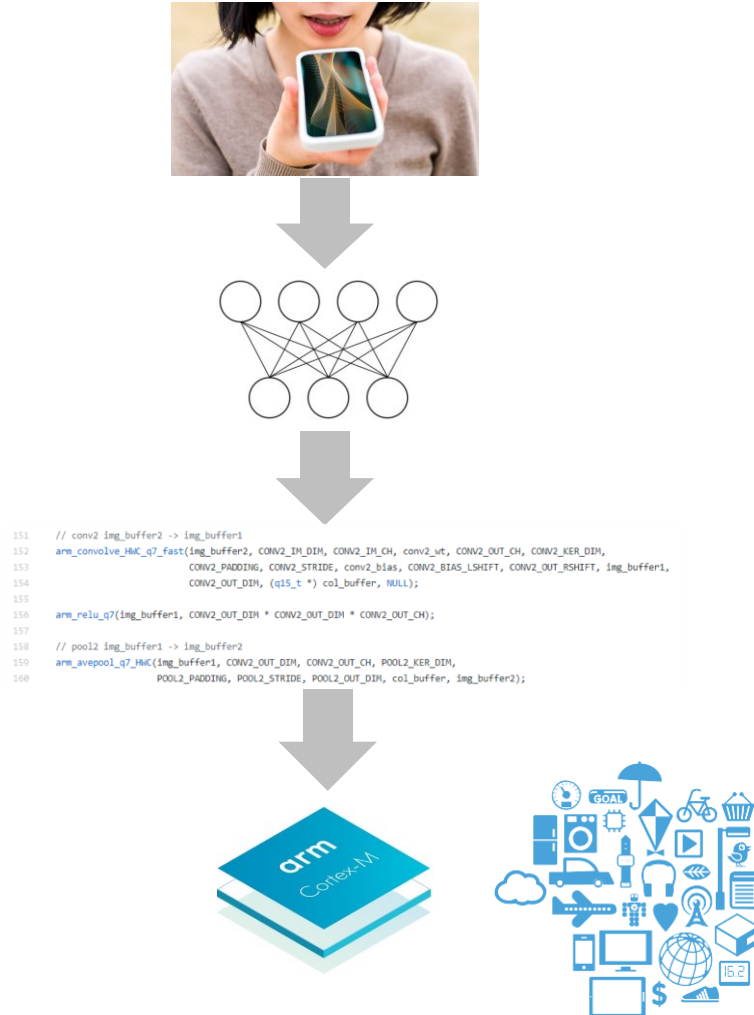
**Limited memory of 128KB of RAM and 1MByte of FLASH in our board  
80MHz single core devices with around 1-2 operations per cycle.**

# Architecture of M4

- CPU with DSP
- NVIC
- Wake-up CI
- Debug
- Code Interface
- FPU (IEEE 754 Compliant)
- MPU
- SRAM and Peripheral Interface



