Big Data Analysis, Machine Learning and Visualization

Work Package 3

Fabio Vitello

INAF

WP3 Objectives

This WP develops a **prototype framework of data analysis, based on Machine Learning (ML) and Visualization tools** exploiting diverse computing platforms and combining them with exascale applications. The **framework will be tailored to the ACO-S community, identifying the use case** best suited to tackle high-performance visualization tools and ML techniques. Furthermore, it copes with observational data coming from challenging large experiments.

Objective 2. Big data processing and visualization, via adopting innovative approaches (e.g. Artificial Intelligence, inference via Bayesian statistics) for the analysis of large and complex data volumes and for their exploration (e.g. in-situ visualization), capable of efficiently exploiting HPC solutions. (WP3, WP4, WP5)

- Action 2.2: To **develop new methods and/or optimize existing prototype Machine Learning applications** for the automated processing of large and complex data, produced on the Exascale systems by the ACO-S community.

- Action 2.3: To **develop and optimize existing solutions for high performance visualization**, addressing on-site and remote visualization for Exascale and post-Exascale systems.

- Action 2.4: To **develop integrated solutions starting from the outcome of the previous actions,** in order to provide a unique and optimized eco-system for Exascale platforms for big and complex data sets.

WP3 Tasks

T3.1: Requirements from AAA community

It will assess the **requirements of the community** and assess corresponding solutions in terms of Machine Learning and Visualization, exploiting the functionalities available in selected tools and/or the development of new functionalities.

