Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

INFN

Systems of Neural Networks in HEP

…and the ICSC

Piergiulio Lenzi - UniFi

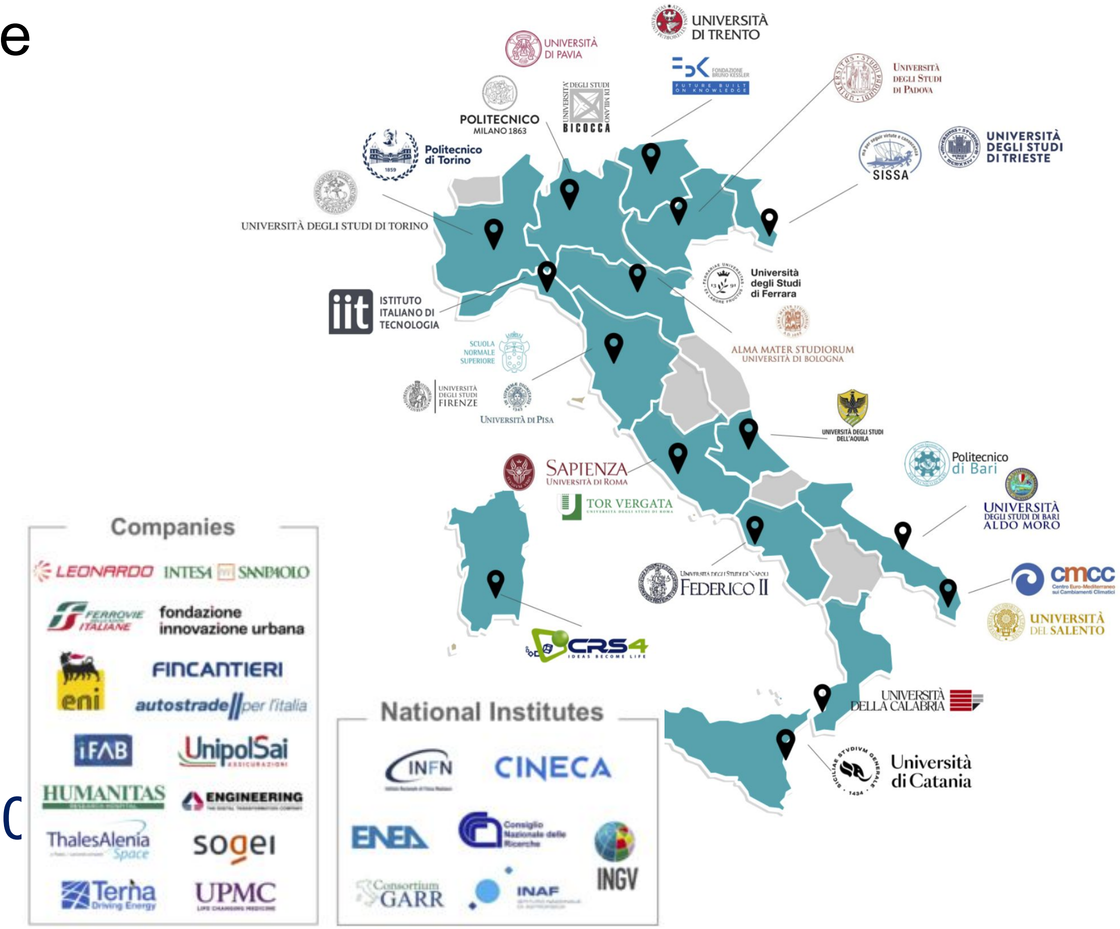
ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

Third ML-INFN Hackathon

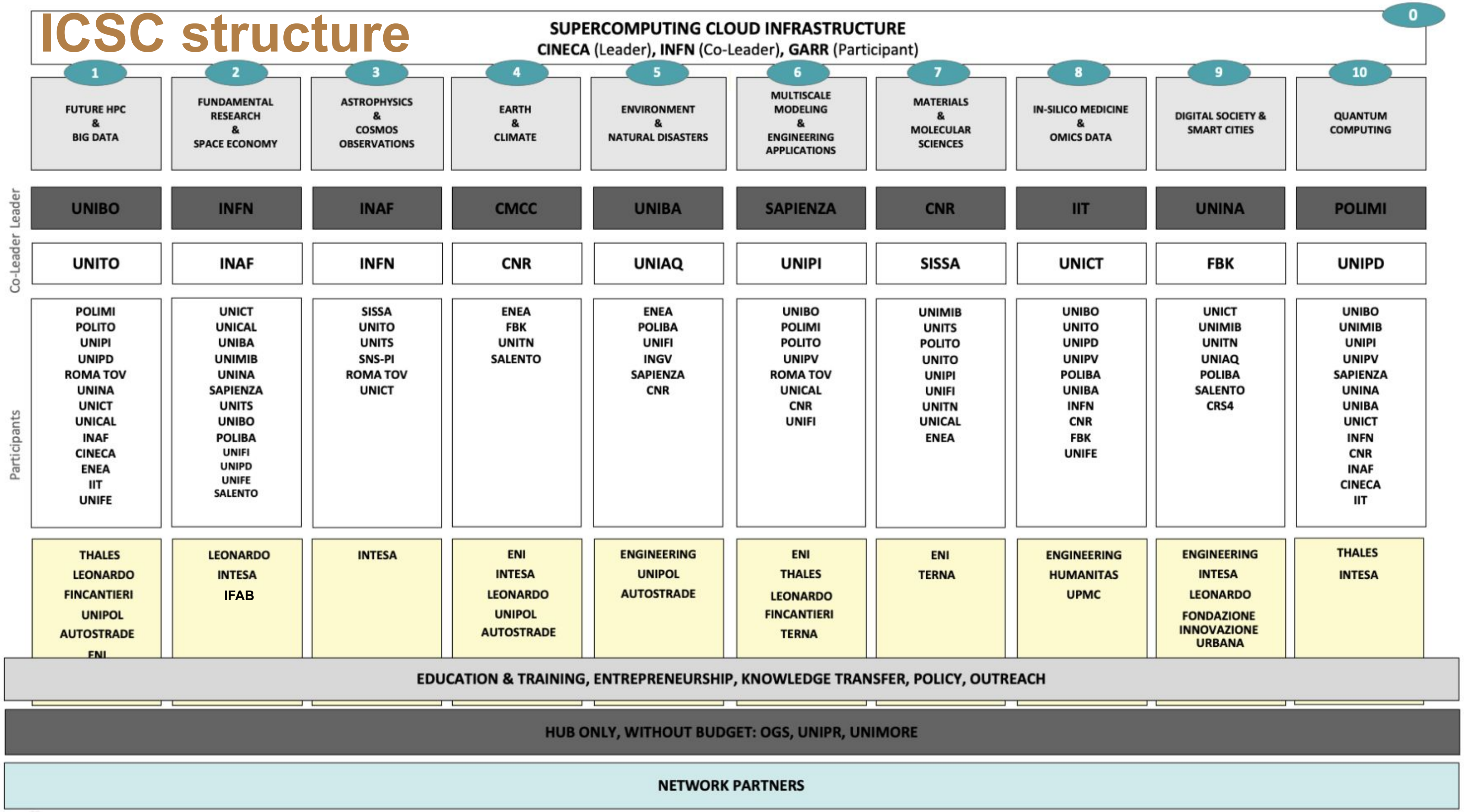
Bari, 22 Novembre 2022

The ICSC (or CN1, or Centro Nazionale 1 or Centro Nazionale HPC)

