Intelligenza Artificiale dall'Edge all'Exascale: lo Spoke 7 del Progetto FAIR

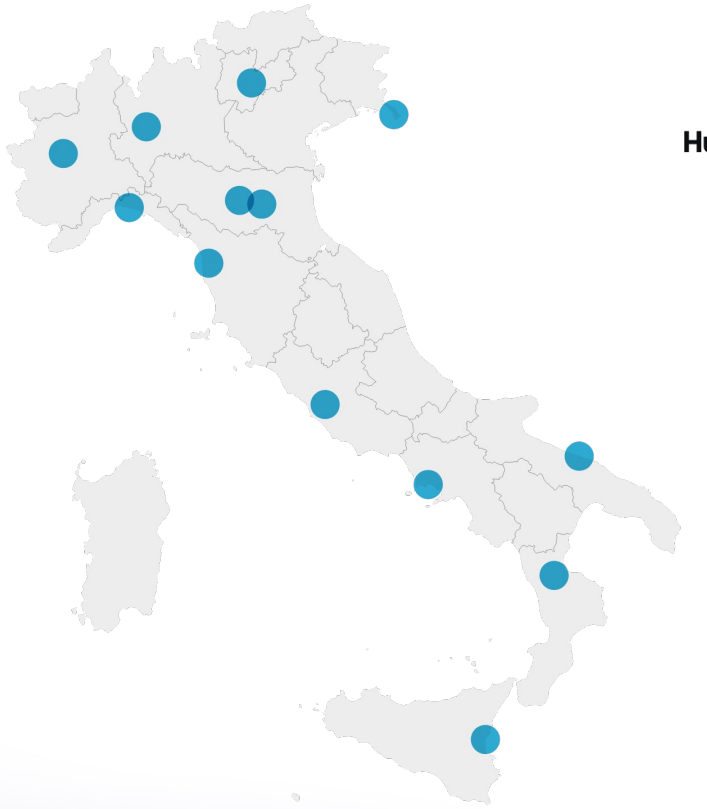
Tatiana Tommasi
Politecnico di Torino

Progetto FAIR: Future Artificial Intelligence Research

Financing program
PNRR M4C2 Investment 1.3 – Partnerships extended to universities, research centers and companies for the financing of basic research projects
Project Code: PE00000013
Start - End date of activities: 02/01/2023 - 01/01/2026
Proposer: National Research Council
Total financing amount: € 114.493.643,75

The **Extended Partnership (EP)** FAIR aims to contribute to addressing the research questions, methodologies, models, technologies and also the ethical and legal rules to build Artificial Intelligence systems capable of **interacting** and **collaborating with humans**, of **perceiving and act** within **constantly evolving contexts**, to be **aware of one's limits** and capable of **adapting to new situations**, to be aware of the perimeters of **security and trust**, and to be attentive to the **environmental and social impact that their creation and execution can result.**

Progetto FAIR: Future Artificial Intelligence Research



1 spoke — Human-centered AI
2 spoke — Integrative AI
3 spoke — Resilient AI
4 spoke — Adaptive AI
5 spoke — High quality AI
6 spoke — Symbiotic AI
7 spoke — Edge-exascale AI
8 spoke — Pervasive AI
9 spoke — Green-aware AI
10 spoke — Bio-socio-cognitive AI

SOGGETTI PUBBLICI
Università
Politecnico di Milano
Politecnico di Torino
Scuola Internazionale Superiore di Studi Avanzati
Scuola Normale Superiore
Università degli Studi di Bari Aldo Moro
Università degli studi di Modena e Reggio Emilia
Università degli Studi di Napoli Federico II
Sapienza Università di Roma
Consiglio Nazionale delle Ricerche
Università della Calabria
Alma Mater Studiorum Università di Bologna
Consiglio Nazionale delle Ricerche
Istituto Nazionale di Fisica Nucleare
Università di Catania
Università di Pisa
Università di Trento

SOGGETTI PUBBLICI
Organismi di Ricerca
Sapienza Università di Roma
Alma Mater Studiorum Università di Bologna
Consiglio Nazionale delle Ricerche
Istituto Nazionale di Fisica Nucleare

SOGGETTI PRIVATI:
Organismi di Ricerca
Università Campus Bio-Medico di Roma
Università Commerciale Luigi Bocconi
Consorzio Interuniversitario Nazionale per l'Informatica
Fondazione Bruno Kessler
Istituto Italiano di Tecnologia

SOGGETTI PRIVATI:
Imprese
Bracco Imaging S.p.A.
Deloitte Risk Advisory S.R.L S.B.
Expert.ai S.p.A.
INTESA SANPAOLO S.P.A.
Leonardo S.p.A.
Lutech S.p.A.
STMicroelectronics s.r.l.