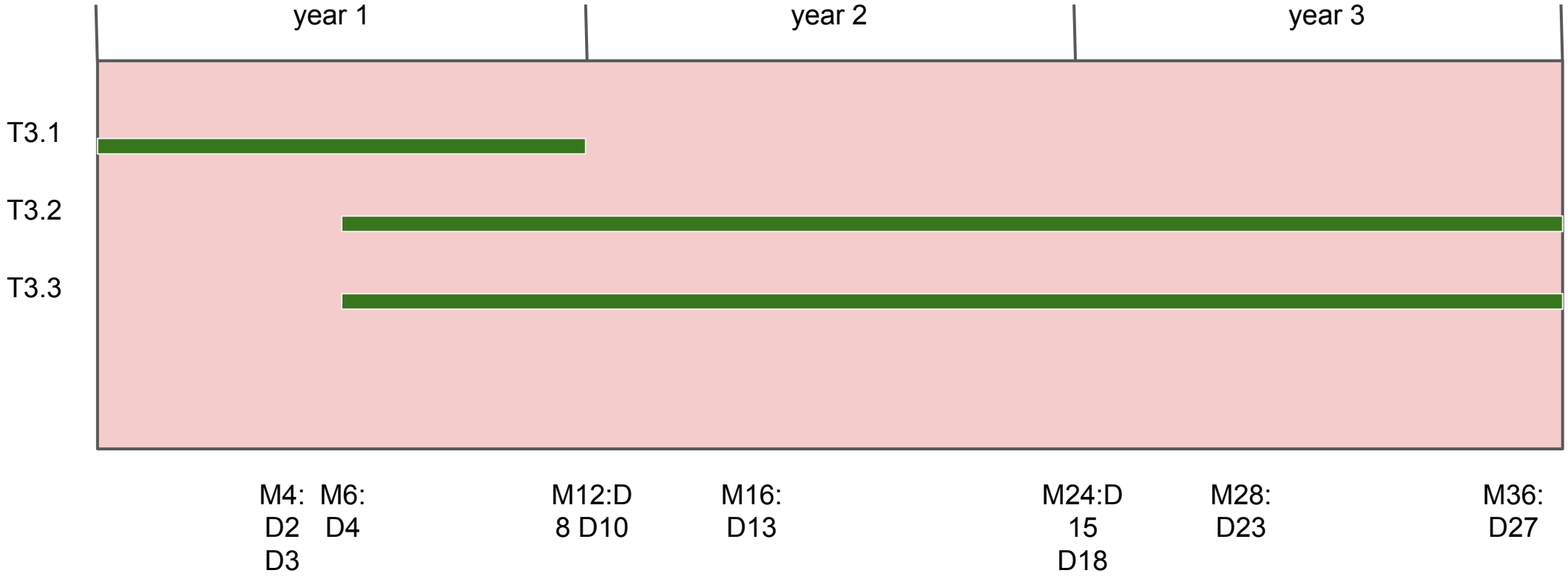
T3.2: Innovative Machine Learning

This task is in charge of **designing, implementing and evaluating ML** components in ACO-S pipelines. Targets include off-line processes for the transformations and enrichment of data, such as classification, segmentation, reduction, and emulation. For each component targeted, the task **will provide an adequate ML model, together with an efficient and scalable implementation**. Finally, its **performance** (both in terms of task outcome and efficient computation) will be assessed, to produce a final evaluation on the advantages and disadvantages of the produced ML solutions. **Together with industrial partners** the task will address *a)* search of patterns using network analysis in very large, noisy datasets, *b)*anomaly detection, *c)* techniques of privacy preserving, deployed in federated learning services (Cybersecurity)

T3.3: HPC/Cloud Visualization Services

The main emphasis will be on fast and **interactive visualization**, using services at the HPC facilities **near the data,** but with high speed delivery of the results to the user's desktop. Such tools will be tested and integrated on the existing open-source database as https://smart-turb.roma2.infn.it/ which will be further developed during this project such as to host a large quantity of geophysical and astrophysical datasets. The ultimate goal is to provide visualization capabilities that are fully interactive and can access and visualize large data sets at a **remote repository**.

