Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

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

Spoke 3 - Astrophyiscis and Cosmos Observations

Innovation Grants
Interoperable Data Lake

Spoke 3 (proponente) e Spoke 2

DALLA RICERCA ALL'IMPRESA

U. Becciani

Interoperable Data Lake
Kickoff Meeting 6 Ottobre 2023
Istituto di Radioastronomia - Bologna

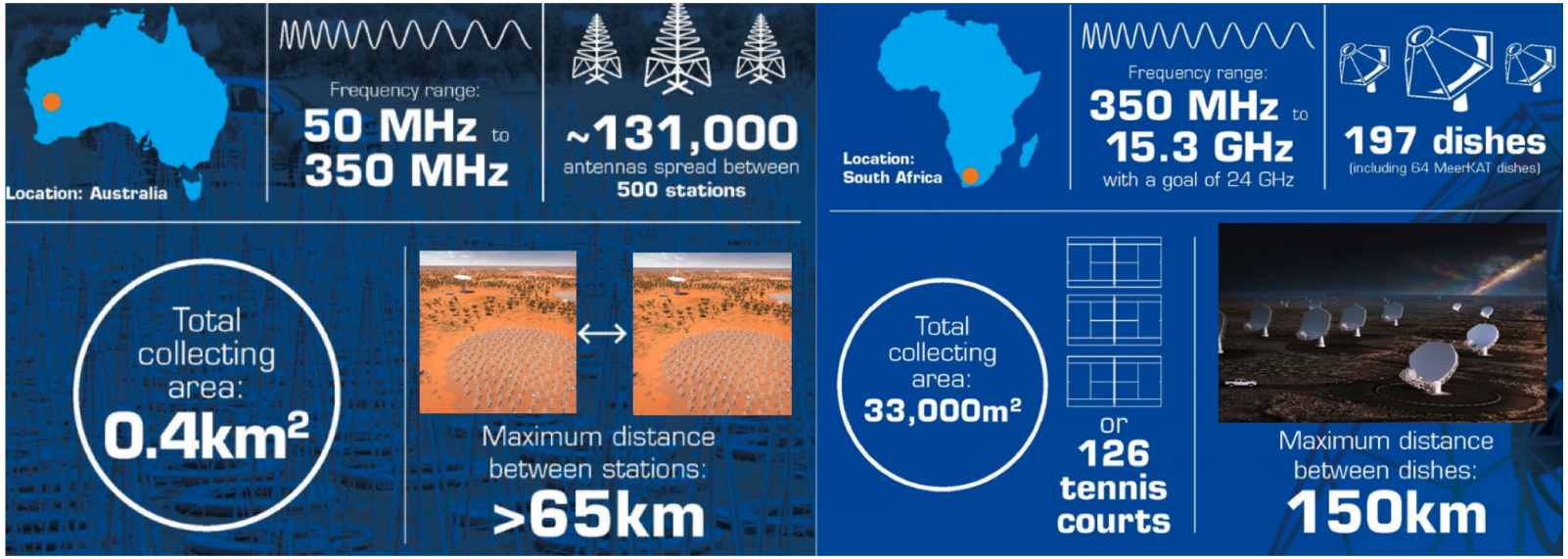
Spoke 3 Leader INAF
Co-Leader INFN

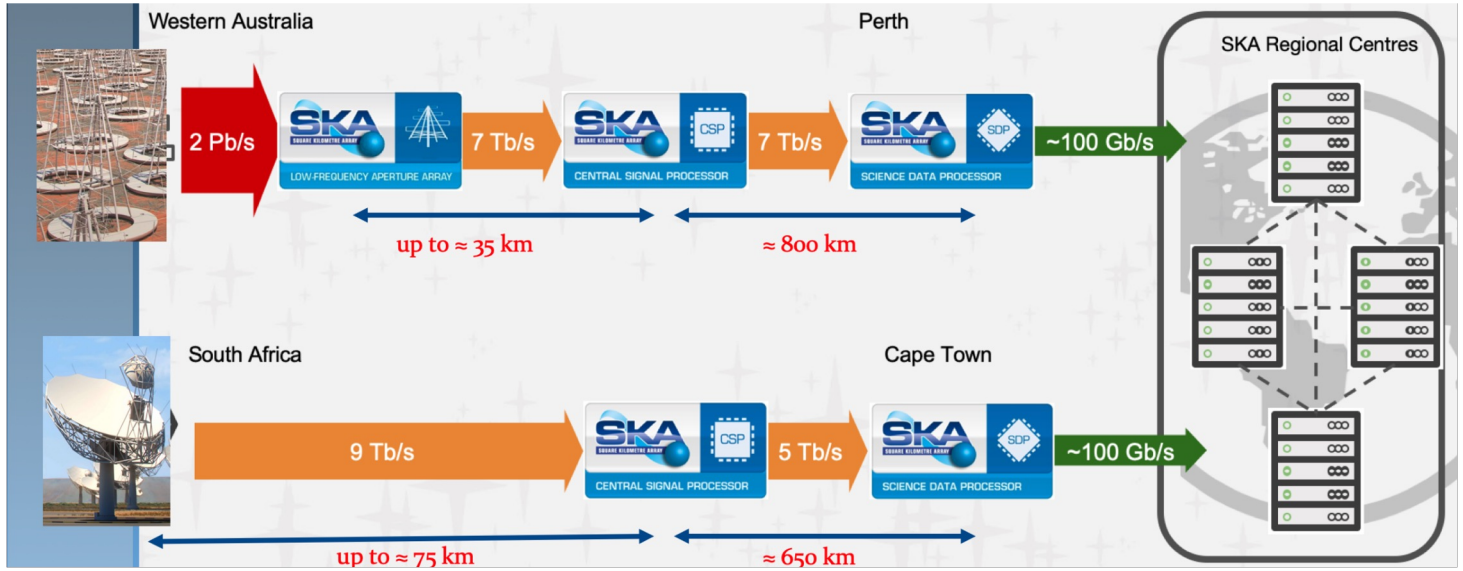
Enti Affiliati
Università Roma Tor Vergata
Università di Trieste
Università di Torino
Università di Catania
Scuola Normale Superiore - Pisa
Sissa - Trieste

