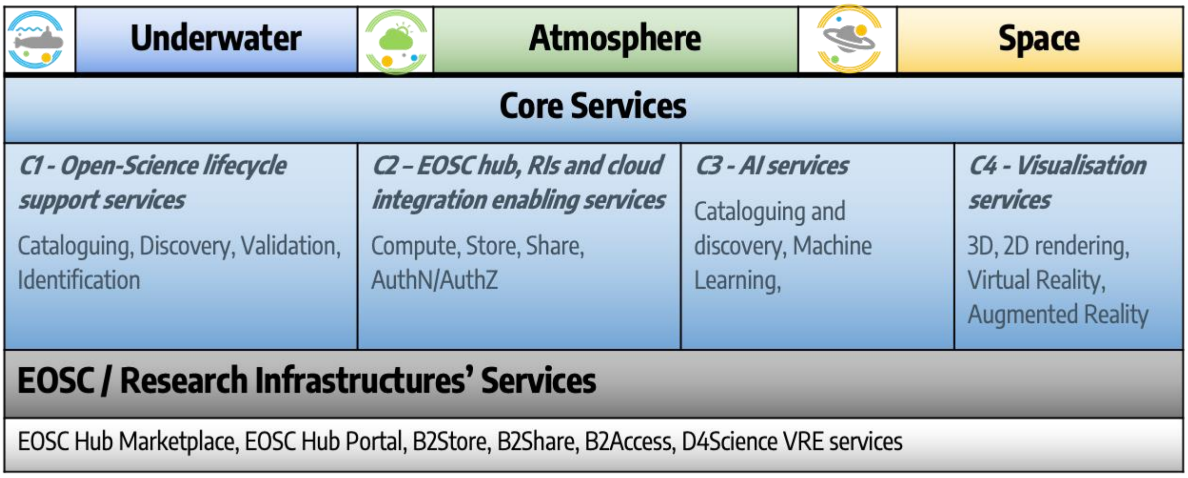
AI services for Open Science

What do we mean by AI services?



Underwater | Atmosphere | Space

Core Services

C1 - Open-Science lifecycle support services
Cataloguing, Discovery, Validation, Identification

C2 – EOSC hub, RIs and cloud integration enabling services
Compute, Store, Share, AuthN/AuthZ

C3 - AI services
Cataloguing and discovery, Machine Learning,

C4 - Visualisation services
3D, 2D rendering, Virtual Reality, Augmented Reality

EOSC / Research Infrastructures' Services

EOSC Hub Marketplace, EOSC Hub Portal, B2Store, B2Share, B2Access, D4Science VRE services

Fig.1: A schematic view of NEANIAS Thematic and Core services

The architecture of the NEANIAS project proposes a view in which thematic services are developed on top of more general purpose Core Services, which enable the usage of resources of the computational environment in which they are deployed, within the context of the European Open Science Cloud.

Artificial Intelligence, in general, refers to a whole discipline, encompassing a variety of different approaches, that can hardly be mapped to a service or set of services, however complex and powerful. Nonetheless, the successes of Machine Learning techniques and the growth of interest in their application in the most diverse fields, a process sometimes referred to as the Deep Learning revolution, has made it clear that *some* AI approaches and techniques could be fruitfully engineered as services of general interest within a cloud environment. There are whole projects aimed at achieving this kind of result, even outside the EOSC context (see, e.g., the Horizon 2020 funded AI4EU[1] initiative), but within NEANIAS we wanted to explore the chance to pursue this objective with the concrete perspective of supporting thematic, open science services, at the same time empowering researchers within the Consortium, as well as users of these services, but also presenting successful case studies of AI applications to open science (and guidance on how to develop additional ones) for further developments within EOSC. AI services developed within NEANIAS, therefore, will be employed firstly within NEANIAS, but they will be made available to the EOSC just like thematic services.

[1] https://www.ai4eu.eu/

More specifically, the first release of the set of core services comprised four AI services, employing both computational resources and other NEANIAS core services, offered and hosted by the GARR cloud infrastructure:

- AI Science Gateway acts as a Jupyter Hub based "entry-point" to support developers in accessing the overall computational environment, and supporting back-end Machine Learning (ML) tools;
- Trained models can be accessed by means of a specific component based on BentoML (C3.2), aimed at supporting scaling from trained models to production-grade services;
- Back-end computation, especially for speeding up learning tasks, can also employ:
    - a GPU oriented distributed ML framework (Horovod) (C3.3), able to exploit opportunities offered by a cloud environment in which multiple GPU resources are available;
    - a different distributed computing framework especially supporting non deep ML approaches (SPARK) (C3.4), able to exploit opportunities offered by a cloud environment in which multiple CPUs are available.