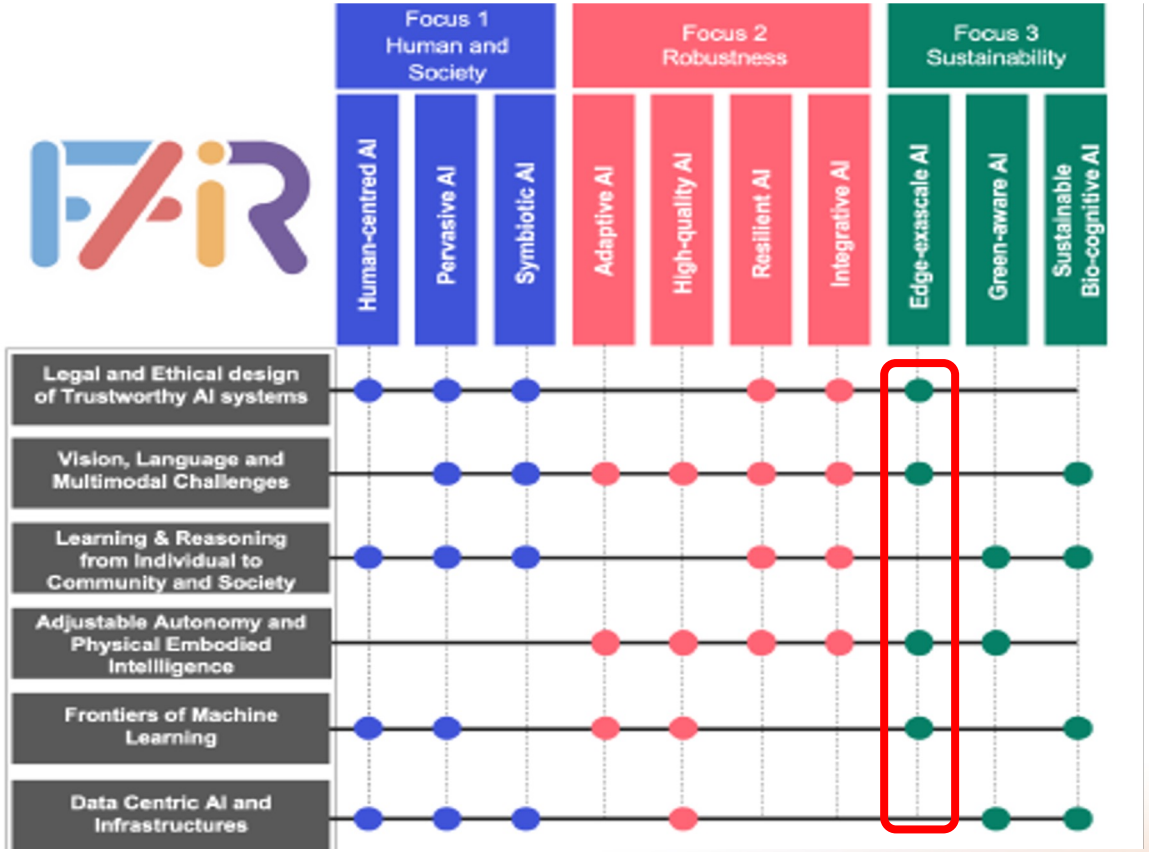
Progetto FAIR: Future Artificial Intelligence Research

spoke 7

Edge-exascale AI

The contribution of Spoke 7 also extends across 4 TPs

- **TP1**: Legal and Ethical Design of Trustworthy AI Systems
- **TP2**: Vision, Language and Multimodal Challenges
- **TP4**: Adjustable autonomy and Physical embodied Intelligence (co-leader)
- **TP5**: Frontiers of Machine Learning



spoke 7

In numbers

8 DEPARTMENTS INVOLVED

DAUIN - DEPT. OF CONTROL AND COMPUTER ENGINEERING
DENERG - DEPT. OF ENERGY
DET - DEPT. OF ELECTRONICS AND TELECOMMUNICATIONS
DIATI - DEPT. OF ENVIRONMENT, LAND AND INFRASTRUCTURE ENGINEERING
DIGEP - DEPT. OF MANAGEMENT AND PRODUCTION ENGINEERING
DISAT - DEPT. OF APPLIED SCIENCE AND TECHNOLOGY
DISEG - DEPT. OF STRUCTURAL, GEOTECHNICAL AND BUILDING ENGINEERING
DISMA - DEPT. OF MATHEMATICAL SCIENCES

Politecnico di Torino 1859

MAI PER SEGUIR VIRTUTE E CANOSCENZA SISSA

11 PhD POSITIONS
9 POSITIONS ASSIGNED

CRITICAL MASS
29 PROFESSORS

PUBBLICATIONS: 74
Journals: 15
Conferences: 59

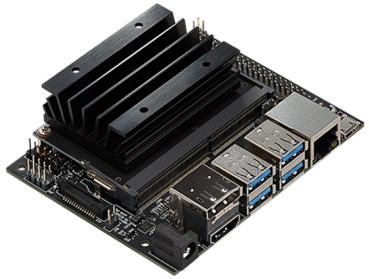
12 ASSISTANT PROFESSORS
WITH TIME CONTRACT
10 POSITIONS ASSIGNED

6 RESEARCH ASSISTANSHIPS
4 POSITIONS ASSIGNED

AI-core foundational venues (14):
CVPR, ICML, IROS, ICRA, ICASSP, INTERSPEECH, …

spoke 7

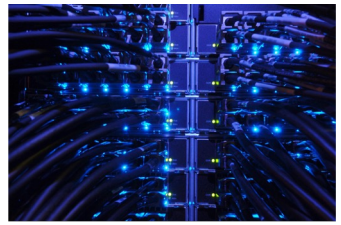
Edge and Exascale AI



Tiny AI

- ✓ Higher energy efficiency
- ✓ Environment friendly
- ✓ Lower hardware requirement
- ✓ Faster training and inference

Parallel AI



Foundation

WP 7.1 | WP 7.3 | WP 7.2 | WP 7.3

Downstream tasks

Computer Vision and Sensing in Extreme Computational Frameworks — WP 7.4 — Perception

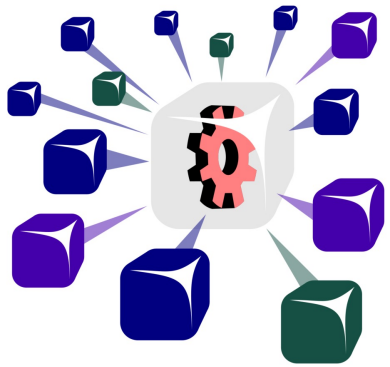
Action — WP 7.5 — Intelligent Systems, Autonomous Robots and Interaction in Extreme Computational Frameworks