# Developing ML Solution on Cortex-M MCUs – 5 Labs for learning



# Microcontroller – STM32 L4 with B-L475E-IoT01A board

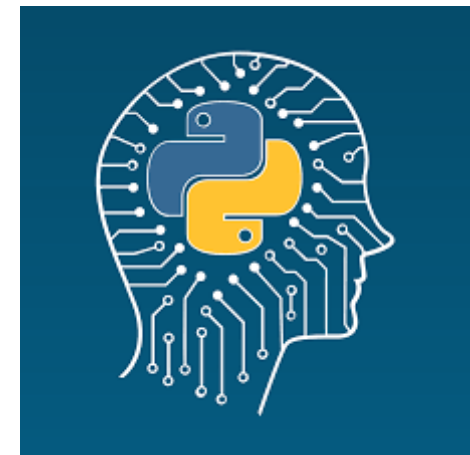
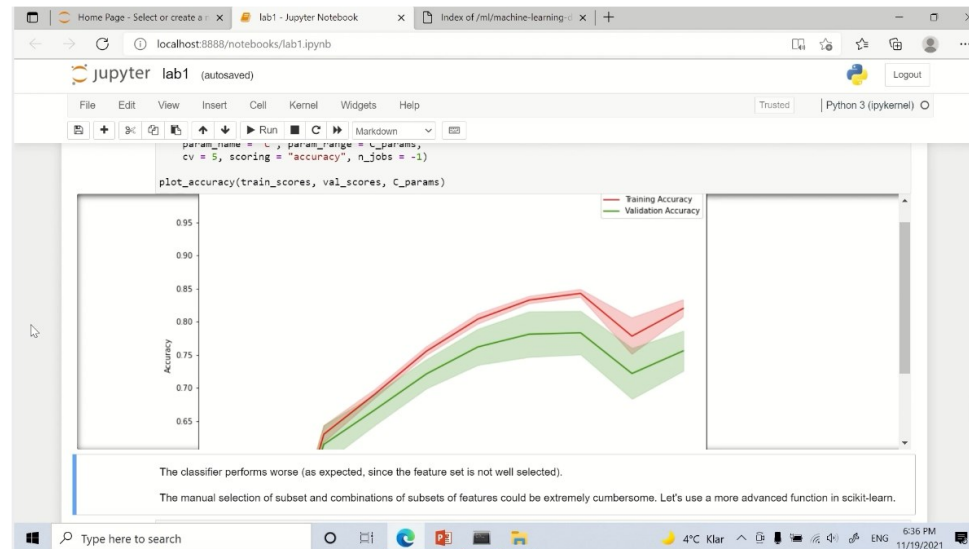
## Key Features

- Ultra-low-power **STM32L4 Series MCUs based on Arm® Cortex®-M4** core with 1 Mbyte of Flash memory and 128 Kbytes of SRAM, in LQFP100 package
- 64-Mbit Quad-SPI (Macronix) Flash memory
- Bluetooth® V4.1 module (SPBTLE-RF)
- Sub-GHz (868 MHz or 915 MHz) low-power-programmable RF module (SPSGRF-868 SPSGRF-915)
- 802.11 b/g/n compliant Wi-Fi® module from Inventek Systems (ISM43362-M3G-L44)
- Dynamic NFC tag based on M24SR with its printed NFC antenna
- 2 digital omnidirectional microphones (MP34DT01)
- Capacitive digital sensor for relative humidity and temperature (HTS221)
- High-performance 3-axis magnetometer (LIS3MDL)
- 3D accelerometer and 3D gyroscope (LSM6DSL)
- 260-1260 hPa absolute digital output barometer (LPS22HB)
- Time-of-Flight and gesture-detection sensor (VL53L0X)
- 2 push-buttons (user and reset)



# LAB 1: LAB 1 - Getting started with Python for machine Learning.

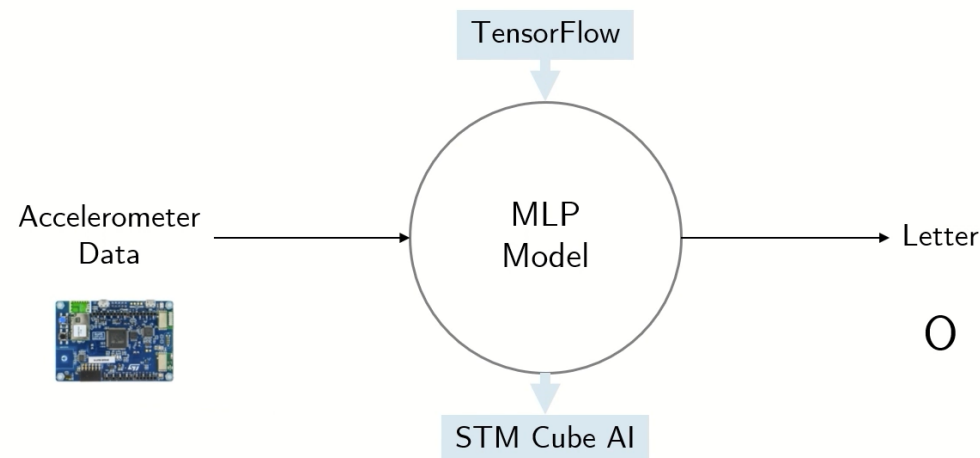
- Utilize Python and Anaconda to work with dataset and Machine Learning training.
- Utilize API for feature extraction, model implementation with support vector machines and feature extraction.
- Project-based lab: Implement a simple model for activity recognition with accelerometer data. Train your model and evaluate it competing in the class-room.





