Pervasive AI

(Continual) Learning in Distributed and Heterogeneous Systems

Tutorial of the ECML-PKDD 2023 Conference
September 22, 2023 – Torino, Italy

Davide Bacciu, Antonio Carta, Patrizio Dazzi, Claudio Gallicchio
University of Pisa, Italy

Fundamentals of Pervasive AI

<u>Davide Bacciu</u>, Antonio Carta, Patrizio Dazzi, Claudio Gallicchio
University of Pisa, Italy

“Predicting and learning continually, into the wild”

Outline

- Scenario and motivation
- Tutorial Organization
- Learning efficiently with streaming data
- Applications & ongoing activities

Scenario & Motivation

An Hyper-connected world in the near future

A near future in which 5G/6G networks have been globally deployed, in which the day-to-day activities of society depends on billions of connected devices having low-latency connectivity.

Applications are accessed mostly from mobile devices that rely on a pervasive computing environment

Pervasive computing environment as an enabler

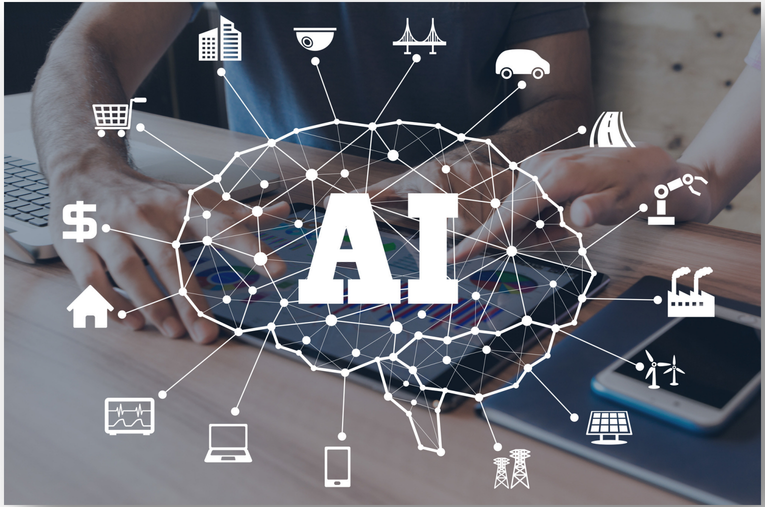
Reverse the paradigm

- FROM: move the data to the data center
- TO: move the data center to the data

Effects

- reduced latency
- larger bandwidth
- (potentially) more control on data

Artificial Intelligence as a transformative phenomenon



- Enabling the delivery of advanced, personalized and automated services
- Facilitating, simplifying and speeding-up human-machine interaction
- Applications with a cyber-physical nature

But also.. AI for networks, computing and data infrastructures

Pervasive Computing

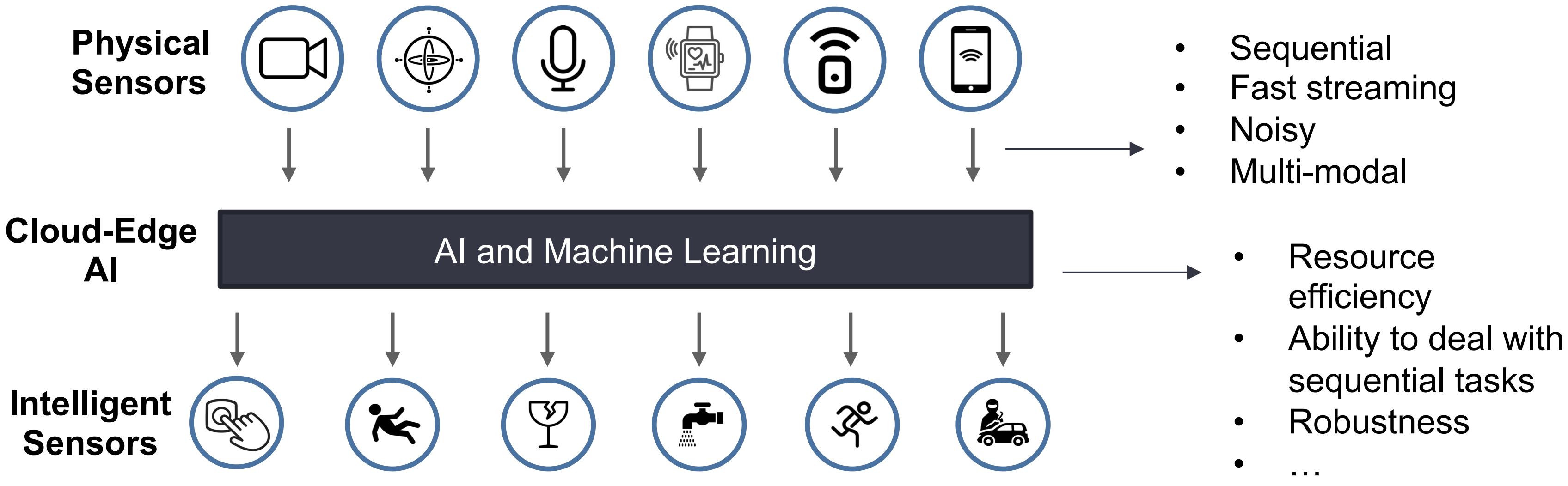
+

Artificial Intelligence

Calls for multi-disciplinary skills and putting different scientific communities in contact

