AI SPRINT
AI IN SECURE PRIVACY-PRESERVING COMPUTING CONTINUUM

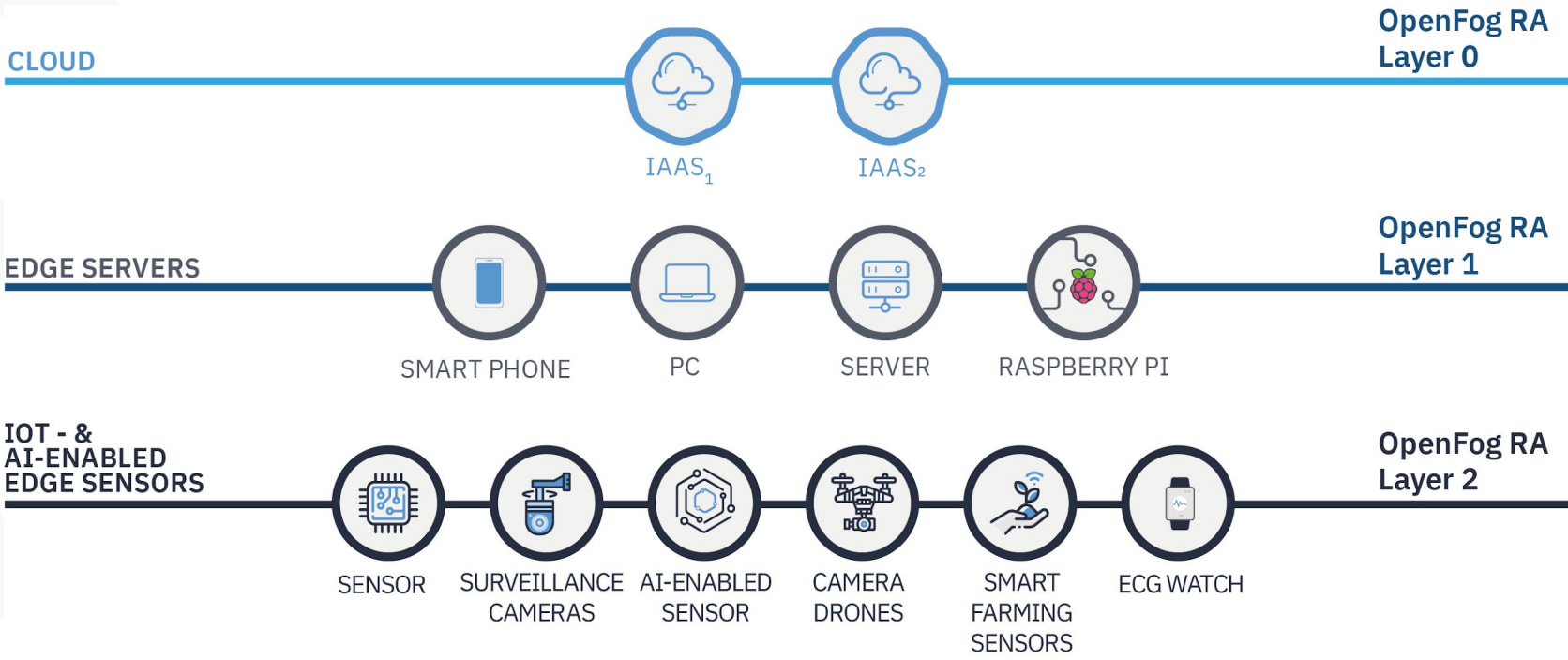
Unleashing the potential of AI in Healthcare and E-Brainscience

Danilo Ardagna, Politecnico di Milano

danilo.ardagna@polimi.it

AI-SPRINT project has received funding from the European Union Horizoon 2020 research and innovation programme under Grant Agreement **No. 101016577.**

- By 2026, AI worldwide market will approach $900 billion (CAGR 18.6%[1]) while edge computing will reach $324 billion (CAGR 13.6%[2])
- AI needs resources at the edge of the network
- New challenges from the infrastructural perspective