Implication

WP 7.6 — Ethical, Legal, Economical and Societal issues in Edge and Exascale AI

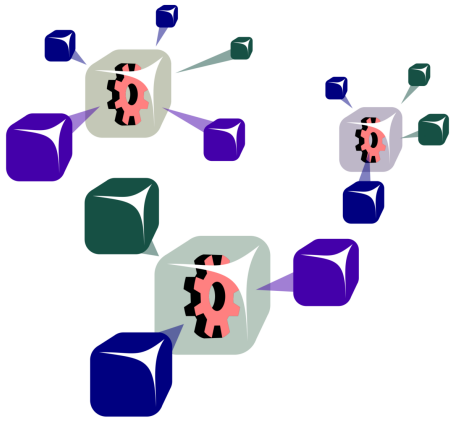
Why Tiny Machine Learning?

At a central node:
Centralized Computing



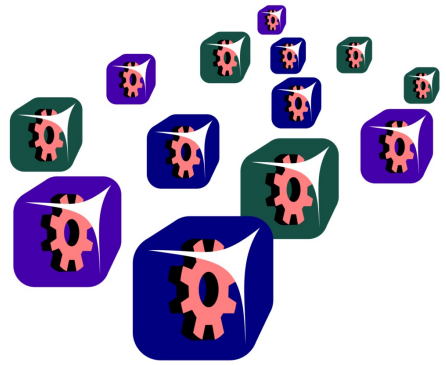
- High-resources
- High bandwidth requirements
- High connection stability requirements
- High delay in the evaluation

Near the sensor node:
Edge Computing



- Low-resources
- Smaller bandwidth requirements
- Smaller connection stability requirements
- Smaller delay in the evaluation

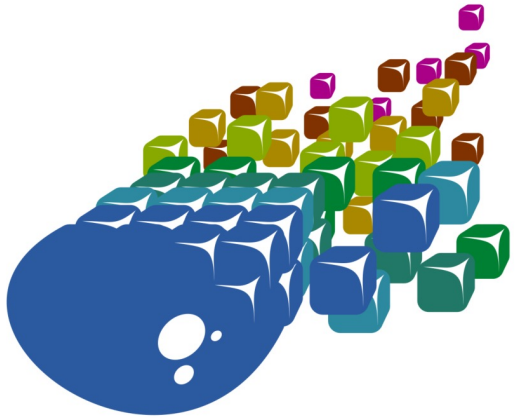
At the sensor node:
Tiny Machine Learning



- Minimum resources
- Strict energy requirements
- No connection/transmission required
- Minimum delay in the evaluation

How? Deep Neural Network Compression

Model quantization

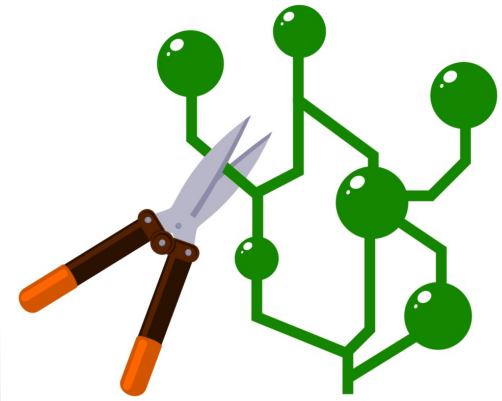


Minimize the number of bits necessary to encode data

- ***Quantization of the activations***
- ***Quantization of the parameters***

- **Post-training quantization**
- **Quantization-aware training**

DNN pruning



Remove parts of the DNN that are redundant

Assign to each removable part in a DNN a score. Then, the parts with the lowest scores are pruned.

- **Post-training pruning**
- **Pruning at initialization**

DNN Pruning at Initialization

NTK Regime: $$f(\boldsymbol{x}, \boldsymbol{\theta}_{t+1}) = f(\boldsymbol{x}, \boldsymbol{\theta}_t) - \alpha \boldsymbol{\Theta}(\boldsymbol{x}, \boldsymbol{x}) \nabla_f \boldsymbol{\mathcal{L}}$$

Neural Tangent Kernel (NTK)

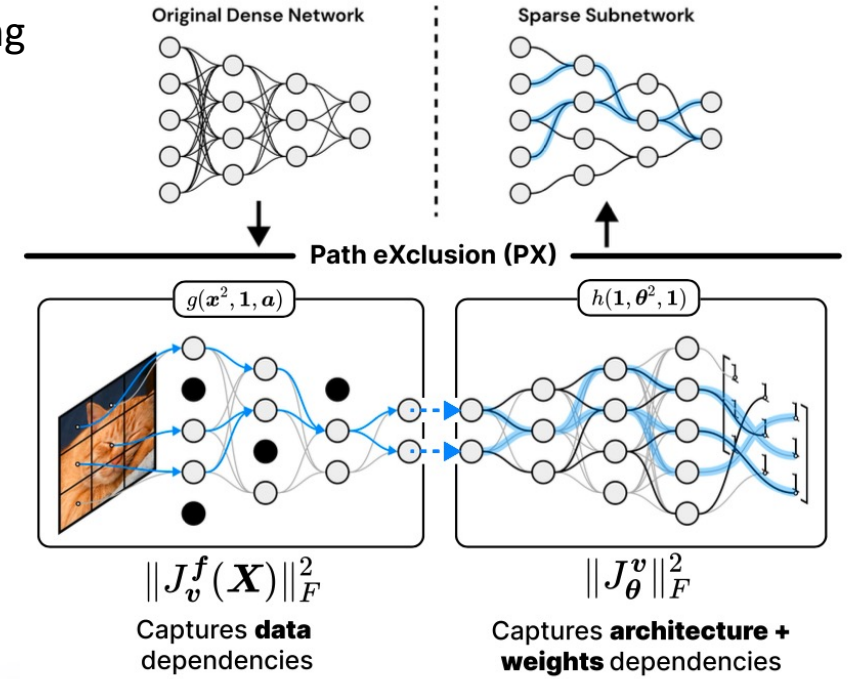
$\|\boldsymbol{\Theta}(\boldsymbol{x}, \boldsymbol{x})\|_F^2$ predicts (at initialization) how $\boldsymbol{\theta}$ will evolve during training