- Funded in the framework of PNRR (Piano Nazionale Ripresa e Resilienza) for **3 years**
- **Financed as one of 5 National Centers** aiming to foster innovation in selected research fields, via the realization of large-scale, state-of-the-art infrastructures and labs
- **CN1 started its activities on Sept 1st 2022**
- Headquarters at the **Bologna tecnopolo**
- In numbers:
  - 34 Universities and Research institutions
  - 15 major Italian companies
  - 1500 researchers committed + 250 PhD and 250 LD to hire


Università di Pavia
Università di Trento
FBK
Università degli Studi di Padova
Politecnico Milano 1863
Università degli Studi di Milano Bicocca
Politecnico di Torino
SISSA
Università degli Studi di Trieste
Università degli Studi di Torino
Università degli Studi di Ferrara
iit Istituto Italiano di Tecnologia
Alma Mater Studiorum Università di Bologna
Scuola Normale Superiore
Università degli Studi Firenze
Università di Pisa
Università degli Studi dell'Aquila
Sapienza Università di Roma
Tor Vergata
Politecnico di Bari
Università degli Studi di Bari Aldo Moro
Federico II
cmcc
Università del Salento
CRS4
Università della Calabria
Università di Catania
Companies
Leonardo
Intesa Sanpaolo
Ferrovie dello Stato Italiane
fondazione innovazione urbana
eni
Fincantieri
autostrade per l'italia
iFAB
UnipolSai
Humanitas
Engineering
ThalesAlenia Space
sogei
Terna
UPMC
National Institutes
INFN
CINECA
ENEA
Consiglio Nazionale delle Ricerche
INGV
Consortium GARR
INAF

ICSC structure



0 SUPERCOMPUTING CLOUD INFRASTRUCTURE
CINECA (Leader), **INFN** (Co-Leader), **GARR** (Participant)

	1	2	3	4	5	6	7	8	9	10
	FUTURE HPC & BIG DATA	FUNDAMENTAL RESEARCH & SPACE ECONOMY	ASTROPHYSICS & COSMOS OBSERVATIONS	EARTH & CLIMATE	ENVIRONMENT & NATURAL DISASTERS	MULTISCALE MODELING & ENGINEERING APPLICATIONS	MATERIALS & MOLECULAR SCIENCES	IN-SILICO MEDICINE & OMICS DATA	DIGITAL SOCIETY & SMART CITIES	QUANTUM COMPUTING
Leader	UNIBO	INFN	INAF	CMCC	UNIBA	SAPIENZA	CNR	IIT	UNINA	POLIMI
Co-Leader	UNITO	INAF	INFN	CNR	UNIAQ	UNIPI	SISSA	UNICT	FBK	UNIPD
Participants	POLIMI POLITO UNIPI UNIPD ROMA TOV UNINA UNICT UNICAL INAF CINECA ENEA IIT UNIFE	UNICT UNICAL UNIBA UNIMIB UNINA SAPIENZA UNITS UNIBO POLIBA UNIFI UNIPD UNIFE SALENTO	SISSA UNITO UNITS SNS-PI ROMA TOV UNICT	ENEA FBK UNITN SALENTO	ENEA POLIBA UNIFI INGV SAPIENZA CNR	UNIBO POLIMI POLITO UNIPV ROMA TOV UNICAL CNR UNIFI	UNIMIB UNITS POLITO UNITO UNIPI UNIFI UNITN UNICAL ENEA	UNIBO UNITO UNIPD UNIPV POLIBA UNIBA INFN CNR FBK UNIFE	UNICT UNIMIB UNITN UNIAQ POLIBA SALENTO CRS4	UNIBO UNIMIB UNIPI UNIPV SAPIENZA UNINA UNIBA UNICT INFN CNR INAF CINECA IIT
	THALES LEONARDO FINCANTIERI UNIPOL AUTOSTRADE ENI	LEONARDO INTESA IFAB	INTESA	ENI INTESA LEONARDO UNIPOL AUTOSTRADE	ENGINEERING UNIPOL AUTOSTRADE	ENI THALES LEONARDO FINCANTIERI TERNA	ENI TERNA	ENGINEERING HUMANITAS UPMC	ENGINEERING INTESA LEONARDO FONDAZIONE INNOVAZIONE URBANA	THALES INTESA