Aziende Aderenti allo Spoke
Intesa SanPaolo
UnipolSai
Sogei
IFAB
Leonardo
Thales Alenia Space Italia S.p.A. (TASI)

Spoke 3 - Astrophysics and Cosmos Observations Main Objectives

Scientific Themes	Main Projects
Radioastronomy	Square Kilometer Array (**SKA**), Low Frequency Array (LOFAR2.0) e **Meerkat+**
Observational Astrophysics e time-domain	Spatial missions: **Euclid** e **Gaia, LSPE and Litebird** Legacy Survey of Space and Time (**Vera Rubin observatory**) Extremely Large Telescope (**ELT**)
High-Energy	**CTA, ASTRI,** FERMI, Dampe, HERD, AMS02, SWGO, etc
Large Scale Simulation	**HPC Theory** (P-GADGET3 -> OpenGADGET, PLUTO, plasma physics simulations, etc..)
Big Data	HPC computing processing, Management and distribution of large dataset in the Datalake, High rate analysis





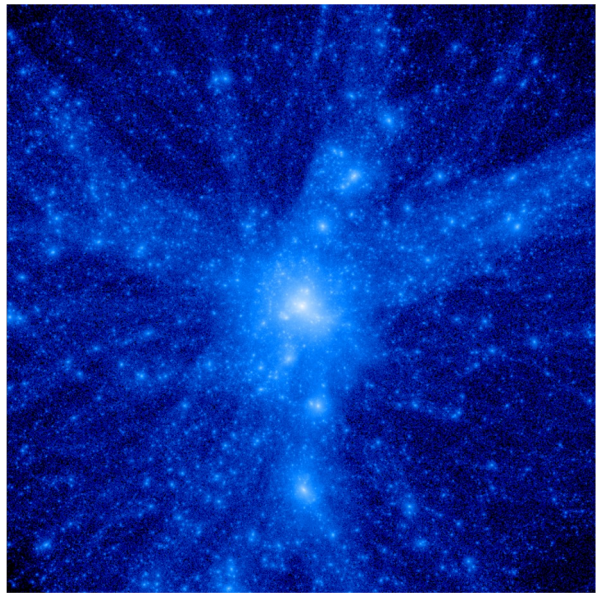
SKA

Big challange and Big Data

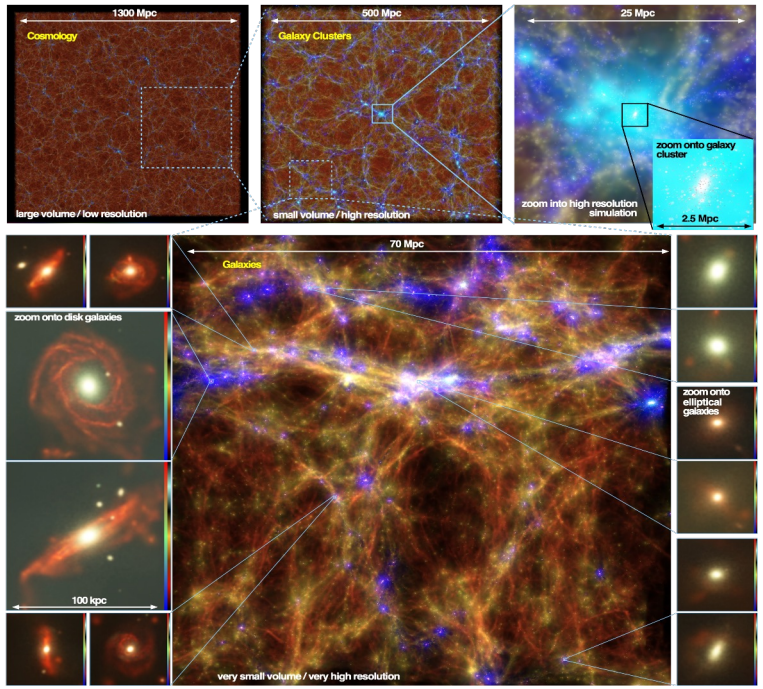
700 PB/yr.