→ **QUADRATIC COST IN** $\|\boldsymbol{\theta}\|_0$, $\|\boldsymbol{x}\|_0$ **!!!**

Based on input-output paths (within NNs)
we define the upper bound: $\|\boldsymbol{\Theta}(\boldsymbol{x}, \boldsymbol{x})\|_F^2 \leq \|\boldsymbol{J}_{\boldsymbol{v}}^{f}(\boldsymbol{x})\|_F^2 \cdot \|\boldsymbol{J}_{\boldsymbol{\theta}}^{\boldsymbol{v}}\|_F^2$

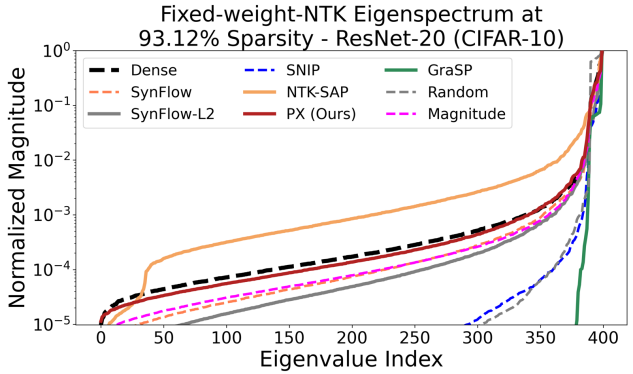
Prune parameters that least influence the
upper bound via Path eXclusion (PX)

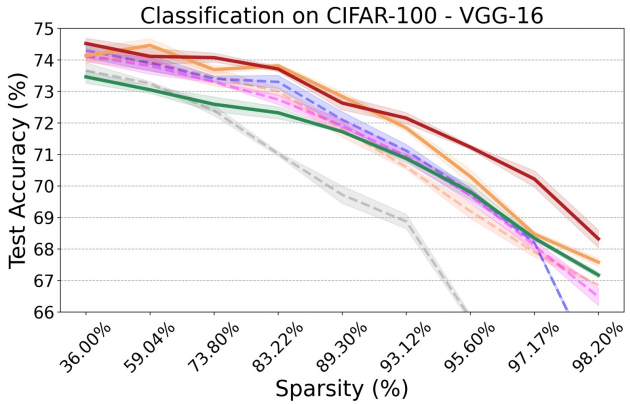
- Resulting subnetwork will evolve as its dense counterpart during training

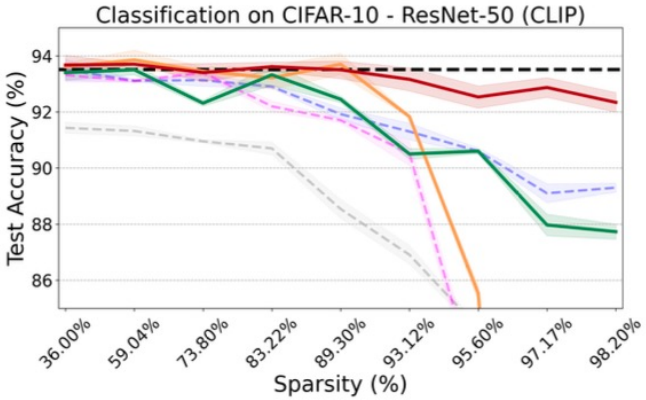


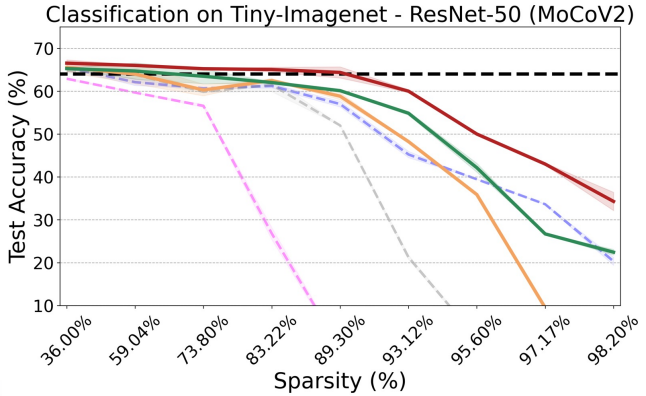
Iurada, L., Ciccone, M., Tommasi, T., Finding Lottery Tickets in Vision Models via Data-driven Spectral Foresight Pruning CVPR 2024