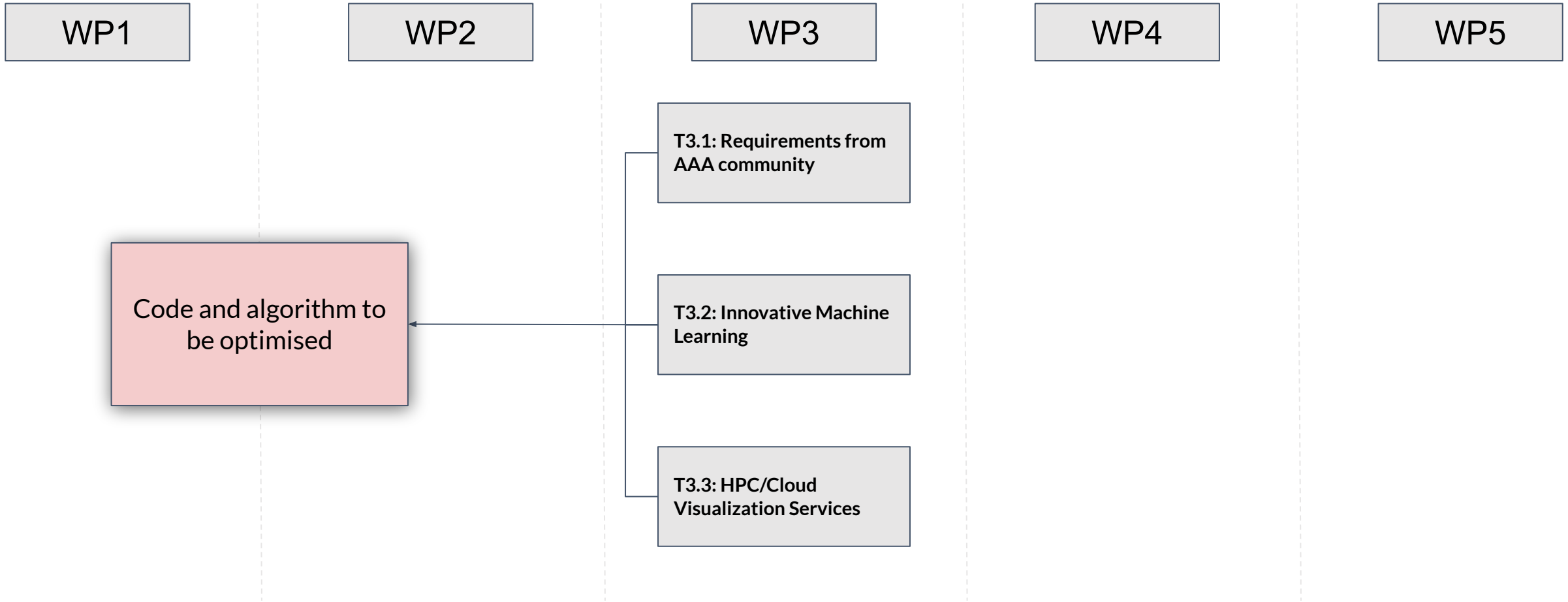
Timeline


year 1
year 2
year 3
T3.1
T3.2
T3.3
M4: D2 D3
M6: D4
M12:D 8 D10
M16: D13
M24:D 15 D18
M28: D23
M36: D27

Deliverables

#ID	Deliverable name	Short description	Type	Dissem. level	Delivery date (months)
D2	Activity status report (M4.1)	Science cases/requirements definition activity status report	Report	Public	M4
D3	Spoke Project Plan (M4.2)	This will be written by all WPs	Report	Public	M4
D4	Activity status report (M6.1)	Science cases/requirements definition activity status report	Report	Public	M6
D8	Activity status report (M12.2)	Spoke's first year activity status report, written by all WPs	Report	Public	M12
D10	Requirements definition document (M12.4)	Spoke requirement definition document: science cases, algorithms and technologies definitions	Report	Public	M12
D13	Requirements definition document (M12.4)	Astrophysics Requirements Definition document	Report	Public	M16
D15	Activity status report (M24.2)	Spoke's second year activity status report, written by all WPs	Report	Public	M24
D18	Identified ML and visualization solutions (M24.5)	Identified ML and visualization solutions implementation on existing/available HPC infrastructure results	Report	Public	M24
D23	Best analysis strategies (M28.2)	Report on best analysis strategies and profiling on HPC infrastructures for astrophysical cases	Report	Public	M28
D27	Final activity report (M36.4)	Spoke's final activity status report, written by all WPs	Report	Public	M36