First data science ready for 2025

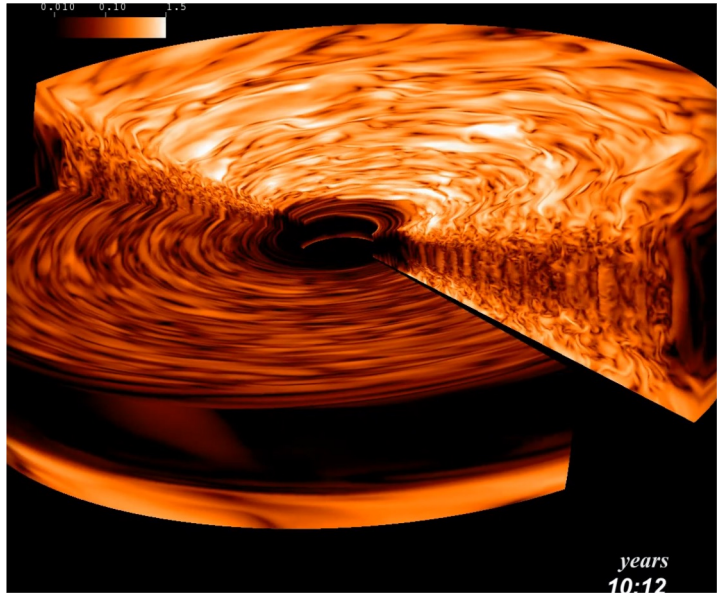
SKA-RC @ Tecnopole and CN



Dark Matter distribution: high-res simulation for cluster of galaxies
OpenGadget code



Gas density map *(Magneticum project)*
OpenGadget code



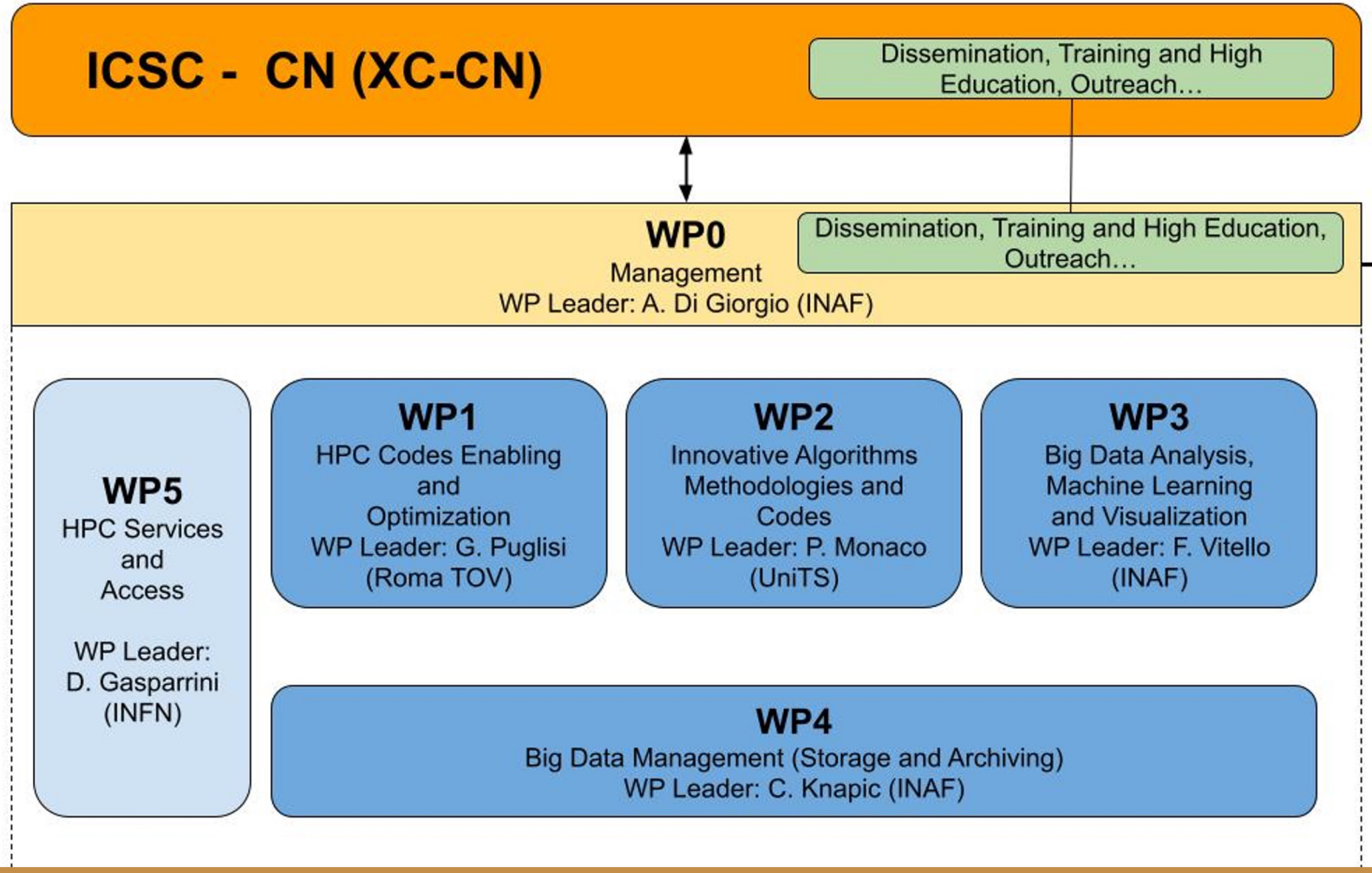
MRI-induced turbulent velocity field in a protoplanetary disk simulations.
[Credits: M. Flock (2011)]
PLUTO Code