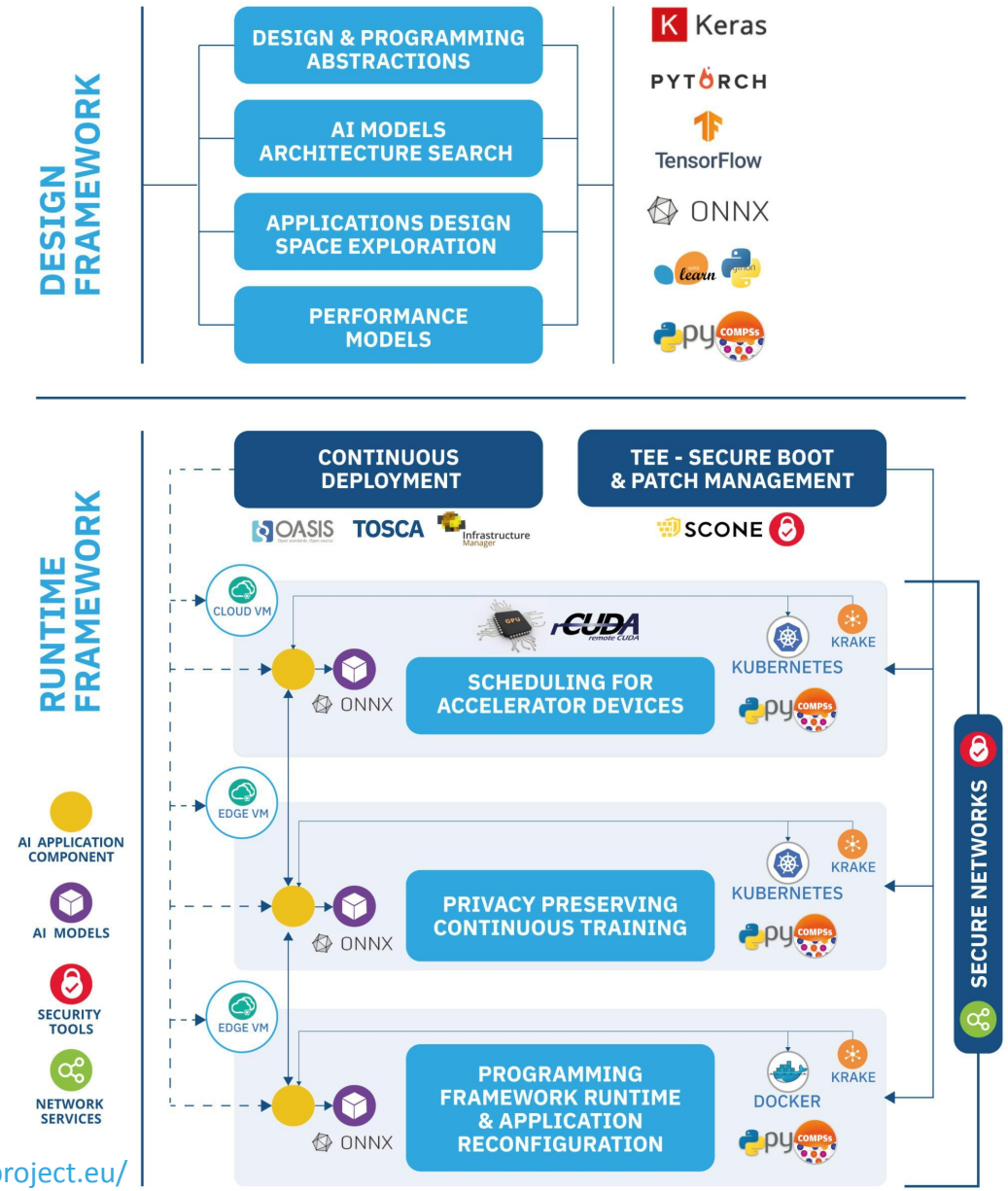


[1]IDC Semiannual Artificial Intelligence Tracker, July 2022

[2]IDC Worldwide Edge Spending Guide, August 2022

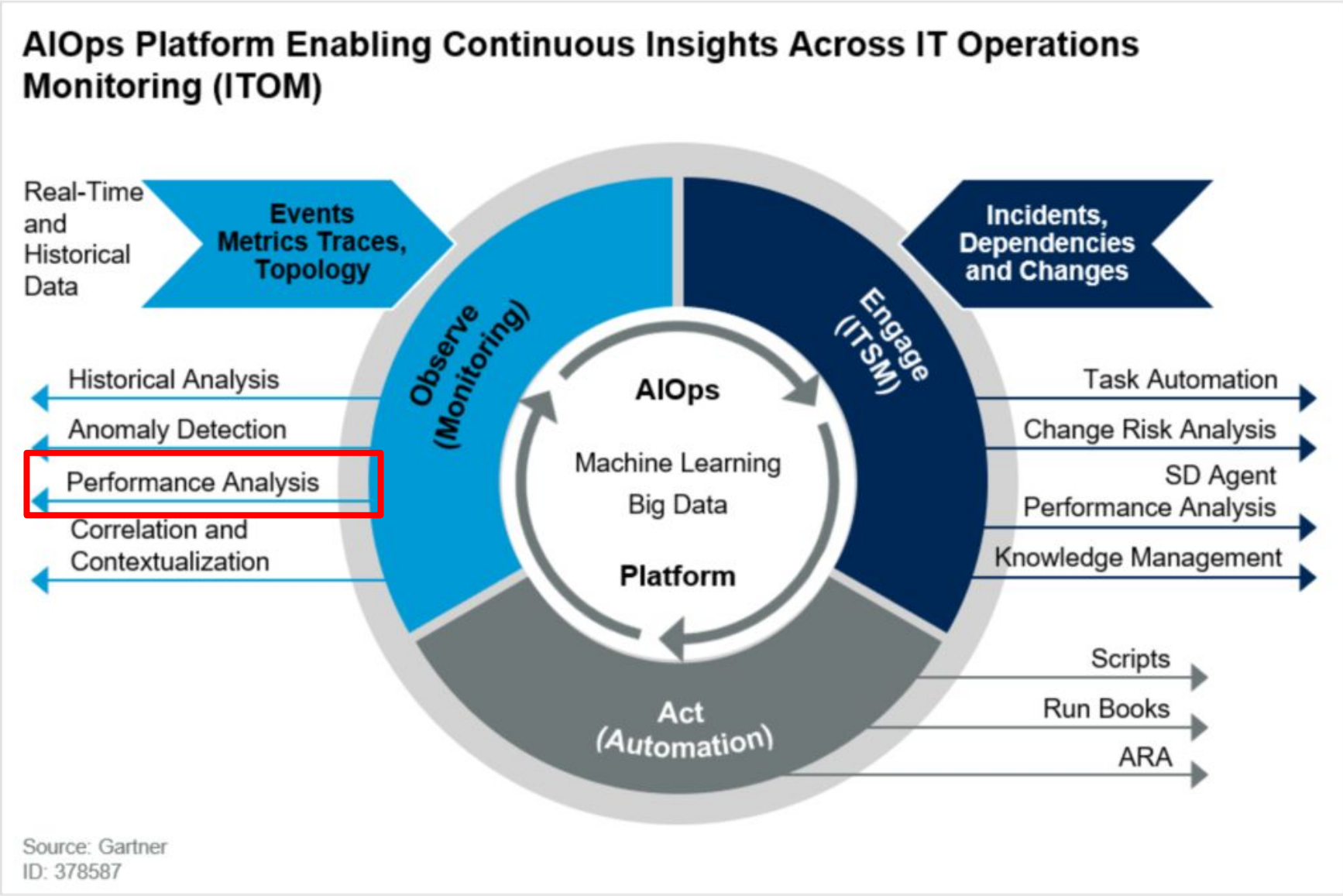
AI-SPRINT promises

- Simplified programming models
- Automated deployment and dynamic reconfiguration
- Secure execution of AI applications
- Highly specialized building blocks for distributed training, privacy preservation and architecture enhancement



H. Sedghani, et. al. Advancing Design and Runtime Management of AI Applications with AI-SPRINT. AIM 2021 Workshop Proceedings

New trends: AIOps