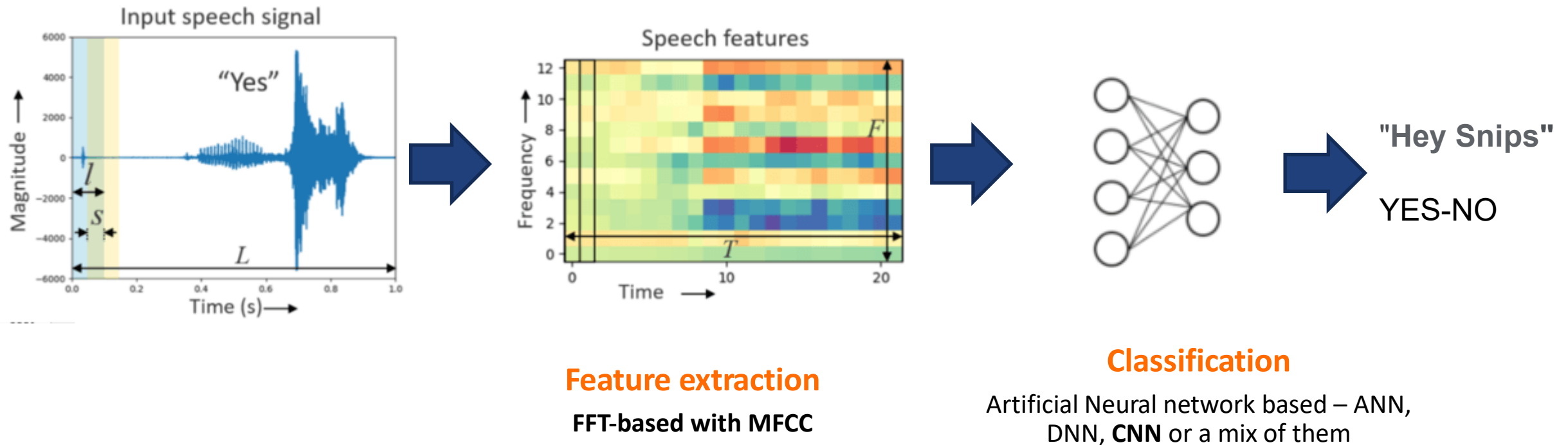
## LAB 2: STM32 CUBE AI and TensorFlow Lite for Microcontrollers, Applications on embedded platforms

- Utilize STM32 CUBE AI and TensorFlow Lite for Microcontrollers.
- Train the first model on Tensor flow and deploy the inference on a microcontroller using CUBE AI.
- Project-based lab: Develop the activity recognition using multi-layer perceptron (MLP)
- Run and evaluate the performance on a STM32 Microcontroller