Data knowledgement of the new instruments goes through cosmological models and simulations that need high resolution

Challenge Simulations: 1.000-5.000 nodes e 0.5 PB- 5 PB

Pre-Exascale and Exascale infrastructures are necessary

Spoke 3 – Struttura dei WP



Project Coordination Board

Spoke Leader: **U. Becciani**
co-Leader: **P. Lubrano**
System Engineer: **G. Taffoni**
Technical Supervisor: **C. Gheller**

- Spoke 0: Infrastructure
- Spoke 1: Future HPC
- Spoke 2: Fun. Res & Space Economy
- Spoke 10: Quantum Comp.

WP1: HPC Codes Enabling and Optimization.

The WP1 activity consists in the selection of a number of codes that require intensive computational resources to face the next generation of scientific challenges and performs their redesign, reimplementation and optimisation in order to effectively exploit state-of-the-art HPC solutions

⇒ **Affiliate Institution involved** (INAF 176 PMs, SNS 48 PMs, Torvergata 72PMs, SISSA 42 PMs,INFN 31 PMs, UniTo 38PMs, UniTs 43 PMS)

⇒ **Main activities description**

- Concluded the Code Selection phase,
- 17 codes to be optimized,
- Currently: the Code redesign optimization strategy,

WP2: Design of innovative algorithms, methodologies and codes towards exascale and beyond.

This WP identifies innovative algorithms and methodologies upgrading their capability to exploit, and scale on, the exascale and post exascale architectures, reintegrating the resulting improved features in codes, workflows and pipelines. The energy impact will also be specifically considered.

⇒ **Affiliate Institution involved:** (INAF 44 PM, UniTS 49 PM, SISSA 47 PM, UniTO 30 PM, UniCT 47 PM, INFN 13 PM, SNS 49 PM, RomaTOV 84 PM)

⇒ **Main activities description:**

Science cases definition,
code/algorithm identification, working in close contact with WP1: 21 codes selected
tracking progress for all participating codes

WP3: Big Data Analysis, Machine Learning and Visualization

The WP3 develops a prototype framework of data analysis, based on Machine Learning (ML) and Visualization tools exploiting diverse computing platforms and combining them with exascale application.

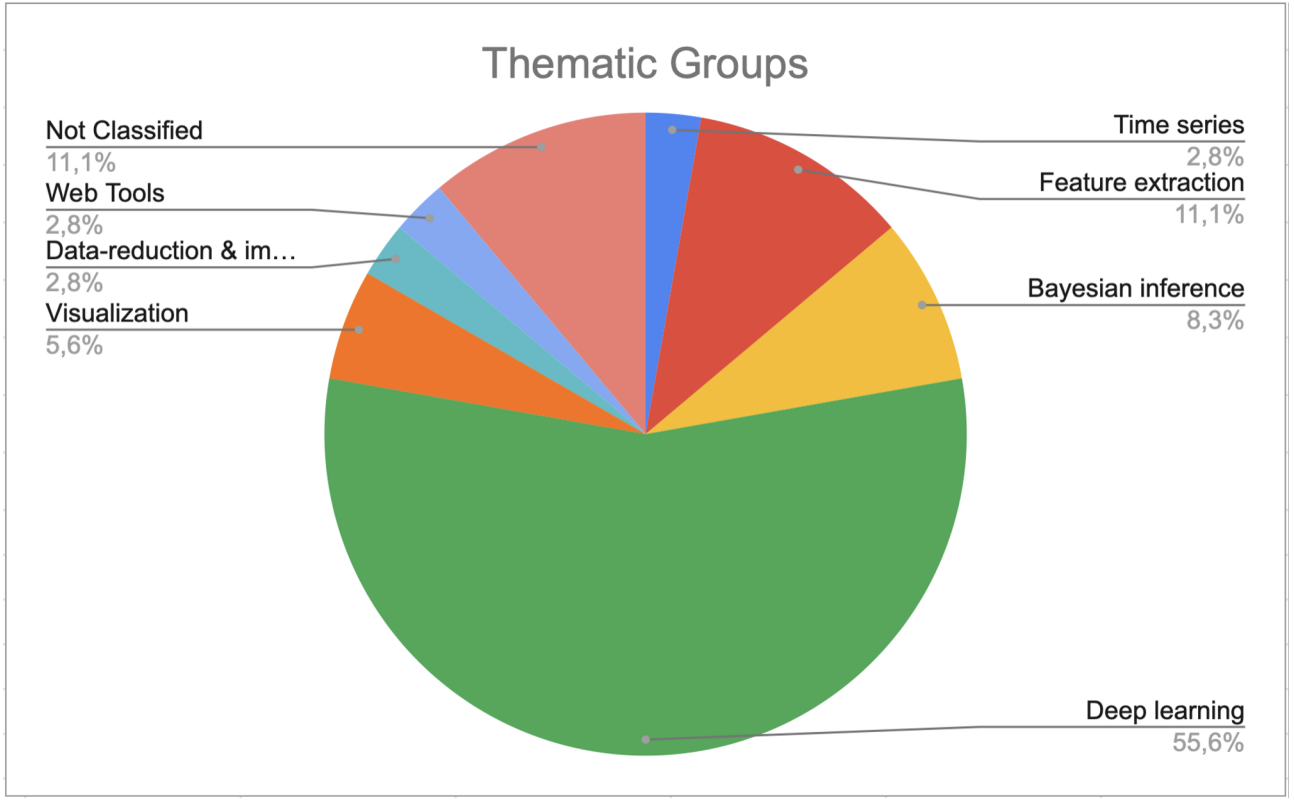
⇒ **Affiliate Institution involved:** (INAF 225 PM, UniTS 25 PM, SISSA 47 PM, Unito 34 PM, UniCT 62,.2 PM, INFN 118 PM, SNS 49PM, ROMTOV 30.5 PM)