A Prototypical Pervasive AI Application

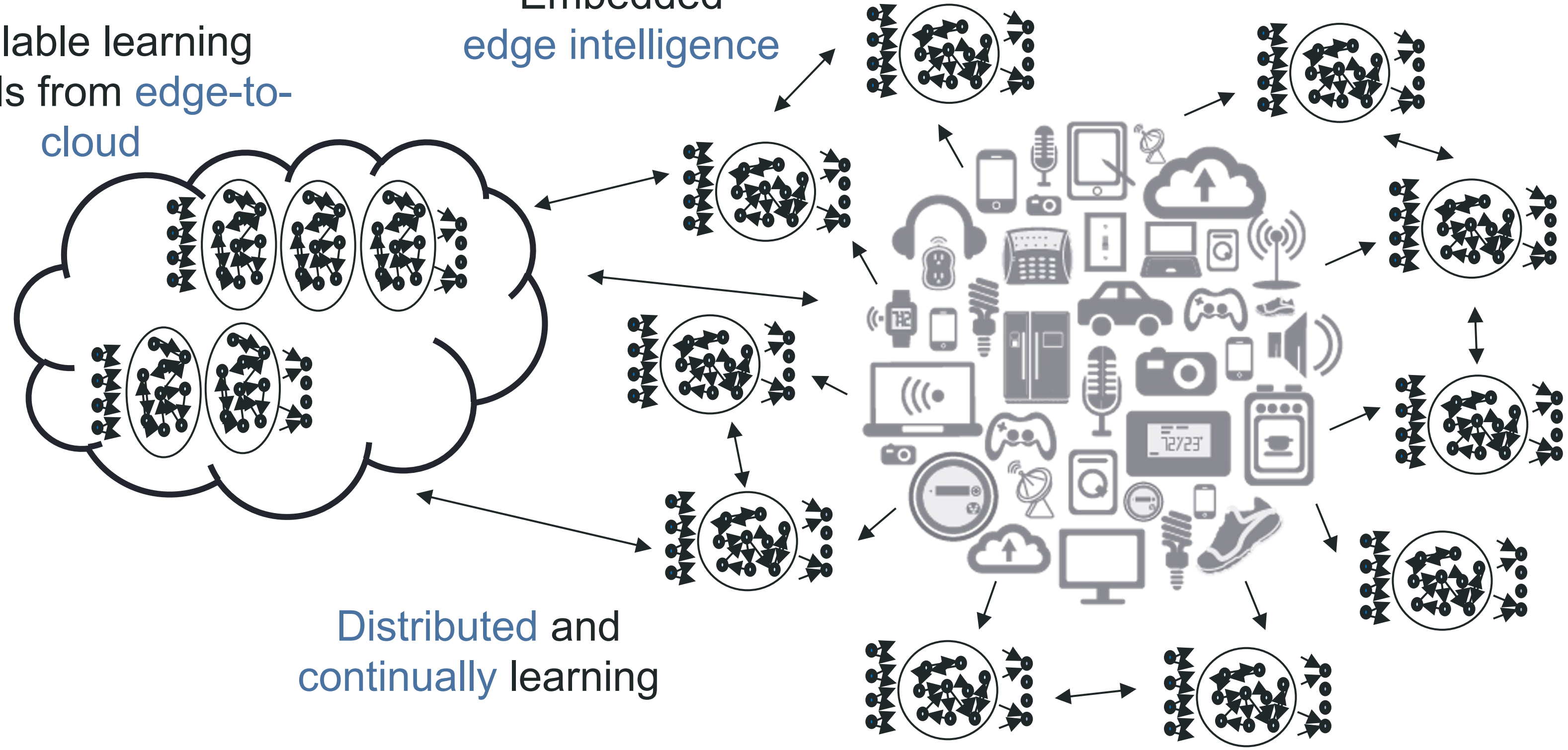
Systems integrating real and virtual AI-based sensors for monitoring of environments, vehicles, things, animals, people



The Pervasive AI Scenario

Scalable learning models from edge-to-cloud

Embedded edge intelligence



Distributed and continually learning

Pervasive AI – Recipe for success

Desiderata for the learning machinery

- Can handle timeseries (recurrent NNs)
- Efficient on fast-streaming data
- Allows embedding (**edge**)
  - Hardware friendly
- Can scale (**cloud**)
  - Communication friendly
- Simple to train continually
- Facilitate distributed training

Safe, dependable, secure, private and interpretable

Tutorial Organization (and Organizers)

Tutorial Outline (I)

- Module 1 – Fundamentals (Davide Bacciu)
    - Pervasive AI: scenario & motivations
    - Learning with streaming data
    - Reservoir computing fundamentals
- Module 2 – Distributed and federated learning solutions and infrastructures (Patrizio Dazzi)
    - Limits of ML on single-machine
    - Distributed, federated and fully-decentralized learning
    - Library time: FedRay

Tutorial Outline (II)

- Module 3 – Pervasive AI with randomized models (Claudio Gallicchio)
  - Advanced reservoir computing architectures
  - Training reservoir computing models: online, unsupervised, federated
  - Neuromorphic computing, learning beyond backpropagation
- Module 4 - Continual Learning (Antonio Carta)
  - Fundamentals of continual learning
  - Continual learning for embedded systems
  - Distributed continual learning
  - Library time: Avalanche

Tutorial Reference Page

http://pai.di.unipi.it/tutorial-on-continual-learning-in-distributed-and-heterogeneous-systems/

Tutorial Organizers

Davide Bacciu (davide.bacciu@unipi.it)

- Associate Professor: deep and generative learning, learning for graphs, continual learning, distributed and embedded learning
- Coordinator of 2 EU project on pervasive AI
- Chair on IEEE NN technical committee and VP of Italian Association for AI

Antonio Carta (antonio.carta@unipi.it)

- Assistant Professor: continual learning and recurrent NNs
- Lead maintainer of the Avalanche continual learning library

Tutorial Organizers

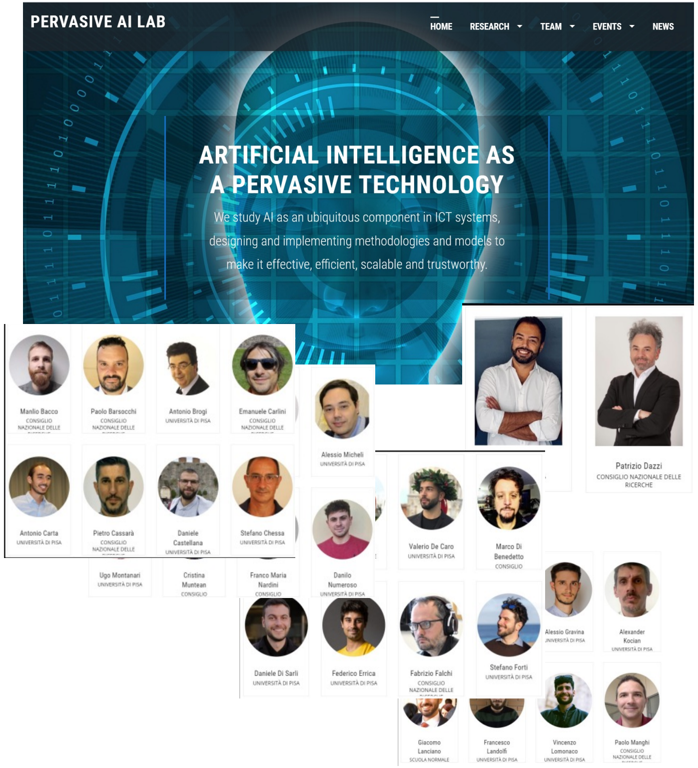
Patrizio Dazzi (patrizio.dazzi@unipi.it)

