POLITECNICO MILANO 1863

The Abdus Salam International Centre for Theoretical Physics

Hardware Architectures for Embedded and Edge AI (from ML to HW and back)

Prof. Manuel Roveri

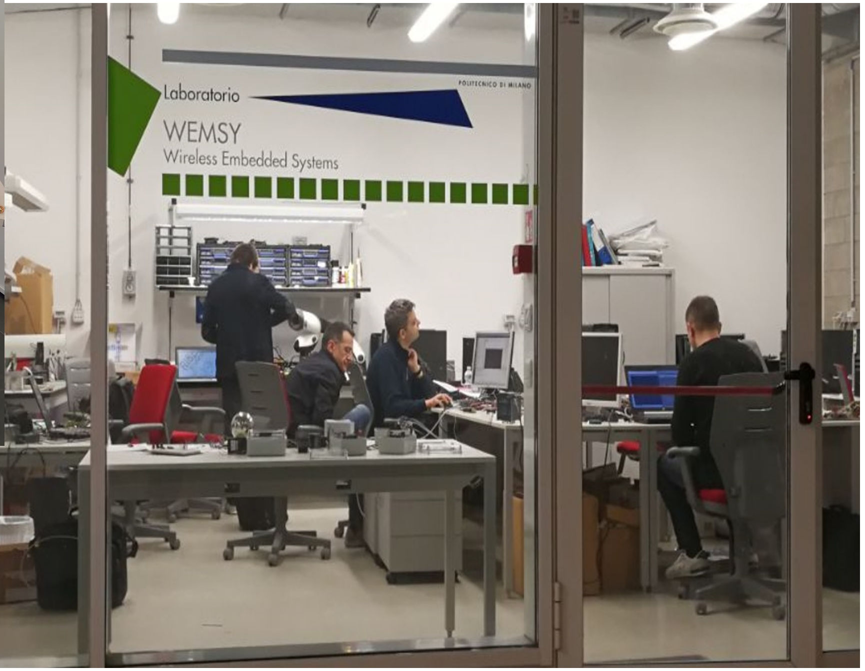
«Workshop on Widening Access to TinyML Network by Establishing Best Practices in Education»

Prof. Manuel Roveri

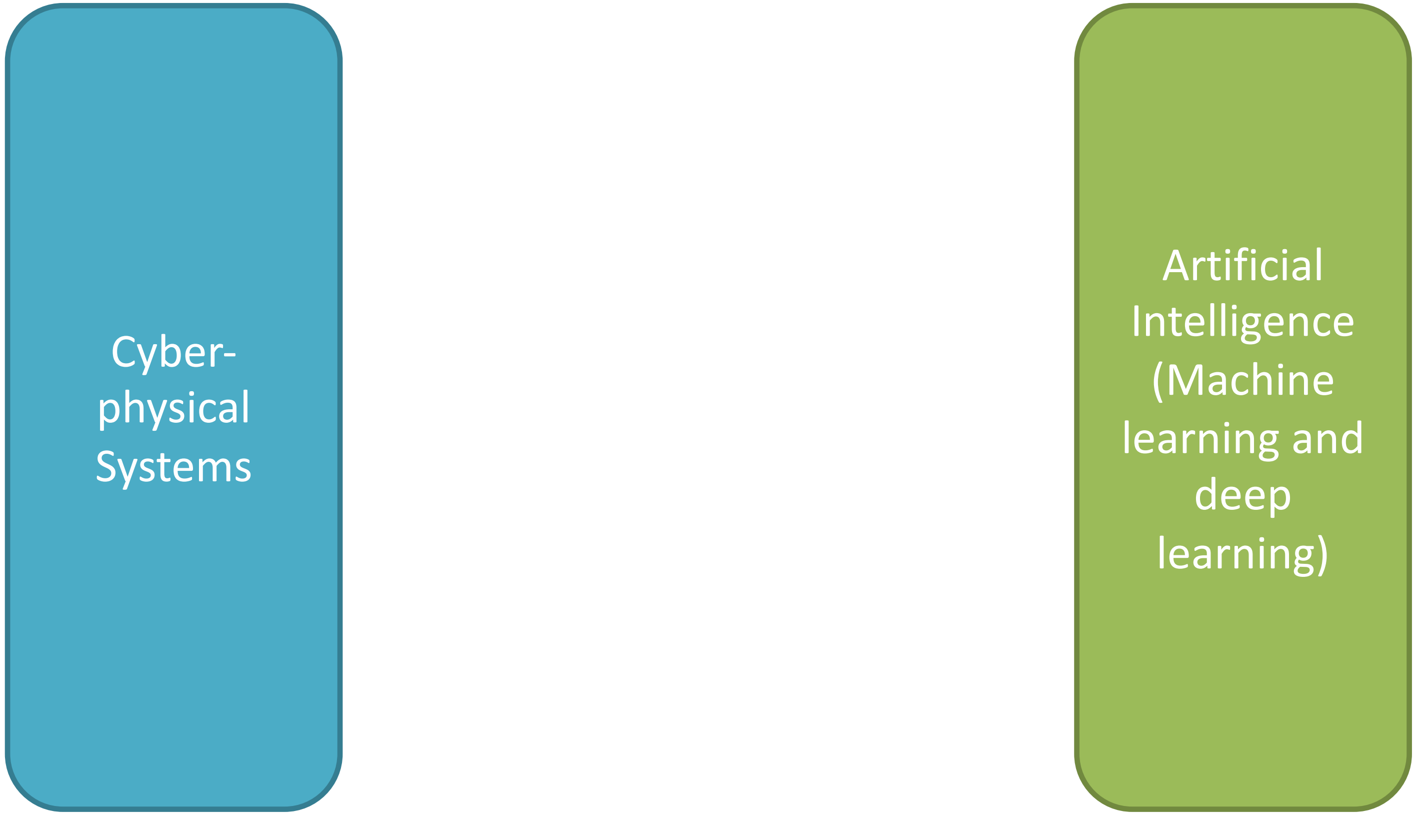
- **Full Professor**

  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

AI-Tech Research Lab @ Politecnico di Milano



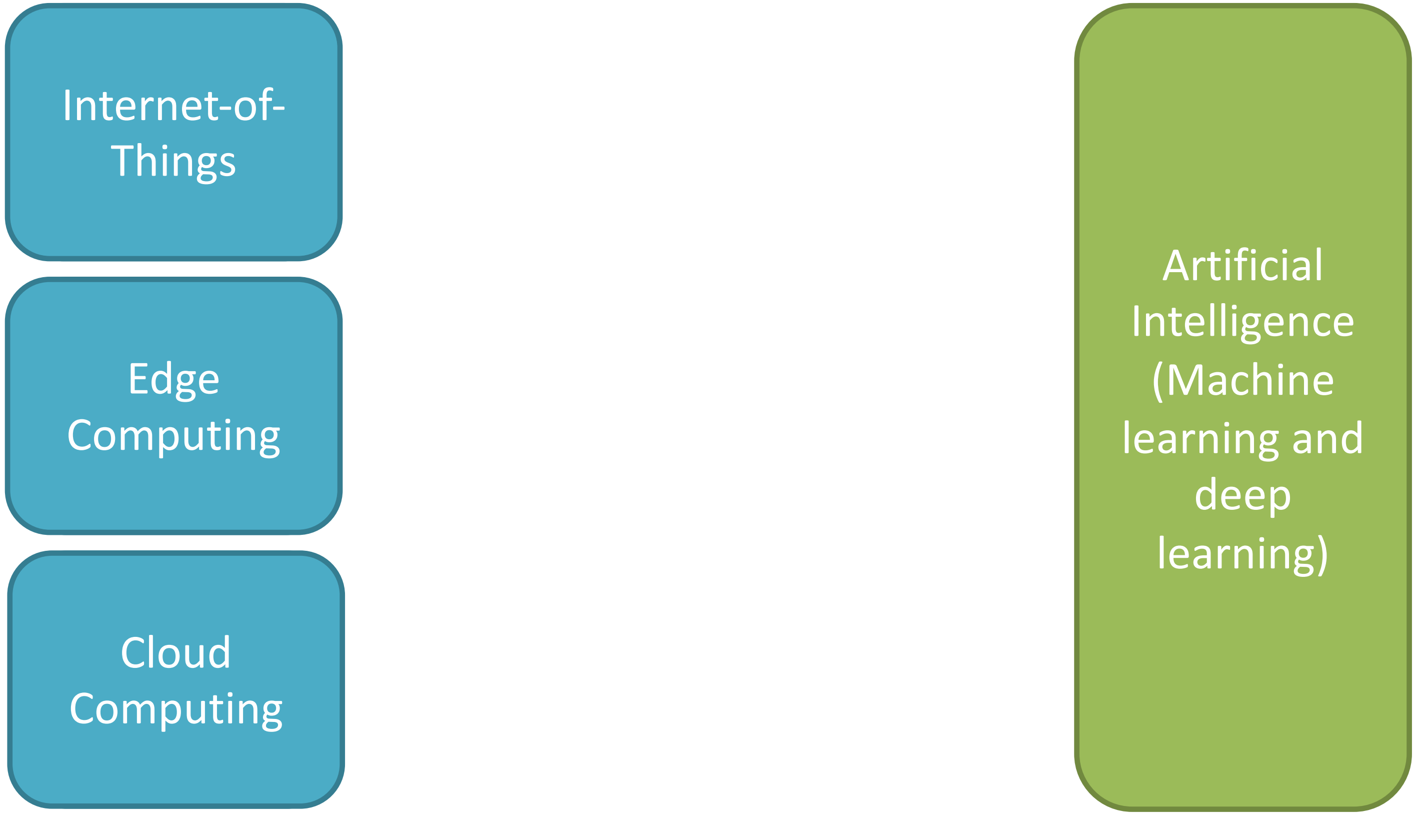
The research activity



Cyber-physical Systems

Artificial Intelligence (Machine learning and deep learning)