According to BSC, "By applying AIOps to these common IT operations responsibilities, ITOps teams can reduce event noise, avoid downtime, optimize application performance, address infrastructure complexity, and keep customers happy"

"Yet, there is no doubt: There is no future of IT Operations that does not include AIOps."

Source: Gartner® Market Guide for AIOps Platforms, Pankaj Prasad, Padraig Byrne, Gregg Siegfried, May 30, 2022

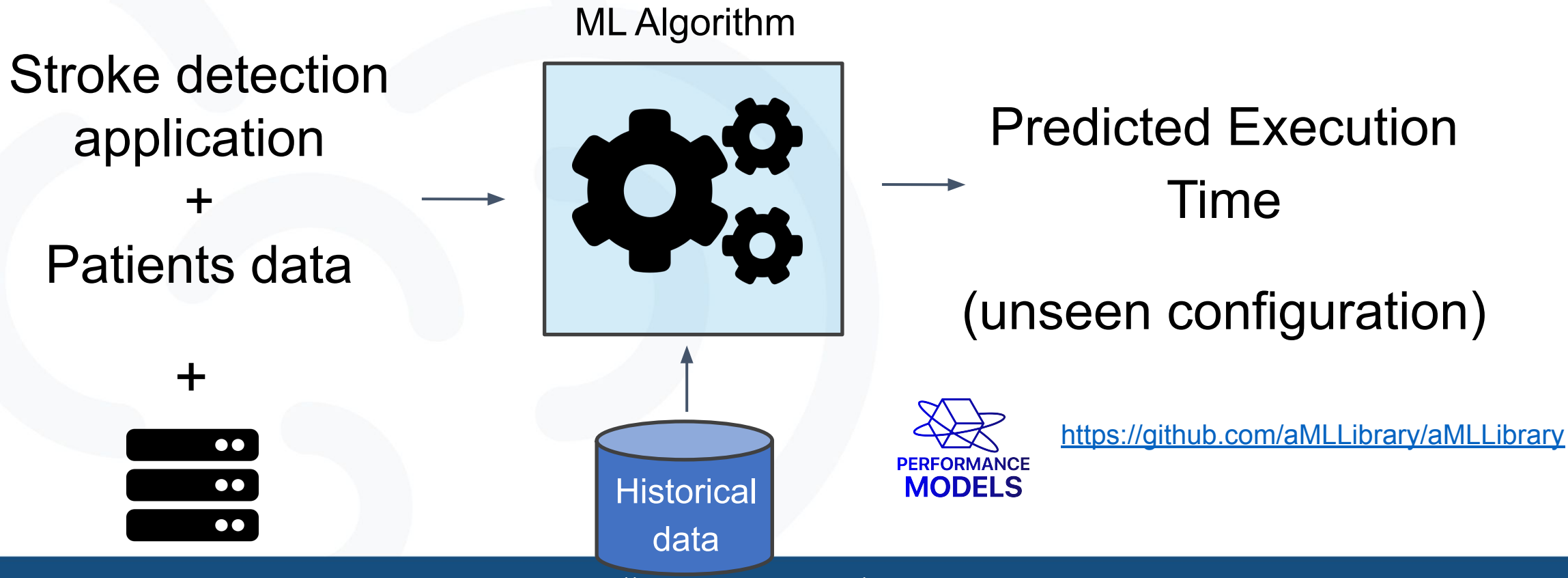
According to IDC, "**By 2026, 90% of Global 2000 CIOs will use AIOps** solutions to drive automated remediation and workload placement decisions that include cost and performance metrics, improving resiliency and agility."