- Assistant Professor: cloud computing, edge computing, federated learning
- Coordinator of 2 EU project on intelligent placement and management of cloud applications

Claudio Gallicchio (claudio.gallicchio@unipi.it)

- Assistant Professor: reservoir computing, neuromorphic computing, learning for graphs, distributed and embedded neural networks
- Founder of the IEEE task forces on Reservoir Computing and on Randomization-Based Neural Networks and Learning Systems

PAILab – The Pervasive AI Laboratory @ Pisa, Italy



PAI LAB

Joint initiative by Dipartimento di Informatica @ UNIPI and Istituto Scienza e Tecnologia dell'Informazione @ CNR

(www) pai.di.unipi.it (email) pai-info@isti.cnr.it

Features

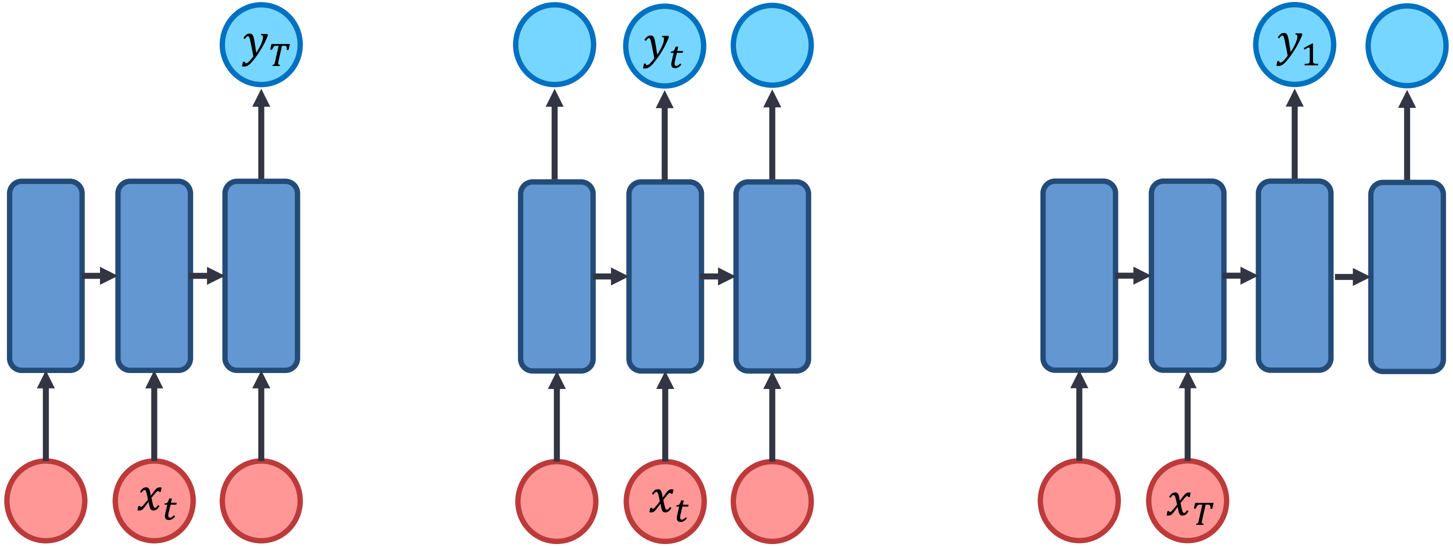
- ~50 members
- Coordinating 3 EU Projects and 1 KA; participation in 2 H2020 projects and 3 industrial projects
- >15M Euro secured grants

Focus

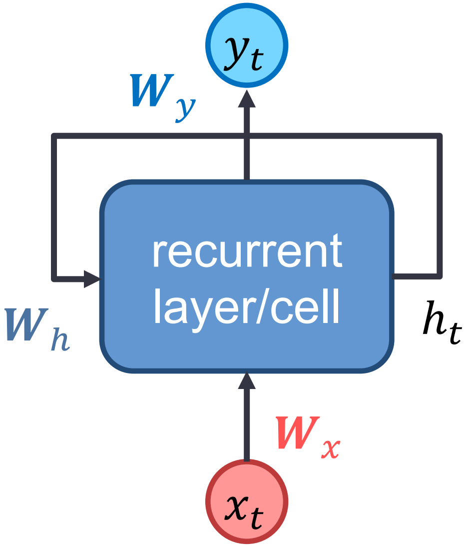
- AI as a ubiquitous component in ICT systems
- Design communication and computing systems to support pervasive AI

Learning (efficiently) with streaming data

Pervasive AI => Sequential Data Processing



Recurrent Neural Networks



State update:

input
previous state

$$\boldsymbol{h}(t) = tanh(\boldsymbol{x}(t)\,\boldsymbol{W}_x + \boldsymbol{h}(t-1)\boldsymbol{W}_h)$$

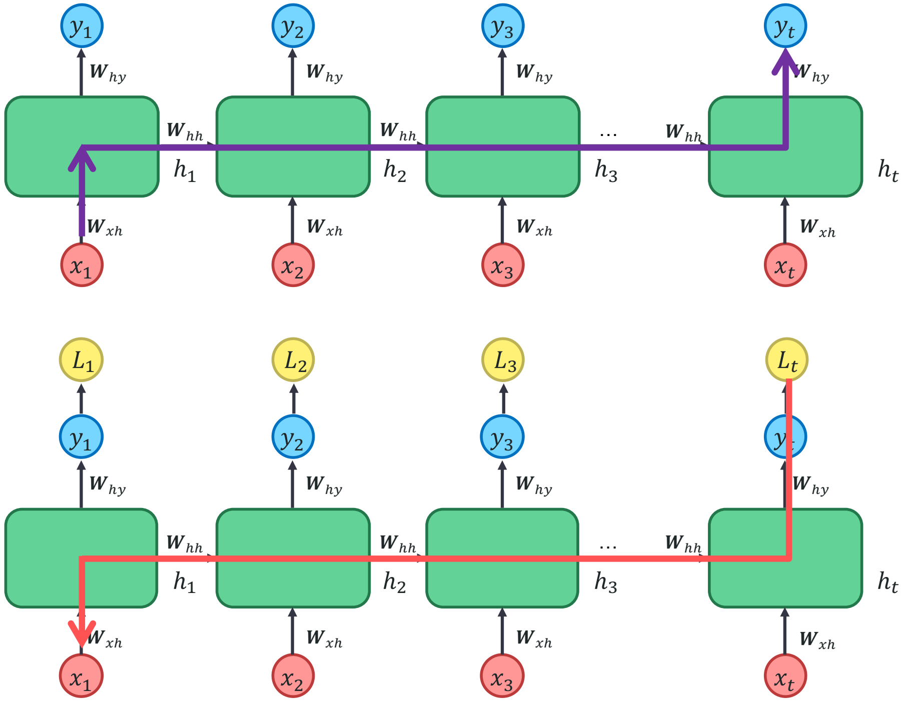
state
input weight matrix
recurrent weight matrix