The research activity



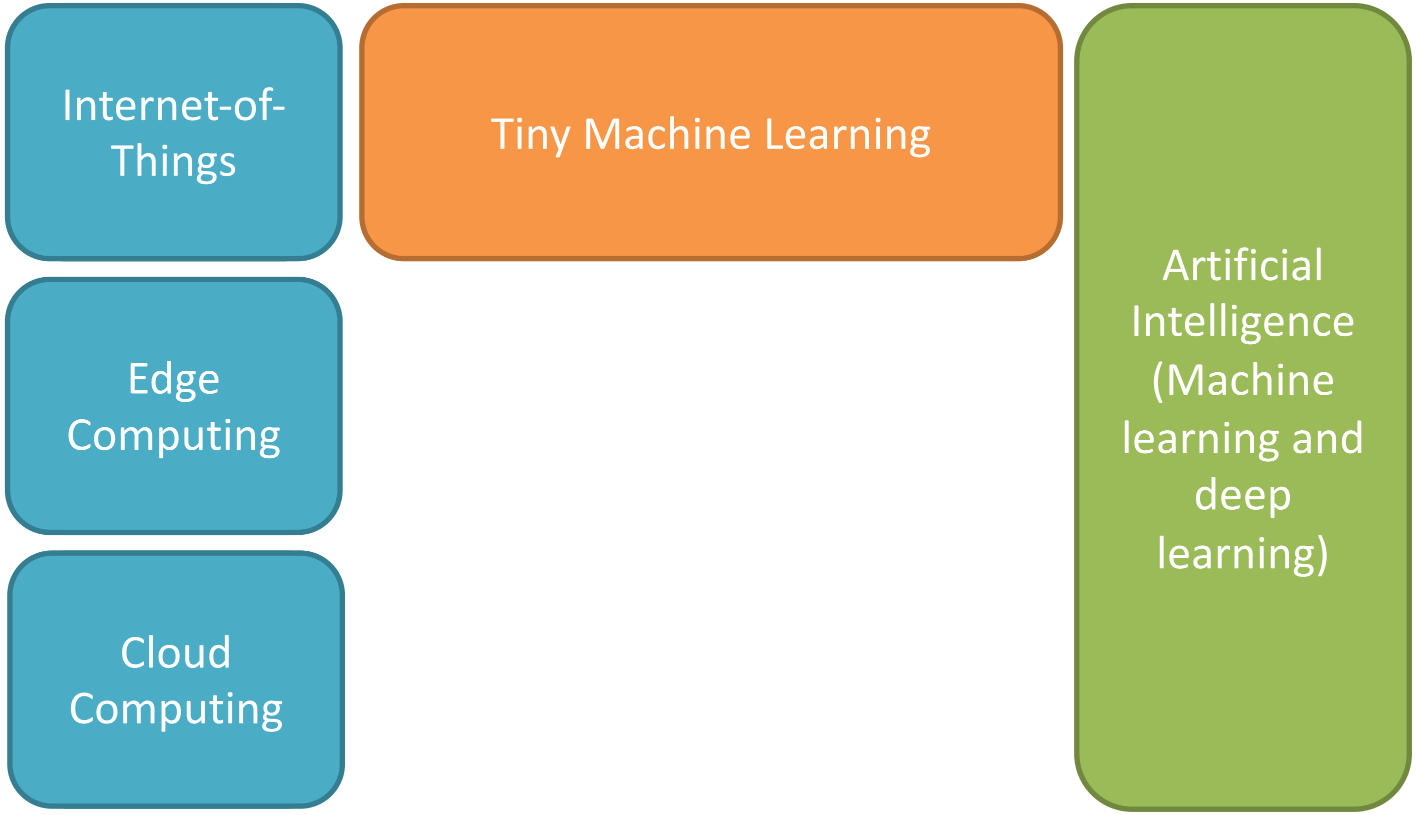
Internet-of-Things

Edge Computing

Cloud Computing

Artificial Intelligence (Machine learning and deep learning)

The research activity



Internet-of-Things

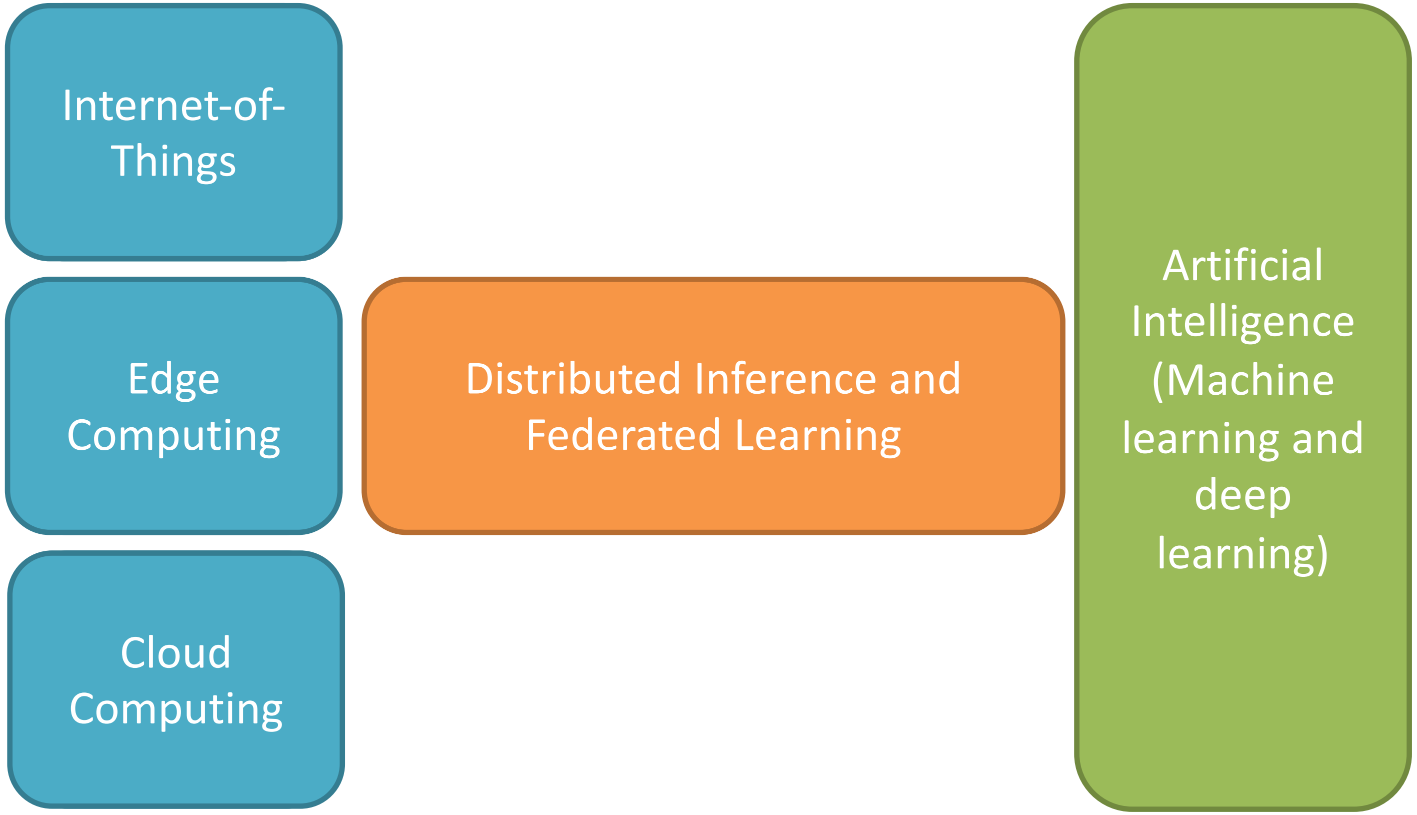
Tiny Machine Learning

Edge Computing

Cloud Computing

Artificial Intelligence (Machine learning and deep learning)