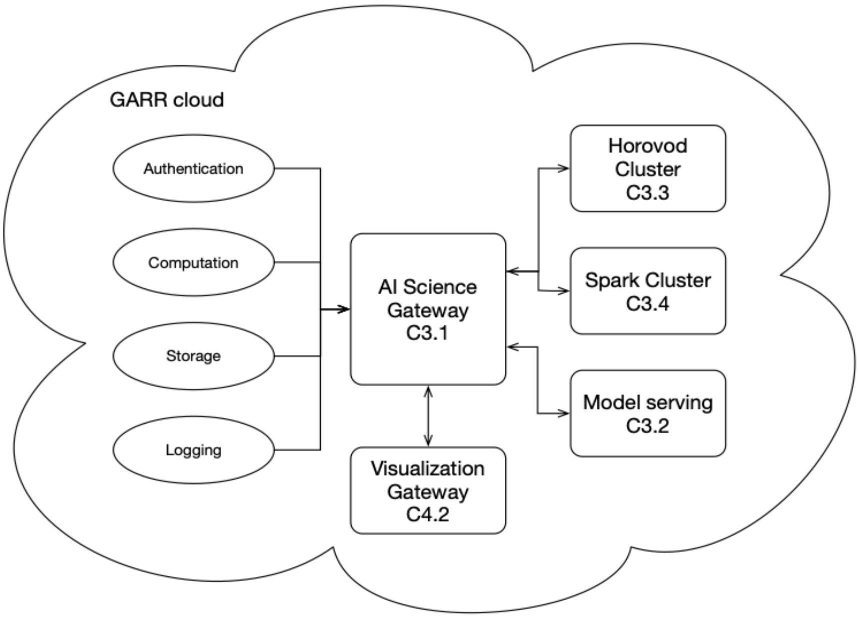
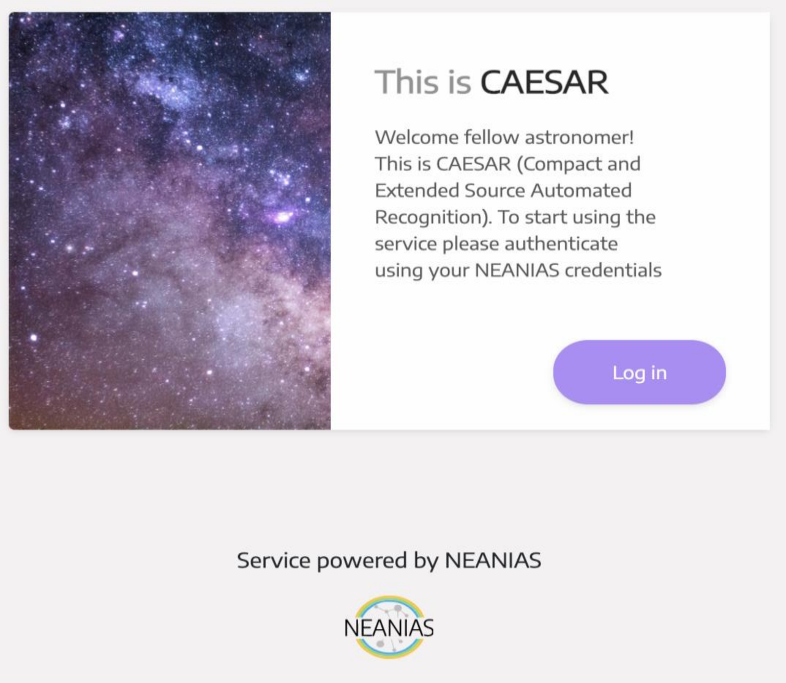


Fig.2: A conceptual view of AI services within the NEANIAS context

The AI services are undergoing further developments, both to improve the provided functionalities, as well as to better fit the EOSC ecosystem, but the first release has demonstrated their functionalities and potential.

How AI is used as of this moment by thematic services

The first thematic service employing NEANIAS AI services, and in particular the AI Gateway, is Space-ML[2]. This service has been developed for source identification, classification and characterization of sources in large-scale radio surveys. It employs a deep convolutional neural network, based on an instance segmentation framework (Mask R-CNN), that supports both the detection and the classification of radio compact sources, radio galaxies with extended morphology and sidelobe imaging artefacts. It can work either as a standalone source finder, or as a classifier stage applied to source finders catalogue outputs of another NEANIAS thematic service (CAESAR).