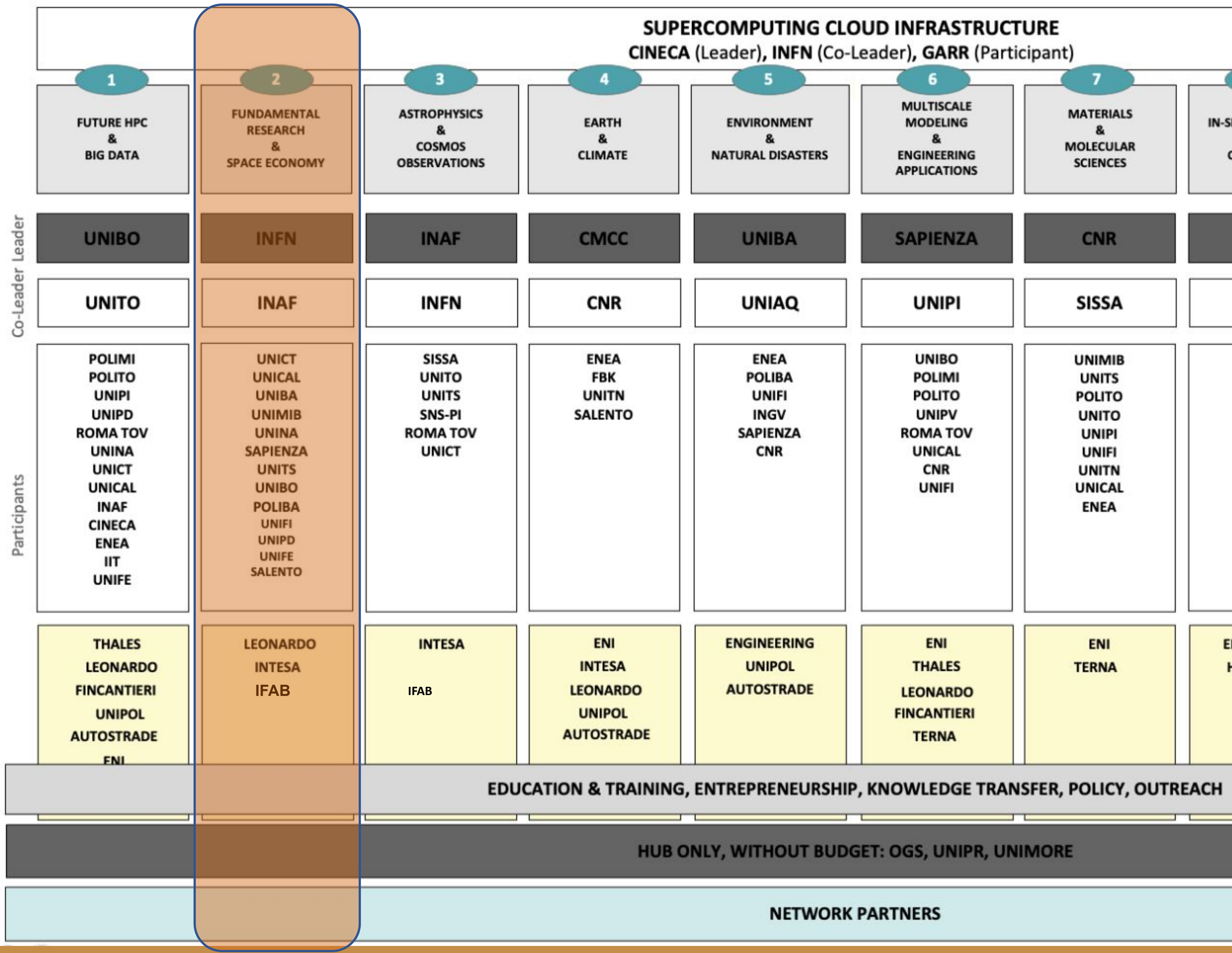
EDUCATION & TRAINING, ENTREPRENEURSHIP, KNOWLEDGE TRANSFER, POLICY, OUTREACH

HUB ONLY, WITHOUT BUDGET: OGS, UNIPR, UNIMORE

NETWORK PARTNERS

Lo Spoke 2 in ICSC: Fundamental research and space economy

- Spoke Leader: INFN
  - Sandra Malvezzi
  - Tommaso Boccali
- Spoke co-Leader: INAF
  - Antonio Stamerra
- 13 Università
- 3 privati confermati



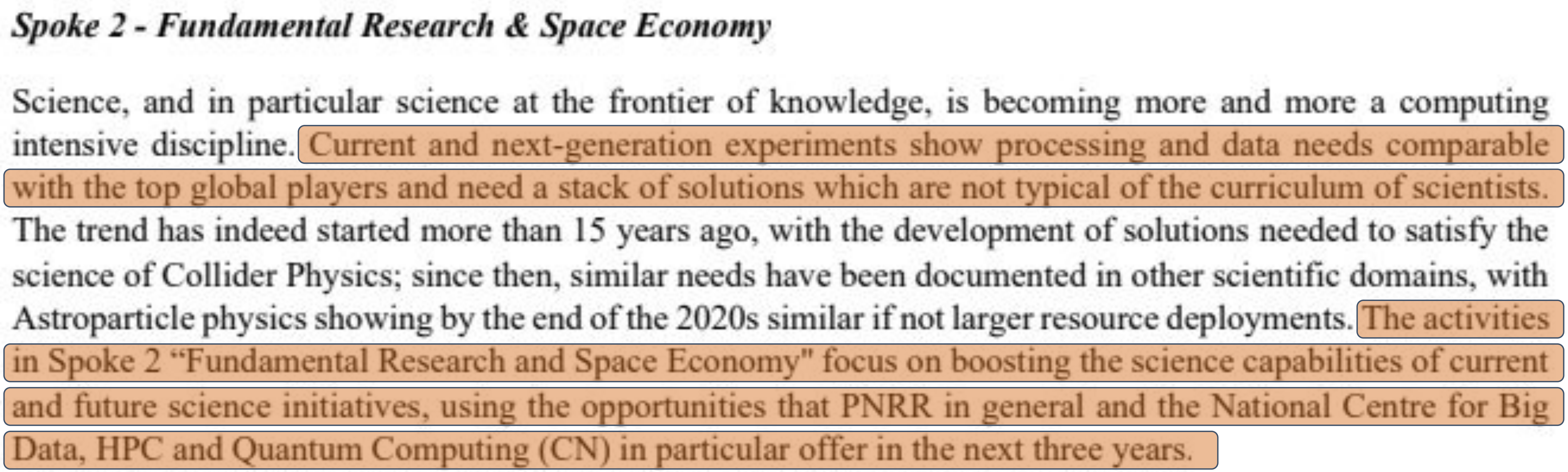
SUPERCOMPUTING CLOUD INFRASTRUCTURE
CINECA (Leader), **INFN** (Co-Leader), **GARR** (Participant)

	1	2	3	4	5	6	7	
	FUTURE HPC & BIG DATA	FUNDAMENTAL RESEARCH & SPACE ECONOMY	ASTROPHYSICS & COSMOS OBSERVATIONS	EARTH & CLIMATE	ENVIRONMENT & NATURAL DISASTERS	MULTISCALE MODELING & ENGINEERING APPLICATIONS	MATERIALS & MOLECULAR SCIENCES	IN-S C
Leader	UNIBO	INFN	INAF	CMCC	UNIBA	SAPIENZA	CNR	
Co-Leader	UNITO	INAF	INFN	CNR	UNIAQ	UNIPI	SISSA	
Participants	POLIMI POLITO UNIPI UNIPD ROMA TOV UNINA UNICT UNICAL INAF CINECA ENEA IIT UNIFE	UNICT UNICAL UNIBA UNIMIB UNINA SAPIENZA UNITS UNIBO POLIBA UNIFI UNIPD UNIFE SALENTO	SISSA UNITO UNITS SNS-PI ROMA TOV UNICT	ENEA FBK UNITN SALENTO	ENEA POLIBA UNIFI INGV SAPIENZA CNR	UNIBO POLIMI POLITO UNIPV ROMA TOV UNICAL CNR UNIFI	UNIMIB UNITS POLITO UNITO UNIPI UNIFI UNITN UNICAL ENEA	
	THALES LEONARDO FINCANTIERI UNIPOL AUTOSTRADE ENI	LEONARDO INTESA IFAB	INTESA IFAB	ENI INTESA LEONARDO UNIPOL AUTOSTRADE	ENGINEERING UNIPOL AUTOSTRADE	ENI THALES LEONARDO FINCANTIERI TERNA	ENI TERNA	E H

