POLITECNICO
MILANO 1863

Hardware Architectures for Embedded and Edge AI

Prof Manuel Roveri – manuel.roveri@polimi.it

Lecture 1 – Introduction to the course

Prof. Manuel Roveri

- **Full Professor**

  Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Italy
  Email: **manuel.roveri@polimi.it**
  Web: http://roveri.faculty.polimi.it

- **Research interests**: **TinyML, IoT and edge computing, privacy-preserving machine and deep learning**
- **Lecturer of « Computing Infrastructures» and «Hardware Architecture for Embedded and edge AI»**
- **Associate Editor** of IEEE Trans. on Artificial Intelligence, Neural Networks, IEEE Trans. on Emerging Tecnologies in Computational Intelligence, IEEE Trans. on Neural Networks and Learning Systems
- Chair of the IEEE CIS **Technical Activities** strategic planning committee and IEEE CIS **Neural Network** Technical Committee
- **Co-Founder of DHIRIA**, a Spin-Off of Politecnico di Milano

The research team

- Massimo Pavan
  (PhD Student)
- Alessandro Falcetta
  (PhD Student)
- Matteo Gambella
  (PhD Student)
- Luca Colombo
  (PhD Student)
- Gabriele Viscardi
  (WemSy Lab Coordinator)
- Diego Riva
  (Research Assistant)
- Francesco Puoti
  (Research Assistant)

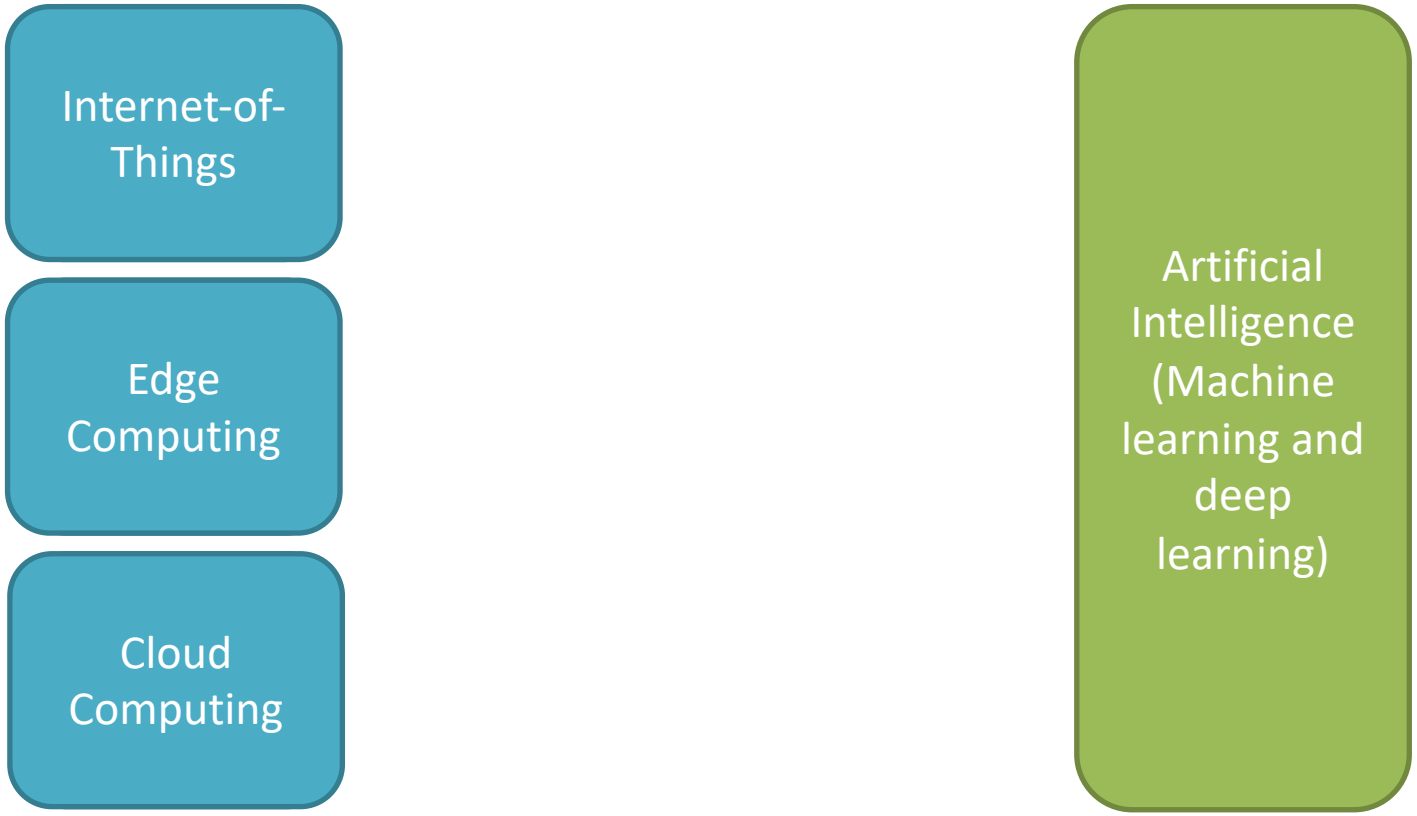
The WemSy Lab @ Lecco Campus of Politecnico di Milano

The research activity

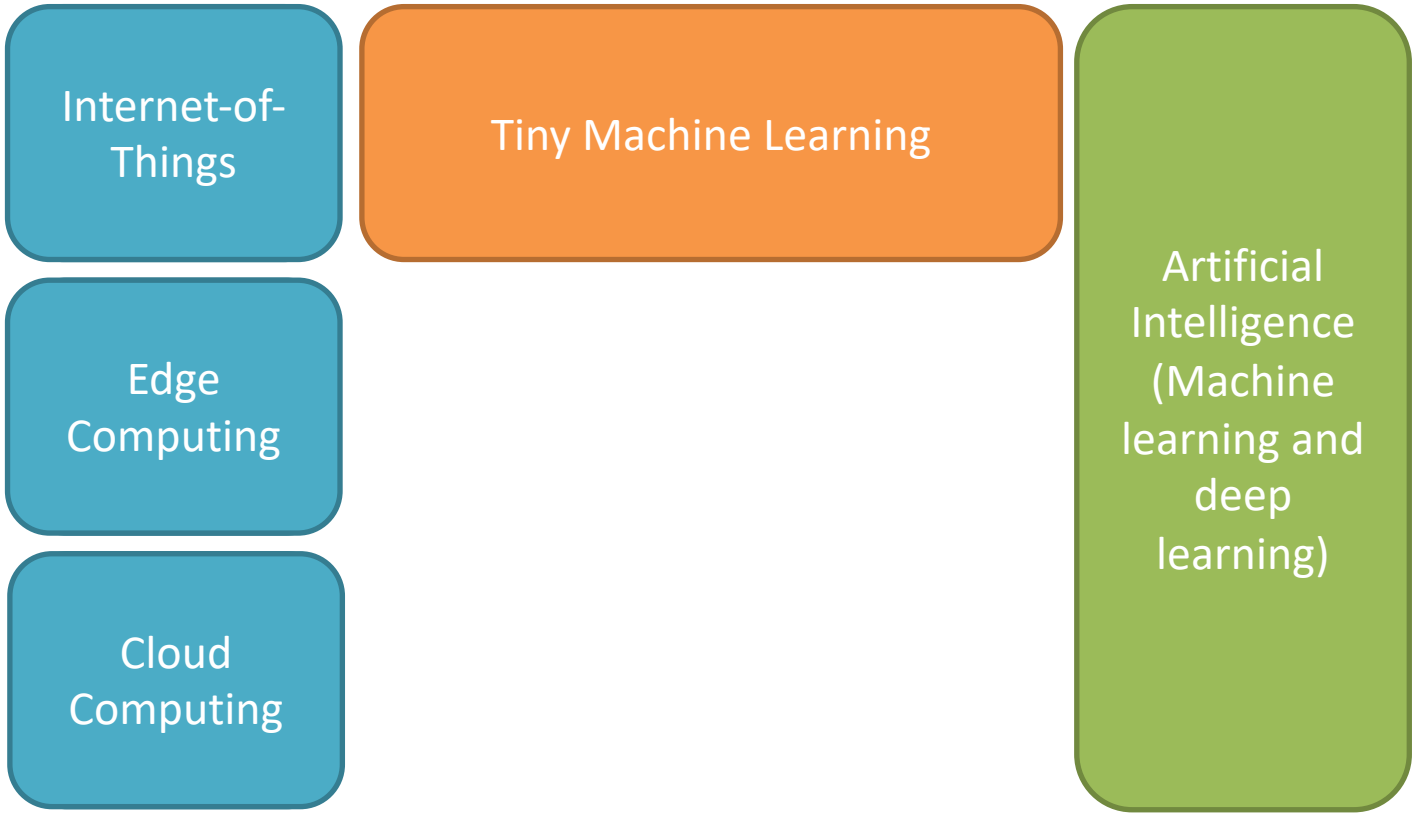
Cyber-physical Systems

Artificial Intelligence (Machine learning and deep learning)