⇒ **Main activities description:**

Science use cases definition,
use case to thematic groups,
collecting ideas for Key Science Cases,

⇒ **Thematic groups**

Time series: 1 use case;
Feature extraction: 4 use cases;
Bayesian inference: 3 use cases;
Deep learning: 20 use cases;
Visualization: 2 use cases;
Data-reduction & imaging: 1 use case;
Web Tools: 1 use case;
Not Classified: 4 use cases.



WP4: Storage and Archive

⇒ **Scientific progress for the A&A community**

Simulated and observed data storage and archiving, provided with well descriptive data models Virtual Observatory and FAIR compliance; publication of the key science data products and observing data. Improvement of features and capabilities of state of the art software solutions for data distribution; Enhancement of ergonomic science portals;

⇒ **Industrial contributions**

Contribution to the interspoke industrial project “Interoperable Data Lake” by INFN + Leonardo + Thales in which will be implied frontiers block-chain paradigms and innovative databases using industrial standards.

⇒ **Infrastructure Requirements**

~10 TB per participating data product type, plus resources for Science Key Projects; Storage space for data lake implementation.

⇒ **Risk & Criticality Assessment of impact**

Lack of manpower expertise in archiving and Big Data Handling (moderate)

Lot of declared effort from scientific counterpart, little real availability (medium)

WP4: Storage and Archive

The WP4 develops a prototype of distributed archive, FAIR and VO compliant to supply Big Data storage of the Scientific products foreseen to come from Key Science projects and use case both about Simulations, Observations and their Analysis.

⇒ **Affiliate Institution involved:** (INAF, UniTS, UniTo, INFN ROMTOV)

⇒ **Main activities description:**

- Science Archive use cases definition, Support in the definition of data models and metadata descriptors, Interoperable platform for data search and retrieval equipped with advanced data management features
- Use cases and suggestions coming from thematic groups,

⇒ **Thematic groups** Web tools; Platforms; Data Models; Work flow;

WP5: HPC service and Access.

The WP5 activity consists in managing, maintaining and deploying an integrated environment providing the tools for the efficient development of the work described in the other WP. WP5 will develop also the main access to data cned the activity of spoke 3

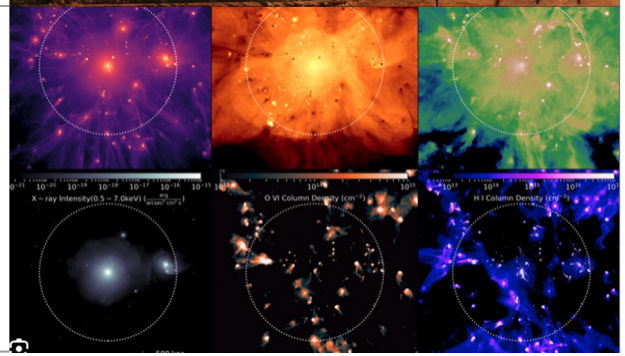
⇒ **Affiliate Institution involved** (INFN 72 PM, INAF 27 PM)

⇒ **Main activities description**: Collection of HW/SW requirements of the spoke, creation of Indico pages for meeting activities, preparation of the public website. configuration of the software repository and version tracking.