EDUCATION & TRAINING, ENTREPRENEURSHIP, KNOWLEDGE TRANSFER, POLICY, OUTREACH

HUB ONLY, WITHOUT BUDGET: OGS, UNIPR, UNIMORE

NETWORK PARTNERS



Spoke 2 - Fundamental Research & Space Economy

Science, and in particular science at the frontier of knowledge, is becoming more and more a computing intensive discipline. Current and next-generation experiments show processing and data needs comparable with the top global players and need a stack of solutions which are not typical of the curriculum of scientists. The trend has indeed started more than 15 years ago, with the development of solutions needed to satisfy the science of Collider Physics; since then, similar needs have been documented in other scientific domains, with Astroparticle physics showing by the end of the 2020s similar if not larger resource deployments. The activities in Spoke 2 "Fundamental Research and Space Economy" focus on boosting the science capabilities of current and future science initiatives, using the opportunities that PNRR in general and the National Centre for Big Data, HPC and Quantum Computing (CN) in particular offer in the next three years.

- Activities are organized in 6 work packages (WP)
  - 3 are driven by scientific needs (have the need to perform a calculation and search for technologies)
  - 3 are driven by technological assets (have the technology and search for applications)

WP 1-2-3

- **WP2.1: Design and development of science-driven tools and innovative algorithms for Theoretical Physics**
  - ***Main Activities***: development of algorithms, codes, and computational strategies for the simulation of physical theories and models, towards pre-Exascale and Exascale architectures. Theoretical research projects in domains already using HPC solutions.
- **WP2.2: Design and development of science-driven tools and innovative algorithms for Experimental High Energy Physics**
  - ***Main Activities***: selection, data reduction, simulation and reconstruction algorithms (either via explicit programming or large scale Machine Learning solutions) for HEP experiments (LHC, Future Colliders, KEK, IHEP, neutrino experiments...), with applications ranging from innovative triggers to distributed analysis techniques.
- **WP2.3: Design and development of science-driven tools and innovative algorithms for Experimental Astroparticle Physics and Gravitational Waves**
  - ***Main Activities***: data reduction, reconstruction and time cross-correlation algorithms, data selection and simulations of astroparticle and gravitational waves experiments, tools for cross-correlations and pattern recognition in multi-messenger physics, including novel implementations using techniques like Machine Learning.

WP 4 - 5 - 6

- **WP2.4: Boosting the computational performance of Theoretical and Experimental Physics algorithms**
  - ***Main Activities***: porting of applications to GPUs and heterogeneous architectures (e.g., scalability of scientific codes and applications on GPU/CPU many-cores clusters, local and remote offloading, mission-critical algorithms on FPGAs, ...). The solutions and tools implemented during the project will be easily extendable to other scientific domains of the Centre and to the industrial partners in the Spoke; moreover, the personnel trained within the Centre will help to spread and boost the application of HPC methodologies to Italian academic and industrial fields, for a comprehensive advancement of the Italian system.
- **WP2.5: Architectural Support for Theoretical and Experimental Physics Data Management on the Distributed CN infrastructure**
  - ***Main Activities***: support for the adaptation of existing applications on the data-lake distributed infrastructure, and via innovative computational models (for example sharing of gauge configurations in lattice field theories, long- term data preservation, streaming access to data, tiered storage solutions, ...). The solutions implemented will be tailored to the needs of the scientific fields, easily extendible not only to the nearby scientific domains in the Centre, but also to all academic and industrial realities where needs to access distributed computing and large amounts of data exist. In particular, the industrial partners in the Spoke have expressed interest in using the same technologies for their specific use cases.
- **WP2.6: Cross–domain Initiatives**
  - ***Main Activities***: optimization and adaptation of widely used software packages on the national Centre infrastructure, like Geant4 or FLUKA or generic high-performance techniques for data access/analysis; statistical and AI-based tools; data-interpretations tools. In the context of the Space Economy Italian Strategy, develop and deploy techniques to access, analyse and process the data from the Mirror Copernicus program, creating the conditions to enable radically innovative services. In particular, enable thorough and continuous observation programs for global and local processes, allowing external partners to operate a large variety of services, including the planning for emergencies, risks and resources.

Kick-off and current activities

- Kick-off of the spoke 2 on Oct 13th (see the slides for more info)
- Kick-off of the CN1 this weekend with national authorities in Bologna

All WPs are currently in an investigative phase, in which **use cases for the new infrastructure are being polled**

Aim and broad brush roadmap:

- Use case identified within year 1
- Testbeds in place by year 2
- Production by the end of the project

“Yes, yes, but what does it do for me?”

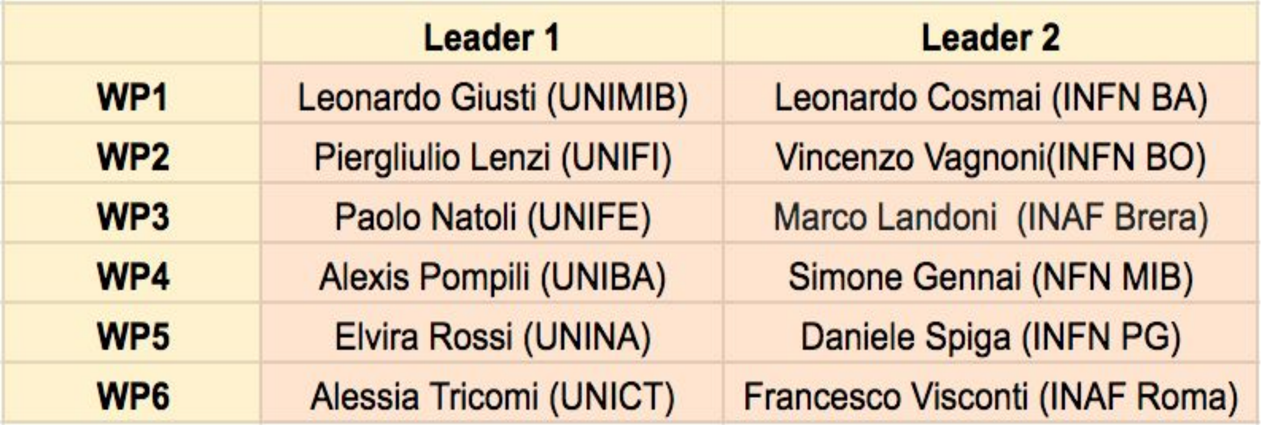
- **Watch out for job opportunities!**
- It gives you a **platform to make requests to**: come with a use case, with a request of resources and a plan, we will try to fit it. (caveat: not something you need “tomorrow”)
- Join the discussion and announcement lists
  - https://lists.infn.it/sympa/info/cn1-spoke2-wp2-all (and similar for other WPs)
- Watch the indico category
  - https://agenda.infn.it/category/1774/