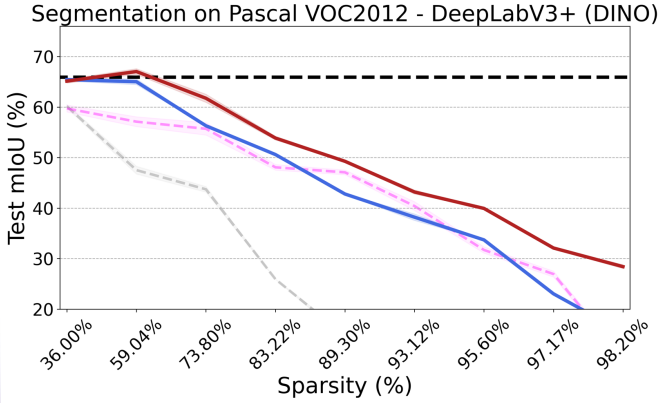
DNN Pruning at Initialization





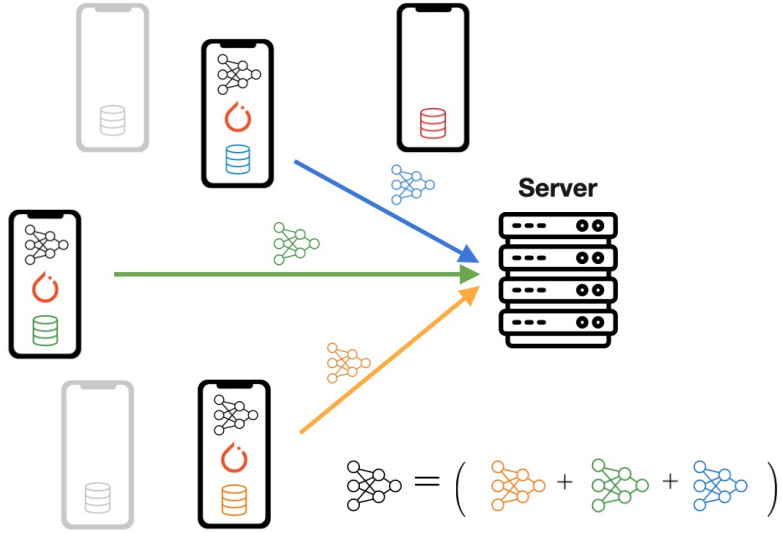






Distributed Computing: Federated Learning

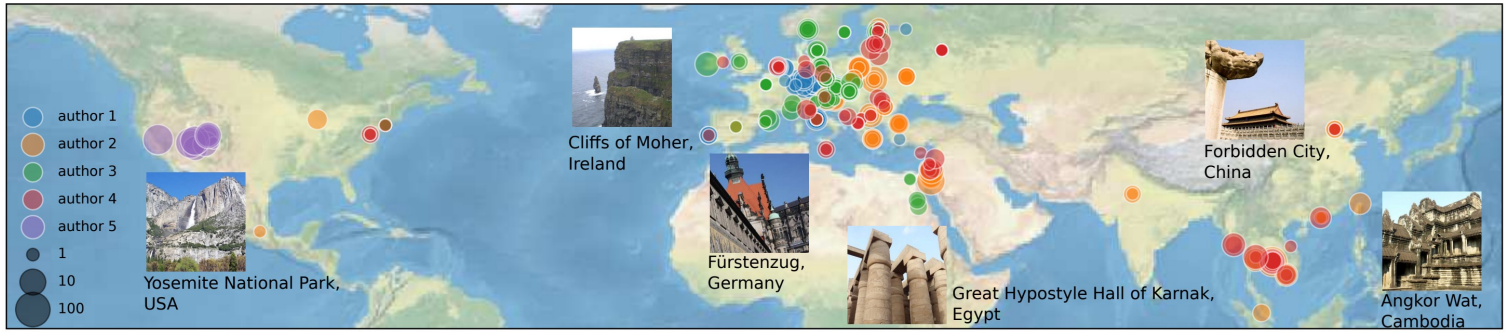
Population of users join a “federation” towards learning a model collaboratively

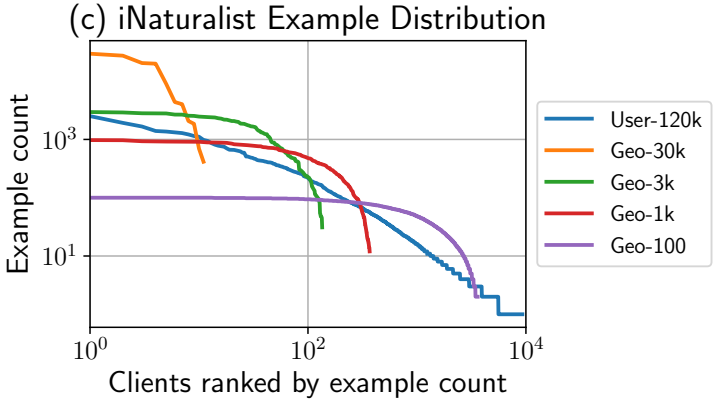


- 🔒 Data never leaves the device (privacy enhancing)
- 🏠 Training on data is local (distributed computation)
- 📡 Communicate model updates
- 💪 Powerful, but many challenges

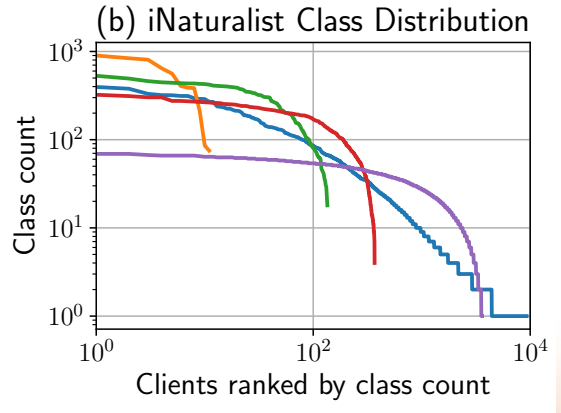
- **Opportunity for safely collecting data while distributing computation**