The research activity



Internet-of-Things

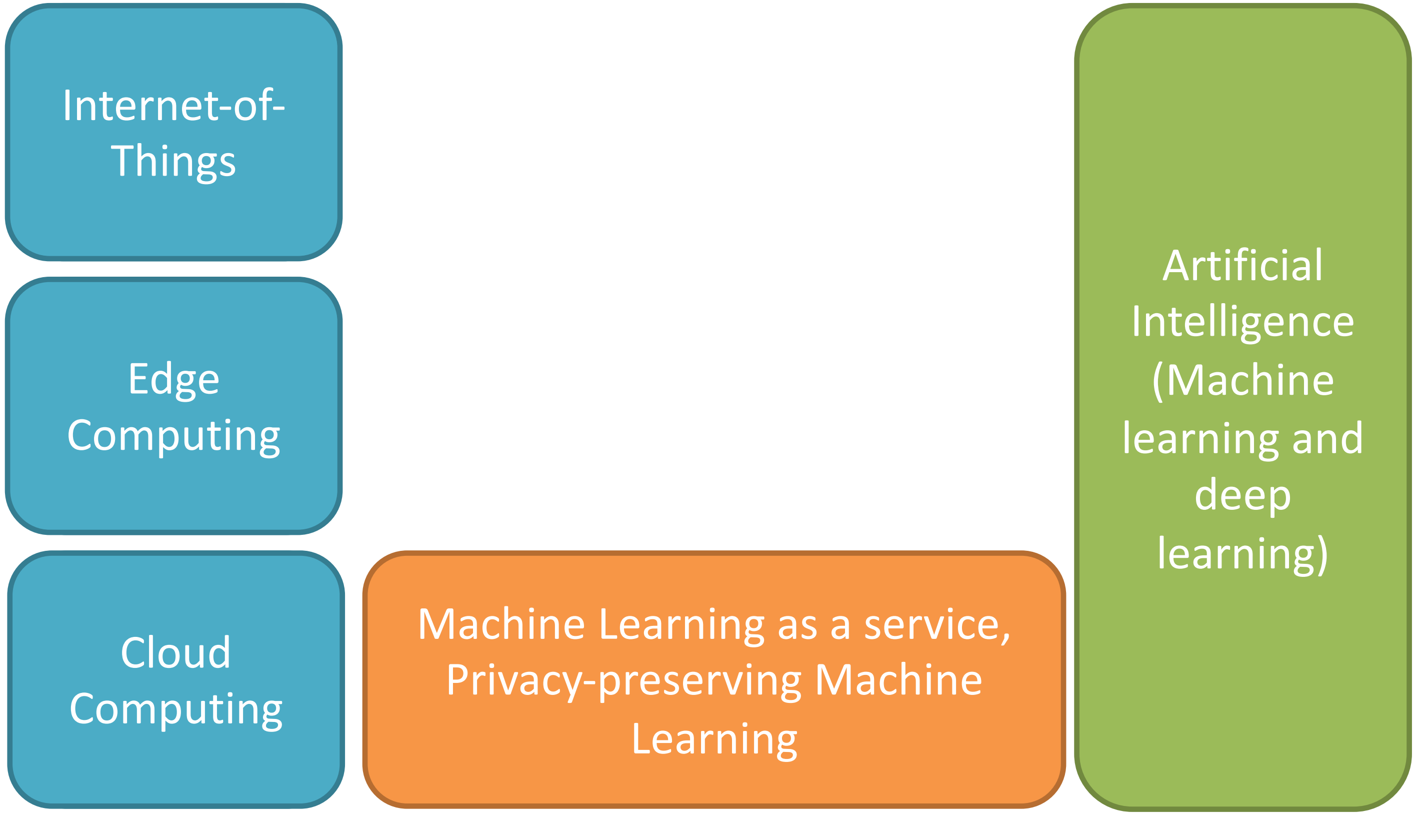
Edge Computing

Cloud Computing

Distributed Inference and Federated Learning

Artificial Intelligence (Machine learning and deep learning)

The research activity



Internet-of-Things

Edge Computing

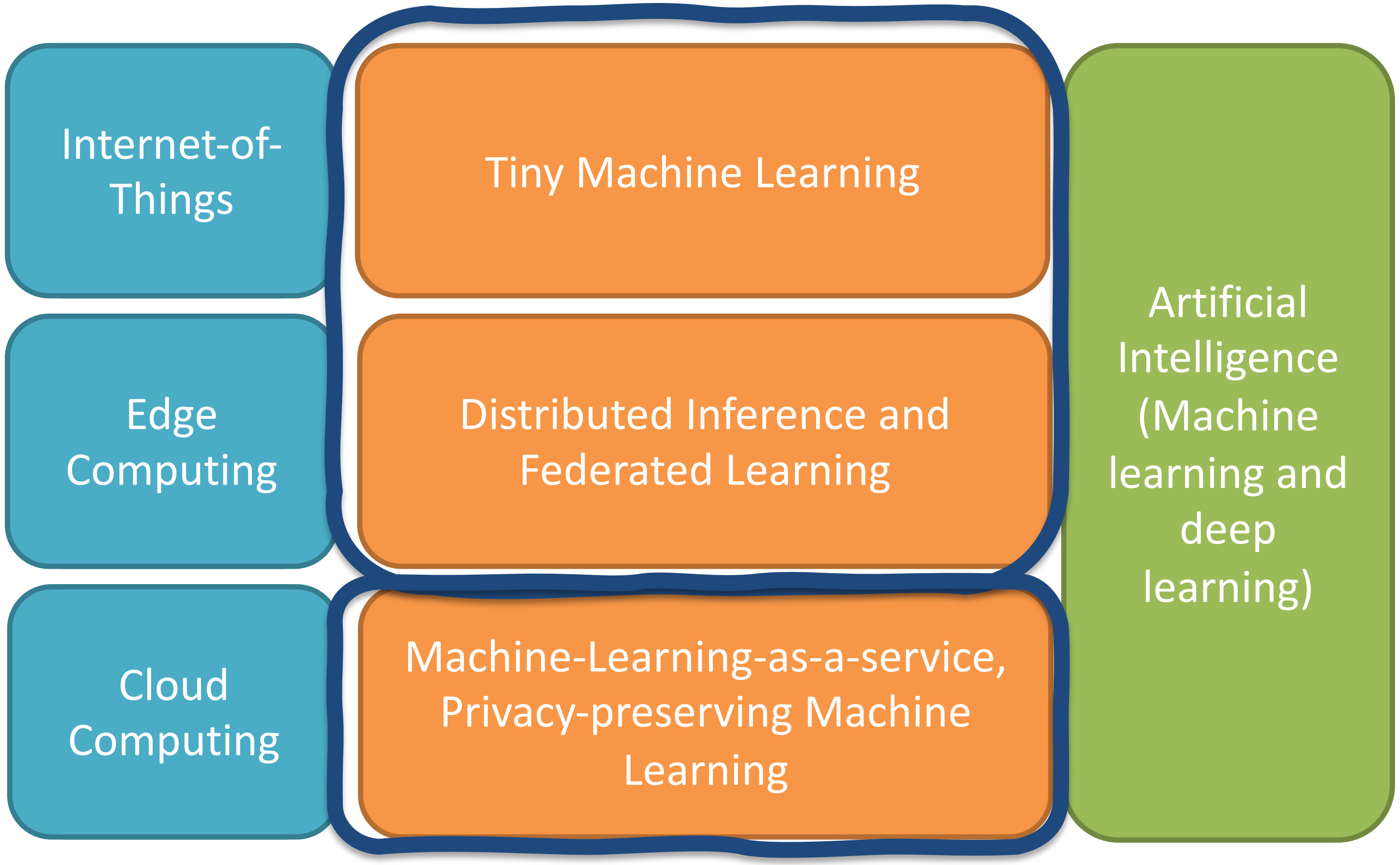
Cloud Computing

Machine Learning as a service, Privacy-preserving Machine Learning

Artificial Intelligence (Machine learning and deep learning)

The research activity

“Hardware Architectures for Embedded and Edge AI” Course



Internet-of-Things

Edge Computing

Cloud Computing

Tiny Machine Learning

Distributed Inference and Federated Learning

Machine-Learning-as-a-service, Privacy-preserving Machine Learning

Artificial Intelligence (Machine learning and deep learning)

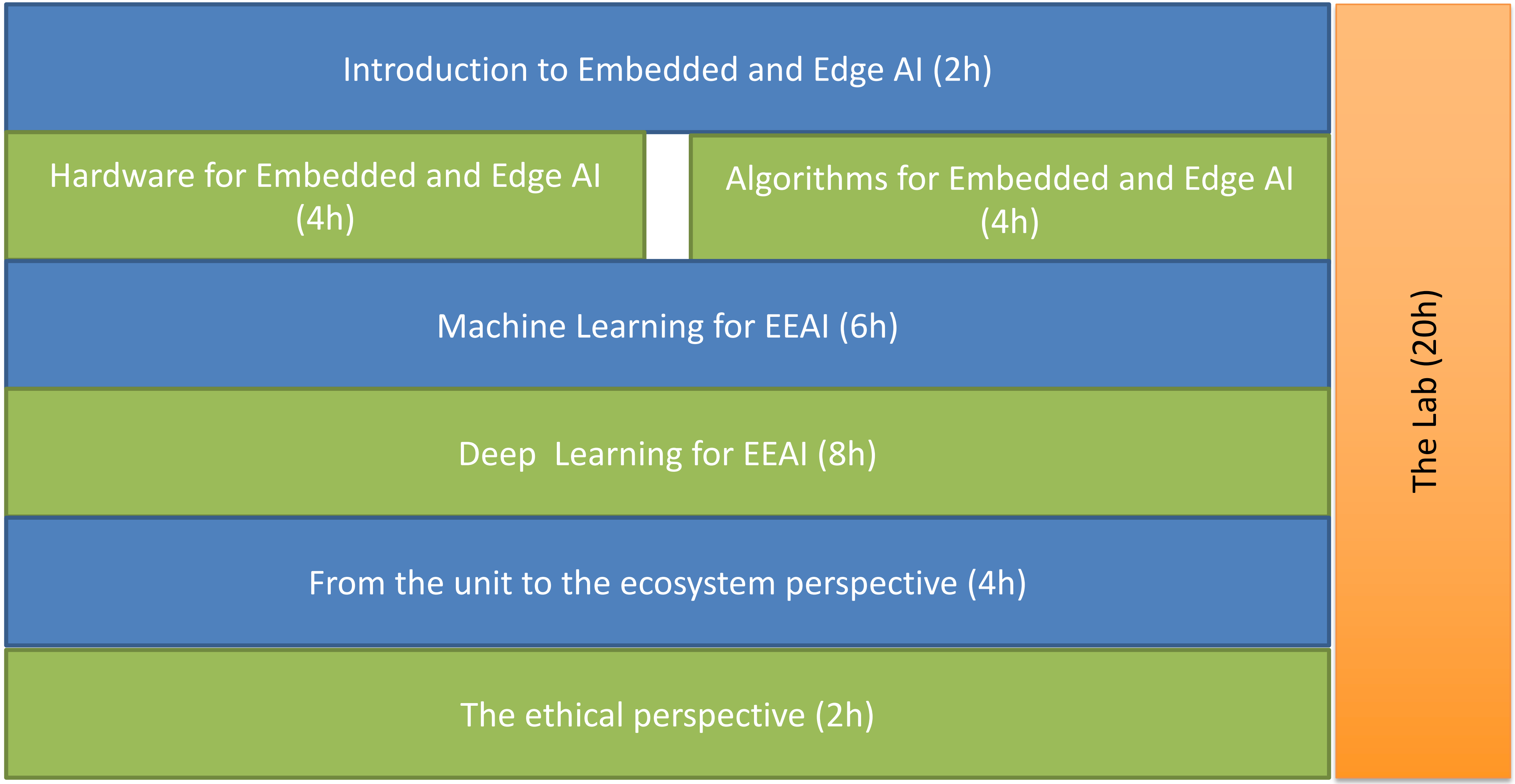
“Computing Infrastructures” Course

“Hardware Architectures for Embedded and Edge AI” Information about the course

Course Details

- Course Title: ″HARDWARE ARCHITECTURES FOR EMBEDDED AND EDGE AI″
- Academic Year 2022/2023
- School of Industrial and Information Engineering
- Master of Science degree - Computer Science and Engineering
- Course Type Mono-Disciplinary Course
- Credits (CFU / ECTS) 5.0
- Course Organization: 30h lectures (M. Roveri) + 20h labs (M. Pavan)
- Number of enrolled students: 62
  - 66% Computer Science
  - 27% Electronics
  - 7% Bio – Control Theory - Telecom