The CN1 is being shaped right now

We all can contribute shape it

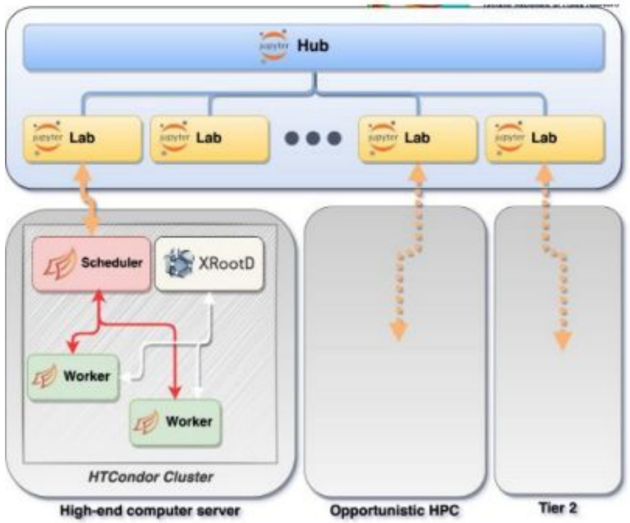
to match the needs of our field.

Spoke 2 leaders



	Leader 1	Leader 2
WP1	Leonardo Giusti (UNIMIB)	Leonardo Cosmai (INFN BA)
WP2	Piergliulio Lenzi (UNIFI)	Vincenzo Vagnoni(INFN BO)
WP3	Paolo Natoli (UNIFE)	Marco Landoni (INAF Brera)
WP4	Alexis Pompili (UNIBA)	Simone Gennai (NFN MIB)
WP5	Elvira Rossi (UNINA)	Daniele Spiga (INFN PG)
WP6	Alessia Tricomi (UNICT)	Francesco Visconti (INAF Roma)

WP2 - how it looks like today

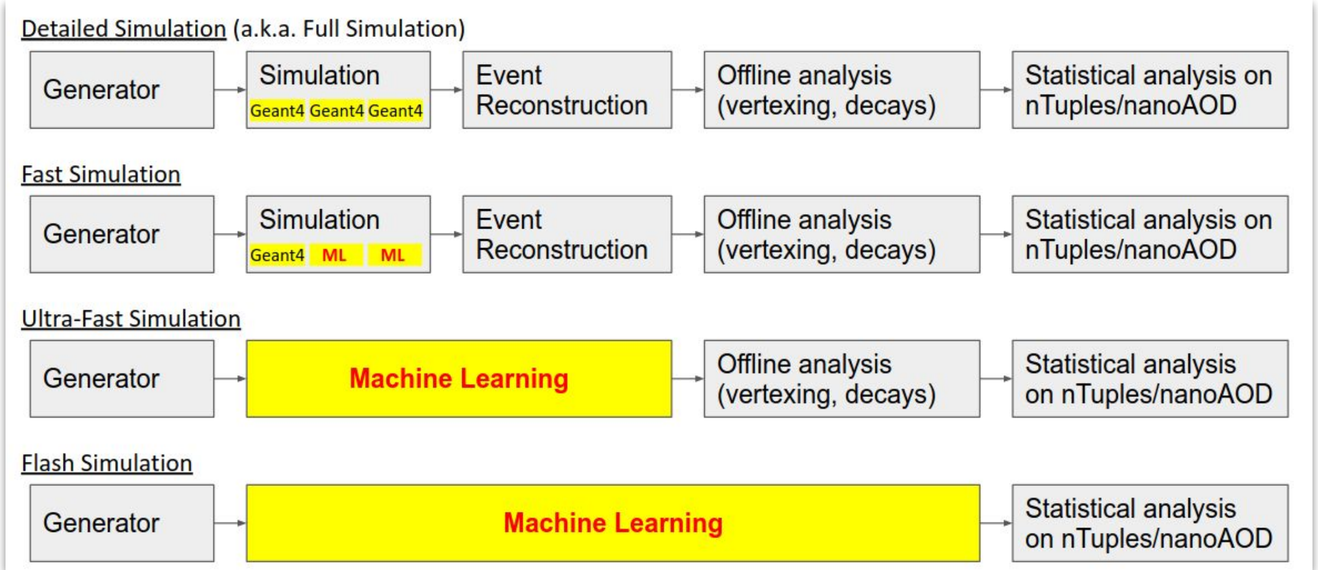


- ~ 75 people expressed interest, i.e. are in the mailing list (**are you?**)
- Expressed Needs are in terms of
  - Algorithms □ Machine Learning, and porting on heterogeneous infrastructures
  - Infrastructures → high throughput distributed analysis

Strong interactions with WP4 and WP5

ML based Simulation - Fast, *Faster*, *Flash* (I)

ML-based fast simulation has strong traction (LCHb, CMS, ATLAS) in Italy → Can profit massively from ICSC



Detailed Simulation (a.k.a. Full Simulation)

Generator → Simulation (Geant4 Geant4 Geant4) → Event Reconstruction → Offline analysis (vertexing, decays) → Statistical analysis on nTuples/nanoAOD

Fast Simulation

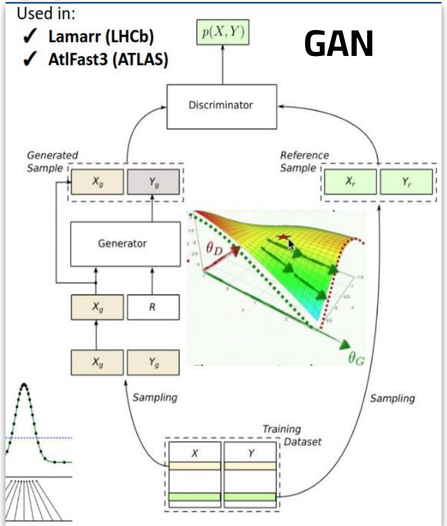
Generator → Simulation (Geant4 ML ML) → Event Reconstruction → Offline analysis (vertexing, decays) → Statistical analysis on nTuples/nanoAOD

Ultra-Fast Simulation

Generator → Machine Learning → Offline analysis (vertexing, decays) → Statistical analysis on nTuples/nanoAOD