Interactions with other wps



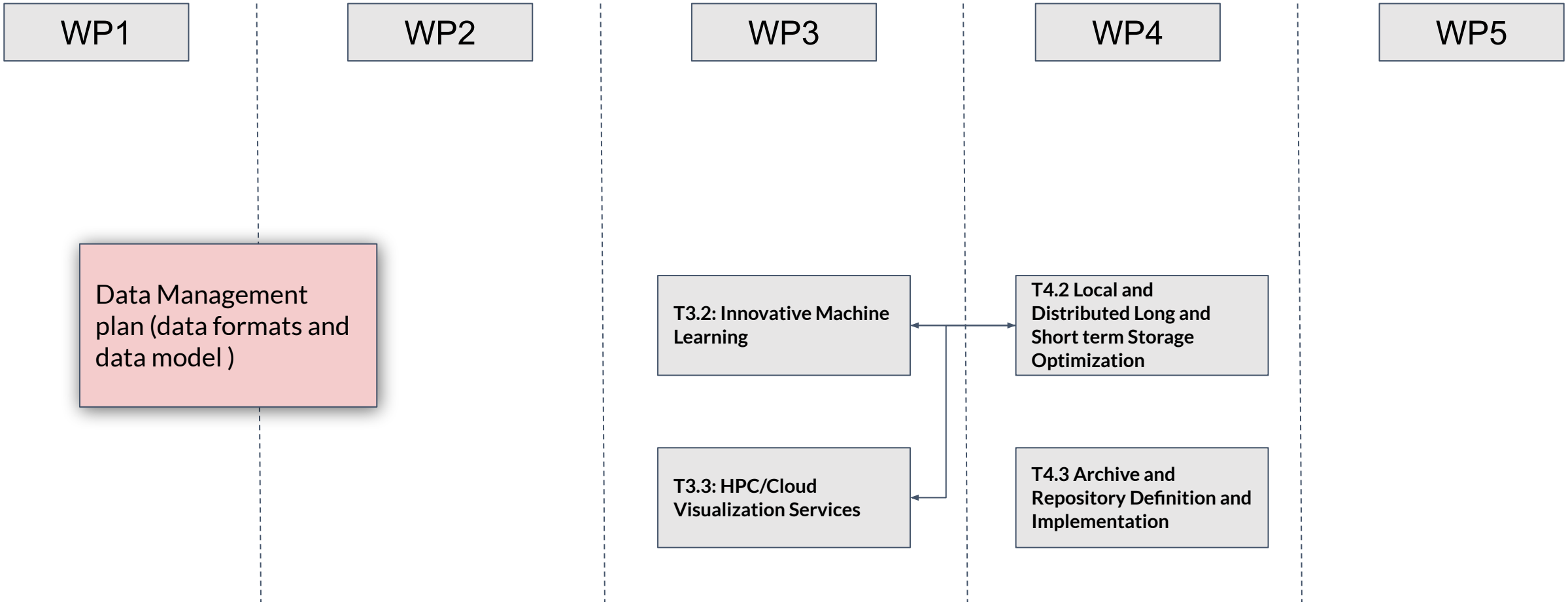
Interactions with other wps

WP1 | WP2 | WP3 | WP4 | WP5

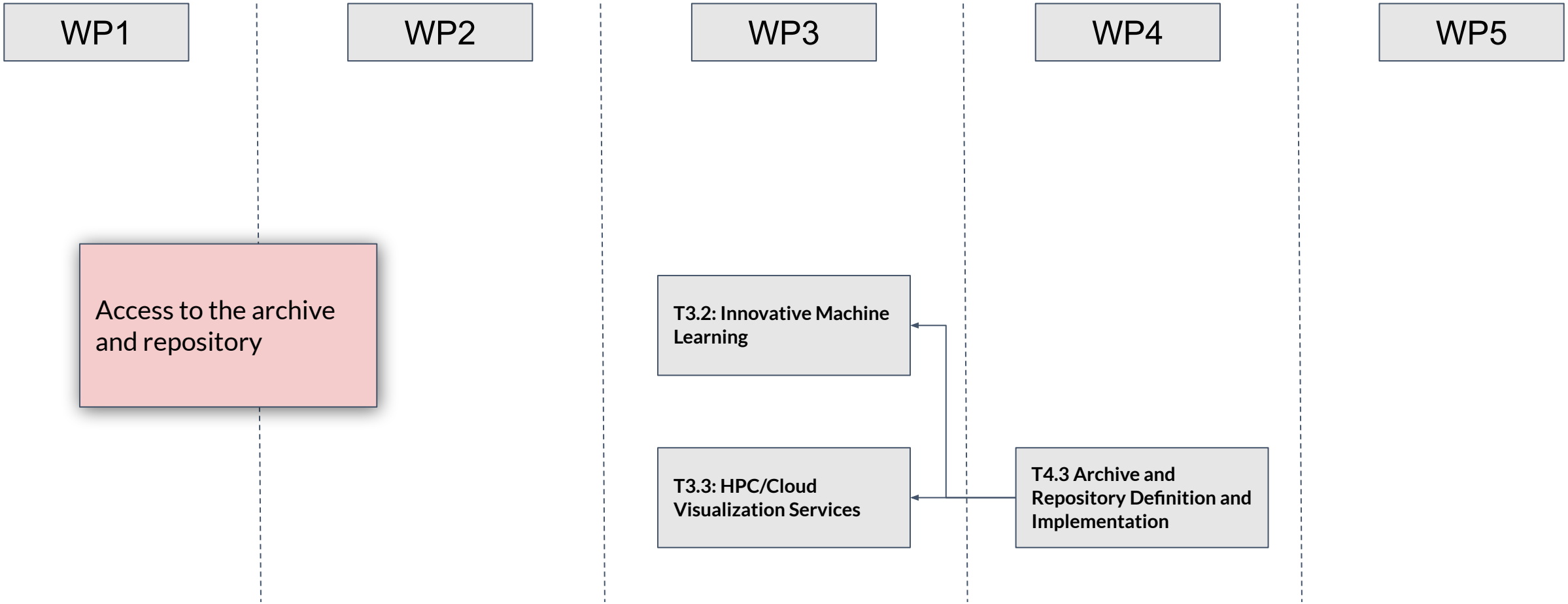
T3.1: Requirements from AAA community → T4.1 **Data management, Standardization and Interoperability**