Output function:

$$\boldsymbol{y}(t) = \boldsymbol{h}(t)\boldsymbol{W}_y$$

output
state
output weight matrix

Propagation Issues



Fading/Exploding memory

- the influence of inputs far in the past vanishes/explodes in the current state
- many (non-linear) transformations

Gradient Propagation

- gradient might vanish/explode through many non-linear transformations
- difficult to train on long-term dependencies

How to tackle propagation issues?

Gated Recurrent Networks

- LSTM/GRU models achieved tremendous success over the years
- This comes at very high cost in terms of
  - Time
  - Parameters
  - Complex training strategies
  - Articulated parameter coupling (due to feedback)

“Backpropagation through time as a way of learning sequences is especially implausible”

Hinton, Geoffrey. "The forward-forward algorithm: Some preliminary investigations.“ *arXiv preprint arXiv:2212.13345*(2022).

Deep Neural Networks

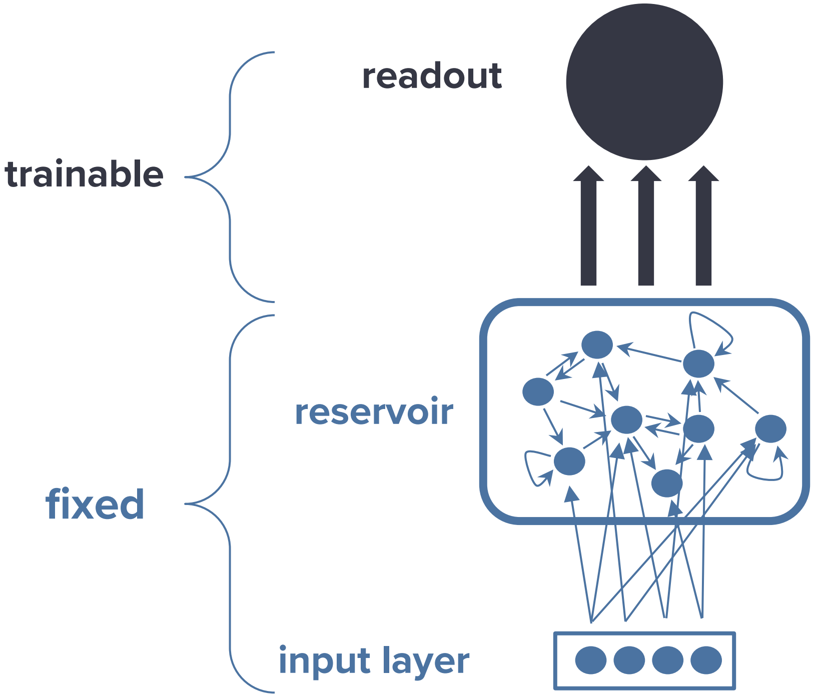
Powerful representations by applying multiple non-linear levels of transformation

Deep Learning = $\boxed{\text{Architectural Biases}}$ + Learning Algorithms

Randomization = Efficiency

- Training algorithms are cheaper and simpler
- Model transfer: do not need to transmit all the weights
- Amenable to efficient embedded and neuromorphic implementations

Reservoir Computing: focus on the dynamical system



randomized

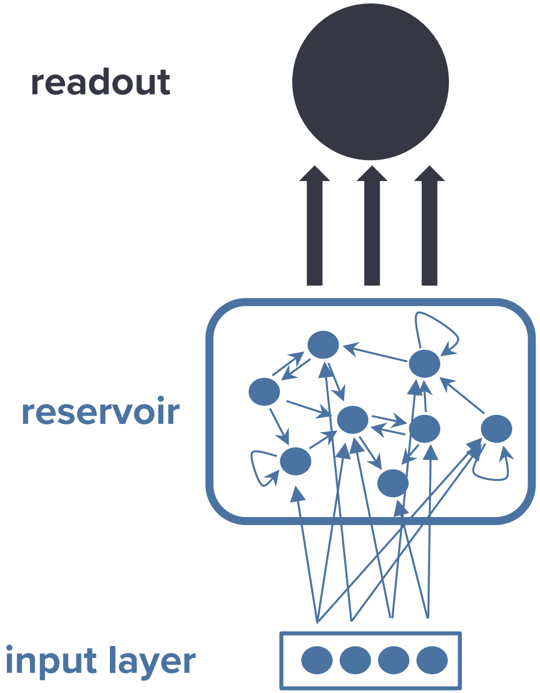
$$\boldsymbol{h}(t) = \tanh(\boldsymbol{x}(t)\underline{W_x} + \boldsymbol{h}(t-1)\underline{W_h})$$

Randomly initialized under stability conditions on the dynamical system

Stable dynamics => Echo State Property

Verstraeten, David, et al. *Neural networks* 20.3 (2007).
Lukoševičius, Mantas, and Herbert Jaeger. *Computer Science Review* 3.3 (2009).