Flash Simulation

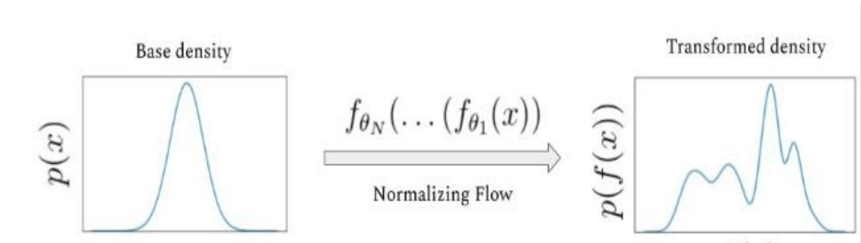
Generator → Machine Learning → Statistical analysis on nTuples/nanoAOD



Used in:
✓ Lamarr (LHCb)
✓ AtlFast3 (ATLAS)

GAN

You'll learn about this in this hackathon



Normalizing Flow/GMM

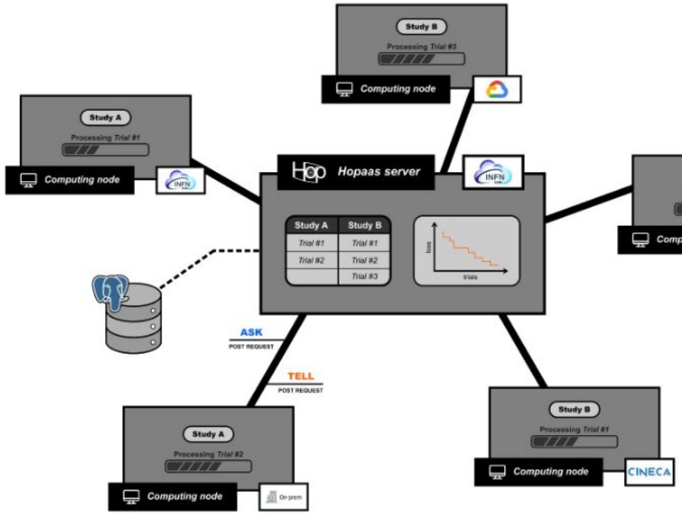
Massive GPU resources needed in **bursts of time**, to shorten the development cycle.

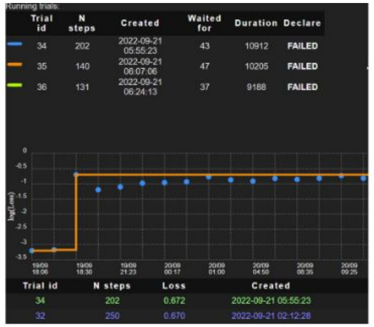
ML based Simulation - Fast, *Faster*, *Flash* (I)

Hyperparameter optimization crucial for this application → ***You'll learn about it in this hackathon***

An extreme use case that could/should be adopted also in other contexts (smaller/simpler networks)

Optimize physics output is a responsibility

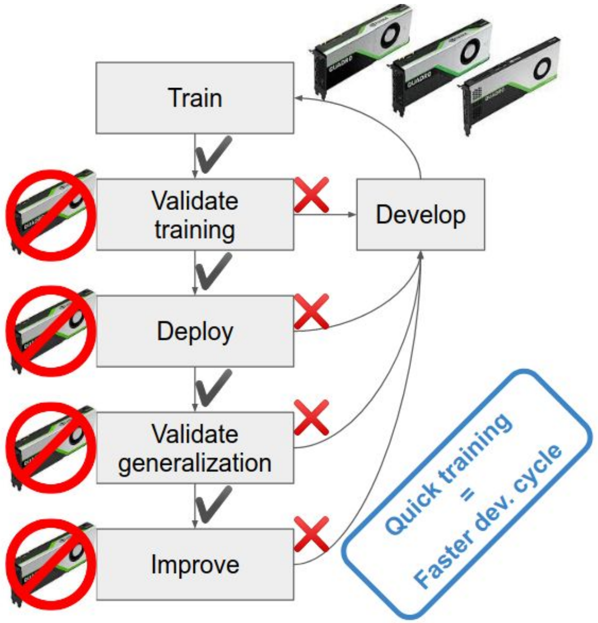



Running trials:

	Trial id	N steps	Created	Waited for	Duration	Declare
🟦	34	202	2022-09-21 05:55:23	43	10912	FAILED
🟧	35	140	2022-09-21 06:07:06	47	10205	FAILED
🟩	36	131	2022-09-21 06:24:13	37	9188	FAILED

Trial id	N steps	Loss	Created
34	202	0.672	2022-09-21 05:55:23
32	250	0.670	2022-09-21 02:12:28

Experimental in-house service hosted on INFN Cloud to coordinate optimization studies across multiple sites via REST APIs.



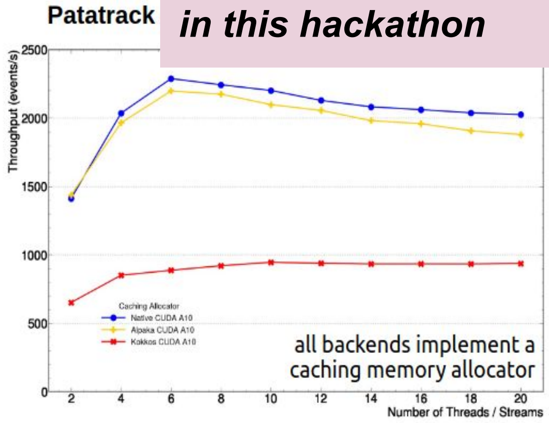
ML based event reconstruction

Trigger infrastructure

Anomaly detection strategies for trigger con uso di FPGA

Porting eterogeneo del codice di trigger su GPU

You'll learn about this in this hackathon



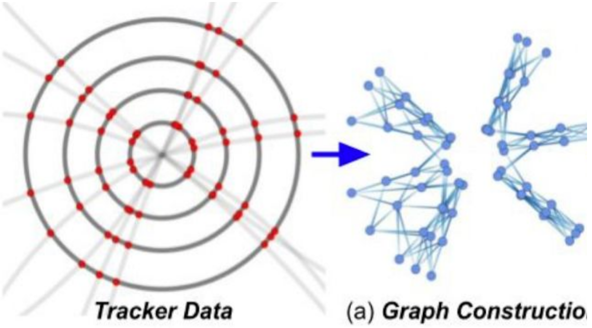
all backends implement a caching memory allocator