Course Organization



Introduction to Embedded and Edge AI (2h)

Hardware for Embedded and Edge AI (4h)

Algorithms for Embedded and Edge AI (4h)

Machine Learning for EEAI (6h)

Deep Learning for EEAI (8h)

From the unit to the ecosystem perspective (4h)

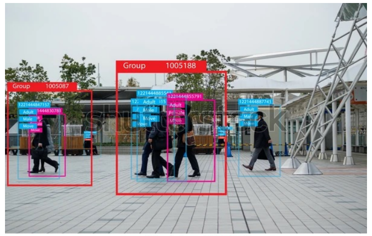
The ethical perspective (2h)

The Lab (20h)

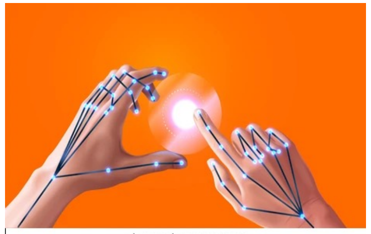
1) Introduction to Embedded and Edge AI (2h)

www.shutterstock.com · 1816734485

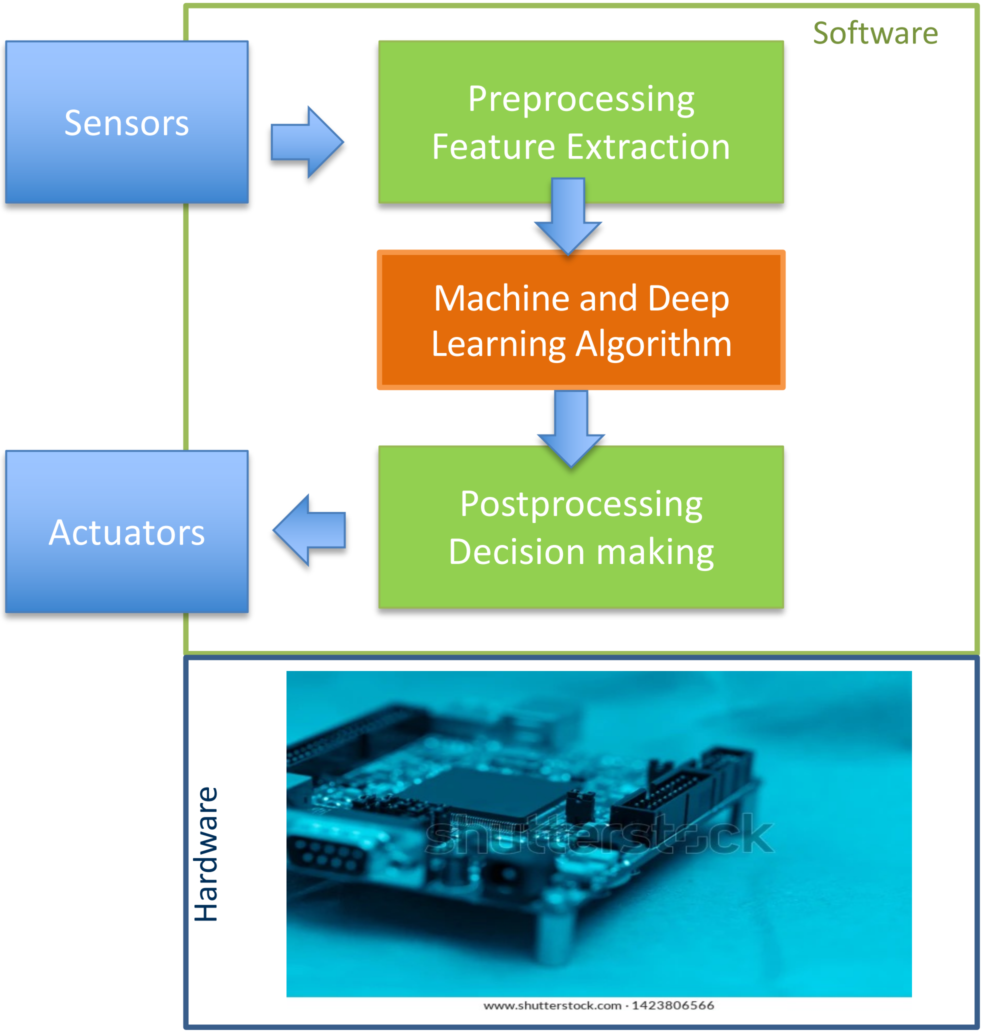
Wake-word detection


www.shutterstock.com · 794528206

Person detection


shutterstock.com · 678755710

Gesture recognition



www.shutterstock.com · 1423806566

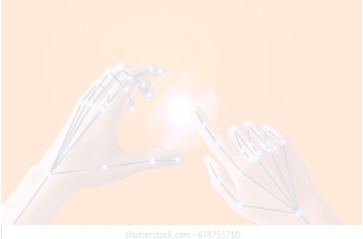
Five Ws in Embedded and Edge Ai:

- Why do we need EEAI?
- What can we do with EEAI?
- Where can we find?
- When do we need it (design)?
- Who is in charge of EEAI code?

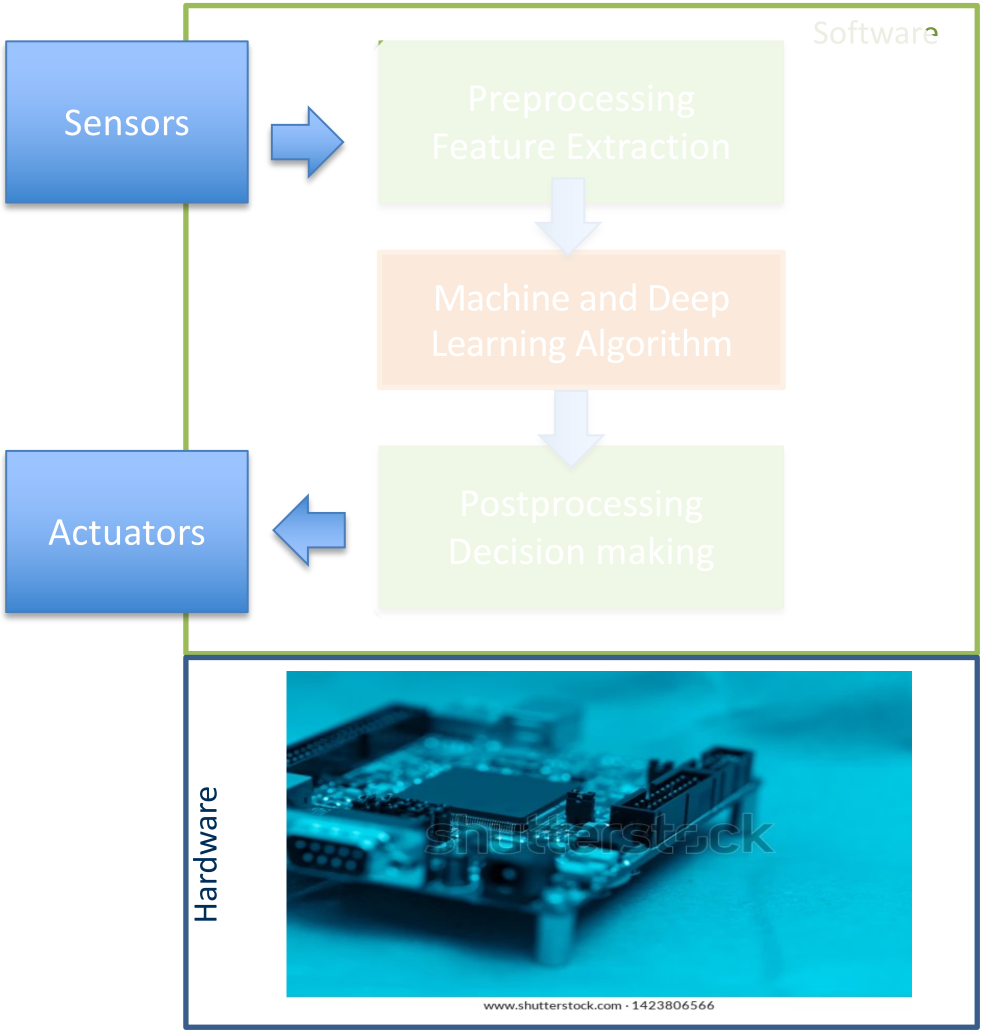
2) Hardware for Embedded and Edge AI (4h)

Wake-word detection

Person detection



Gesture recognition



- **Sensors and signals**: the TinyML perspective (ts, audio, image, video)
- **Sensors (+ application)**:
  - Acoustic and vibration
  - Visual and scene
  - Motion and position
  - Force and tactile
  - Optical and electromagnetic
  - Environmental and chemical

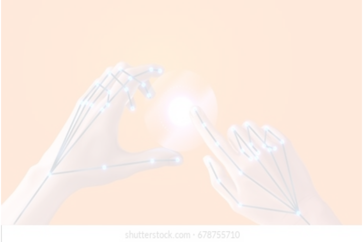
MPUs	**Low-end MCUs**
High-end MCUs	**SoCs**