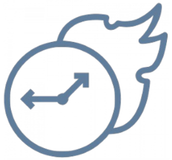
Echo State Networks (ESNs)



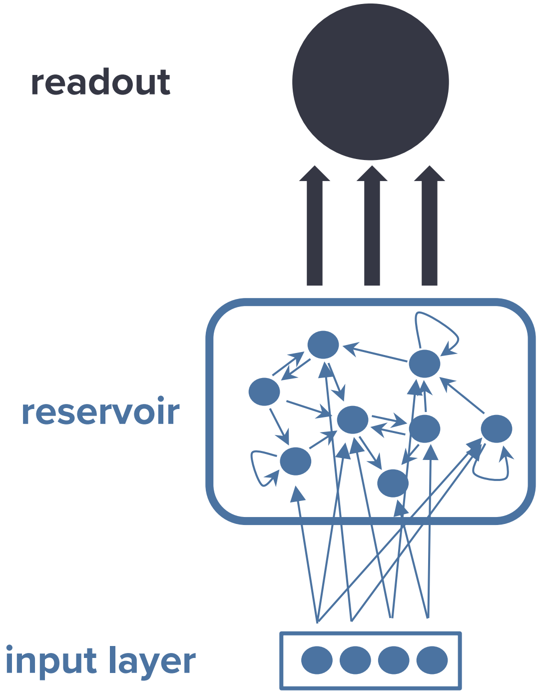
Reservoir

$$\boldsymbol{h}(t) = \tanh(\boldsymbol{x}(t)W_x + \boldsymbol{h}(t-1)W_h)$$

- large layer of recurrent units
- sparsely connected
- randomly initialized (ESP)
- untrained



Echo State Networks (ESNs)



Readout

$$\boldsymbol{y}(t) = \boldsymbol{h}(t)\boldsymbol{W}_y$$

- linear combination of the reservoir state variables
- can be trained in closed form

ESN Property – In Practice

Initialization of $\boldsymbol{W}_h$:

1. Generate a random matrix $\mathbf{W_r}$, whose elements are drawn e.g. from a uniform distribution on [-1,1]
2. Scale by the desired spectral radius

$$\boldsymbol{W}_h \leftarrow \mathbf{W_r} \frac{\rho_{desired}}{\rho(\mathbf{W_r})}$$

- Note that now $\rho(\boldsymbol{W}_h) = \rho_{desired}$ (choose a value < 1)
- The spectral radius is a key hyper-parameter of the reservoir

Training the readout (offline)

$$\mathbf{H} = \begin{bmatrix} | & & | \\ \mathbf{h}(1) & \dots & \mathbf{h}(\mathrm{T}) \\ | & & | \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} | & & | \\ \mathbf{d}(1) & \dots & \mathbf{d}(\mathrm{T}) \\ | & & | \end{bmatrix}$$

- Closed form solution of the least squares problem by direct methods

$$\min_{\mathbf{W_y}} \left\| \mathbf{W_y}\, \mathbf{H} - \mathbf{D} \right\|_2^2$$

- Moore-Penrose pseudo-inversion

$$\mathbf{W_y} = \mathbf{D}\, \mathbf{H}^{+} = \mathbf{D}\, \mathbf{H}^{T} (\mathbf{H}\mathbf{H}^{T})^{-1}$$

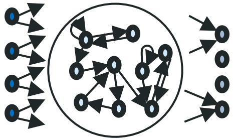
- Ridge-regression

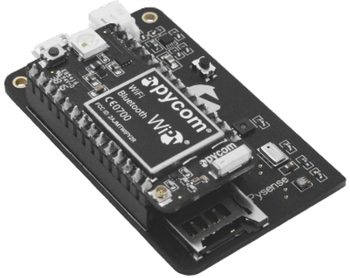
$$\mathbf{W_y} = \mathbf{D}\, \mathbf{H}^{T} (\mathbf{H}\mathbf{H}^{T} + \lambda\, \mathbf{I})^{-1}$$

Later on, you will see more opportunities offered by the reservoir-type architecture in a pervasive context

Applications & Ongoing Activities

Intelligent Sensors



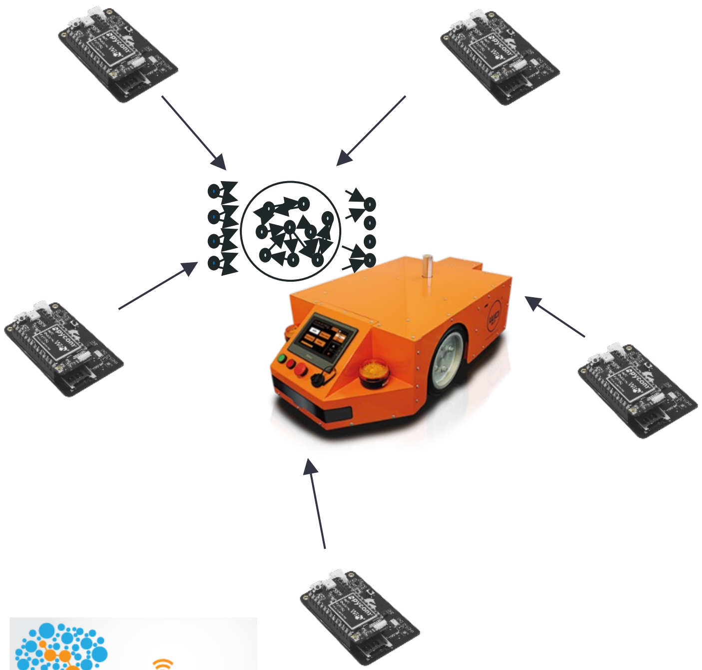


Neural learning in 8Kb of memory + deployment over-the-air

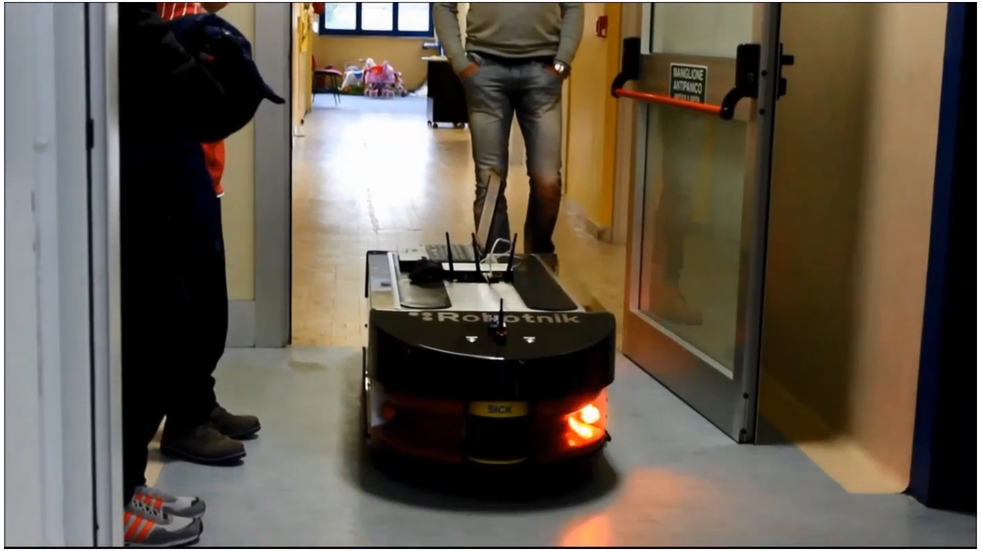
Automating diagnosis (from 30mins to 10sec)