Provide relevant information on I/O specifics extracted from requirements

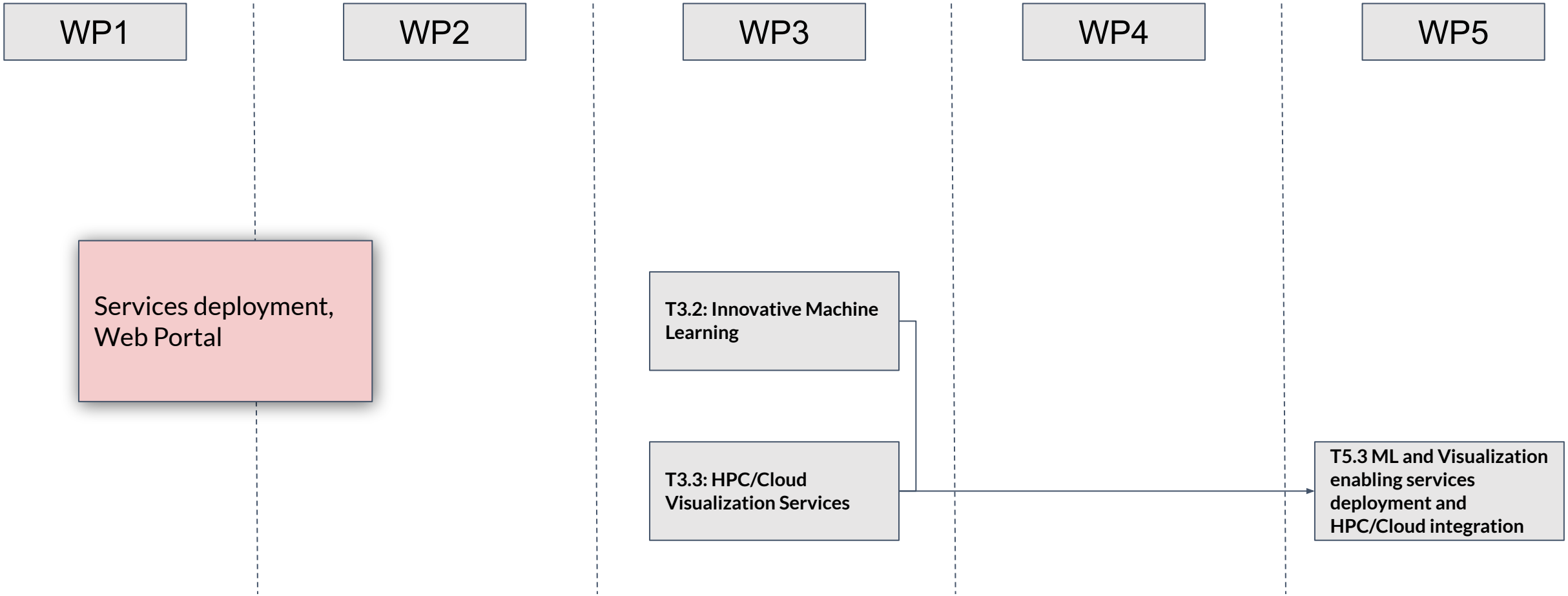
Interactions with other wps



Interactions with other wps



Interactions with other wps



Partner Efforts

Partner	Task	Personale di riferimento	PM A1	PM A2	PM A3	PM
INAF	T3.1, T3.2, T3.3	Fabio Vitello, Claudio Gheller, Ugo Becciani, Andrea Possenti, Aldo Bonomo, Carmelita Carbone, Deborah Busonero, Simone Riggi	41	92	92	**225**
UNITS	T3.2, T3.3	Matteo Costanzi	9	8	8	**25**
SISSA	T3.2, T3.3	Matteo Viel, Roberto Trotta, Nicoletta Krachmalnicoff	11	18	18	**47**
UNITO	T3.1, T3.2, T3.3	Susanna Terracini	10	12	12	**34**
UNICT	T3.1, T3.2	Maria Letizia Pumo, Alessandro Lanzafame, Giuseppe Nunnari, Vincenza Carchiolo, Manicò, Mangioni, Malgeri	16,9	22,8	22,5	**62,2**
INFN	T3.2	Fabio Gargano, Nicola Mazziotta, Matteo Duranti, Valerio Formato, Stefano Della Torre, Sara Cutini, Massimiliano Lattanzi, Martina Gerbino, Melissa Pesce Rollins	43	44	31	**118**
SNS	T3.1, T3.2	Andrei Mesinger	16	17	16	**49**
ROMTOV	T3.2	Giuseppe Puglisi, Luca Biferale, Domenico Marinucci, Marina Migliaccio, Piermarco Cannarsa, Nicola Vittorio, Pasquale Mazzotta	8,5	10,5	11,5	**30,5**