- Computing continua have extensive list of possible configurations that can be allocated for an application
  - Type of processing node, # cores, etc

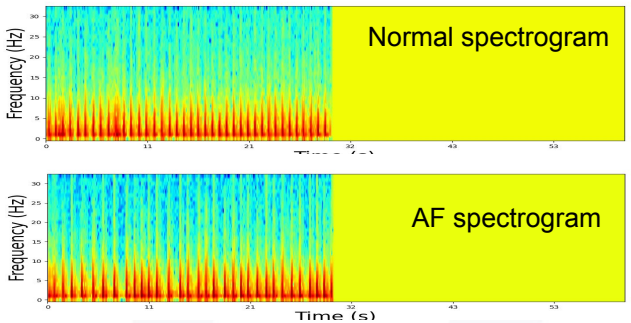
How can we predict the execution time of an application running on a target edge configuration?

- First step of a bigger capacity planning solution:
  1. Predict execution time given an amount of resources available
  2. Optimize computation resources given target deadline (predicted by solution in step 1)

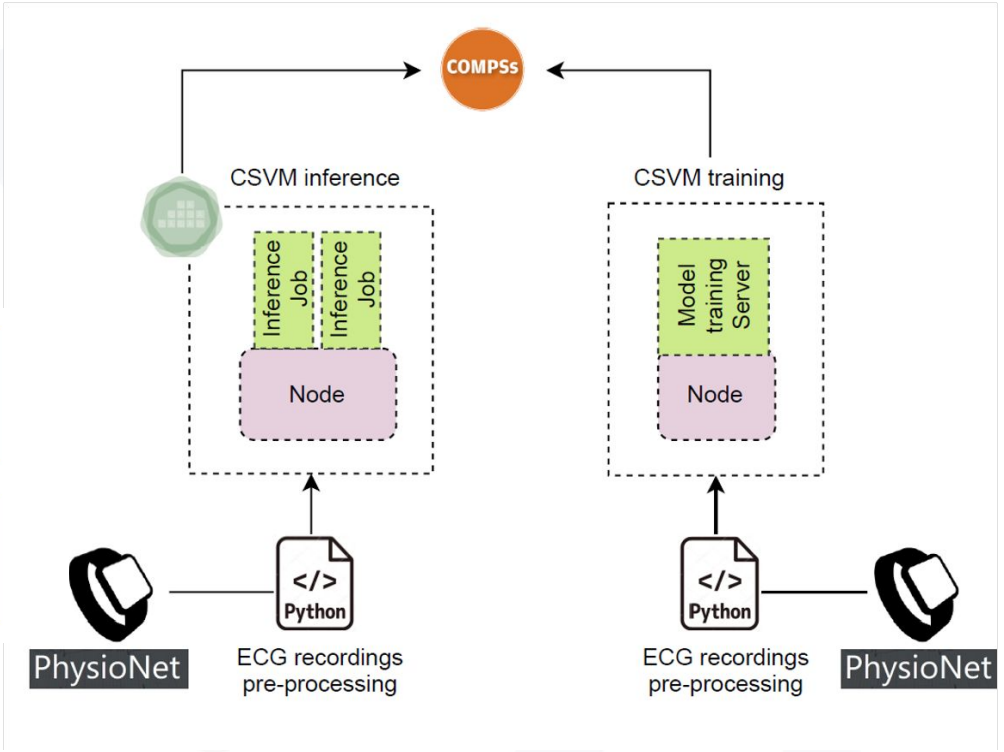
- Use Machine Learning (ML) methods to predict execution time
  - Regression task
  - Extract input features from logs of previous executions



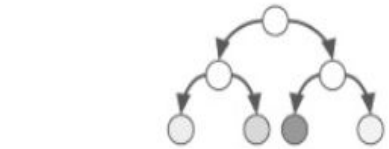
Preliminary results on the stroke detection application