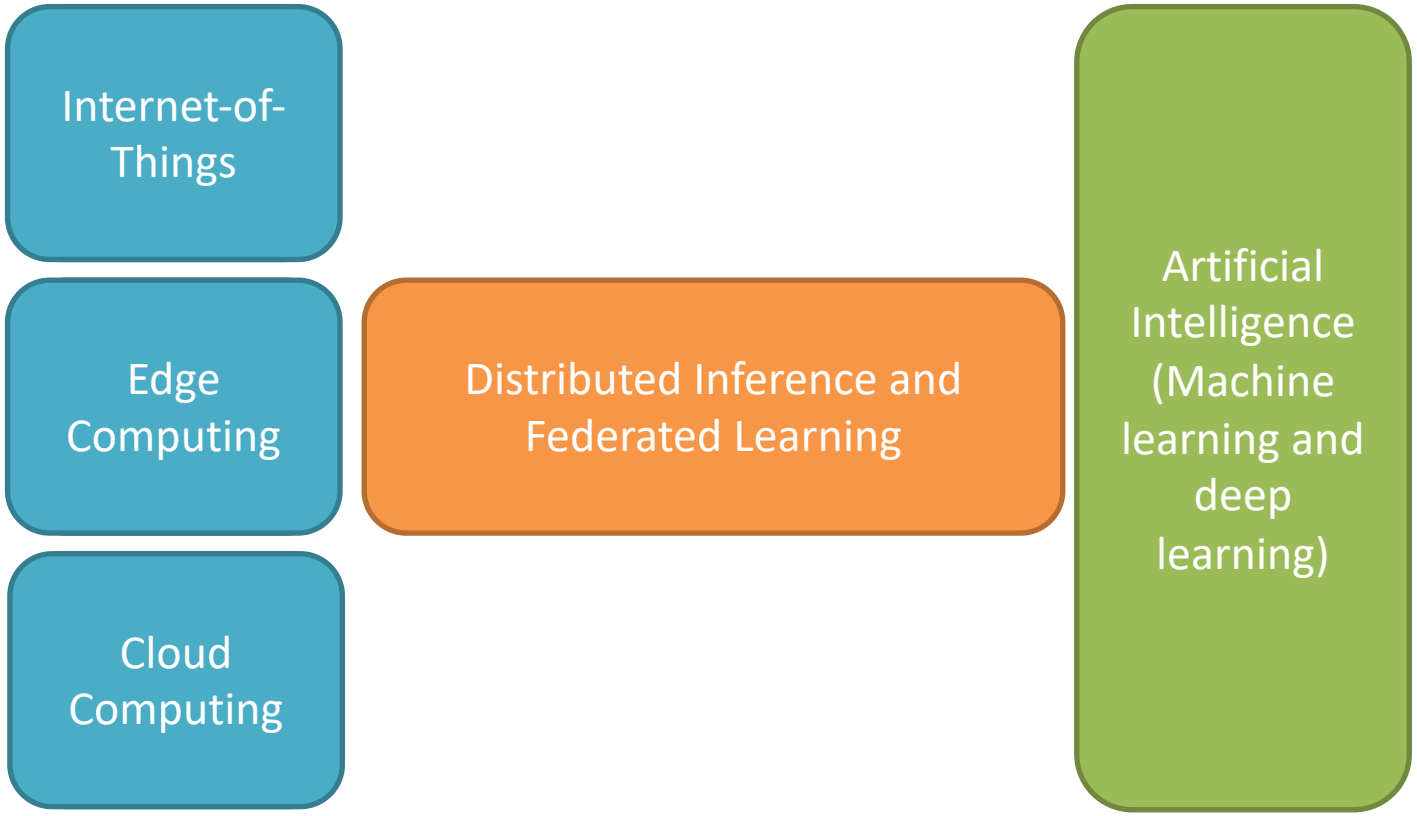
The research activity



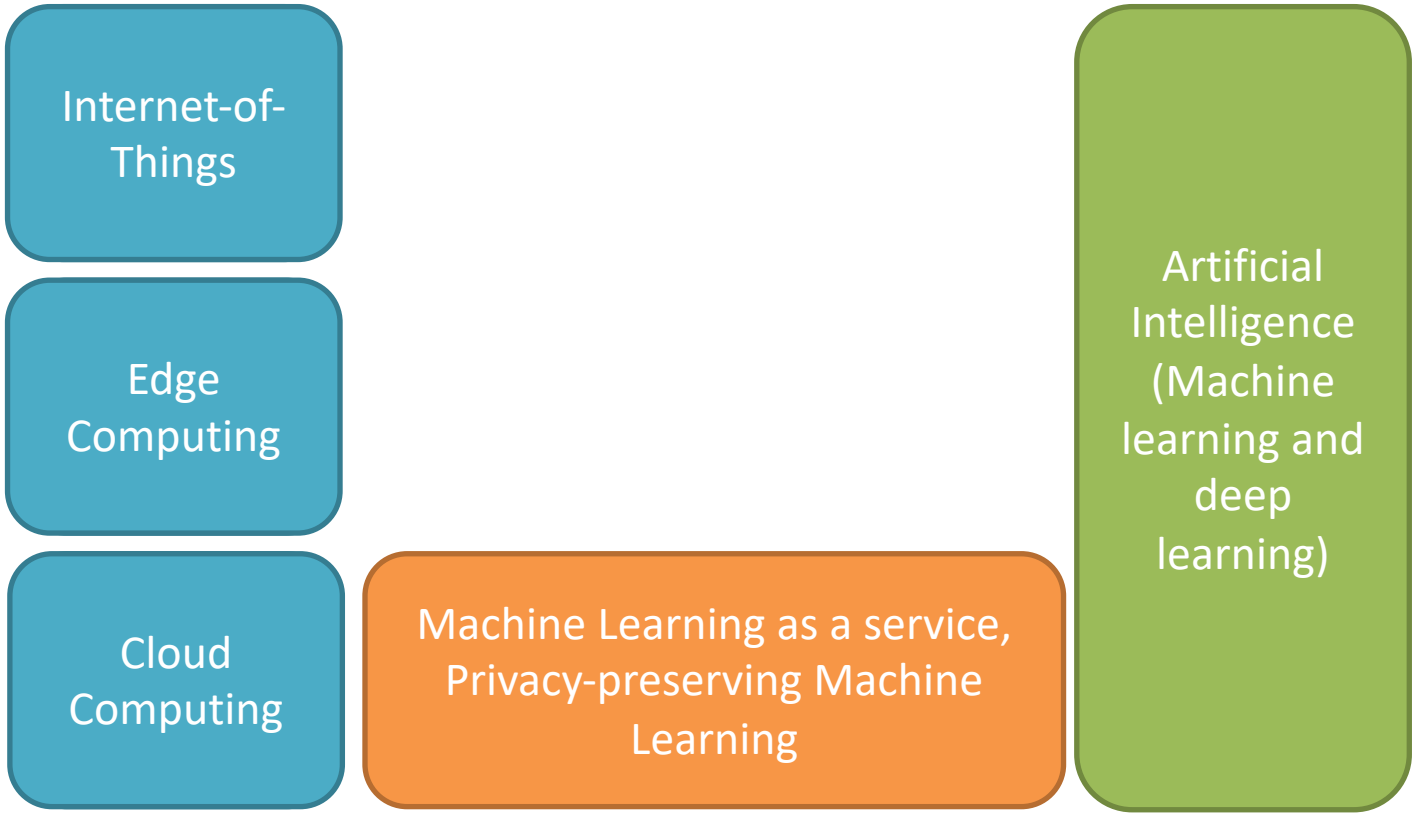
The research activity



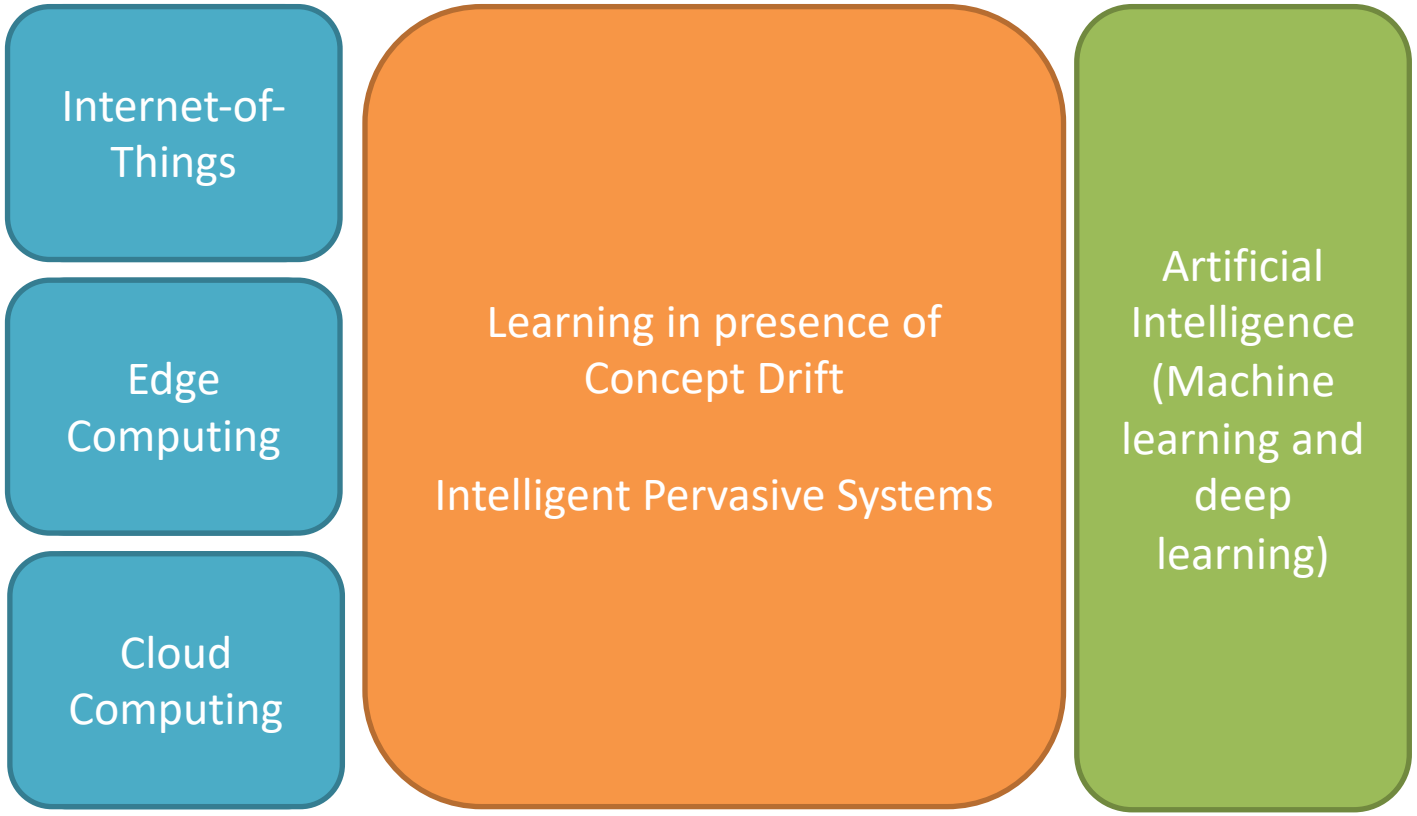
The research activity



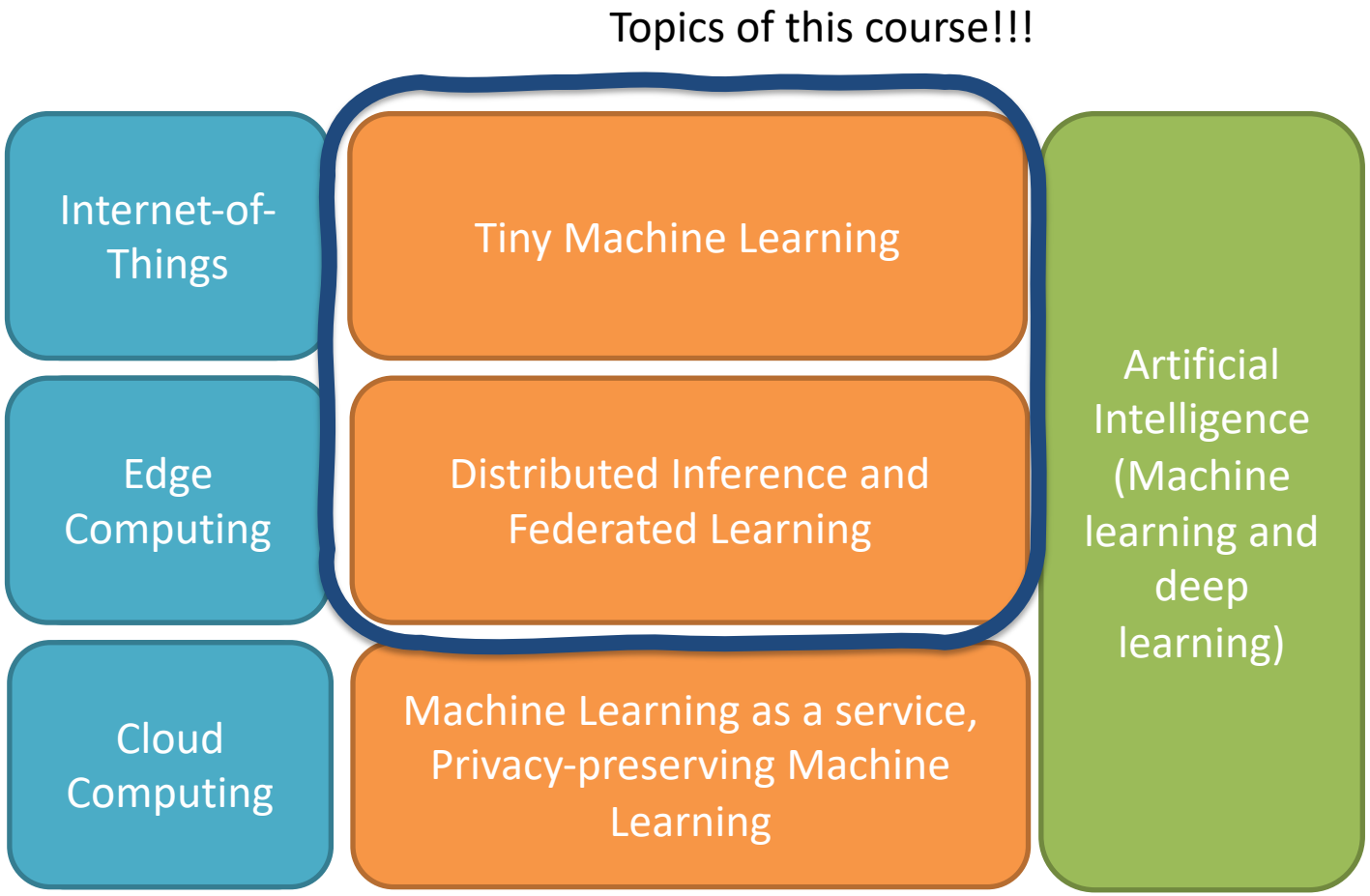
The research activity



The research activity



The research activity



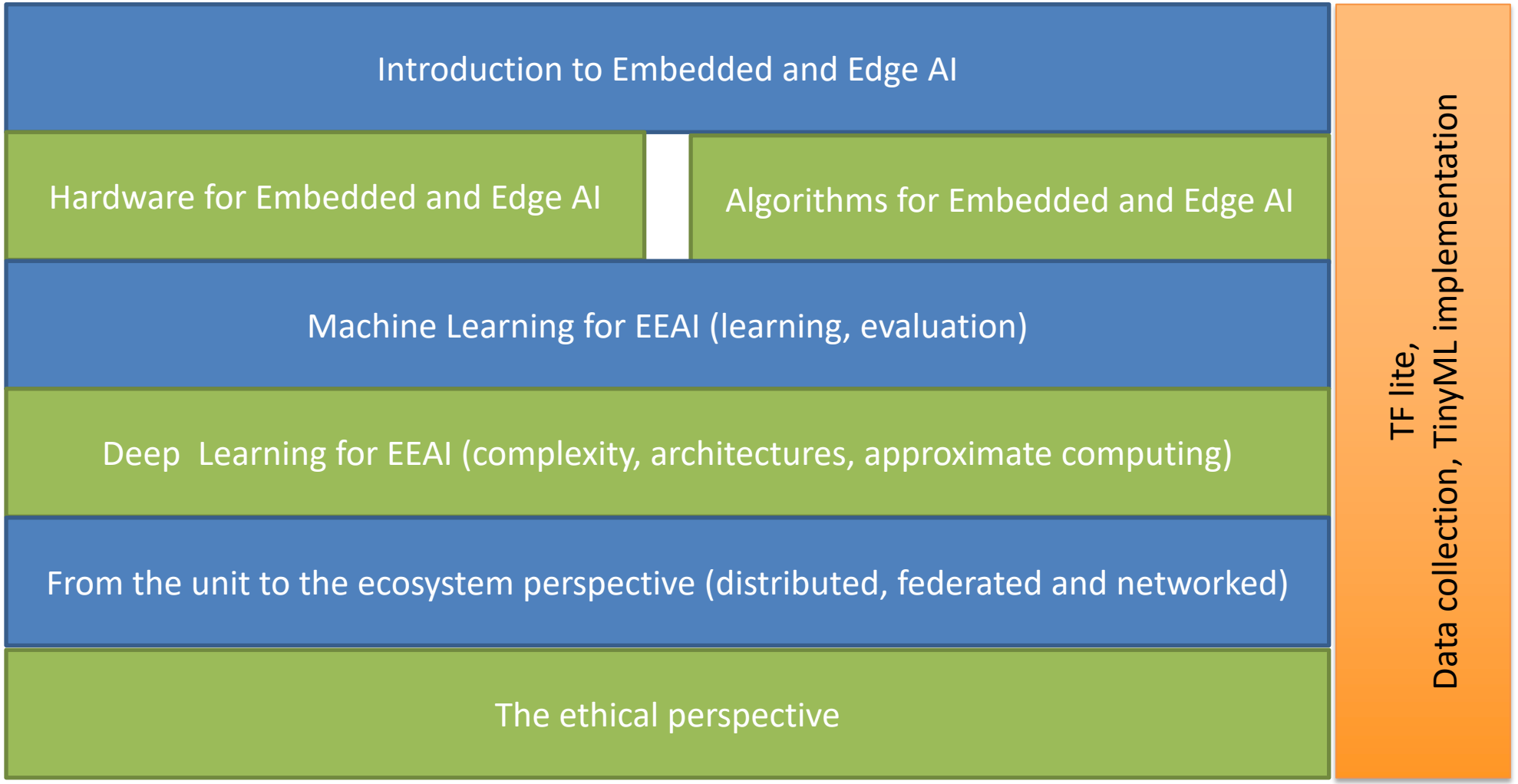
Information about the course

Lecturers

Manuel Roveri

Massimo Pavan

Topics


Introduction to Embedded and Edge AI
Hardware for Embedded and Edge AI
Algorithms for Embedded and Edge AI
Machine Learning for EEAI (learning, evaluation)
Deep Learning for EEAI (complexity, architectures, approximate computing)
From the unit to the ecosystem perspective (distributed, federated and networked)
The ethical perspective
TF lite,
Data collection, TinyML implementation