Partner Efforts

Partner	Task	Personale di riferimento	PM A1	PM A2	PM A3	PM
INAF	T3.1, T3.2, T3.3	Fabio Vitello, Claudio Gheller, Ugo Becciani, Andrea Possenti, Aldo Bonor one Riggi	41	92	92	**225**
UNITS	T3.2, T3.3	Matte	9	8	8	**25**
SISSA	T3.2, T3.3	Matte f	11	18	18	**47**
UNITO	T3.1, T3.2, T3.3	Susan	10	12	12	**34**
UNICT	T3.1, T3.2	Maria Letizia Pumo, Alessandro Lanzafame, Giuseppe Nunnari, Vincenza Carchiolo, Manicò, Mangioni, Malgeri	16,9	22,8	22,5	**62,2**
INFN	T3.2	Fabio Gargano, Nicola Mazziotta, Matteo Duranti, Valerio Formato, Stefano Della Torre, Sara Cutini, Massimiliano Lattanzi, Martina Gerbino, Melissa Pesce Rollins	43	44	31	**118**
SNS	T3.1, T3.2	Andrei Mesinger	16	17	16	**49**
ROMTOV	T3.2	Giuseppe Puglisi, Luca Biferale, Domenico Marinucci, Marina Migliaccio, Piermarco Cannarsa, Nicola Vittorio, Pasquale Mazzotta	8,5	10,5	11,5	**30,5**

Total:
590,7 PM

Partner Efforts

Partner	Task	Personale di riferimento	PM A1	PM A2	PM A3	PM
INAF	T3.1, T3.2, T3.3	Fabio Vitello, Claudio Gheller, Ugo Becciani, Andrea Possenti, Aldo Bonor one Riggi	41	92	92	**225**
UNITS	T3.2, T3.3	Matte	9	8	8	**25**
SISSA	T3.2, T3.3	Matte f	11	18	18	**47**
UNITO	T3.1, T3.2, T3.3	Susan	10	12	12	**34**
UNICT	T3.1, T3.2	Maria Letizia Pumo, Alessandro Lanzafame, Giuseppe Nunnari, Vincenza Carchiolo, Manicò, Mangioni, Malgeri	16,9	22,8	22,5	**62,2**
INFN	T3.2	Fabio lerio Forma no Lattanzi, Marti	43	44	31	**118**
SNS	T3.1, T3.2	Andre	16	17	16	**49**
ROMTOV	T3.2	Giuse Marina Miglia uale Mazzotta	8,5	10,5	11,5	**30,5**

Total:
590,7 PM

Industrial Partners:
ISP, Sogei,
Leitha - Unipol Sai

WP3 Topics


analyses experiments find reconstruction systematic other astrophysical
dynamical precursors neutral automated plasma machine
structures visualise satellite those highly amounts algorithms
networks visualization traces multi signal advantage
learning feature necessary characterization applications high
generate connections celestial particles or data based
cosmological way capable search methods both are parameter codes gaia its
expected galaxy parametric induced space computing mission neural taking big
them surveys generation predict perform their charged
development series as time extractions
inference pipeline dimensional trackers cosmic exploit is
acceleration from simulations exploration gpu systems
identification mechanics create framework implementation analyse large on machines
ska develop paradigm such reasoned