Memory, computation, energy, cost

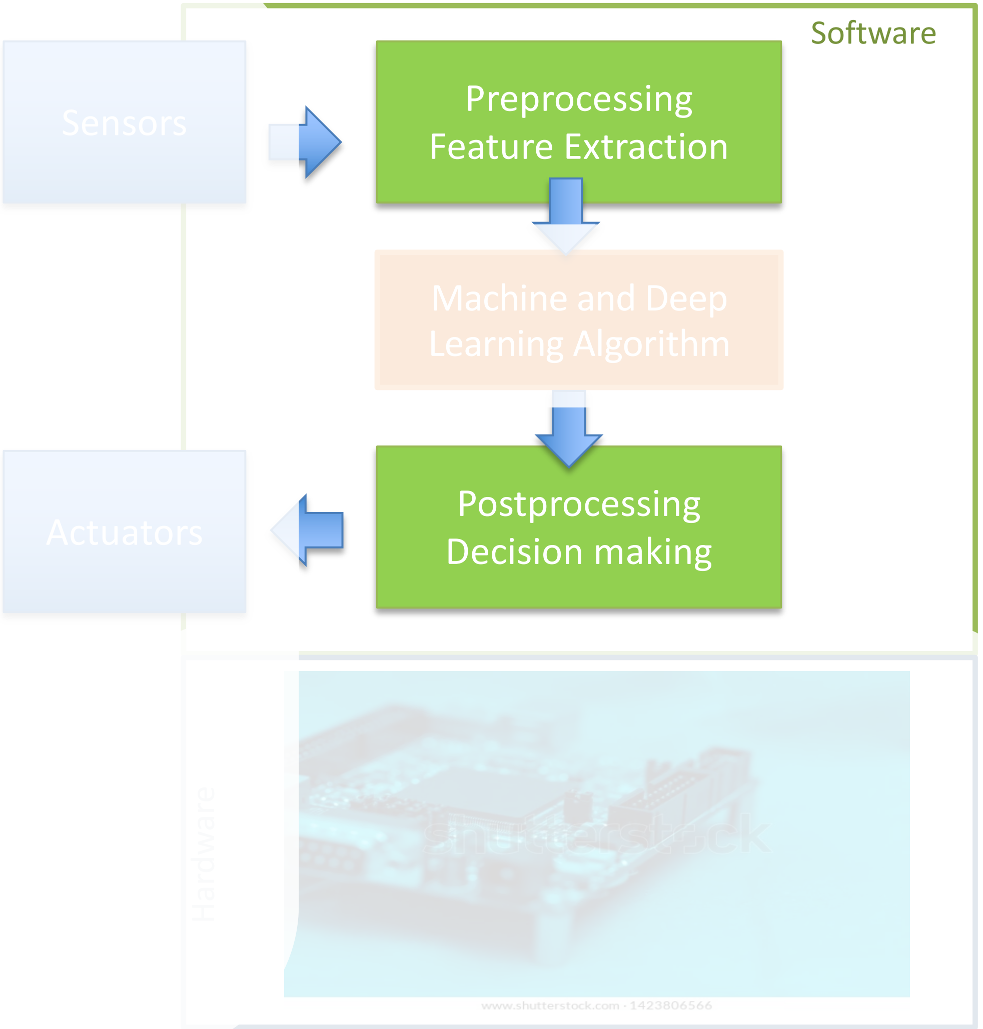
3) Algorithms for Embedded and Edge AI (4h)

Wake-word detection

Person detection



Gesture recognition


Software
Sensors
Preprocessing Feature Extraction
Machine and Deep Learning Algorithm
Postprocessing Decision making
Actuators
Hardware
www.shutterstock.com · 1423806566

Chopping, Windowing

Reconstruction of missing data

Resampling

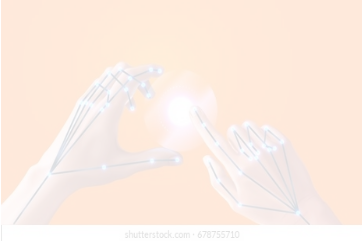
Filtering

Feature extraction

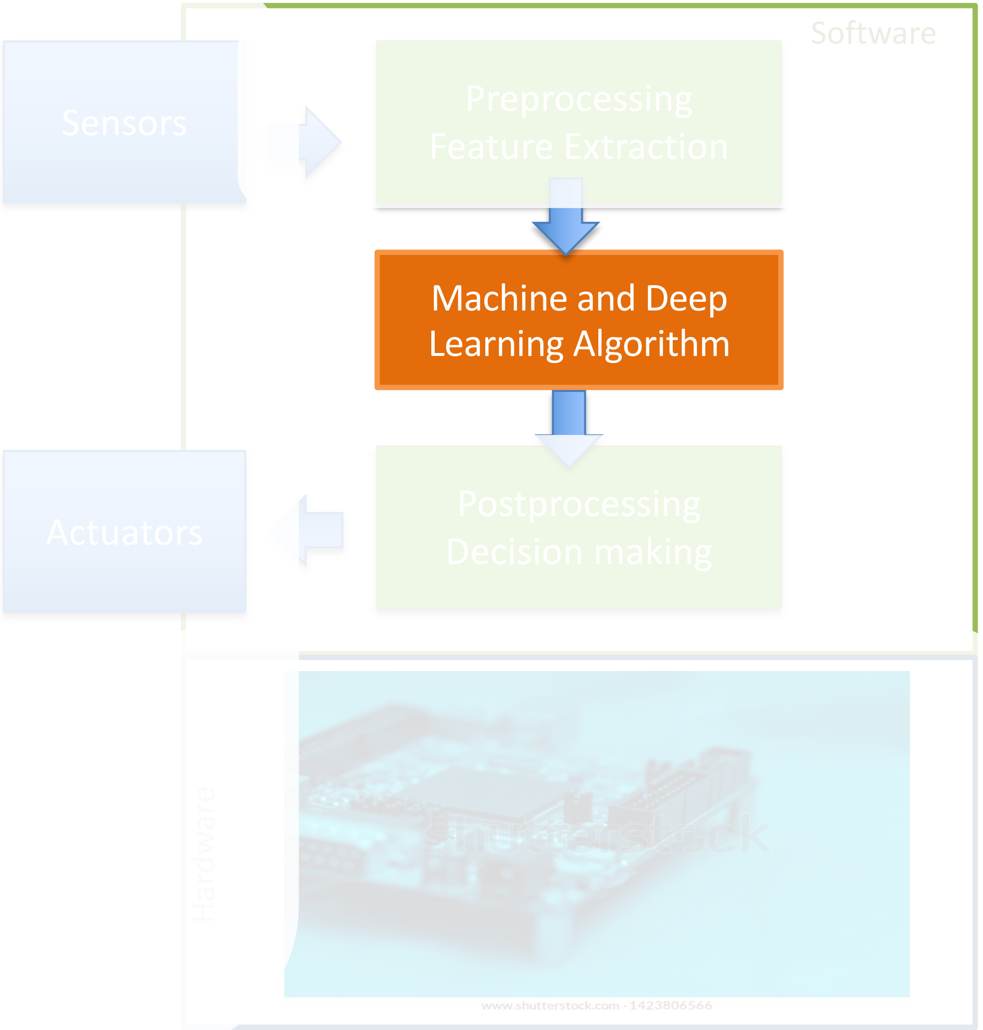
4) Machine Learning for Embedded and Edge AI (6h)