Large-scale collaborative data collection



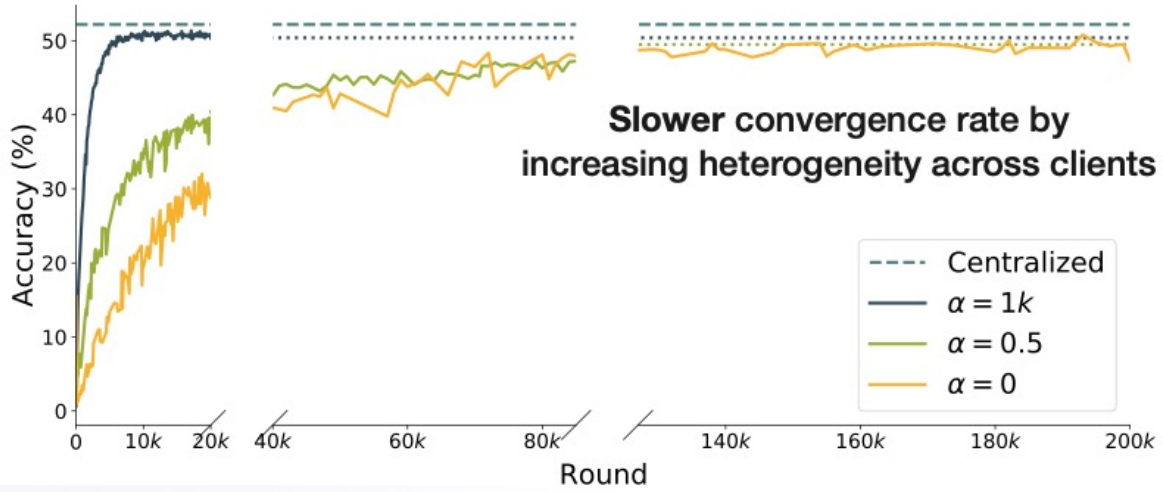


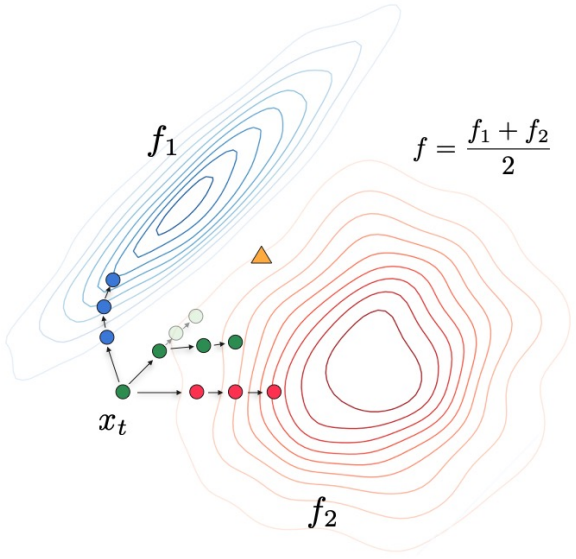
Diversity among decentralized sources



Challenge of Statistical Heterogeneity

- Samples collected on the edge correlate to specific **user habits**, **preferences**, **location** (**Domain** Shift, **Label** Skewness, **Size** Imbalance)
- **Clients drift** from global solution by specializing on local data.
- Training convergence is severely affected by client drift





Client Drift Problem

Research Objectives

⚡ Improving **generalization** performance — close the gap with centralised training

☄️ **Speeding up convergence** by reducing communication rounds

👀 Be considerate with **communication cost**

🧿 **Enabling real-world** vision applications

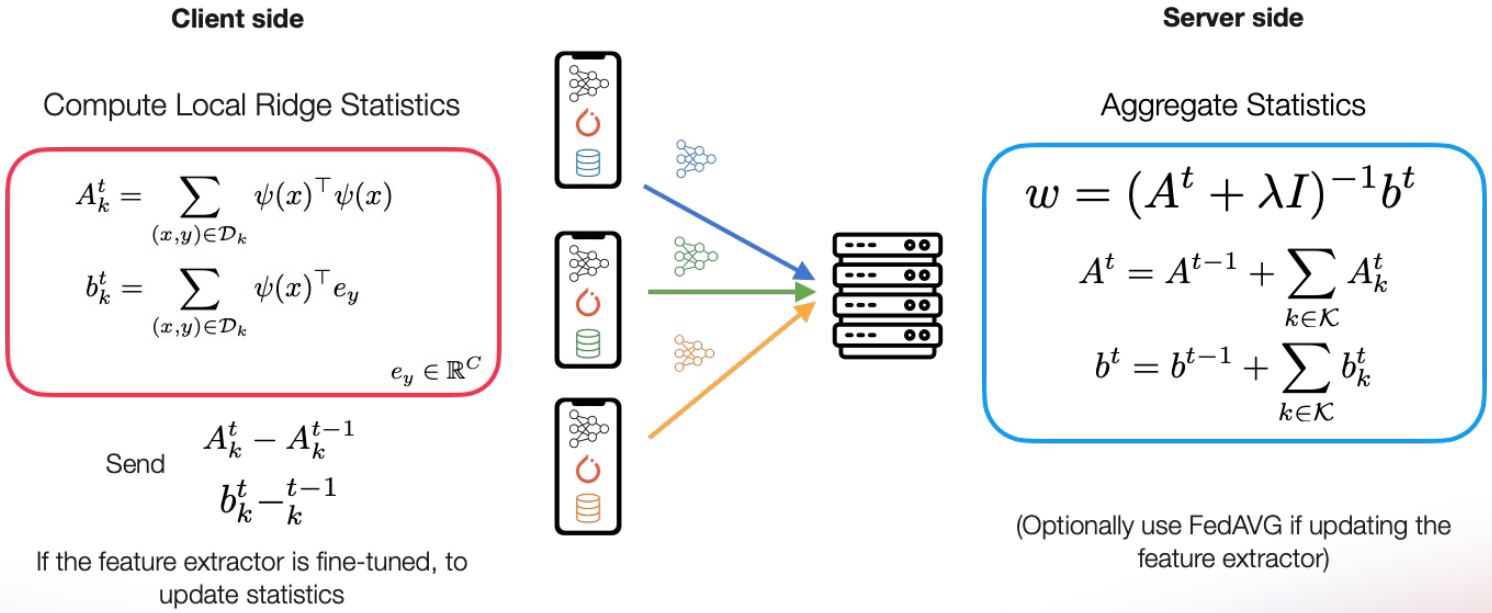
Proposed Solutions

Focus on the last layers of the networks that are most affected by the specific client properties and propose **a different learning objective for classification**.