Machine Learning topics

- **Identification** and **characterization** of planetary signals in the presence of correlated noise.
- AI and machine / deep learning and Bayesian inference (eg MCMC) for the **estimation of astrophysical parameters** from Big Data or to detect **correlations**.
- Development of innovative machine learning techniques for efficient and effective **automated data processing of SKA time-domain data** (e.g. pulsar / transient discovery).
- Development of tools for **synthetic data generation** and interpretability of neural networks. Expected applications are simulation acceleration (eg cosmological super-resolution simulations) and **missing data recovery** (eg astrophysical foregrounds removal).
- Optimization of **neural network** architectures for data analysis, thanks to generative networks enabling approaches of **data augmentation** (eg exploitation of additional information from the latent data-generation process). Expected applications are represented by unsupervised classification (eg SED fitting), modelling acceleration, likelihood-free inference from big data-sets.
- Automated **source extraction** and **characterization** for SKA and precursors data.

Machine Learning topics

- Development of Machine Learning algorithms applied to t**ime series to study the systematics** introduced in a cosmic microwave background instrument (CMB) due to cosmic rays for which it is necessary to create and predict a time series of the induced signal;
- Development of **Machine Learning** based algorithms for the **identification of charged and neutral particles** in satellite experiments and for the **reconstruction of traces** in trackers for space applications.;
- Implementation of the **Parametric Machines paradigm** to search for the dynamical structures and their connections in celestial mechanics;
- **Feature extractions** from high-dimensional galaxy surveys data;
- Development of codes based on **machine-learning** or **other inference methods** for **exploration of highly-dimensional parameter space**.

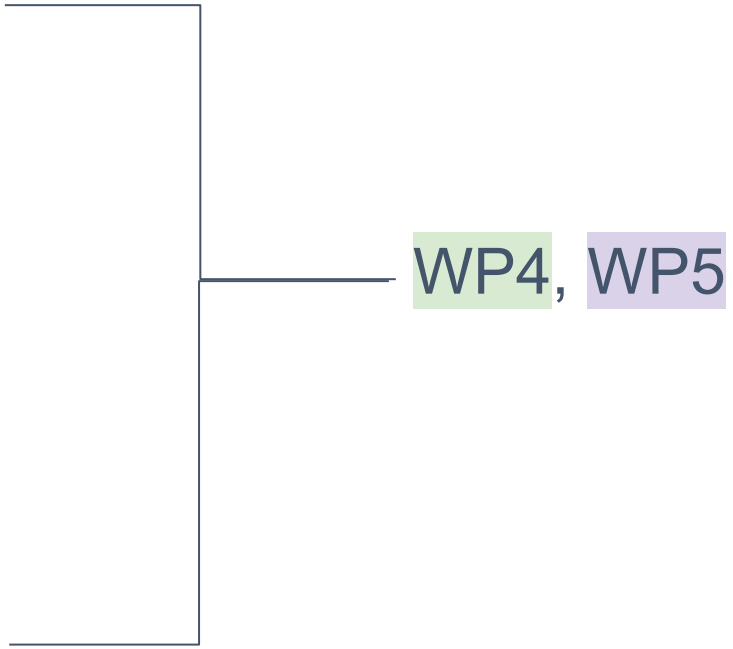
Visualisation topics

- Develop pipeline, capable to **exploit multi-GPU systems, to visualise and analyse** both data from the cosmological and astrophysical plasma simulations and from SKA and its precursors.
- **Visualization** of large amounts of data such as those of the Gaia mission
- Perform a systematic reasoned search for the dynamical structures in Celestial Mechanics and find way to **visualise** them.
- Develop a framework to perform analyses and generate **visualization** taking advantage of GPU computing.