Suggested books and slides

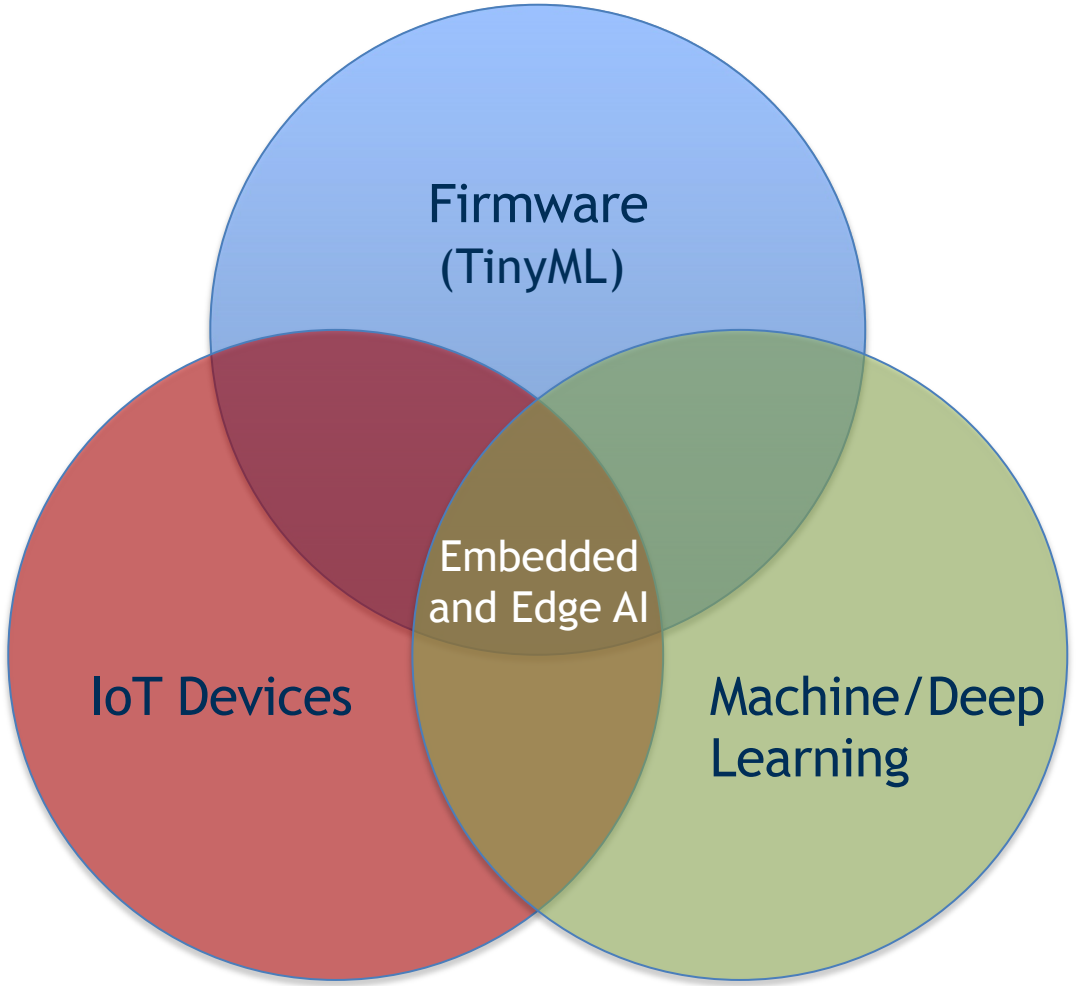
- "AI at the Edge", Daniel Situnayake, Jenny Plunkett, O'Reilly Media, Inc.
- "Tinyml: Machine learning with tensorflow lite on arduino and ultra-low-power microcontrollers", Warden, Pete, and Daniel Situnayake, O'Reilly Media, 2019.
- Slides
- Selected papers

- Please do refer to the WeBeep site of the course

Exam

- The exam will consist in **two parts**:
  1. **Written exam** (16 points) comprising questions (closed/open) about the topics of the course
  2. **Project** (16 points):
     - Your own idea with our own hardware
     - Max 2 people
     - Delivered at the exam dates
     - Code + presentation
     - Evaluation will take into account:
       - The "market" perspective (5 points)
       - The "technological" perspective (6 points)
       - The "ethical" perspective (5 points)

Let's enter into the course topics …

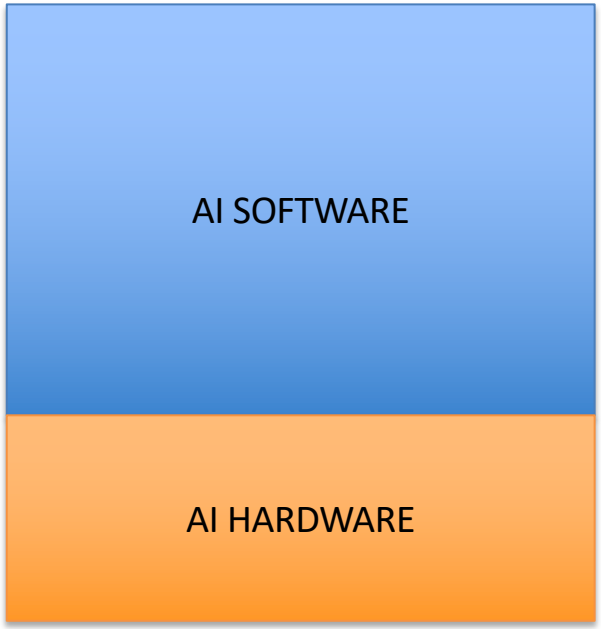

Firmware
(TinyML)
Embedded
and Edge AI
IoT Devices
Machine/Deep
Learning

A step back …

"Artificial Intelligence is the area of computer science that studies the development of **hardware and software systems endowed with abilities typical of human beings**. Such systems are able to autonomously pursue a given purpose by making decisions that, until then, were usually made by human beings "

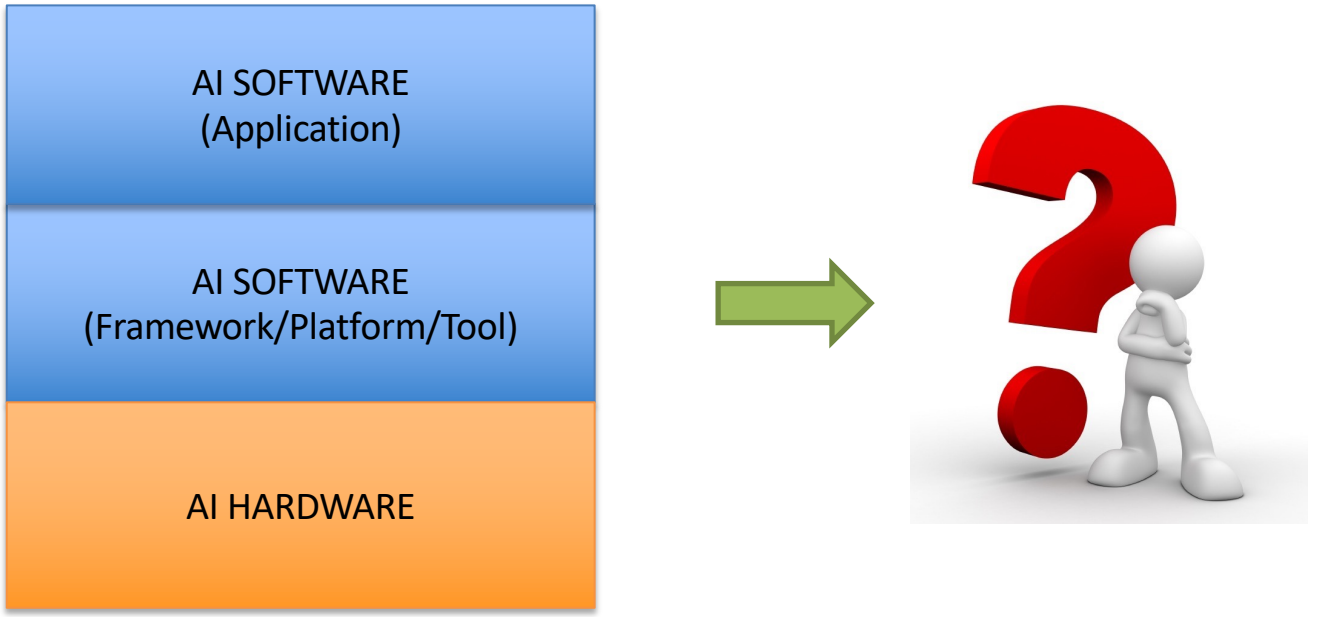
Osservatorio Artificial Intelligence,
Politecnico di Milano, 2018