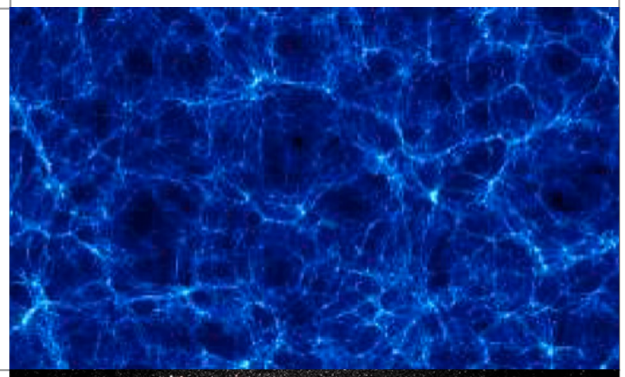
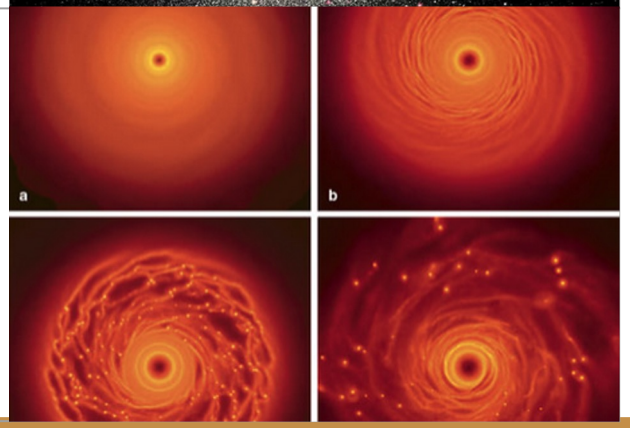
Spoke 3 - Key Science Projects - Flagships Selection

Key Science Projects represent the Spoke 3 flagship applications, main scientific and technological outcomes of the WPs R&D activities.

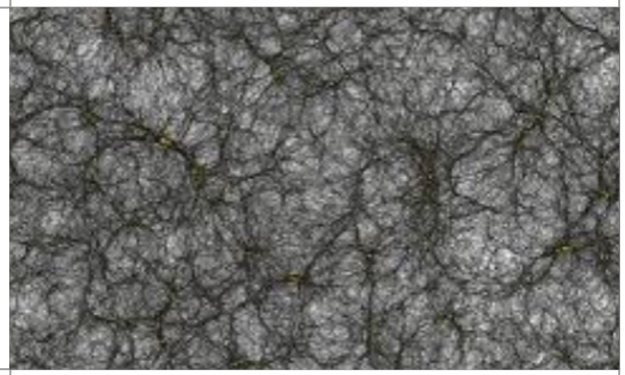
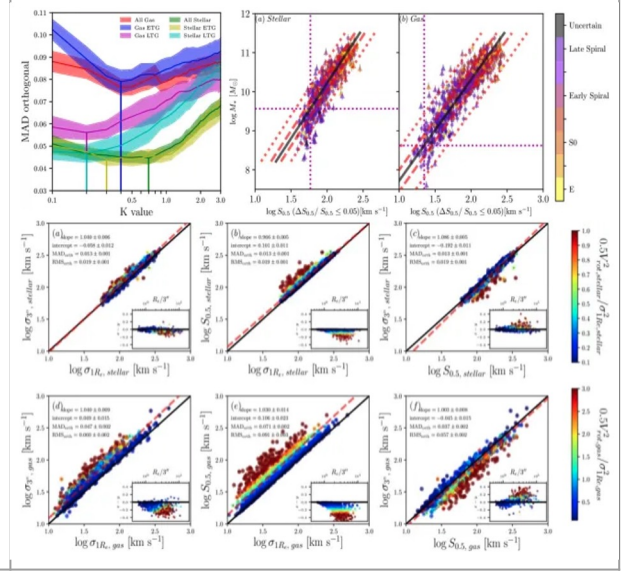
A number of projects have been proposed and are currently under evaluation.

Key Science Project	Short Description / Objectives	WPs	
1. SKA (and Pathfinders) Regional Center HPC Services	Leverage Data Lake Technologies developed in WP4 and IDL for SKA and LOFAR Develop and deploy HPC data processing services for SKA and LOFAR Integrate remote visualization services	WP1, WP2, WP3, WP4	
2. EAGER: Evolution of gAlaxies and Galaxy clustErs in high-Resolution cosmological simulations	High-resolution cosmological simulations of galaxy and galaxy cluster formation performed with a state-of-the-art code developed in Spoke3 A variety of complex physical processes are included.	WP1, WP2	

Spoke 3 - Key Science Projects - Flagships Selection (2)

Key Science Project	Short Description / Objectives	WPs	
3. SLOTH: Shedding Light On dark matter wiTH cosmological simulations	Gain insight into DM properties and structure formation in the early Universe. Key also to interpret JWST observations. Exploits all the 3500 nodes of the booster partition of Leonardo@CINECA. Each of the two runs would require about 4 days wall-clock time.	WP1, WP2	
4. Revealing the populations of compact remnants in dense stellar systems	Novel STEDDAS code to run direct N-body simulations of dense stellar systems (e.g., globular clusters) with up-to-date stellar evolution physics and unprecedented accuracy and performance.	WP1, WP2, WP3, WP4	
5. OPAL: simulating the Origins of Planets for ArieL	Simulate the origins of planets and their bulk compositions from the interplay of the planet formation process with the chemical and physical evolution of circumstellar disks. Build an extensive syntethic dataset of atmospheric compositions to test the official processing and retrieval pipelines of the Ariel space mission.	WP1, WP2, WP4	