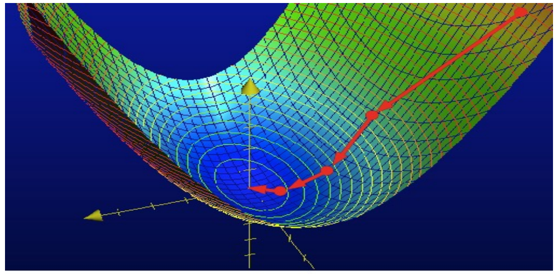
Wake-word detection

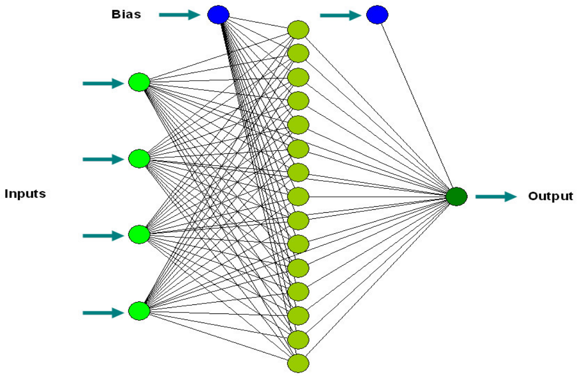
Person detection

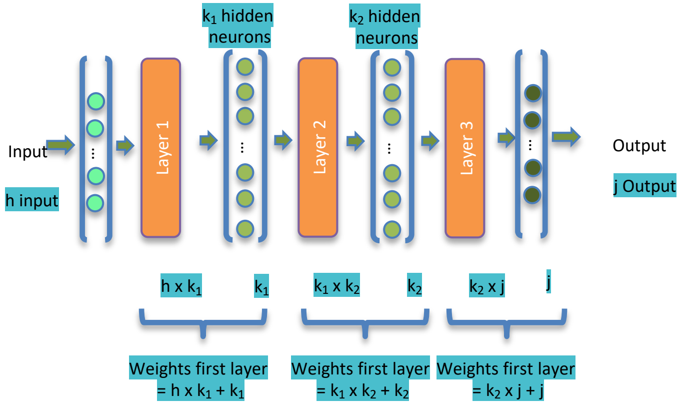


Gesture recognition



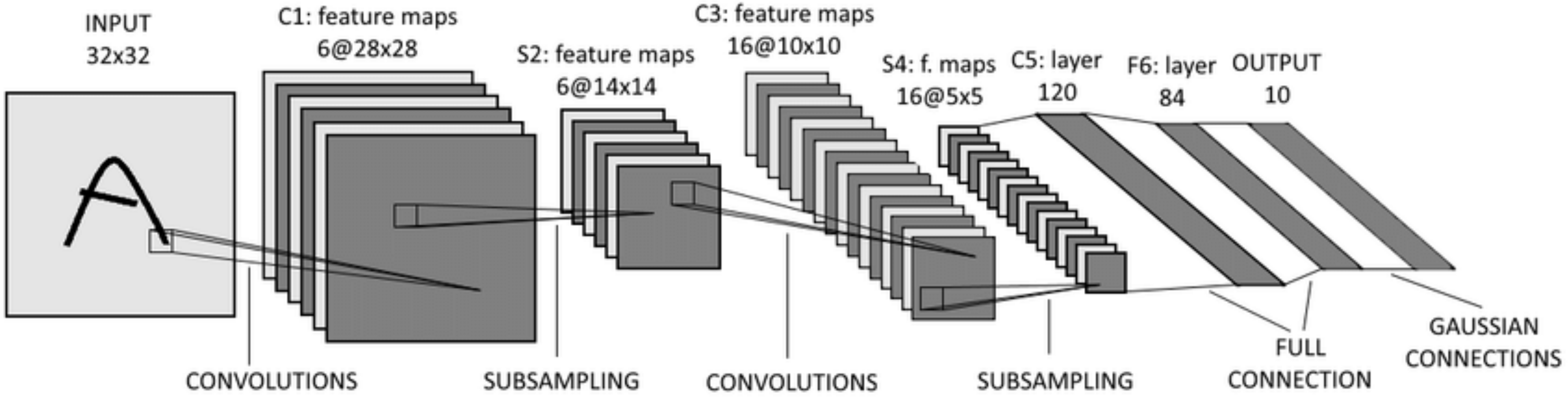






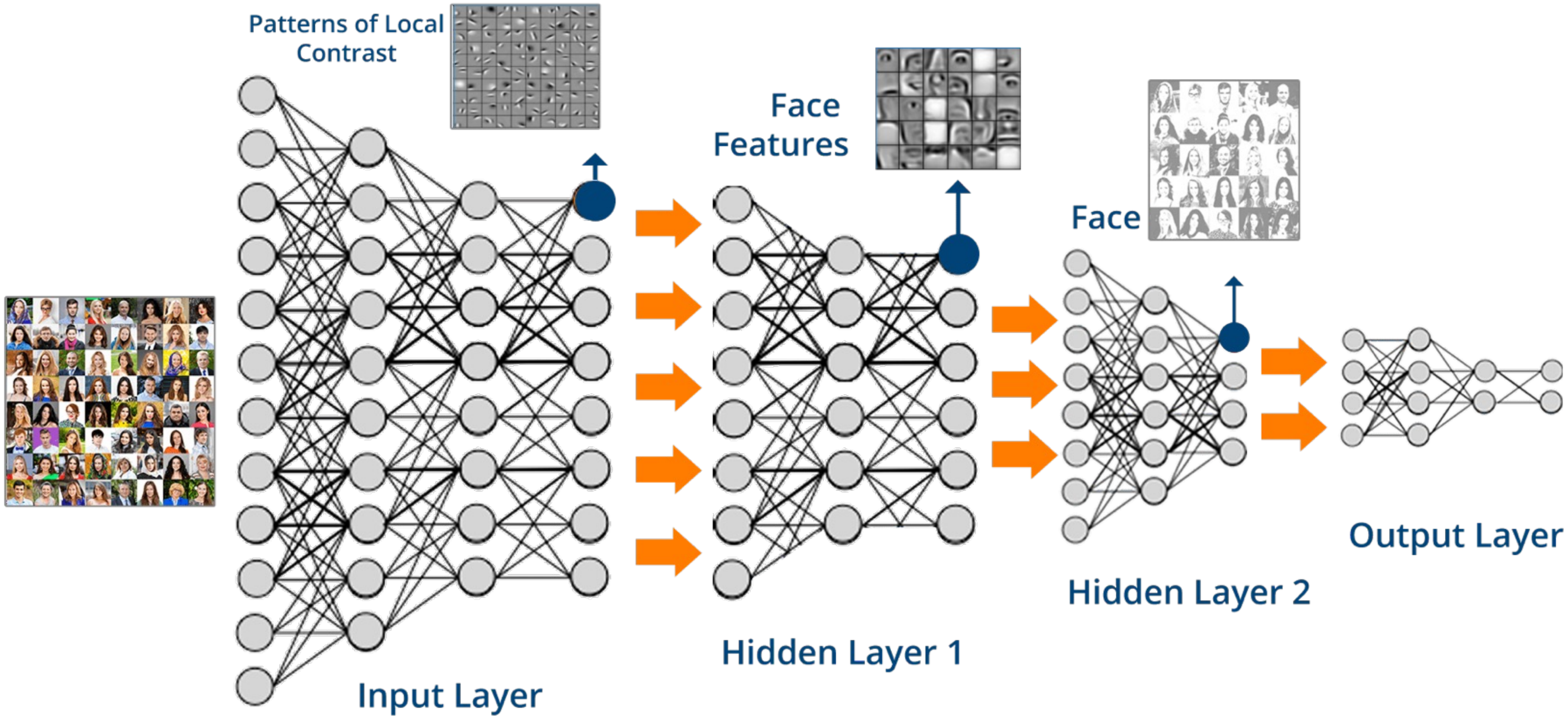
5) Deep Learning for Embedded and Edge AI (8h)

4KB | 19KB | 5KB | 6KB | 1.6KB | 0.5KB | 0.3KB | …KB =≈ 37KB



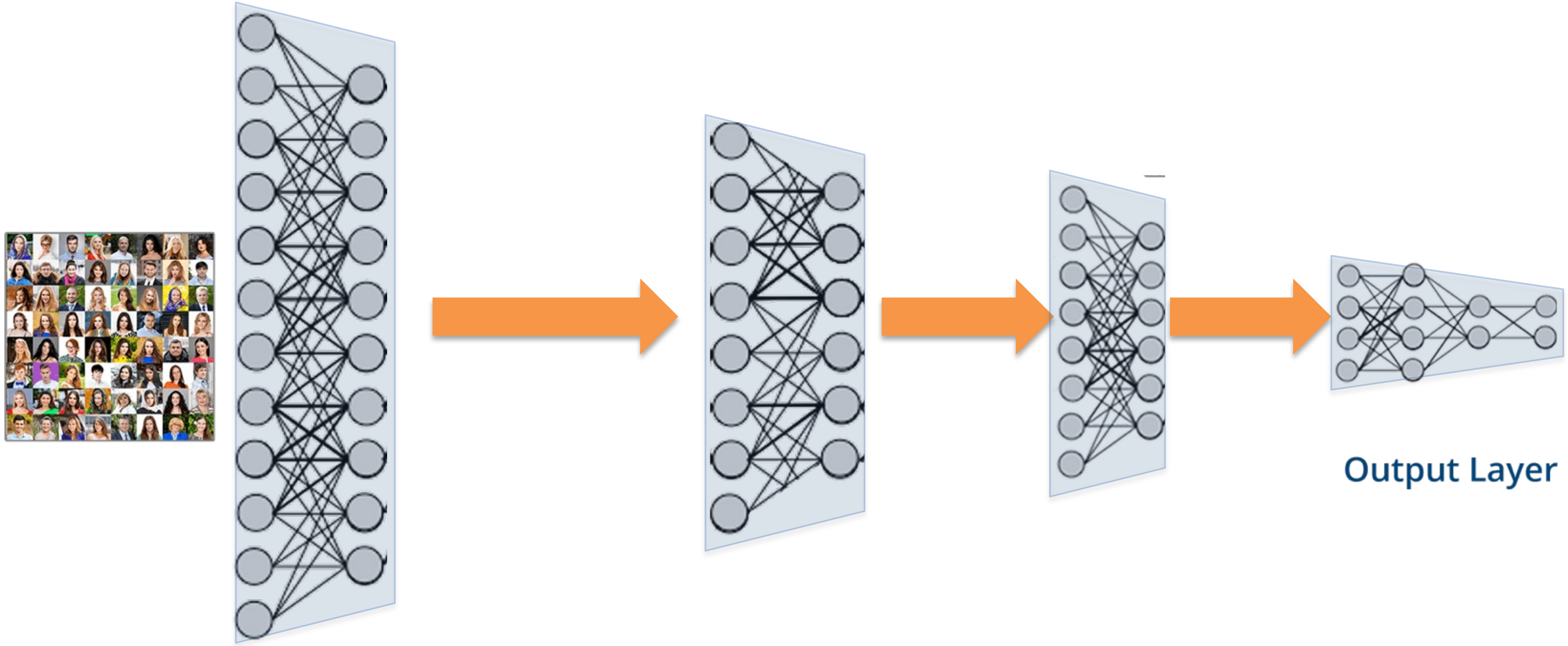
	6 5x5 Filters	16 5x5 Filters	FC1 400->120	FC2 120->84
#Weights	6x5x5x1+6 **=156**	16x5x5x6+16 **=2416**	400x120+120 **=48120**	120x84+84 **=10164**
#MAC	6x5x5x28x28 = **117K**	16x5x5x6x10x10 **=240K**	400x120 **=48K**	120x84 ≈ **10K**

5) Deep Learning for Embedded and Edge AI (8h)