«… hardware and software systems endowed with abilities typical of human beings …»

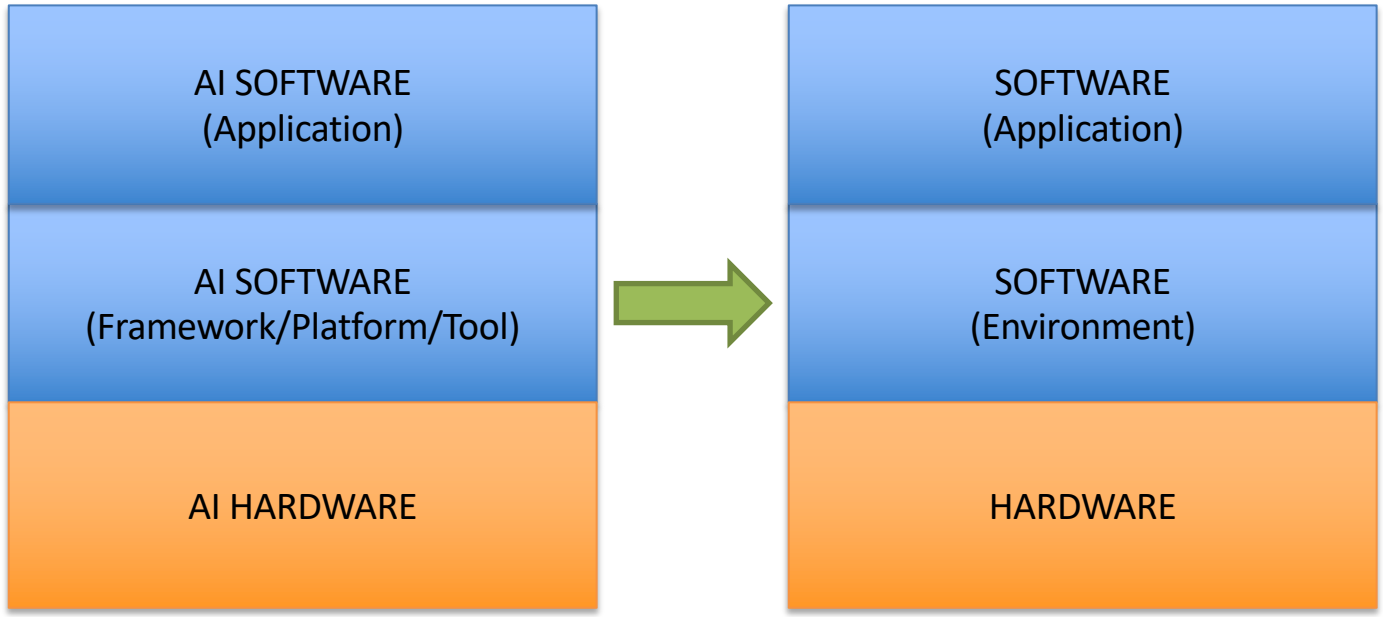


21

«… hardware and software systems endowed with abilities typical of human beings …»



An IT perspective for the AI



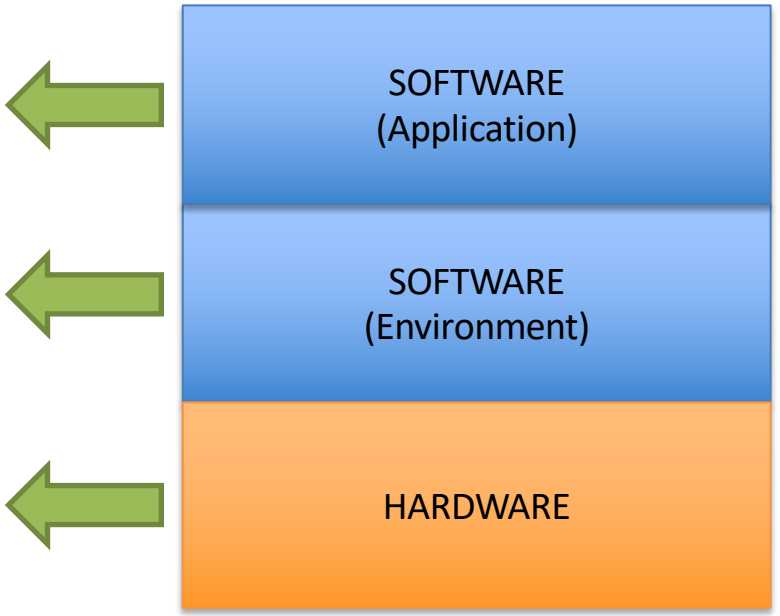
23

An IT perspective for the AI

The application, i.e., the reason why this system exists

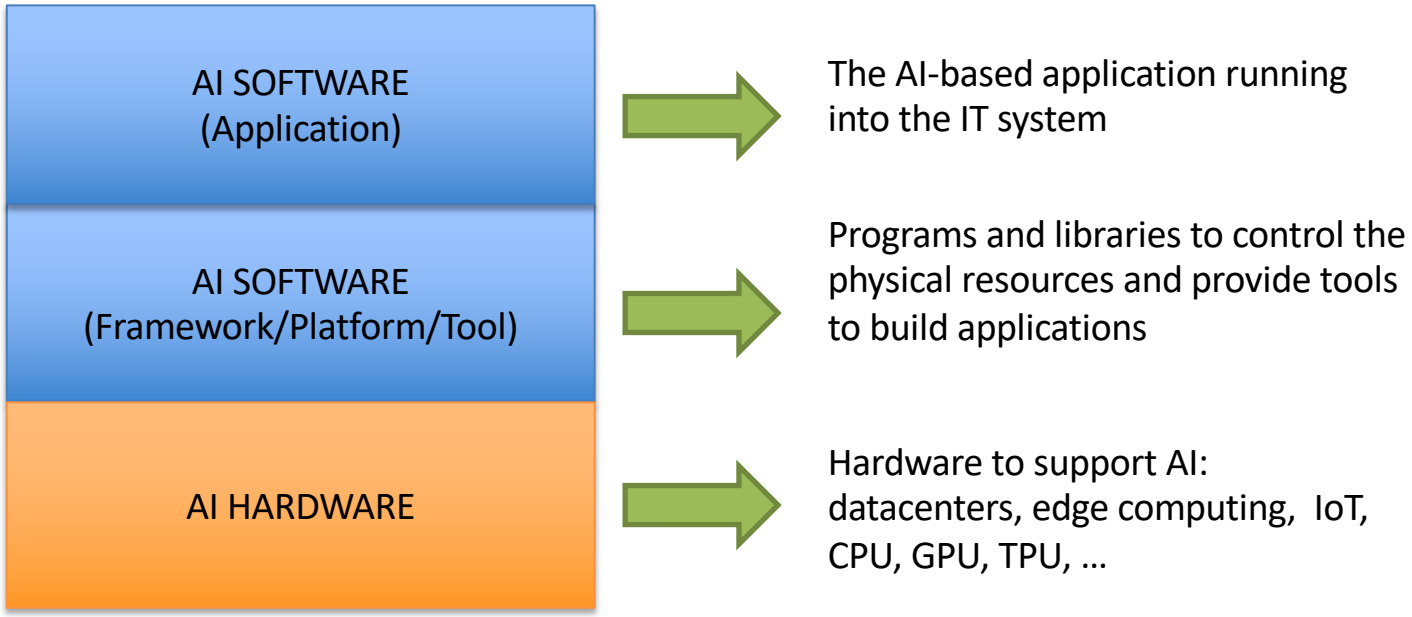
Programs and libraries to control the physical resources and provide tools to build applications

Physical resources of the system (computation, storage, input/output, etc..)