# LAB 3: Speech Recognition with Mel-frequency cepstrum Coefficient and CNN

- Develop an CNN model for audio processing

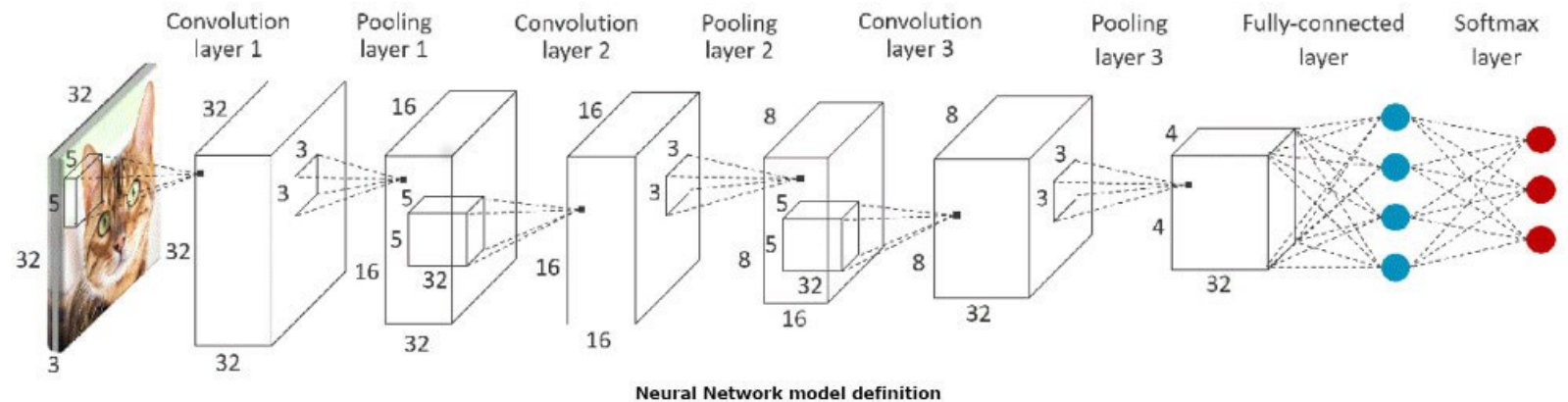
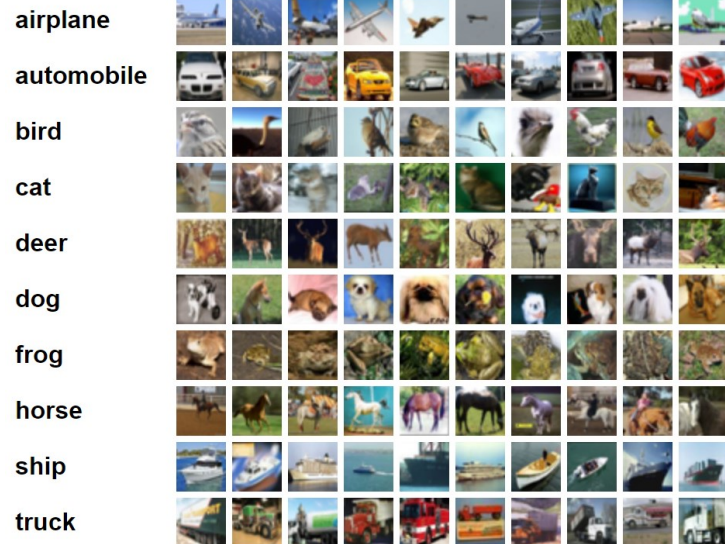


# LAB 4: Image processing with ARM Cortex-M, CMSIS-NN And CIFAR-10 Database.

- Develop a simple model for CIFRA 10 dataset using CUBE AI.

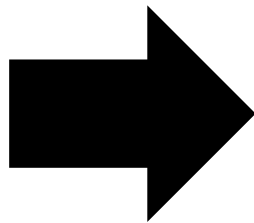
## CIFAR-10

Here are the classes in the dataset, as well as 10 random images from each:



## LAB 5: Practical evaluation of the machine learning models in terms of energy, latency, memory

- How to measure the performance of ML model
- Optimize the ML model to improve the latency and reduce energy consumption.



TensorFlow Lite

	Original Model	Quantized Model	Difference (%)
Inference Time	1133.439 ms ( 90675110 cycles )	318.464 ms ( 25477095 cycles )	71.9%
Memory Size (RAM)	82.16 KB	25.82 KB	68.6%
Energy Consumption	43 * 1.133 mJ	43 * 0.318 mJ	71.9%

\* Power: 43mW (1 Watt = 1 J/s)

# Conclusion

- We learn what is Artificial intelligence and Machine learning
- Machine Learning is a Data science
- Different types of ML
  - This course focuses on Supervised Machine Learning
- Internet of Things is pushing ML at the edge.
  - Microcontroller is one of the most popular processors at the edge
- Steps to bring ML on Microcontrollers
- We will focus on STM32 microcontroller with ARM Cortex-M4
- Lab-oriented course