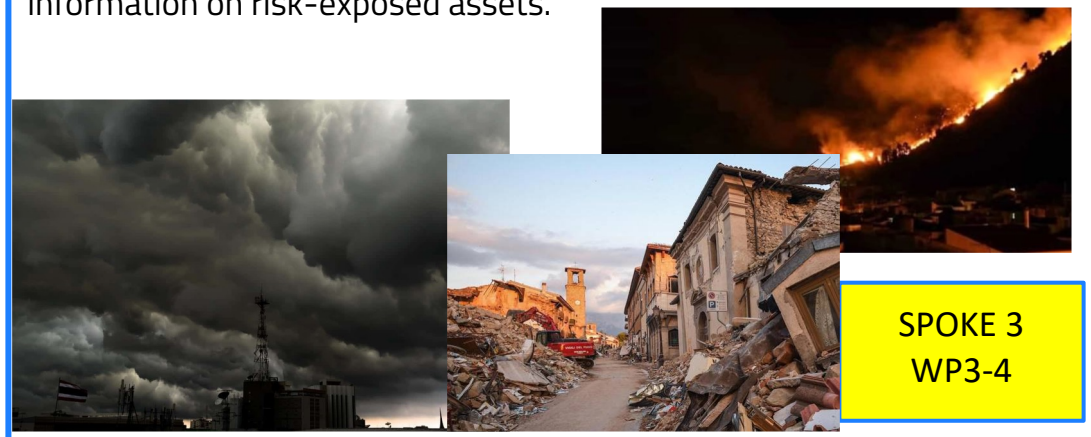
Spoke 3 - Key Science Projects - Flagships Selection (3)

Key Science Project	Short Description / Objectives	WPs	
6. EuMocks	To produce the largest set of cosmological simulations ever produced, aimed at representing the Universe sampled by the Euclid Wide Survey. These simulations will be a crucial ingredient to control errors in the cosmological results of Euclid, starting from Data Release 1 (late 2025).	WP1, WP2	
7. Multi-wavelength inference from the first billion years	Sampling empirical galaxy scaling relations, construct a large database of IGM and galaxy lightcones spanning the first billion years (Cosmic Dawn and Epoch of Reionization) Using bespoke telescope models from (e.g. SKA-low, ROMAN grism, ELT, Subaru) construct mock observations Apply simulation-based inference on multi-wavelength data to infer the properties of the unseen galaxies and cosmology	WP1, WP3, WP4	

Spoke 3 - Innovation Grants projects

1 - HaMMon (Hazard Mapping and vulnerability Monitoring)
Private companies proponents: UnipolSai and Sogei
Spoke Proponent 3. Participants Spokes: 0,1,2,4 and 5.

The project aims to develop tools and methodologies to be used in different industrial contexts for the **quantification of the impacts of extreme natural events on the Italian territory.** I The activities will involve intensive ***use of scientific visualization and artificial intelligence technologies***, especially for assessing and extracting meaningful information on risk-exposed assets.

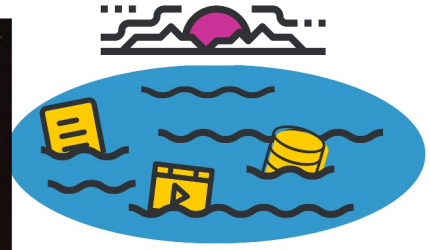


SPOKE 3
WP3-4

Budget 1.9 MEuros. 5 Spokes involved, 13 Agencies and Academic Institutions .. and more (IFAB - Leonardo - XC)
25 FTEs (aprox)
Starting TRL 5 - Target TRL 8

2 - IDL (Interoperable Data Lake)
Private companies proponents: Leonardo and Thales
Spoke Proponent 3. Participant Spoke:2

This project addresses the ***challenges of big data data*** gathering, sharing, knowledge, safety, governance, Intellectual property, training and source,. ***A federated data lake will embed services for the extraction of actionable insights from these large amounts of data and extend the use of space applications to new research lines..***
GAIA-X industrial standard *could be used to implement a Space dataspace*.



SPOKE 3
WP4

Budget 755 KEuros.
2 Spokes involved, 2 Agencies and Academic Institutions
12 FTEs (aprox)
Starting TRL 4 - Target TRL 5

Spoke 3 - Innovation Grants projects

3 - Serial Code Porting on HPC & QC
Private companies proponents: Sogei
Spoke Proponent 3. Participant Spoke: 10.

The goal of the project is to **rewrite Machine Learning (ML) algorithms**, available in their Open Source versions, in a suitable language **for HPC** computation as well as on **quantum computation** *to prevent cyber attacks* (ref. KeepCalm Project).

Budget 268 KEuros.
2 Spokes involved, 2 Agencies and Academic Institutions
4-5 FTEs (aprox)
Starting TRL 4 - Target TRL 6

SPOKE 3
WP1

4 - Fraud Detection
Private companies proponents: Intesa Sanpaolo
Spoke Proponent 10. Participants Spokes: 2, 3

In finance, fraud detection is an extremely important form of anomaly detection. Some examples are **identifying fraudulent credit card transactions and financial documents**. *Comparison between Quantum vs HPC/Cloud solutions*

Budget 546 KEuros.
2 Spokes involved, 6 Agencies and Academic Institutions
10 FTEs (aprox)
Starting TRL 3 - Target TRL 6

SPOKE 3
WP1-2-3

5 - Time series in the banking sector
Private companies proponents: Intesa Sanpaolo
Spoke Proponent 3