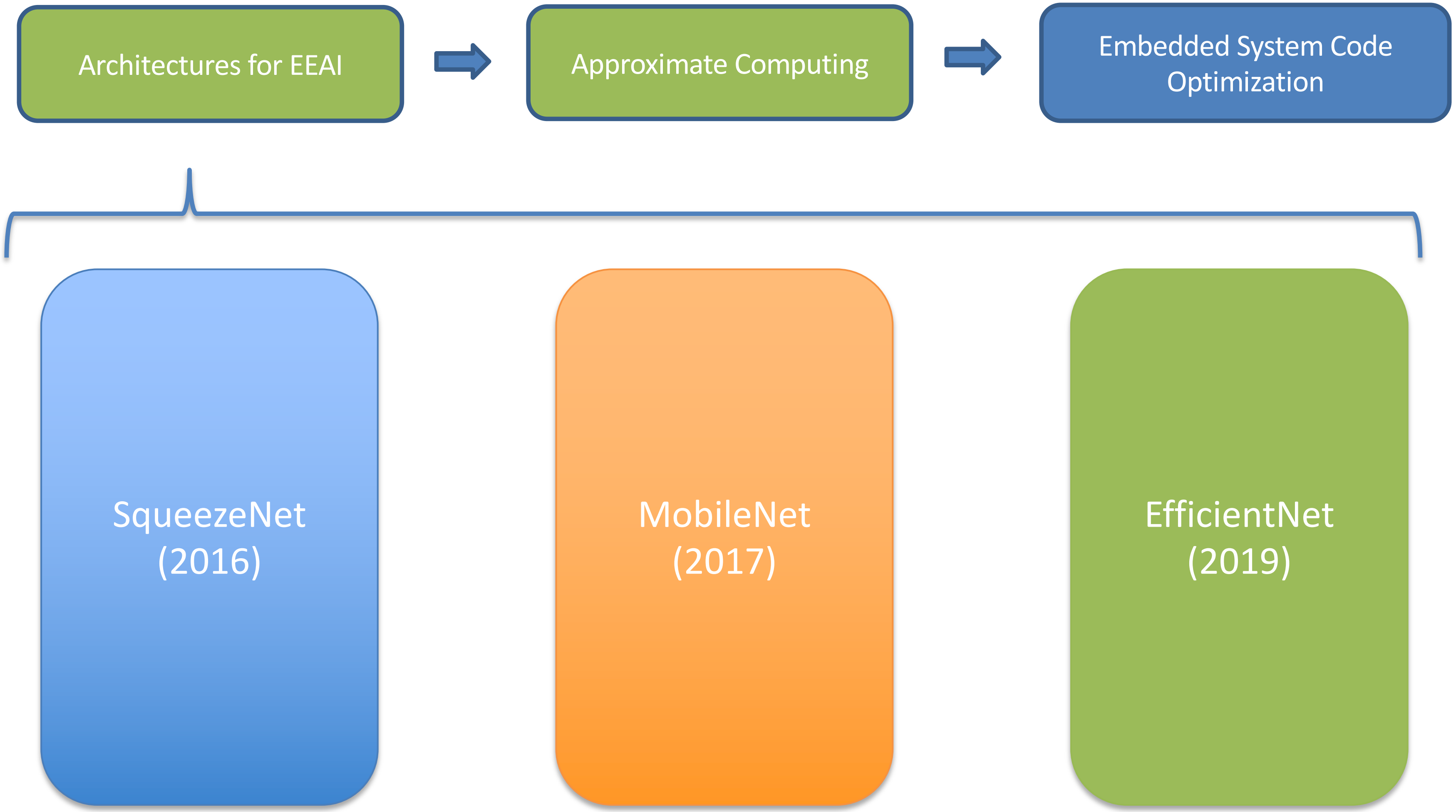


5) Deep Learning for Embedded and Edge AI (8h)

Architectures for EEAI → Approximate Computing → Embedded System Code Optimization



5) Deep Learning for Embedded and Edge AI (8h)



Architectures for EEAI → Approximate Computing → Embedded System Code Optimization

SqueezeNet (2016)

MobileNet (2017)

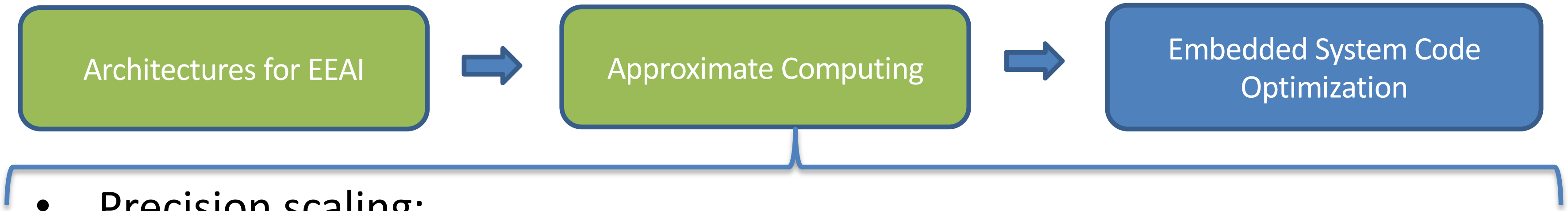
EfficientNet (2019)

Cit. Scholar 12/22: 7K

Cit. Scholar 12/22: 16K

Cit. Scholar 12/22: 9.5K

5) Deep Learning for Embedded and Edge AI (8h)

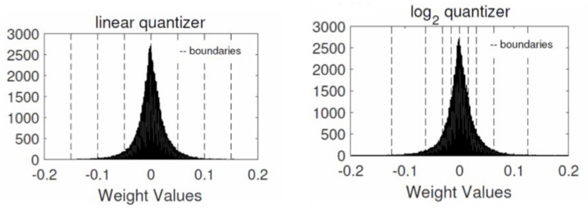


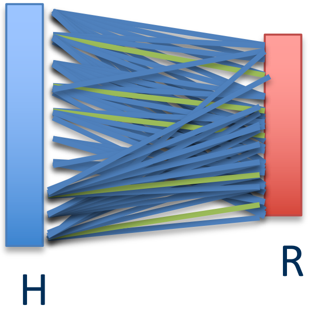
Architectures for EEAI

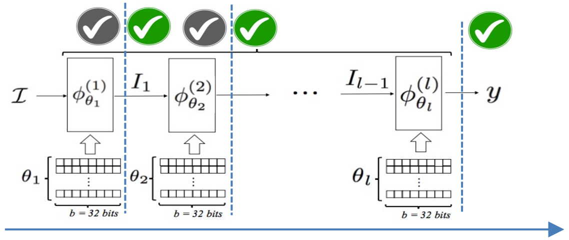
Approximate Computing

Embedded System Code Optimization

- Precision scaling:
  - ✓ **Quantization mechanisms**
  - ✓ **Implementation**
  - ✓ **Learning quantized models (PTQ, QAT)**
- Task dropping:
  - ✓ **network pruning**
  - ✓ **network architecture design**
  - ✓ **transfer learning**
  - ✓ **knowledge distillation**
- Early-exit Neural Networks:
  - ✓ **Architectures and EECs**
  - ✓ **Learning EENNs**