FED3R: Recursive Ridge Regression

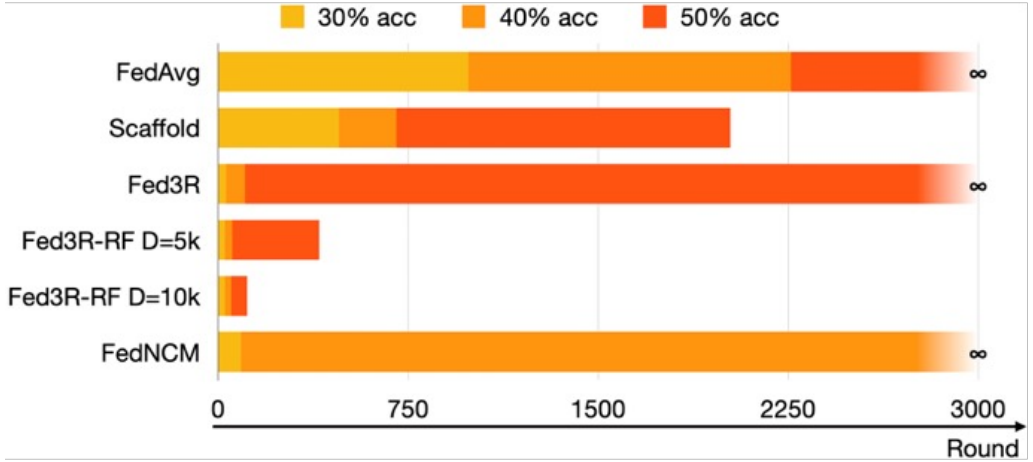
Substitute Classification with Ridge Regression, use it in a one-vs-all formulation

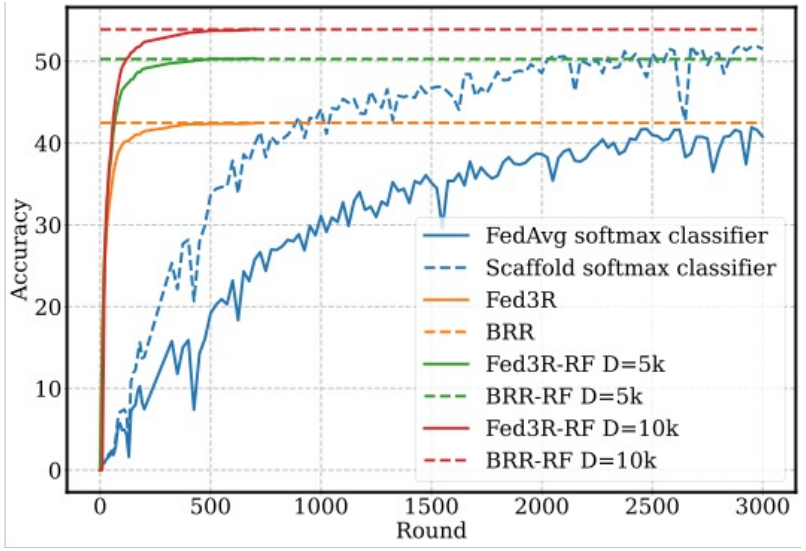
$$W^* = [w_1 \dots w_K] = \arg\min_W \frac{1}{|D|} \sum_{(x,y)\in D} \left\| W^\top \psi(x) - \text{OneHot}(y) \right\|^2 + \lambda \|W\|^2$$



Fanì, E., Camoriano, R., Cabuto, B., Ciccone, M. FED3R: Recursive Ridge Regression for Federated Learning with strong pre-trained models. ICML 2024