Vertexing e tracking

Graph NN per pattern recognition

NN per 4D vertexing

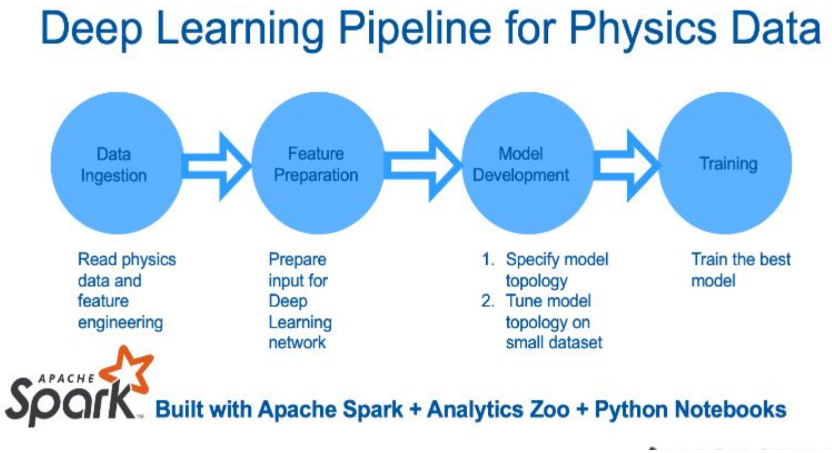
You'll learn about this in this hackathon



Particle ID

Varie iniziative nel contesto di esperimenti attuali e futuri

Pipeline di deep learning



Deep Learning Pipeline for Physics Data

Data Ingestion → Feature Preparation → Model Development → Training

Read physics data and feature engineering

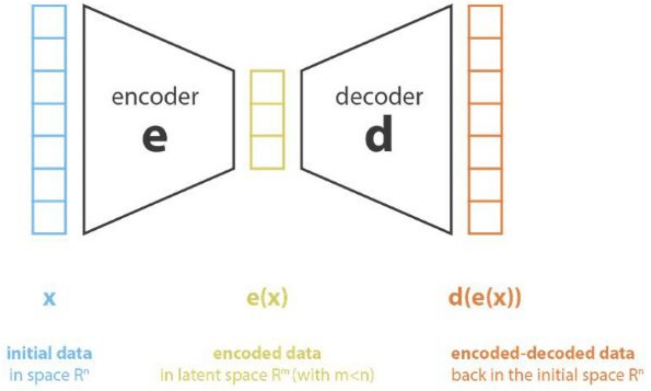
Prepare input for Deep Learning network

1. Specify model topology
2. Tune model topology on small dataset

Train the best model

APACHE Spark Built with Apache Spark + Analytics Zoo + Python Notebooks

ML based event classification

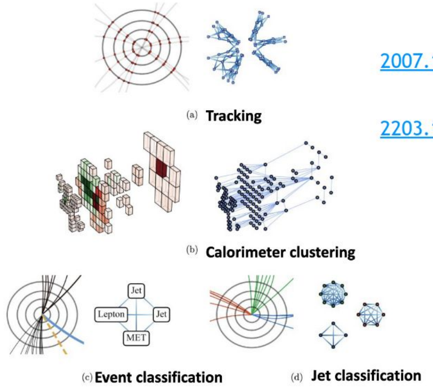

encoder
e
decoder
d
x
e(x)
d(e(x))
initial data in space R^n
encoded data in latent space R^m (with m<n)
encoded-decoded data back in the initial space R^n

Bread & butter in WP2: two main research lines

- S/B **fully supervised**
  - With domain adaptation → ***You'll learn about it in this hackathon***
- Anomaly detection
- Innovative architectures
  - Parametric NN
  - Adversarial ***You'll learn about it in this hackathon***
    - With industrial partners
  - Graph NN ***You'll learn about it in this hackathon***

All of these come with hyperparameter optimization needs

Data structure in HEP



2007.13681

2203.12852

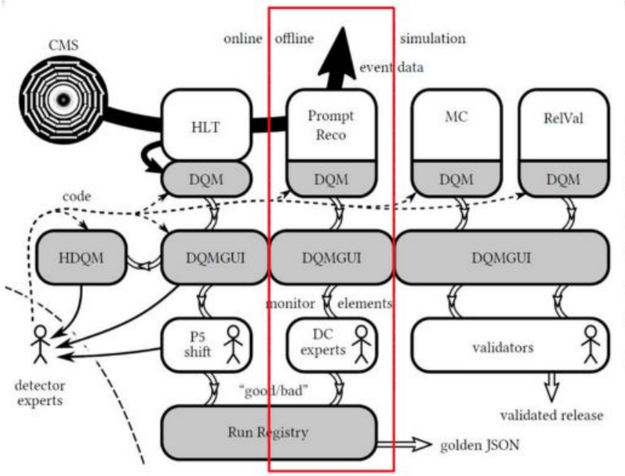
ML based cross domain initiatives

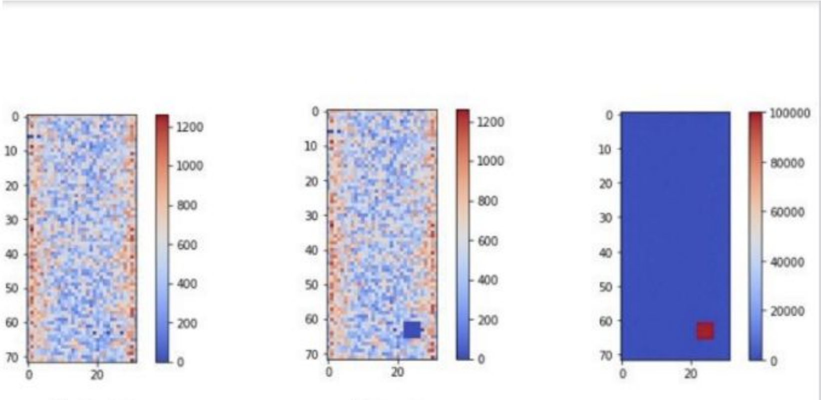
Various initiatives active in Italy that are of shared interest with industru

- Explainable AI → ***You'll learn about it in this hackathon***
- Data Quality monitoring

HEP is an ideal *playground* for XAI due to the MC truth availability

DQM leverages temporal series, uncommon in HEP, can use transformers? → ***You'll learn about it in this hackathon***