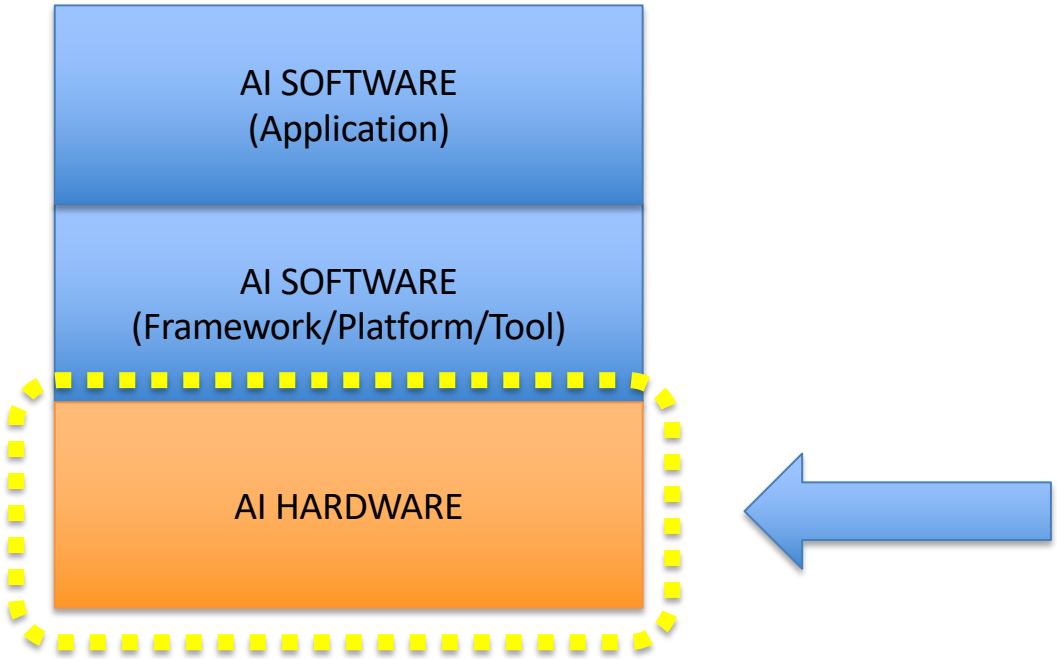
24

An IT perspective for the AI



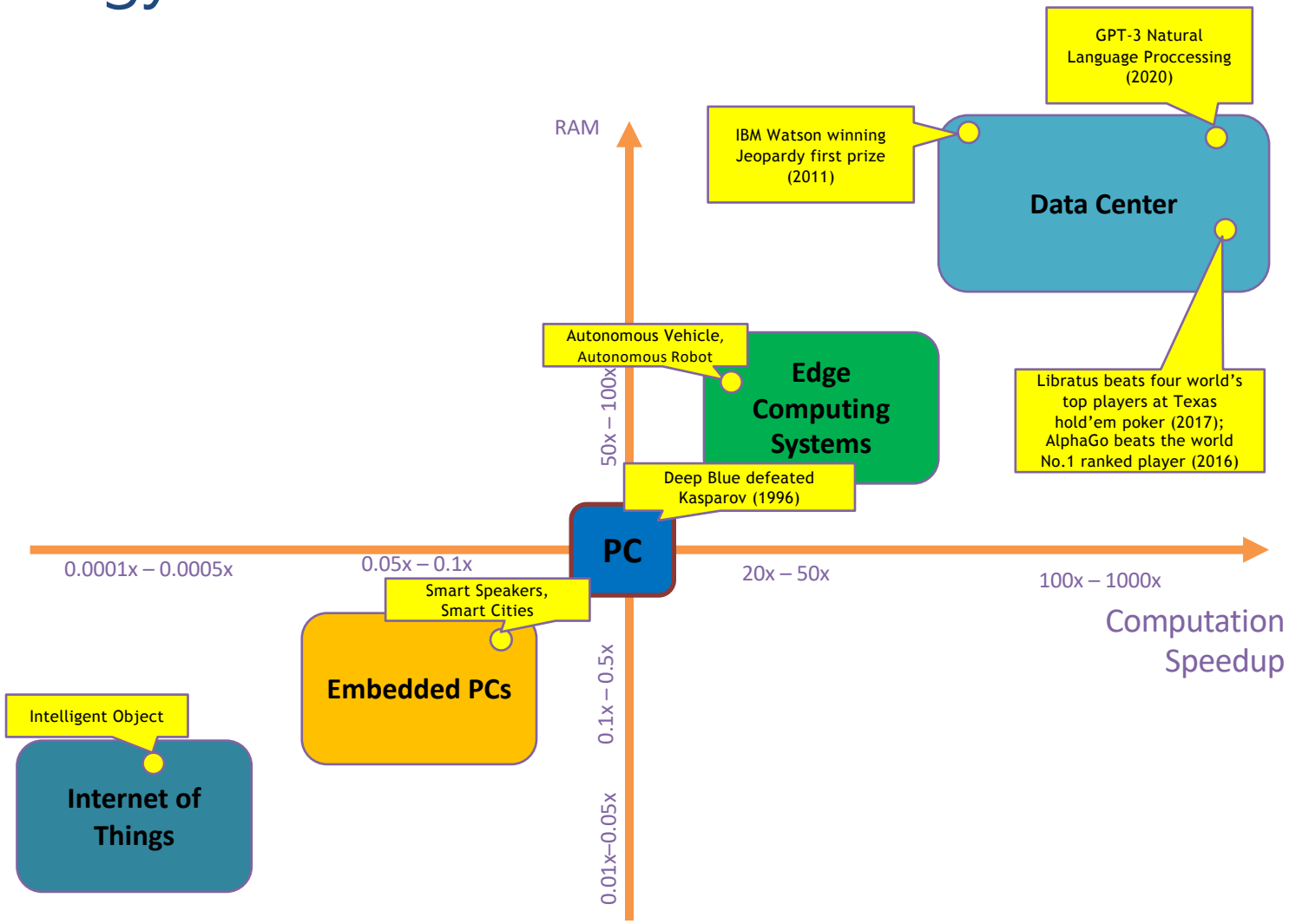
25

«… hardware and software systems endowed with abilities typical of human beings …»

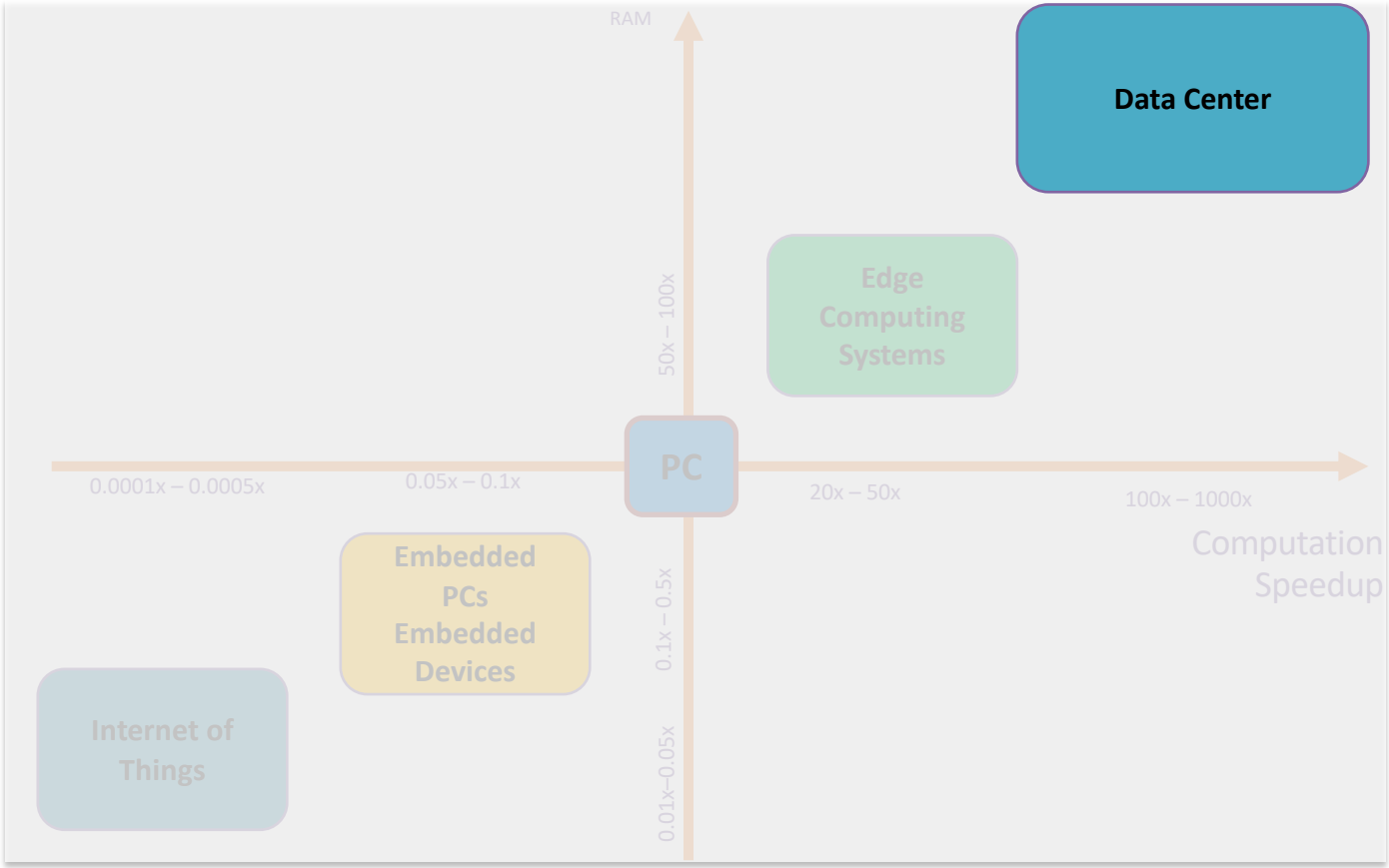


Which are the AI Hardware platforms?

AI and Technology



Examples of Computing Infrastructures



Data Centers: a technological perspective

The Pionen White Mountains is a Swedish data center. This center is located in Stockholm and is one of the largest data centers in the world.