FED3R: Recursive Ridge Regression



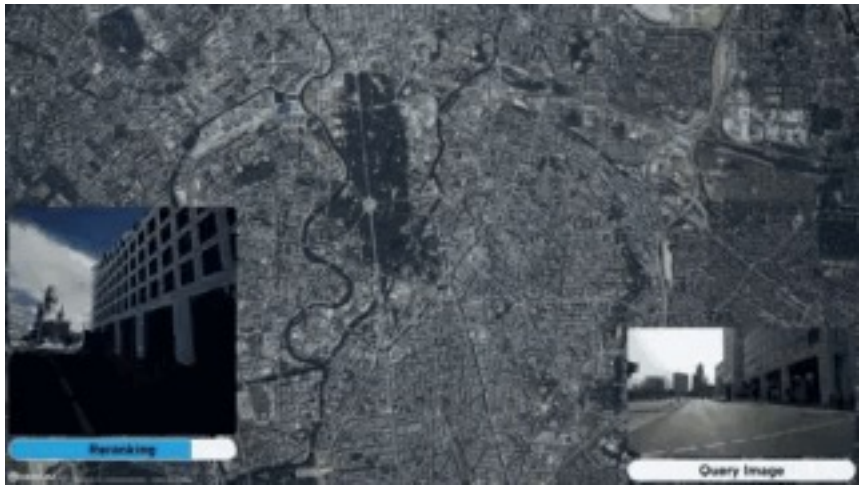


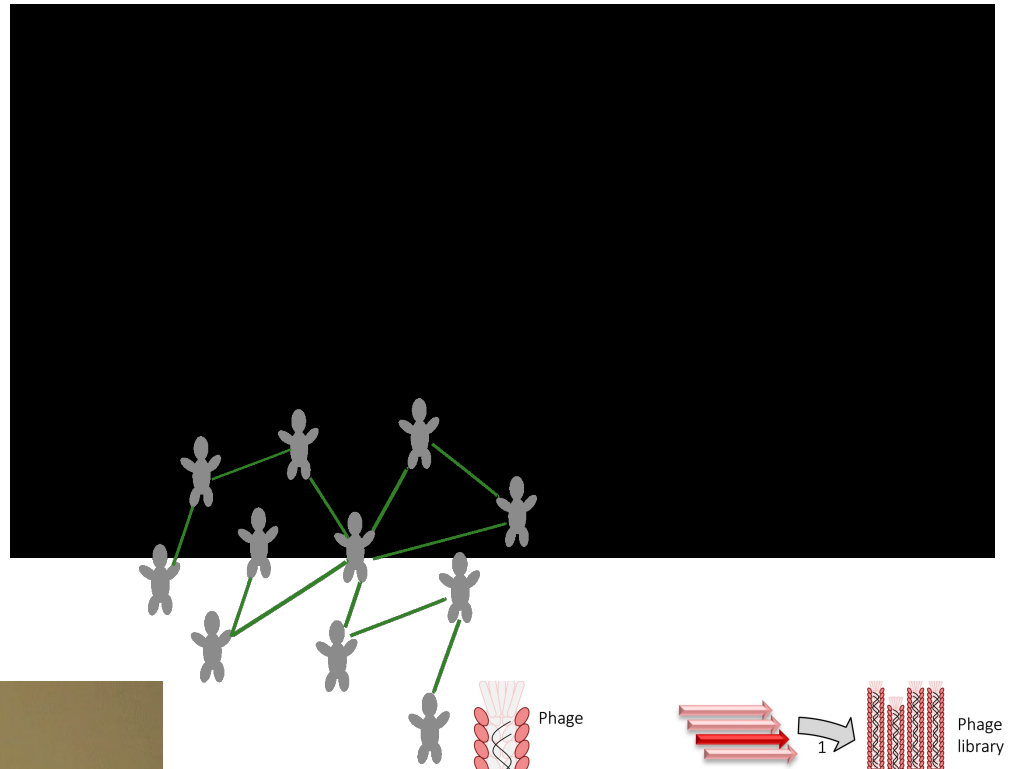
Faster convergence with a very straightforward solution

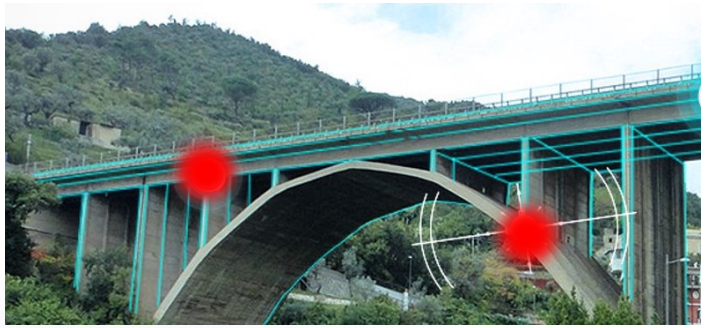
Up to 44x and 18x times faster than FedAvg and Scaffold, respectively

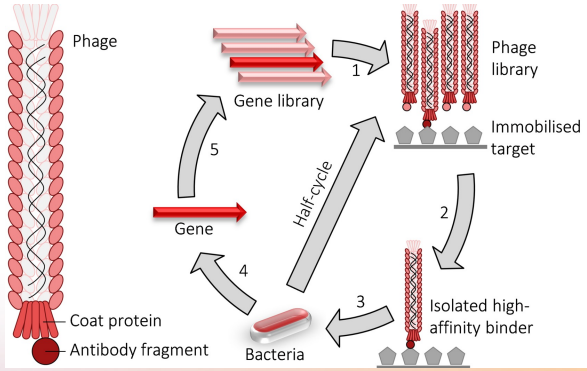
Fanì, E., Camoriano, R., Cabuto, B., Ciccone, M. FED3R: Recursive Ridge Regression for Federated Learning with strong pre-trained models. ICML 2024

Applications









Resources

4 GPU NVIDIA RTX 3090
2 Camera RGB-D Intel Realsense
1 NVIDIA JETSON AGX ORIN DEVELOPMENT KIT - 945-13730-0005-000
2 NVIDIA JETSON TX2 DEVELOPER KIT - 945-82771-0005-000
2 The NVIDIA® Jetson Nano™ Developer Kit 945-13450-0000-100 | Jetson Nano Developer Kit Embedded System Development Boards and Kits
1 Jetson cluster Jetson Mate Cluster Advanced - Carrier Board with 4 Jetson Xavier NX SoMs for GPU Cluster/ Server
1 Penguin edge Penguin Edge™ IFC6640 SBC from Penguin Solutions™

5 workstations (nr. 2 RTX 4090 24 GB)
2 workstations (nr. 2 2x Titan RTX (24 GB)
3 workstations (nr. 4 4x GTX 1070 (8 GB)
3 workstations (nr. 2 2x RTX 2080Ti (11 GB))
1 workstation (Titan RTX (24 GB), RTX A5000)
1 workstation (RTX 3090 (24 GB), RTX A5000)
1 server 16 GTX 1080 (8 GB)

+ FAIR investment @PoliTo of 250k€ for HPC
+ Collaborations with CINECA and Leonardo Supercomputer

Thanks for your Attention

Tatiana Tommasi
tatiana.tommasi@polito.it

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.