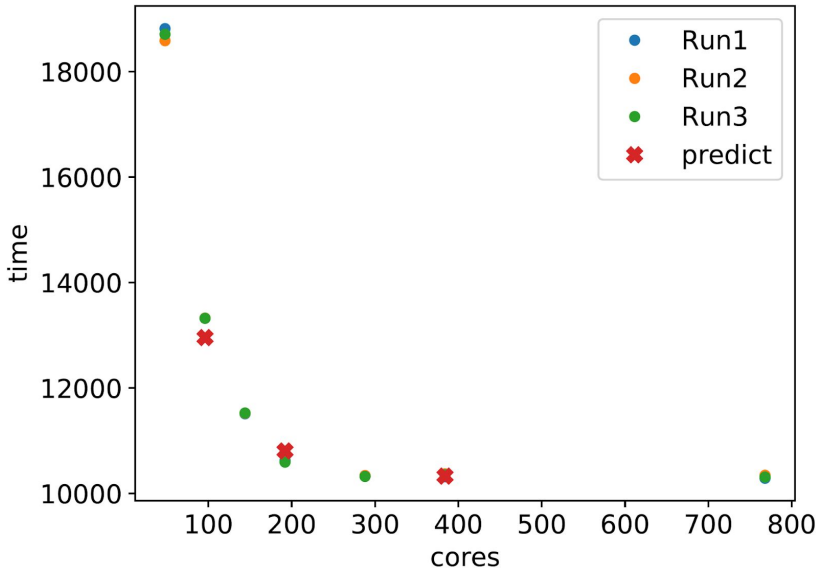


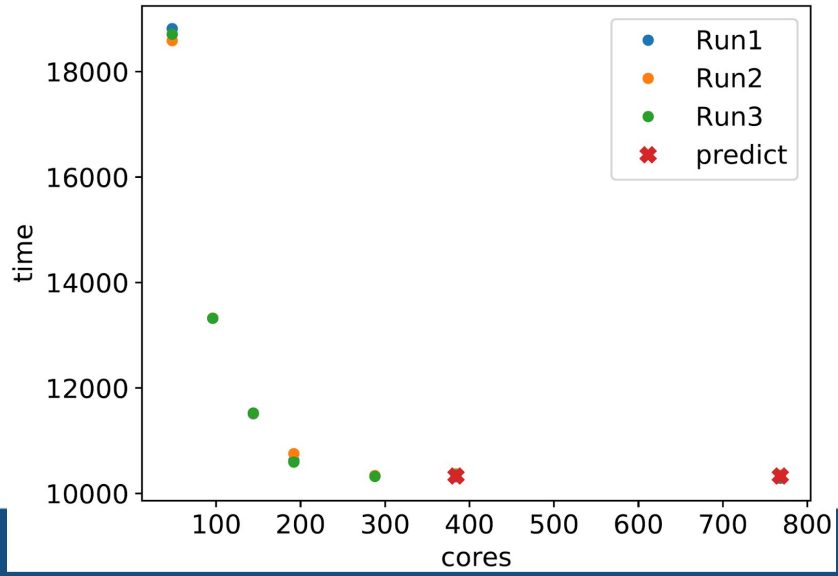
- Time
- Frequency
- Amplitude



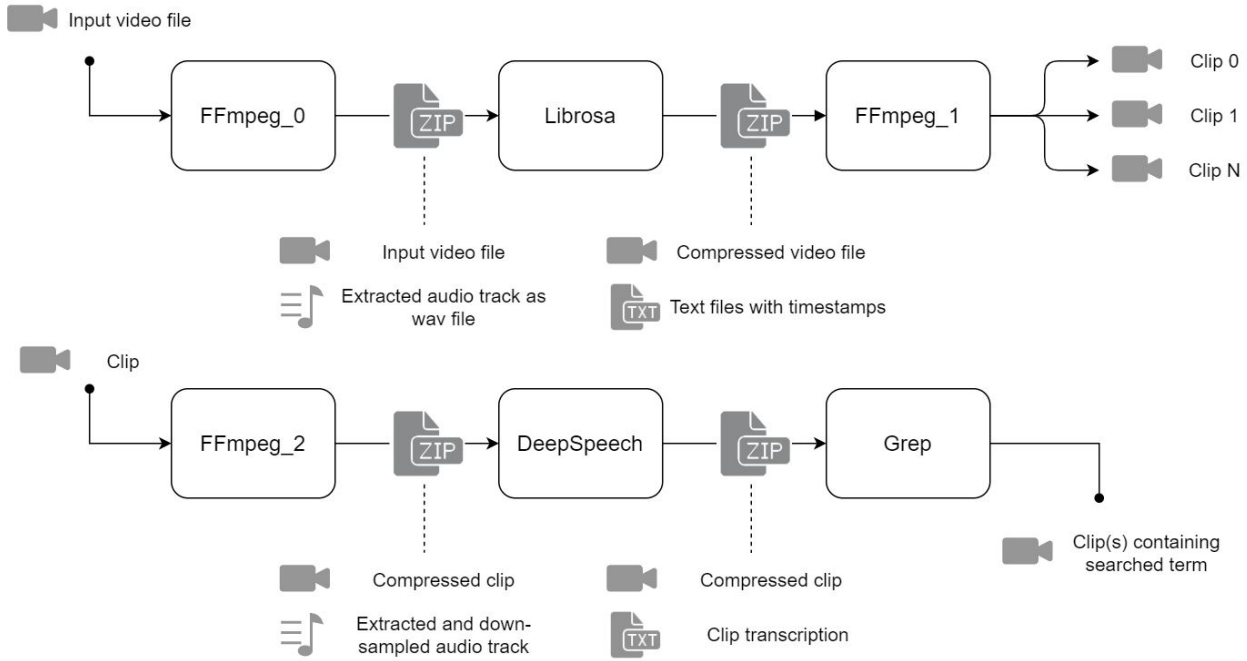
CSVM



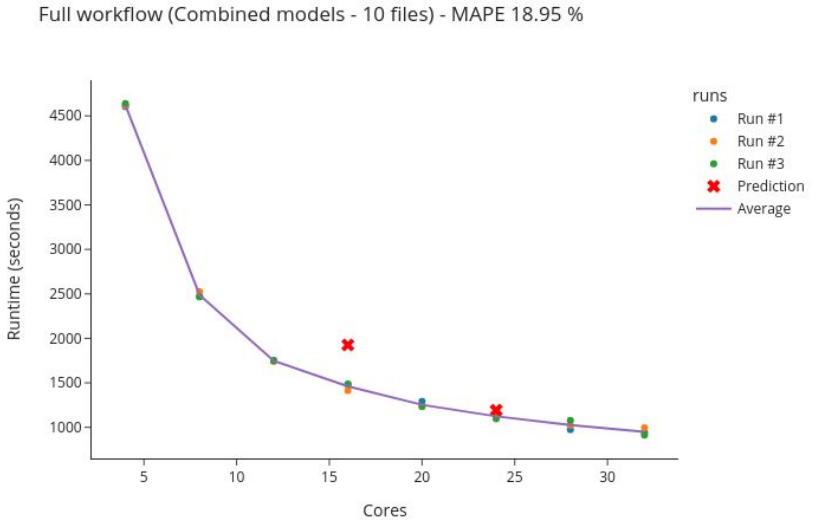


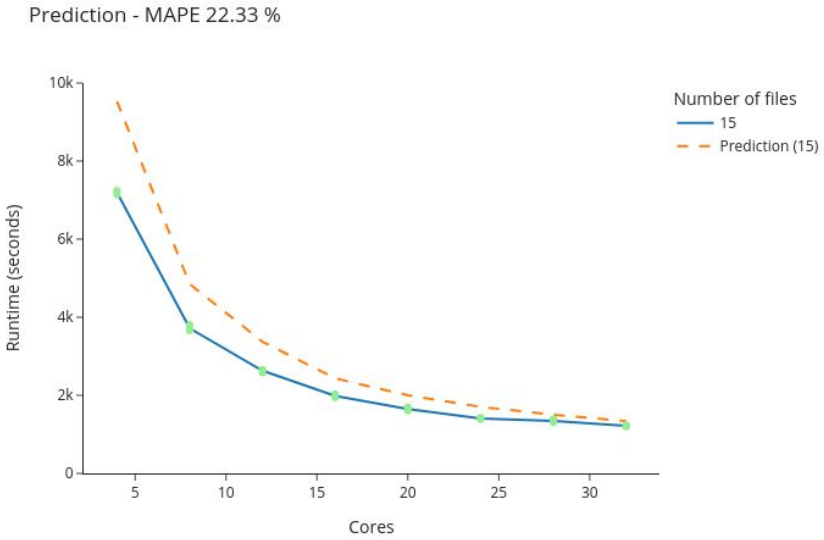


OSCAR & OSCAR-P



Search for a sentence in a video and return clips where that sentence is said