Transform device function through intelligence

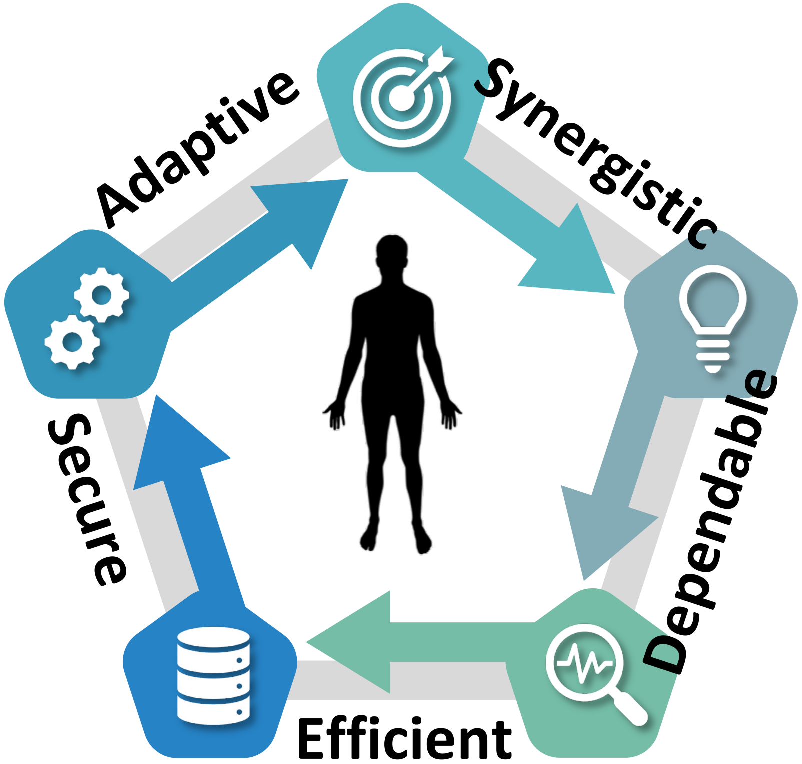
Hybrid Sensor-Robotic Systems



Localizing a trolley robot in hospitals by wireless sensors



Human-in-the-loop in Pervasive AI



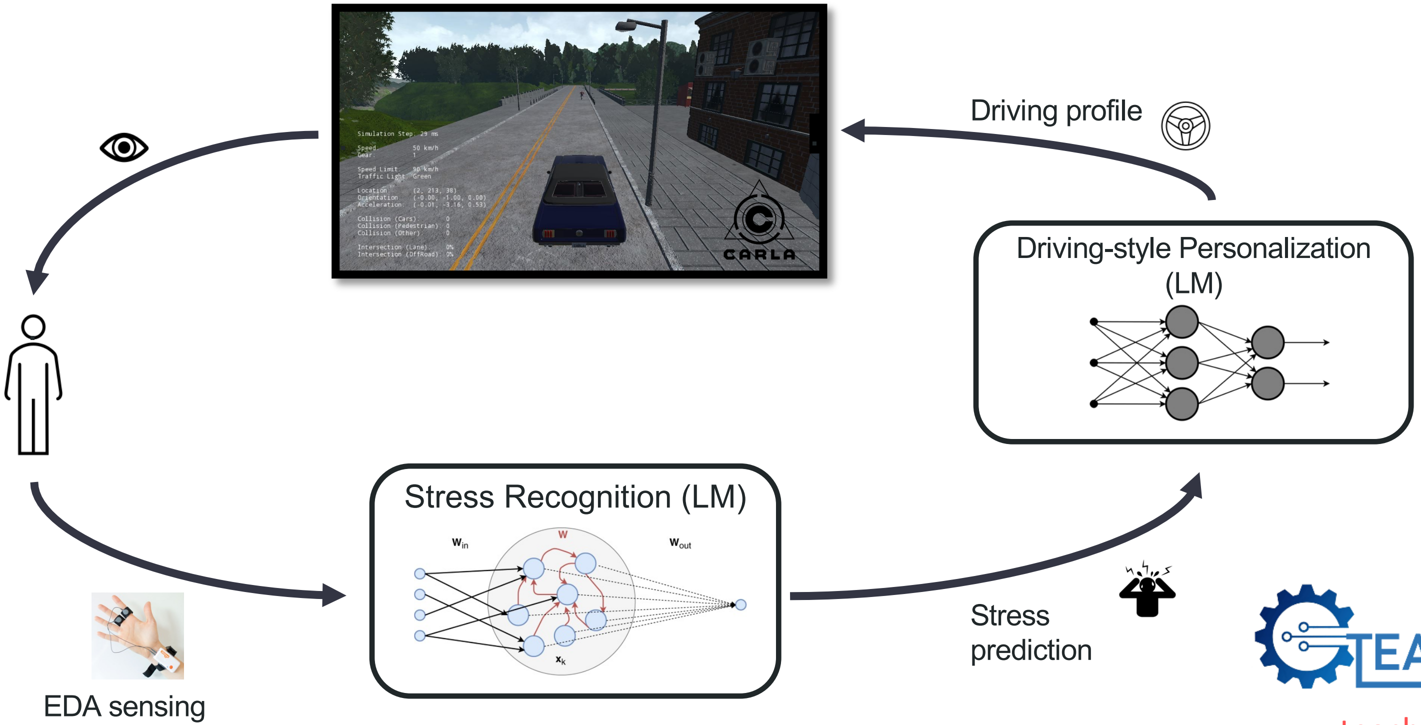
A human-centric perspective on autonomous CPSoS applications