Server for Processing

Server for Storage

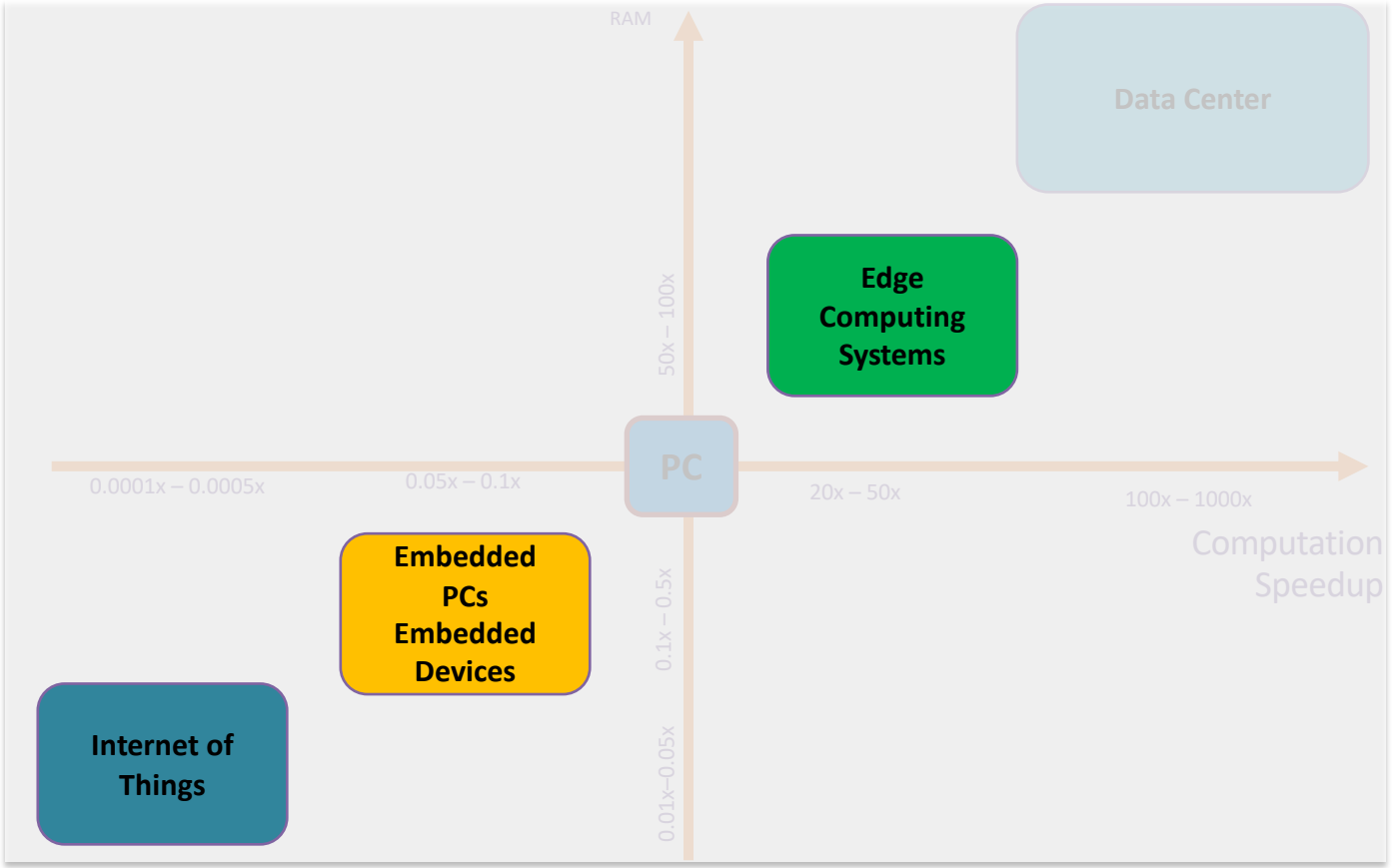
Server for Communication

Data Centers: advantages and disadvantages

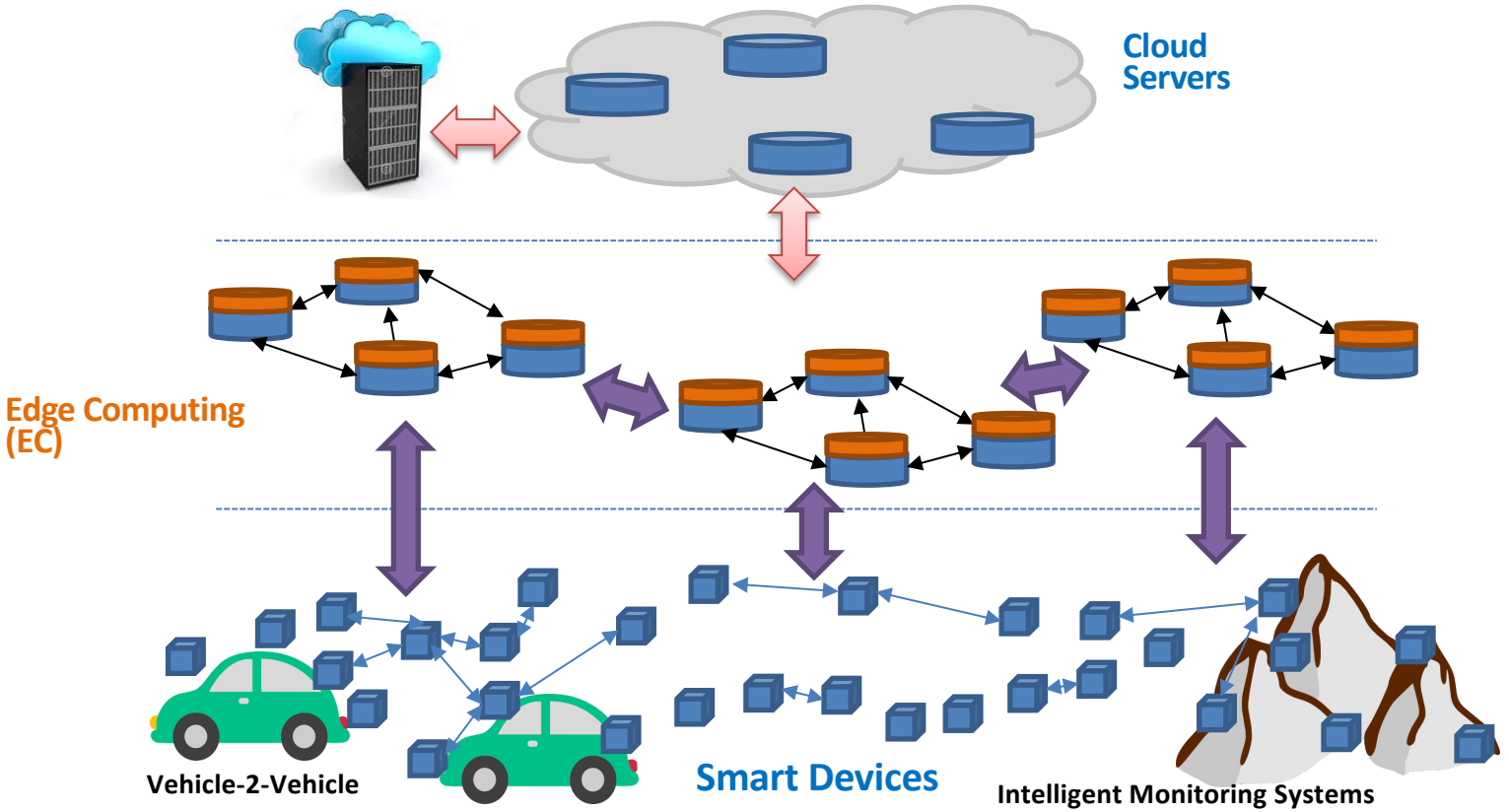
- ✓ Lower IT costs
- ✓ High performance
- ✓ Instant software updates
- ✓ "Unlimited" storage capacity
- ✓ Increased data reliability
- ✓ Universal document access
- ✓ Device Independence

- ➢ Require a constant Internet connection
- ➢ Do not work well with low-speed connections
- ➢ Hardware Features might be limited
- ➢ Privacy and security issues
- ➢ High Power Consumption
- ➢ Latency in making decision

Edge Computing, PC Embedded and IoT



Edge/Fog Computing Systems



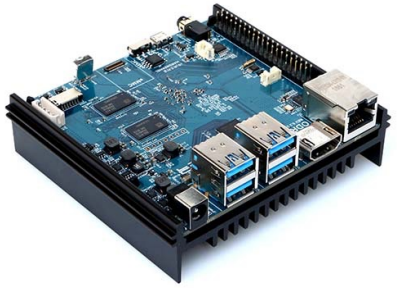
IoT Gateways	SYS-E50-9AP-WIFI	SYS-E100-9S-E	SYS-E300-8D	SYS-5018D-FN8T
				