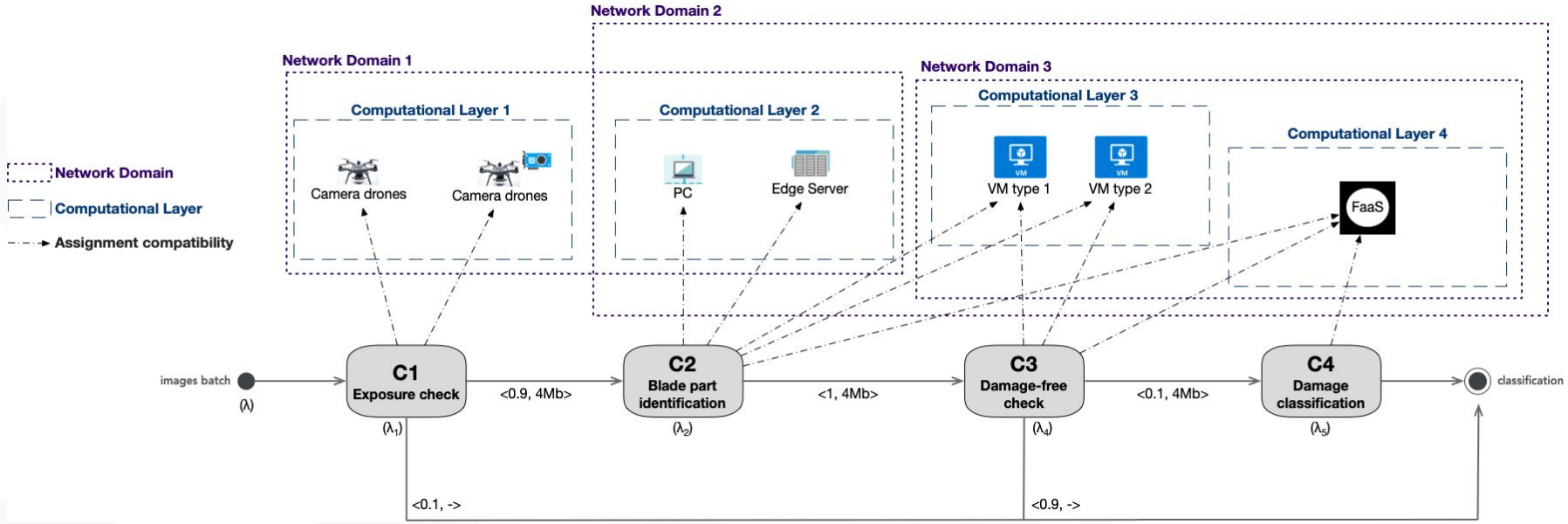


https://github.com/grycap/oscar

https://gitlab.polimi.it/ai-sprint/OSCAR-P

Optimal deployment and …

- Constraints:
  - Component latency
  - Application throughput
  - …

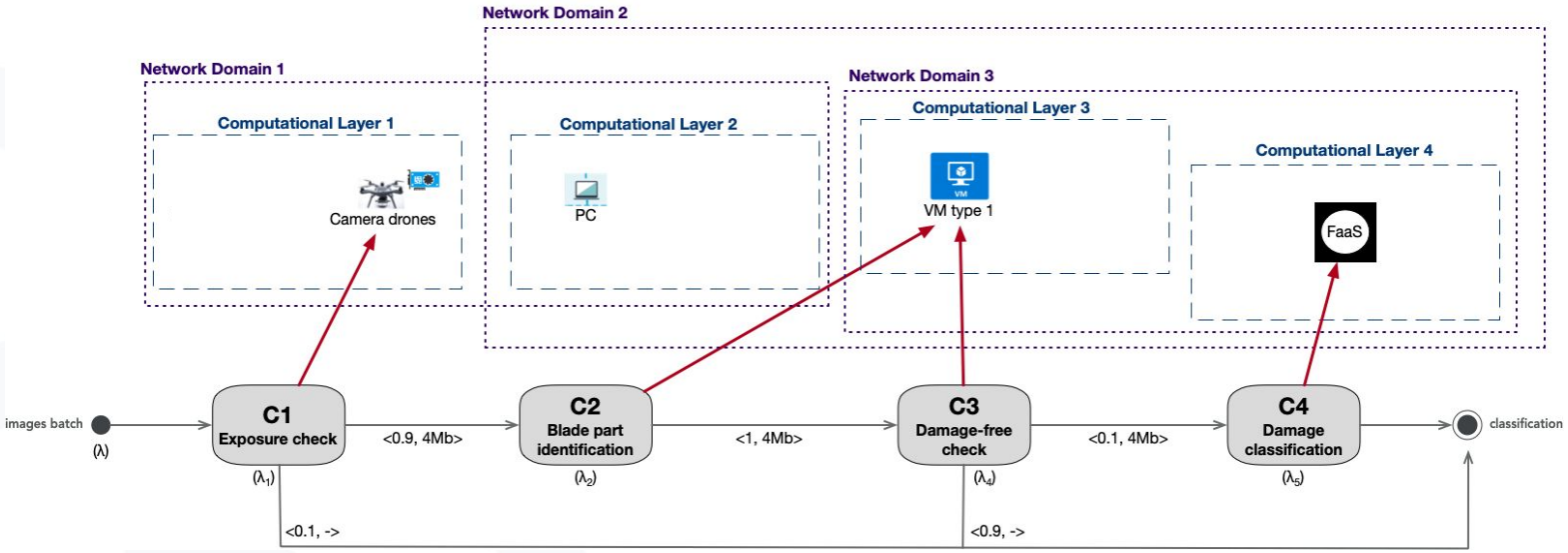


https://gitlab.polimi.it/ai-sprint/space4ai-d

H. Sedghani, F. Filippini, D. Ardagna. A Random Greedy based Design Time Tool for AI Applications Component Placement and Resource Selection in Computing Continua. IEEE Edge 2021.

… and dynamic reconfiguration

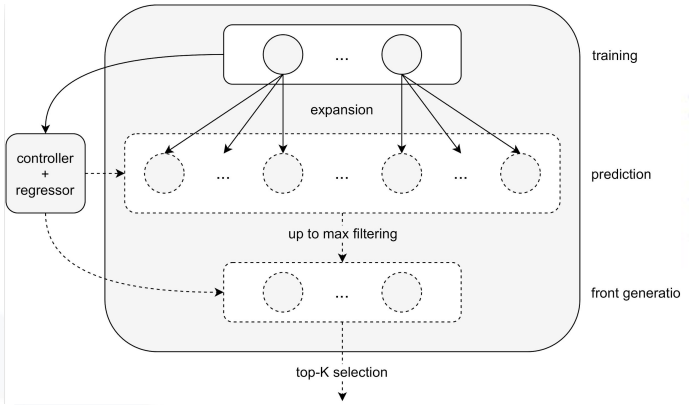
- Components migration
- Changing the deployment
- Scaling out / in cloud VMs