Paradigmatic shift needing support at computing and system level

TEACHING

www.teaching-h2020.eu

CPSoS Applications with human-in-the-loop



www.teaching-h2020.eu

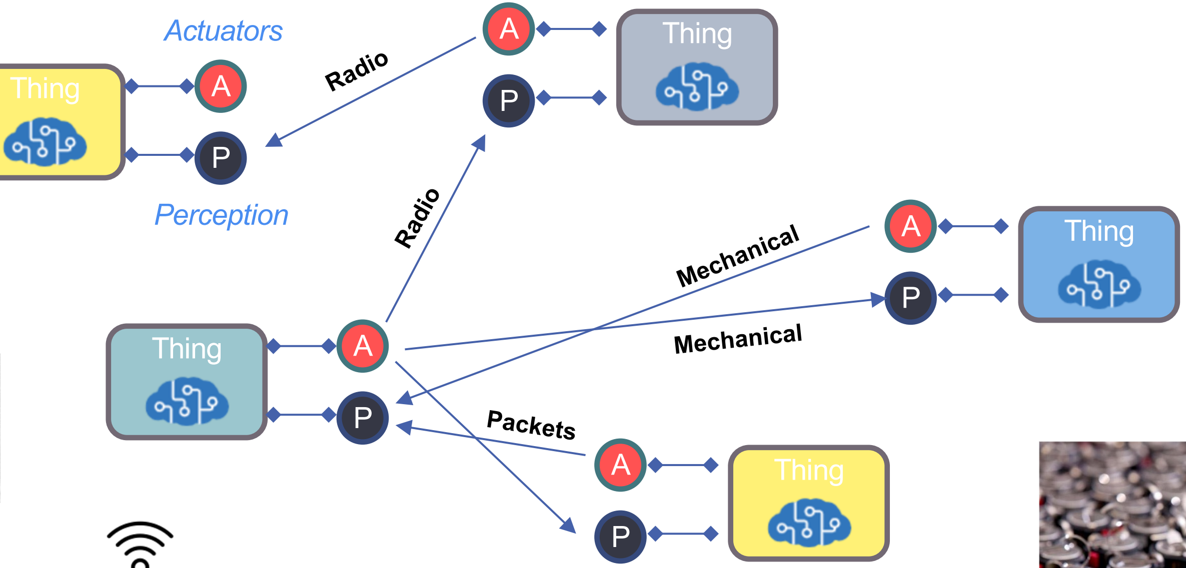
Facilitating Human-Centric Distributed AI

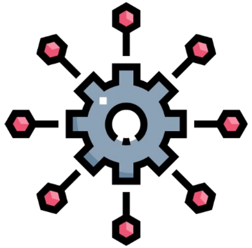
Primitives to mine human-reactions and leverage them for continual application adaptation

TEACHING

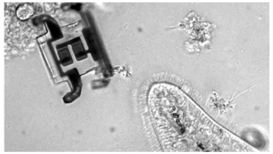
www.teaching-h2020.eu

EIC-EMERGE – A World of Simple Interacting Elements





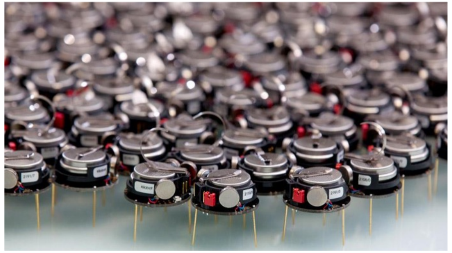
microservices



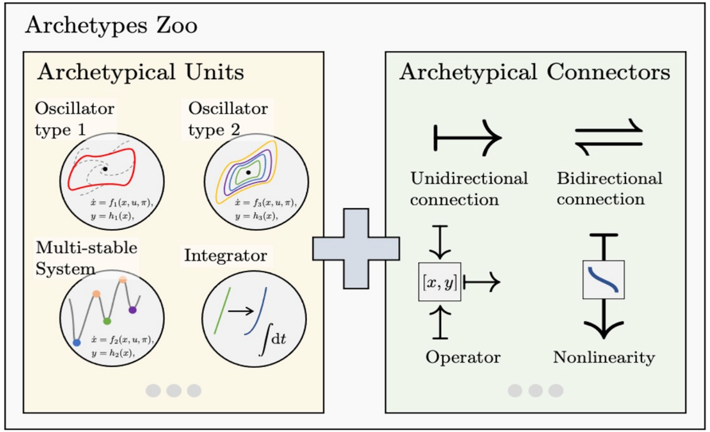
nanodevices

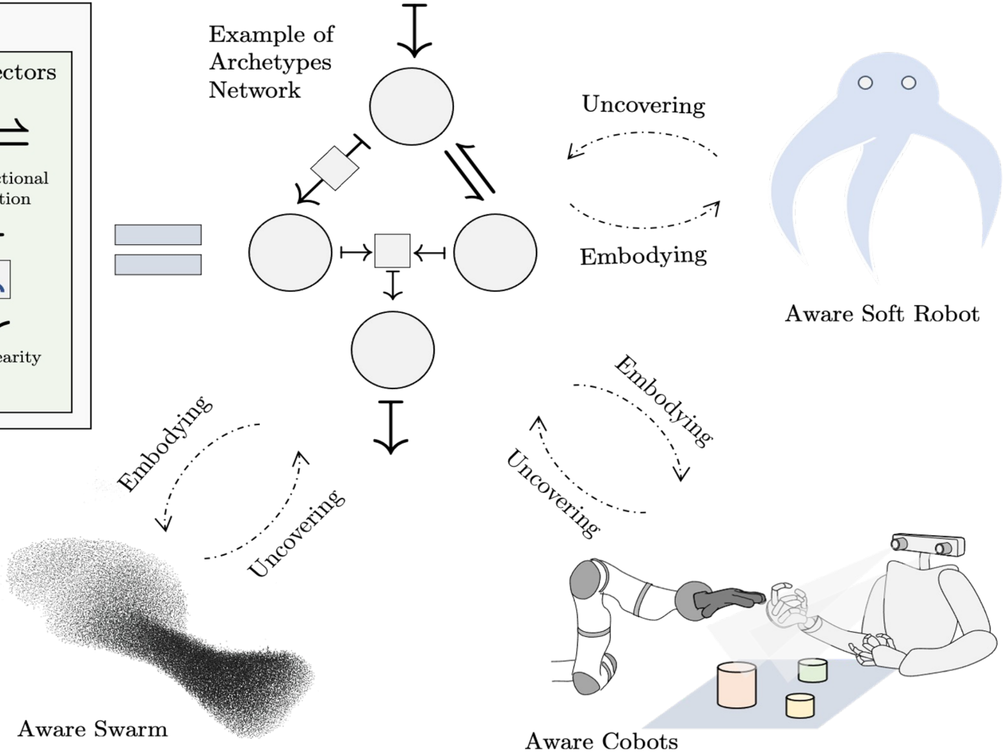
IoT

robots



(Neural & Physical) Computing with Dynamical Systems





Archetype Computing System engine to run dynamical systems enriched with **lifelong and evolutionary learning**