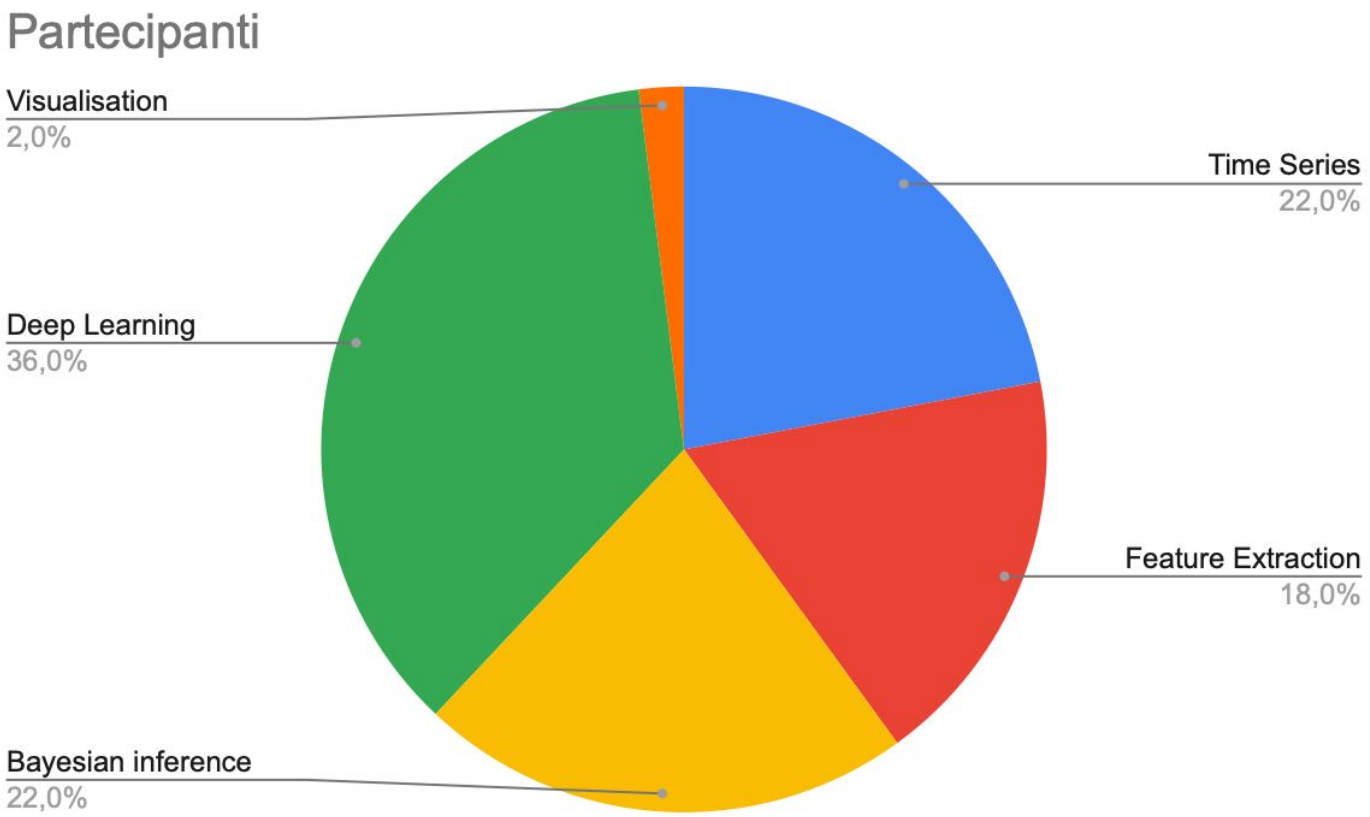
Working Groups (proposal)

Working groups are the place where we share knowledge

1. Eulerian Codes (WP1, WP2)
2. Lagrangian Codes (WP1, WP2)
3. Time series (WP1, WP2, WP3)
4. Feature extraction (WP3)
5. Bayesian inference(WP2 , WP3)
6. Deep learning (WP1,WP3)
7. Visualization (WP3)
8. Data-reduction & imaging (WP1, WP2)
9. Semi-numerical codes (WP1, WP2)
10. Web-tools (WP4, WP5)
11. Platforms (WP4, WP5)
12. Data-model (WP1, WP2, WP3 ,WP4, WP5)



Working Groups



https://docs.google.com/spreadsheets/d/1_J9qliSigUn2alYe4tJ6fYKmlCPRRLTQISsMVDkGi0U/edit#gid=926843366

Conclusions

- Interaction with the other WP is going smoothly;
- The proposed Working Groups structure can be changed;
- Industrial partner have a **strong interest** in WP3 activities;
- Define internal WP organization (deputy wp leader, WG lead);

Thank you!

cn1-spoke3-wp3@inaf.it

fabio.vitello@inaf.it