Processor/Cache				
CPU	**Intel® Atom® processor E3940** Single socket FCBGA 1296 9.5W, 4C	**7th Generation Intel® Core i5-7300U Processor** Single Socket FCBGA 1356 System-on-Chip CPU TDP support 15W	**Intel® processor D-1518,** 2.2GHz; CPU TDP support 35W FCBGA 1667: 4 Cores, 8 Threads / 6MB	**Intel® Xeon® processor D-1518** 2.2GHz; CPU TDP support 35W FCBGA 1667: 4 Cores, 8 Threads / 6MB
System Memory				
Memory Capacity	Up to 8GB Unbuffered non-ECC SO-DIMM DDR3L-1866MHz, in 1 DIMM socket	Up to 32GB Unbuffered non-ECC SO-DIMM, DDR4-2133MHz, in 2 DIMM slots	4x DDR4 DIMM sockets Supports up to 128GB DDR4 ECC RDIMM Supports up to 64GB DDR4 ECC/non-ECC UDIMM	4x DDR4 DIMM sockets Supports up to 128GB DDR4 ECC RDIMM Supports up to 64GB DDR4 ECC/non-ECC UDIMM
Memory Type	DDR3L up to 1866MHz	DDR4 up to 2133MHz	2133/1866/1600MHz ECC DR4 ECC RDIMM and ECC/Non-ECC UDIMM	2133/1866/1600MHz ECC DDR4 ECC RDIMM and ECC/Non-ECC UDIMM
DIMM Sizes	8GB, 4GB, 2GB	16GB, 8GB, 4GB	32GB, 16GB, 8GB, 4GB	32GB, 16GB, 8GB, 4GB
Memory Voltage	1.35 V	1.2 V	1.2 V	1.2 V

- ✓ High computational capacity
- ✓ Distributed computing
- ✓ Privacy and security
- ✓ Reduced Latency in making a decision

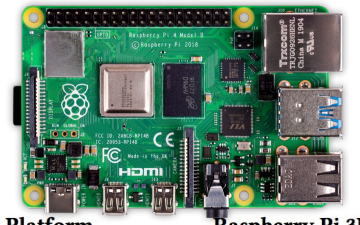
- ➢ Require a power connection
- ➢ Require connection with the Cloud

Embedded PCs

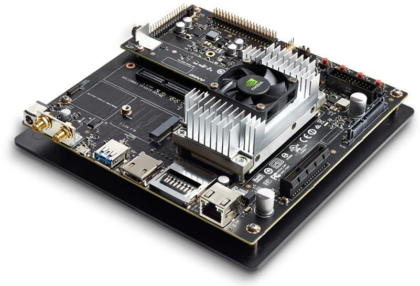
Ordoid



Raspberry



Platform	Raspberry Pi 3B
CPU	ARM11 @ 1.2 GHz
RAM	1024 MB



Jetson TX2

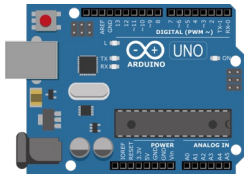
Google Coral

- ✓ Pervasive computing
- ✓ High performance unit
- ✓ Availability of development boards
- ✓ Programmed as PC
- ✓ Large community

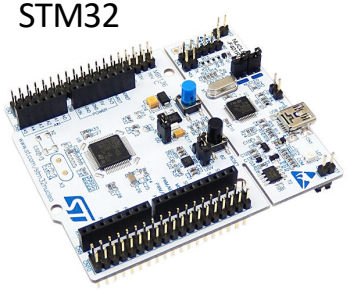
- ➢ Pretty high power consumption
- ➢ (Some) HW design has to done

Internet-of-Things

Arduino



STM32



	STM32 L1 Series	STM32F4 Series
Domain	Ultra Low-Power	High-Performance
Flash Memory (kB)	32 to 512	64 to 2048
RAM Memory (kB)	4 to 80	32 to 320
CPU	ARM® Cortex®-M3	ARM® Cortex®-M4
Frequency (MHz)	32	84 to 180
Supply Voltage (V)	1.65 to 3.6	1.71 to 3.6
Supply Current (μA)	0.28 (0.28) to 230	1.1 (140) to 282

- ✓ Highly Pervasive
- ✓ Wireless connection
- ✓ Battery Powered
- ✓ Low costs
- ✓ Sensing and actuating

- ➢ Low computing ability
- ➢ Constraints on energy
- ➢ Constraints on memory (RAM/FLASH)
- ➢ Difficulties in programming

This is exactly where Embedded and Edge Computing comes into play …

Internet-of-Things

Computing everywhere

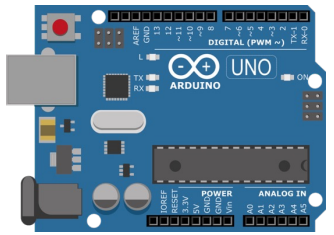
Edge Computing

Move (intelligent) processing as close as possible to data generation units …

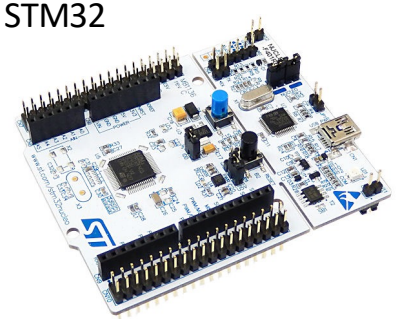
- ✓ Increase autonomy
- ✓ Reduce decision-making latency
- ✓ Reduce transmission bandwidth
- ✓ Increase energy-efficiency
- ✓ Security and Privacy
- ✓ Incremental/Adaptive Learning
- ✓ Ecosystem of units

- ➢ Low computing ability
- ➢ Constraints on energy
- ➢ Constraints on memory (RAM/FLASH)
- ➢ Complexity in design and development
- ➢ Strong connection between HW, SW and ML

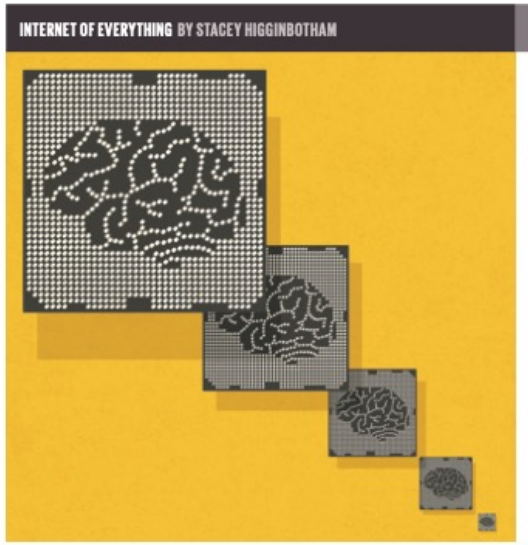
Arduino



STM32



Domain	STM32 L1 Series Ultra Low-Power	STM32F4 Series High-Performance
Flash Memory (kB)	32 to 512	64 to 2048
RAM Memory (kB)	4 to 80	32 to 320
CPU	ARM® Cortex®-M3	ARM® Cortex®-M4
Frequency (MHz)	32	84 to 180
Supply Voltage (V)	1.65 to 3.6	1.71 to 3.6
Supply Current (μA)	0.28 (0.28) to 230	1.1 (140) to 282



INTERNET OF EVERYTHING BY STACEY HIGGINBOTHAM

MACHINE LEARNING ON THE EDGE

IN FEBRUARY, a group of researchers from Google, Microsoft, Qualcomm, Samsung, and half a dozen universities will gather in San Jose, Calif., to discuss the challenge of bringing machine learning to the farthest edge of the network, specifically microprocessors running on sensors or other battery-powered devices.

The event is called the Tiny ML Summit (ML for "machine learning"), and its goal is to figure out how to run machine learning algorithms on the tiniest microprocessors out there. Machine learning at the edge will drive better privacy practices, lower energy consumption, and build novel applications in future generations of devices.

As a refresher, at its core machine learning is the training of a neural network. Such training requires a ton of data manipulation. The end result is a model that is designed to complete a task, whether that's playing Go or responding to a spoken command.

Many companies are currently focused on building specialized silicon for machine learning in order to train networks inside data centers. They also want silicon for conducting inference—running data against a machine learning model to see if the data matches the model's results—at the edge. But the goal of the Tiny ML community is to take inference to the smallest processors out there—like an 8-bit microcontroller that powers a remote sensor.

To be clear, there's already been a lot of progress in bringing inference to the edge if we're talking about something like a smartphone. In November 2019, Google open-sourced two versions of its machine learning algorithms, one of which required 50 percent less power to run, and the other of which performed twice as fast as previous versions of the algorithm. There are also several start-ups such as Flex Logix, Greenwaves, and Syntiant tackling similar challenges using dedicated silicon.

But the Tiny ML community has different goals. Imagine including a machine learning model that can separate a conversation from background noise on a hearing aid. If you can't fit that model on the device itself, then you need to maintain a wireless connection to the cloud where the model is running. It's more efficient, and more secure, to run the model directly on the hearing aid—if you can fit it.

Tiny ML researchers are also experimenting with better data classification by using ML on battery-powered edge devices. Jags Kandasamy, CEO of Latent AI, which is developing software to compress neural networks for tiny processors, says his company is in talks with companies that are building augmented-reality and virtual-reality headsets. These companies want to take the massive amounts of image data their headsets gather and classify the images seen on the device so that they send only useful data up to the cloud for later training. For example, "If you've already seen 10 Toyota Corollas, do they all need to get transferred to the cloud?" Kandasamy asks.

On-device classification could be a game changer in reducing the amount of data gathered and input into the cloud, which saves on bandwidth and electricity. Which is good, as machine learning typically requires a lot of electricity.

There's plenty of focus on the "bigger is better" approach when it comes to machine learning, but I'm excited about the opportunities to bring machine learning to the farthest edge. And while Tiny ML is still focused on the inference challenge, maybe someday we can even think about training the networks themselves on the edge. ■

POST YOUR COMMENTS AT
spectrum.ieee.org/tinymachinelearning-jan2020

ILLUSTRATION BY Dan Page

Embedded and Edge Computing Systems: Applications