EMERGE Emergent awareness from minimal collectives

Università di Pisa

TU Delft

University of Bristol

LMU Ludwig-Maximilians-Universität München

Da Vinci Labs

Funded by the European Union

Funded by the European Union under Grant Agreement 101070918

eic-emerge.eu

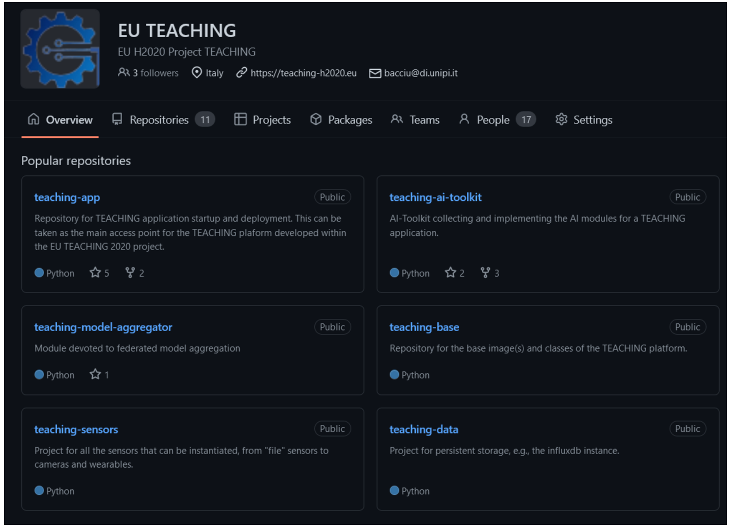
@eic_emerge

company/eic-emerge

@eic_emerge

TEACHING Toolkit for Pervasive AI

- Facilitates development and execution of distributed AI apps
- Micro-service based architecture
- Data stream acceleration (multi-core, GPU, FPGA)
- Ready-made implementation of reservoir computing models for distribution over a network of devices (cloud-edge)



https://github.com/EU-TEACHING/

FedRay - An R&D-oriented framework for Federated Learning

High-level programming support for

- Building federated learning nodes
- Seamless topology-aware communication among nodes
- Instantiating federated learning processes with arbitrary topologies

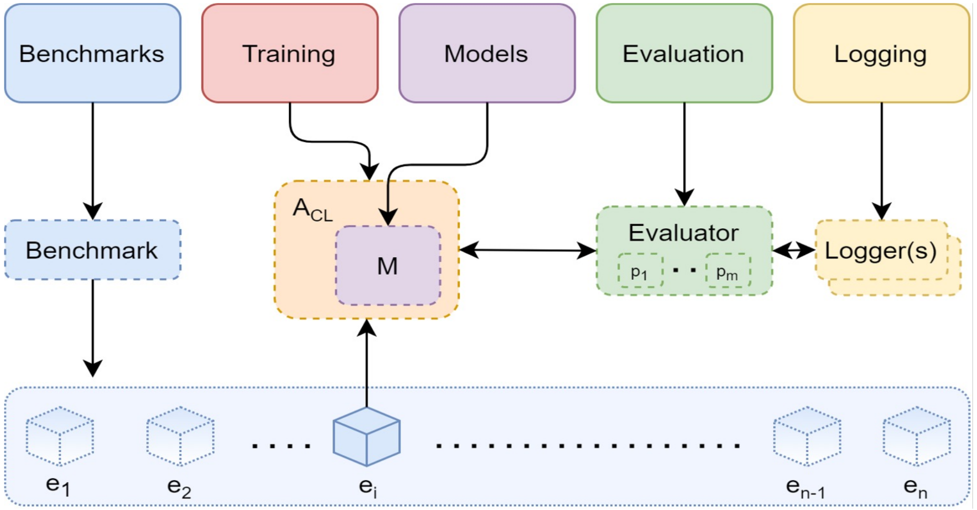
- Implementations of federated reservoir computing algorithms
- Leverages Ray to ease multiprocessing and scalability on clusters

```python
@fedray.remote
class IncFedClient(FedRayNode):

    def build(self, dataset, seq_length: int, batch_size: int, **kwargs) -> None:
        self.device = 'cuda'
        self.dataset = dataset
        self.loader = DataLoader(dataset)

    def run(self):
        reservoir: Reservoir = self.receive().body['model']
        A, B = compute_ridge_matrices(self.loader, reservoir)
        self.send('ridge_matrices', {'A': A, 'B': B})
        readout = solve_ab_decomposition(A, B)
```

https://github.com/vdecaro/fedray/

Avalanche: an End-to-End Library for Continual Learning

- Developed and supported by ContinualAI
- >1.1k stars on GitHub
- Tutorial & lectures
- Ready-made strategies and your own recipes




```
strategy = Replay(model, optimizer,
                  criterion, mem_size)
for train_exp in scenario.train_stream:
    strategy.train(train_exp)
    strategy.eval(scenario.test_stream)
```

https://avalanche.continualai.org

Conclusions

Conclusions

- Pervasive computing as a blue-print for smart-X
- Artificial intelligence pervasively permeating ICT applications
- Pervasive AI needs a convergence of communities
  - Codesign of AI methods and Computing system
- Focus on social and environmental sustainability

WHICH DIRECTIONS IN AI?

- RESERVOIR COMPUTING AND LEARNING WITH ADAPTIVE DYNAMICAL SYSTEMS
- NEUROMORPHIC HARDWARE
- CONTINUAL/LIFELONG LEARNING
- LEARNING FROM EXPERIENCES RATHER THAN FROM DATA ALONE
- SAFE AND DEPENDABLE AI

Acknowledgements


Antonio Carta


Andrea Cossu


Patrizio Dazzi


Valerio De Caro


Vincenzo Lomonaco


Claudio Gallicchio

TEACHING EMERGE

EUROPEAN UNION

Supported by the European Union under GAs n. 871385 & 101070918

Università di Pisa

PAI LAB