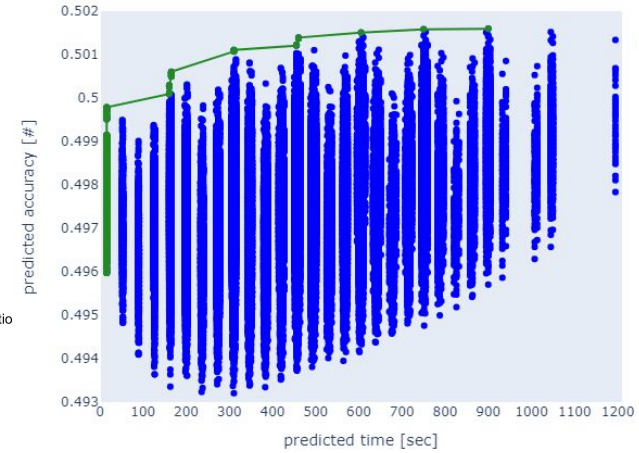


SPACE 4AI-R

AI-SPRINT: Opportunities Beneath the Challenges

- Novel AI automatic design patterns to
  - Design new architectures optimized for the Computing Continuum
  - Reduce resource demand for hyperparameter tuning
  - Go beyond AI designer intuition





POPNAS

https://gitlab.polimi.it/ai-sprint/popnas-v2

E. Lomurno, S. Samele, M. Matteucci, D. Ardagna. Pareto-Optimal Progressive Neural Architecture Search. ACM Workshop on NeuroEvolution@Work 2021.

A. Falanti, E. Lomurno, S. Samele, D. Ardagna, M. Matteucci. POPNASv2: An Efficient Multi-Objective Neural Architecture Search Technique. IEEE WCCI-IJCNN 2022, Padova, Italy.

Thanks for your attention….

COORDINATOR

POLITECNICO MILANO 1863

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

TECHNISCHE UNIVERSITÄT DRESDEN

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

GREGOIRE

Beck et al. work. together.

CLOUD & HEAT

7bulls.com

AirFusion

Trust-IT Services
Communicating ICT to markets

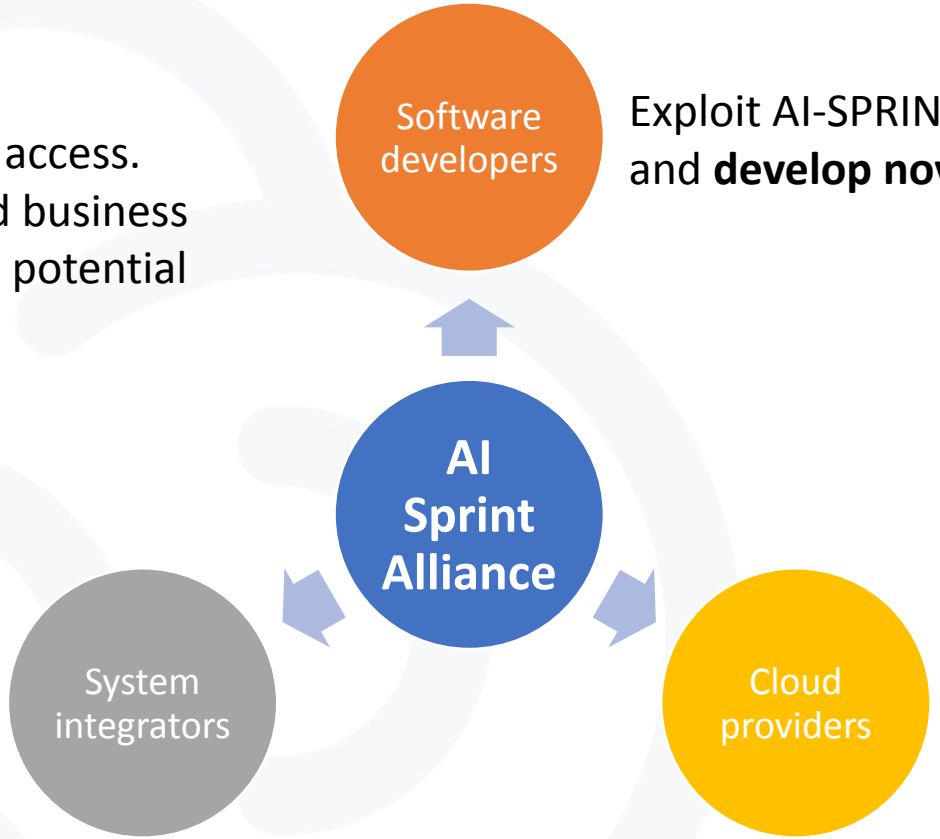
IDC

…. any questions?

https://www.ai-sprint-project.eu/

AI-SPRINT Alliance and Adopter Acceleration Club

June 2022
- Set up and launch of the Alliance
- Collaboration set-up with customised access.
- Regular briefings on technological and business innovations, including market outlook, potential competitors and adopters.



Exploit AI-SPRINT technology to **design** and **develop novel AI applications**

Offer **new services** through the **integration** of **AI-SPRINT technology stack**

Include AI-SPRINT technology **into their service catalogue** to offer easy to use **design frameworks**

https://ai-sprint-project.eu/ai-sprint-alliance