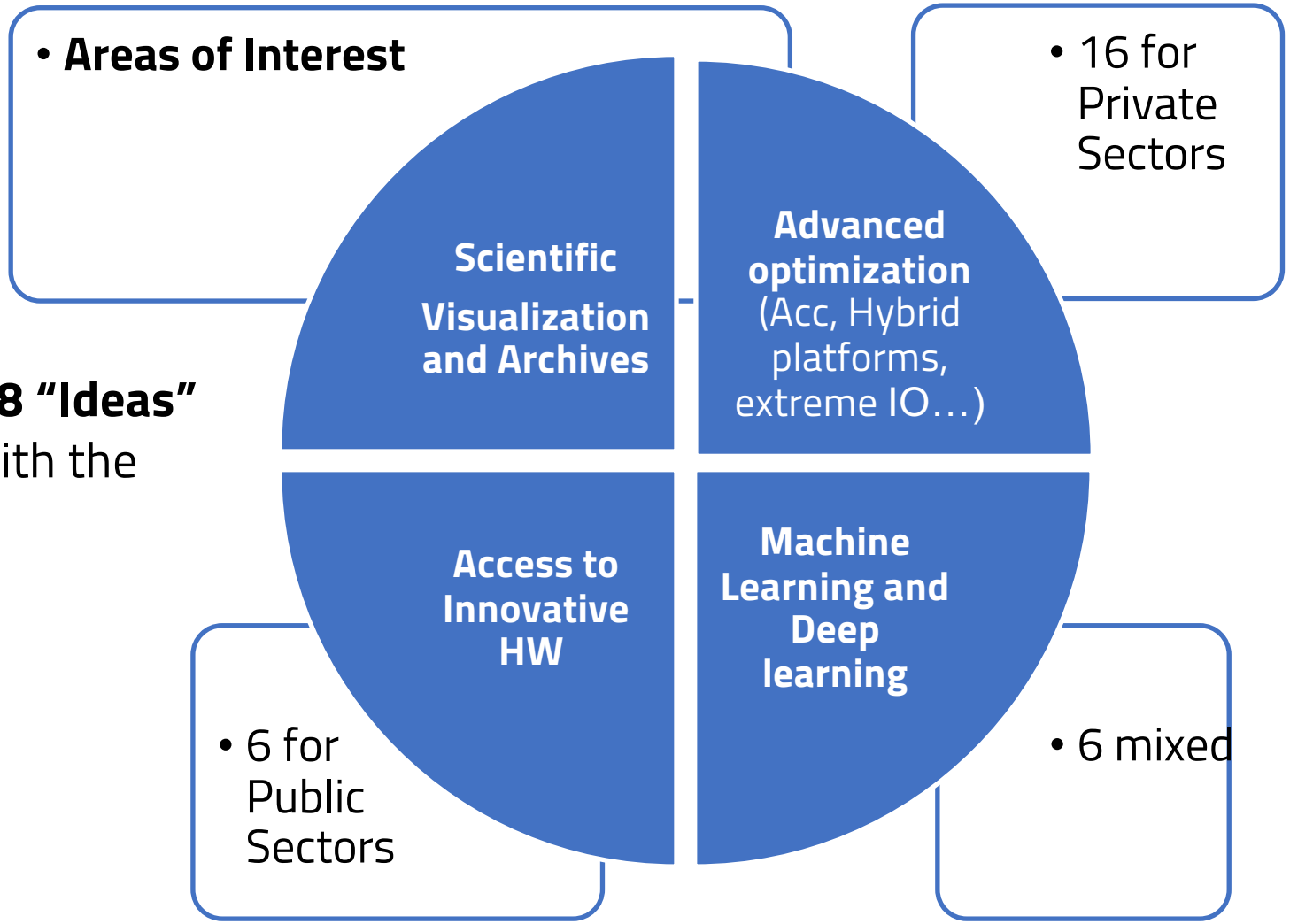
The aim of the project consists in **applying ML techniques** to solve problems in the banking sector, usually described as **time series** (TS).
Banking case: *data quality, corporate credit risk, churn analysis.* The dataset of Data Quality will be the most important.

Budget 180 KEuros.
1 Spoke involved, 2 Agencies and Academic Institutions
4 FTEs (aprox)
Starting TRL 3 - Target TRL 6

SPOKE 3
WP1-2-3

Cascade Funding – Work in progress (some work with small progress so far…)

- **WPs and Partners collected 28 "Ideas"**
- On line document to discuss with the SPOKE Assembly



Azioni

- **Affiliazione delle Aziende agli Spoke**
- **Partecipazione dei rappresentati nell'assembly**
- Attivazione reale dell'Industrial Board dello Spoke

- **Partecipazione attiva nei WP degli Spoke**
- **Report da far rifluire del lavoro degli Spoke**
- **Policy delle pubblicazioni di Spoke 3**

- **Verifica periodica con gli Spoke nei meeting in presenza e remoto (!?)**
- **Apertura ad altri partner/Stakeholder… studiare quali passi fare…**
- ***Rendicontazione***

- **Problemi ? Da discutere insieme… e correttivi da apportare**