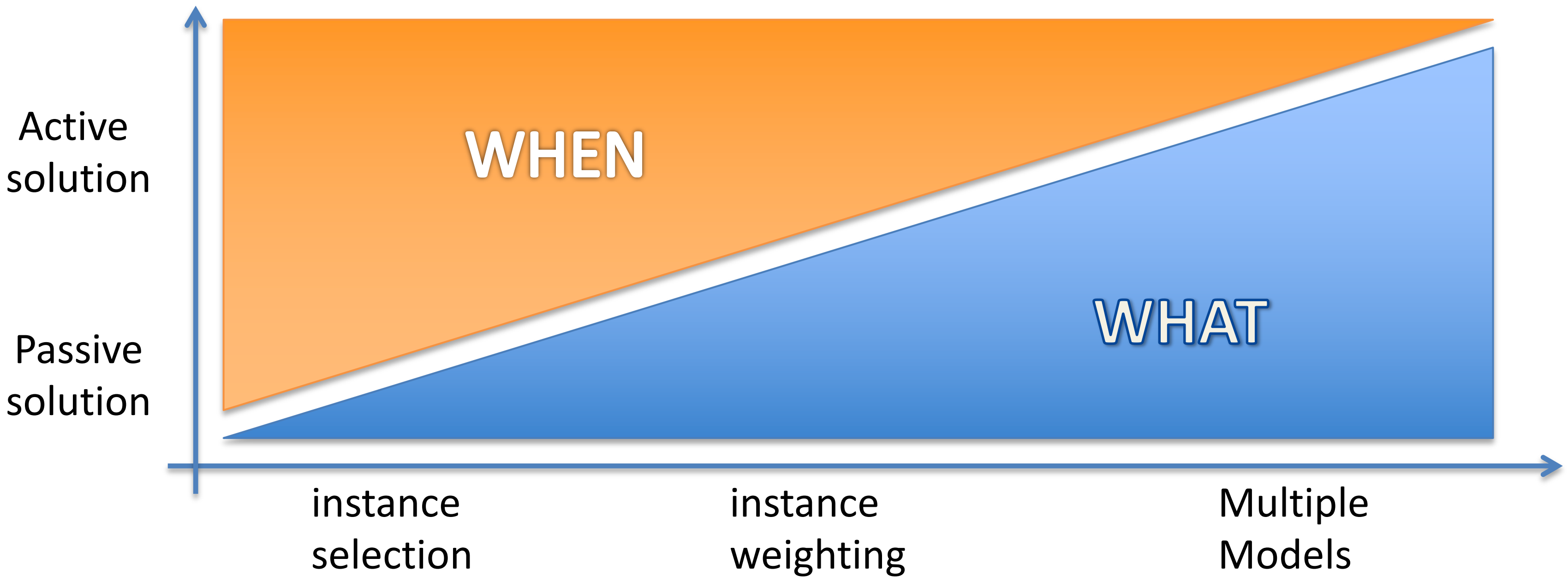




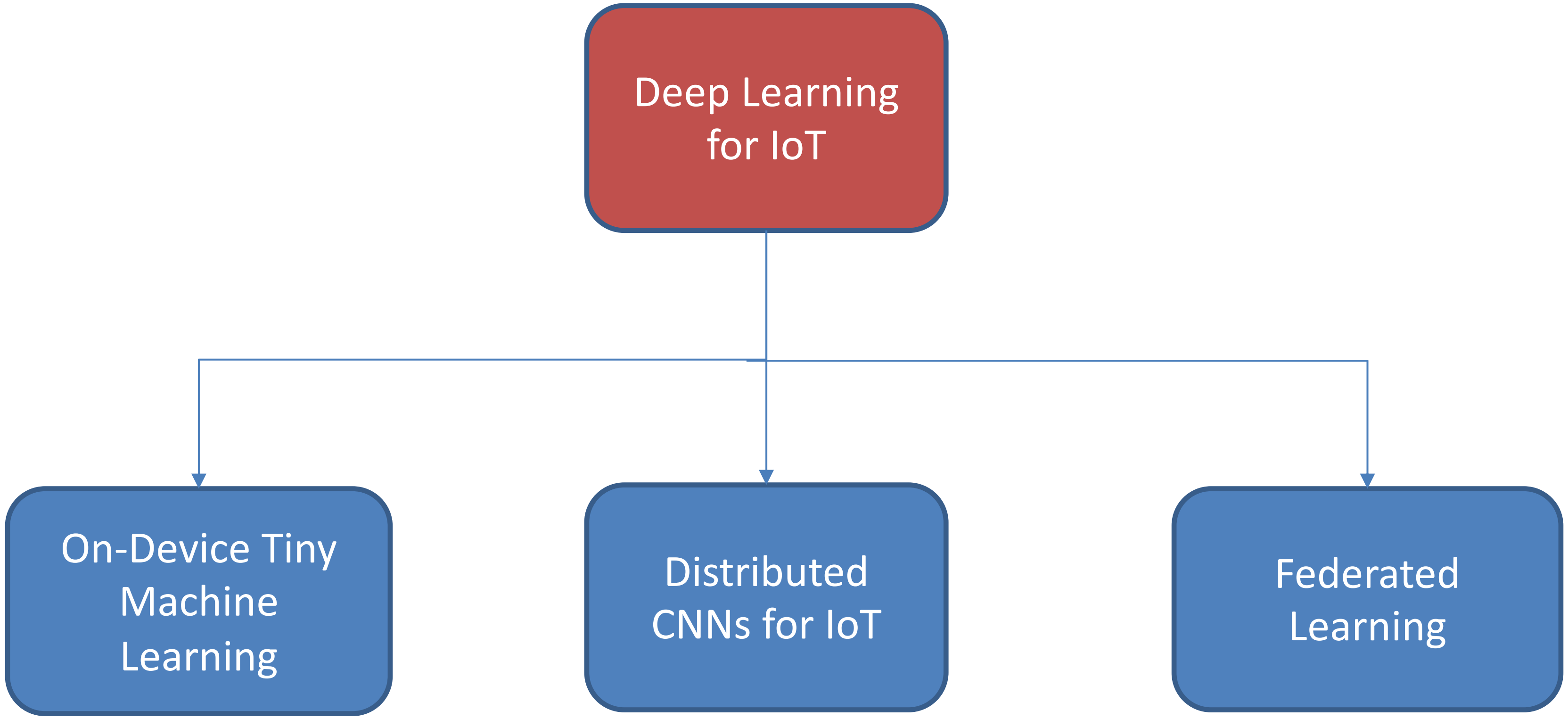


5) Deep Learning for Embedded and Edge AI (8h)

Adaptive mechanisms for Embedded and Edge AI



6) From the unit to the ecosystem perspective (4h)



Deep Learning for IoT

On-Device Tiny Machine Learning

Distributed CNNs for IoT

Federated Learning

7) The ethical perspective (2h)

The Economist

The Fed: more rate pain awaits

Alibaba and the 40 officials

Why dodgy money loves London

How to save America's Supreme Court

MAY 7TH–13TH 2022

The quantified self

How data can help you live longer

TQ

→ May 7th 2022

3 The new wearables

5 Measuring your health

7 Apps and data

8 Delivering treatment

11 Medical research transformed

The Economist
Technology Quarterly:
The quantified self

Measuring up

"Ethics of Design and Values: Solutions and Trade-offs in H-IoT and Beyond"
Prof. Viola Schiaffionati – Prof. Stefano Canali

The labs (20h)

1. Intro/review on Embedded Systems
2. Intro/review on Deep Learning with Tensorflow

Introduction and review of needed background

3. TF lite optimizations: quantization and pruning
4. Training Keyword Spotting – Microphone
5. Deploying Keyword Spotting – Microphone
6. Training and Deploying Visual Wake Word – Camera

Core TinyML Lectures

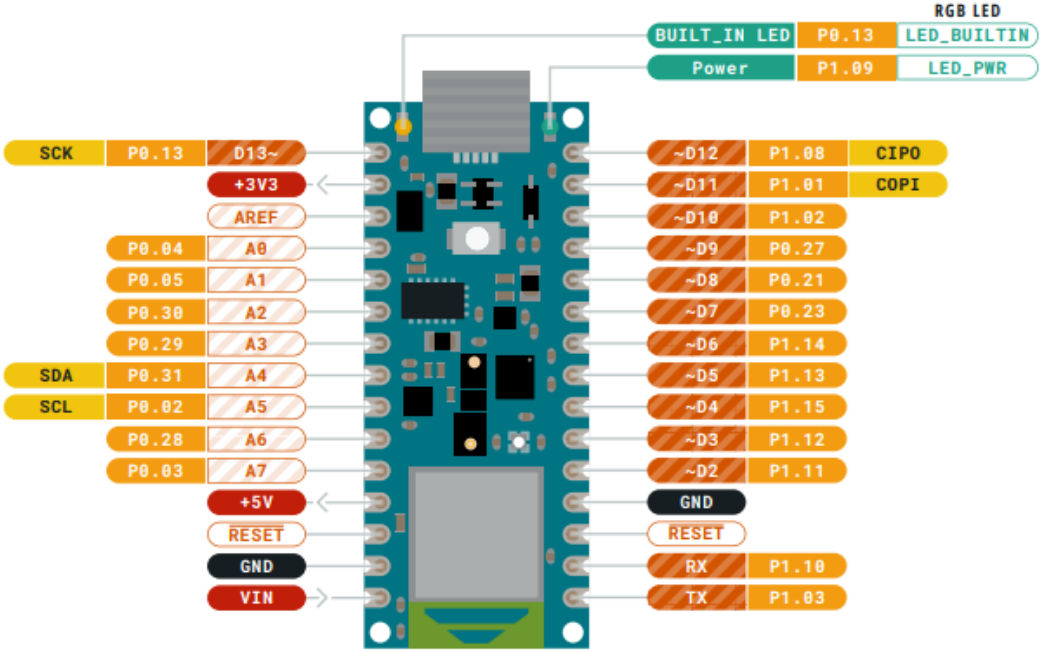
7. Data Collection and Engineering with Edge Impulse – Camera
8. Options for deployment: TFLM and Edge Impulse
9. Training and Deploying Anomaly detection – Accelerometer

Innovative Tools and topics for the future of TinyML

Tools employed: Google Colab, Edge Impulse, Arduino IDE, TFLM

TinyML kit - The Arduino Nano 33 BLE sense

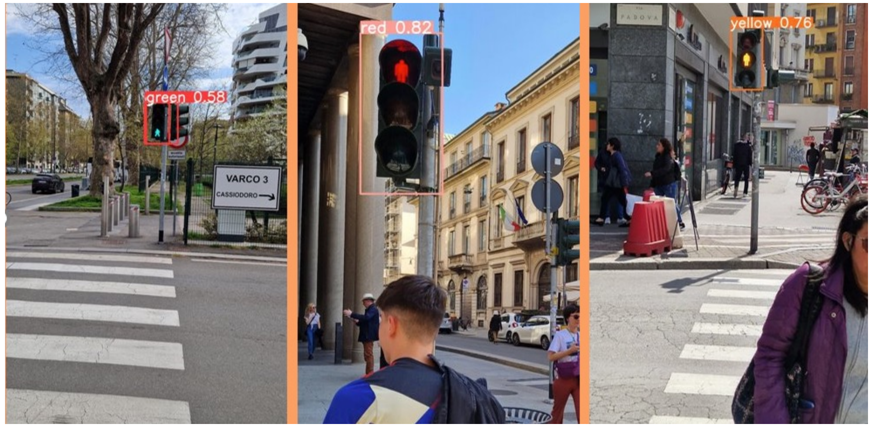
- ■ Ground
- 🟥 Power
- 🟩 LED
- ■ Internal Pin
- 🟫 SWD Pin
- ▨ Digital Pin
- ▧ Analog Pin
- □ Other Pin
- 🟧 Microcontroller's Port
- 🟨 Default

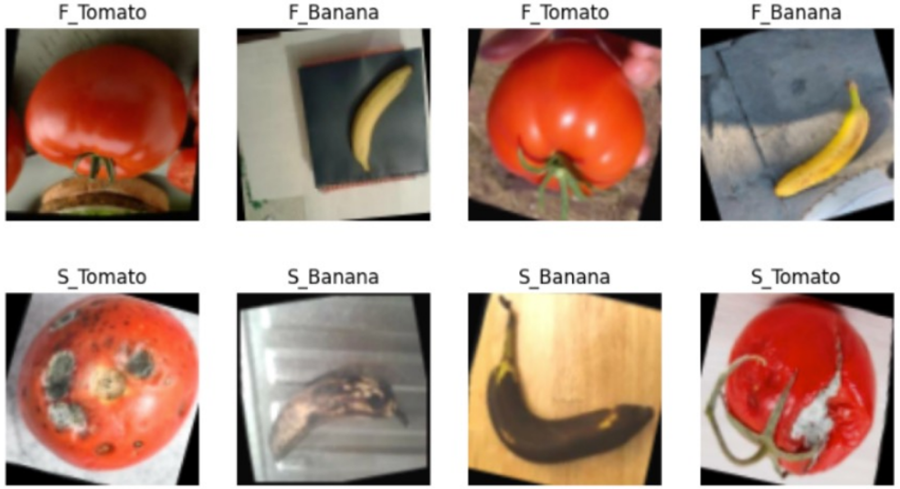


Exam

- The exam will consist in <u>two parts</u>:
  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)

Selected projects of the course





Challenges and opportunities

- Heterogeneity of the students backgrounds
- Fast evolution of the technology
- Keep the correct trade-off between ambition and implementability in the students' projects

- Strong connection between research activity and teaching
- The presence of a "physical" lab to carry out the projects
- Combining theory with implementation
- Strong technical aspects with ethical flavor