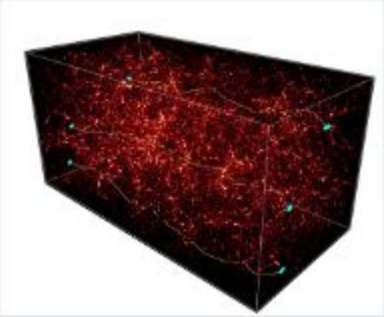




(a) Good Image (b) Dead Image (c) Hot Image

WP 3, i.e. the WP2 parallel for Astroparticle/GW

- Very preliminary assessment of use cases involved:
  - Event search and/or classification (Cherenkov, GW, DM, …)
    - Largely GPU based
    - A few requests for FPGA
    - ML techniques often proposed
  - Simulation/reduction pipelines for cosmological surveys
    - Strongly CPU/HPC based
    - Some I/O + bandwith requests
  - Request for CPU/GPU flops for existing codes (not necessarily connected to algorithm development)
  - Code devolopment:
    - Cross-analysis of datasets, joint analysis of heteregenous data
    - Machine learning development appears "almost everywhere"
    - Porting existing codes on GPU: wide spectrum of readiness
  - Request for improving data accessibility



Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

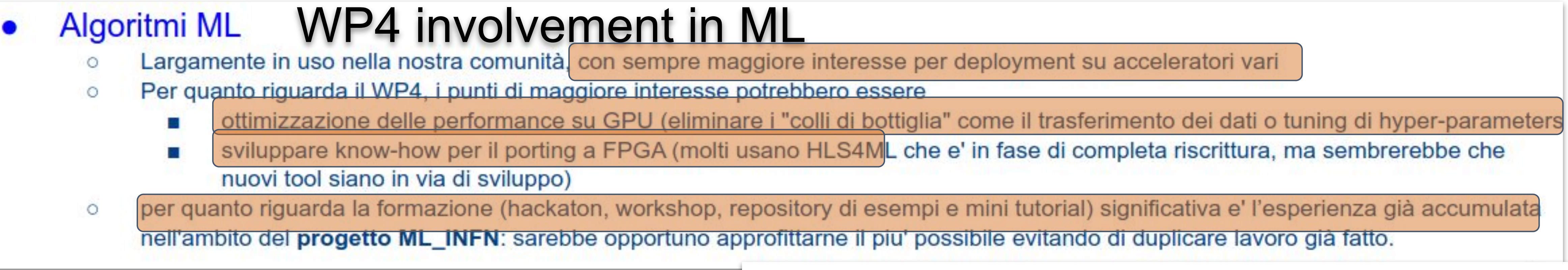
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

INFN

WP 4 and 5

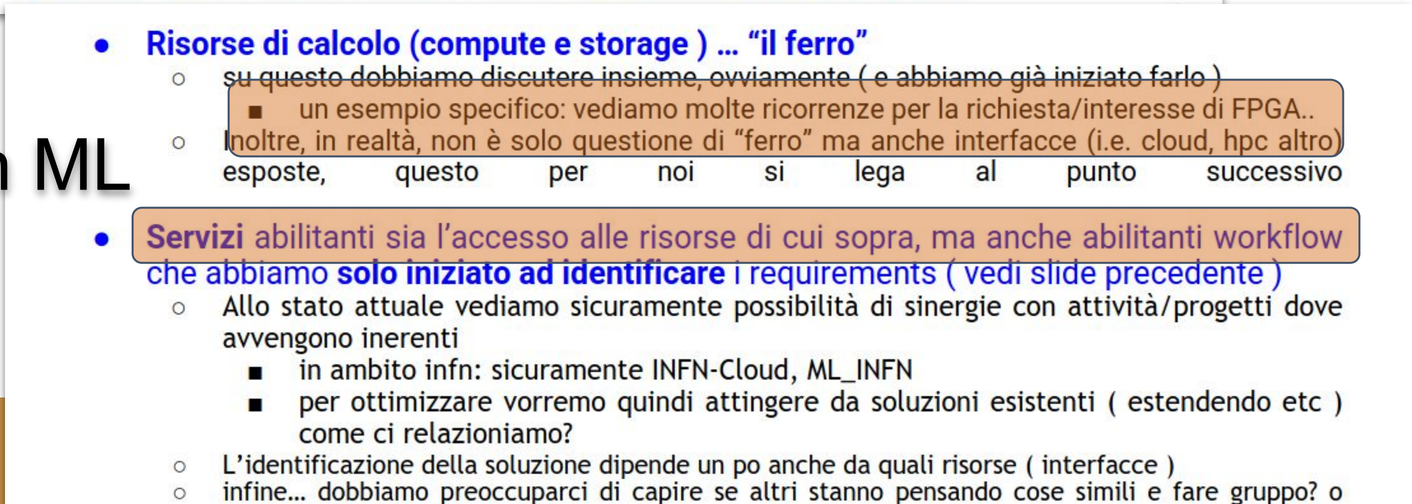
In a nutshell, WP4 proposes new technologies (e.g. GPU, FPGA etc…), WP 5 scales them up

WP4 involvement in ML



- Algoritmi ML
  - ○ Largamente in uso nella nostra comunità, con sempre maggiore interesse per deployment su acceleratori vari
  - ○ Per quanto riguarda il WP4, i punti di maggiore interesse potrebbero essere
    - ■ ottimizzazione delle performance su GPU (eliminare i "colli di bottiglia" come il trasferimento dei dati o tuning di hyper-parameters
    - ■ sviluppare know-how per il porting a FPGA (molti usano HLS4ML che e' in fase di completa riscrittura, ma sembrerebbe che nuovi tool siano in via di sviluppo)
  - ○ per quanto riguarda la formazione (hackaton, workshop, repository di esempi e mini tutorial) significativa e' l'esperienza già accumulata nell'ambito del **progetto ML_INFN**: sarebbe opportuno approfittarne il piu' possibile evitando di duplicare lavoro già fatto.

WP5 involvement in ML



- **Risorse di calcolo (compute e storage) … "il ferro"**
  - ○ su questo dobbiamo discutere insieme, ovviamente (e abbiamo già iniziato farlo)
    - ■ un esempio specifico: vediamo molte ricorrenze per la richiesta/interesse di FPGA..
  - ○ Inoltre, in realtà, non è solo questione di "ferro" ma anche interfacce (i.e. cloud, hpc altro) esposte, questo per noi si lega al punto successivo
- **Servizi** abilitanti sia l'accesso alle risorse di cui sopra, ma anche abilitanti workflow che abbiamo **solo iniziato ad identificare** i requirements (vedi slide precedente)
  - ○ Allo stato attuale vediamo sicuramente possibilità di sinergie con attività/progetti dove avvengono inerenti
    - ■ in ambito infn: sicuramente INFN-Cloud, ML_INFN
    - ■ per ottimizzare vorremo quindi attingere da soluzioni esistenti (estendendo etc) come ci relazioniamo?
  - ○ L'identificazione della soluzione dipende un po anche da quali risorse (interfacce)
  - ○ infine… dobbiamo preoccuparci di capire se altri stanno pensando cose simili e fare gruppo? o

Concluding remarks

Una panoramica delle “condizioni al contorno”

E ora buon lavoro!