This is CAESAR

Welcome fellow astronomer! This is CAESAR (Compact and Extended Source Automated Recognition). To start using the service please authenticate using your NEANIAS credentials

Service powered by NEANIAS

Fig.3: NEANIAS SPACE CAESAR Service login page

The first release of this service does not exploit the Horovod framework for performing machine learning in a distributed multi-GPU setting (C3.3) or the model serving service through BentoML (C3.2), but this could be one of the first thematic services exploiting these opportunities.

How AI services will be used in additional thematic services

Training neural networks is a resource consuming task, but is not the only one that could benefit from distributed computing. Fetching large amounts of data and performing relatively simple data manipulation operations, maybe in order to prepare them as inputs to pre-trained neural networks, are common tasks for example for remote Earth-observing platforms[3]. C3.4 service and Spark represent a useful

[2] https://thematic.dev.neanias.eu/SPACE/space-ml.html

[3] See, e.g., Lan, H., Zheng, X., & Torrens, P. M. (2018). Spark Sensing: A Cloud Computing Framework to Unfold Processing Efficiencies for Large and Multiscale Remotely Sensed Data, with Examples on Landsat 8 and MODIS Data. Journal of Sensors, 2018.

resource supporting this type of task, granting the access and the possibility to exploit big data stored in cloud storage platforms like Amazon S3 or other kinds of cloud oriented data repositories.

An additional task that could exploit the functionalities offered by the C3.4 service is represented by the possibility of performing in a distributed way operations like hyperparameter optimization: machine learning algorithms, in fact, are generally characterized by the need of specifying specific operating parameters that have a significant impact on the achieved performance, both in terms of effectiveness and efficiency. One of the available examples of usage of the service describes this kind of use case.

Under the hood

NEANIAS implies EOSC, and EOSC means cloud: to deliver services in a cloud environment, the best resource manager that works with docker images (the most widely adopted means of packaging for cloud ready services and micro-services) is Kubernetes (K8s). Through K8s we can deploy multiple applications granting performance and dependency isolation, and making the best usage of available computational resources. Nonetheless, this is a relatively recent scenario, and this kind of deployment is still not so widely investigated and documented. For instance, the deployment of a low-level Spark engine on K8S is not trivial, and the Spark community has recently developed a Spark Operator that can manage custom resources defined for Spark. Those resources are SparkApplication objects that could be dynamically deployed and monitored by the managers of these services via the K8S API.

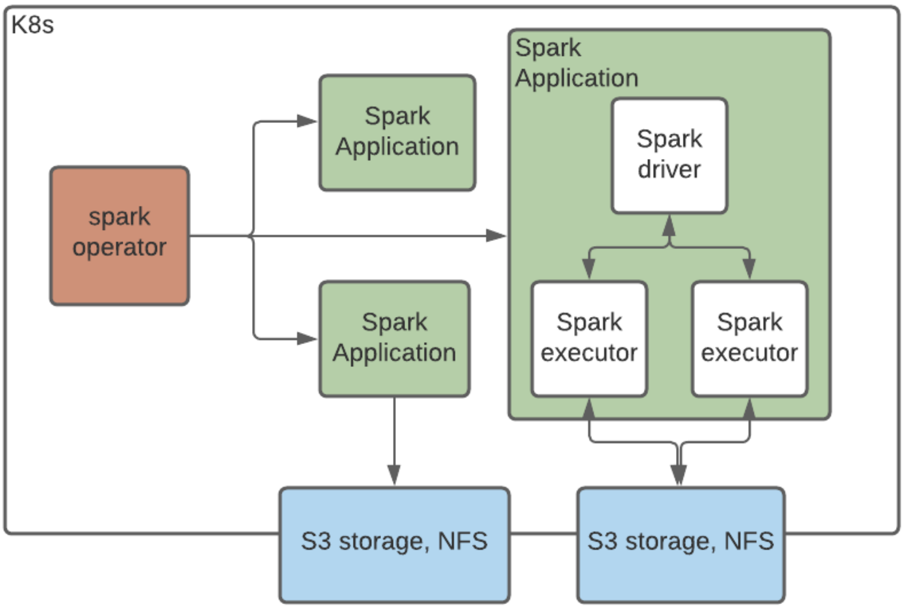


Fig.4: Deployment of the C3.4 AI service on Kubernetes.

How AI services might be proposed in EOSC per se

The development and experimentation of these services within the NEANIAS context is crucial for defining a proper strategy for scaling and offering these enabling services (and we hope to have given you an idea of why it can be reasonable to define them as core services) to the overall EOSC environment. The final goal is to offer services that could be adopted by researchers outside NEANIAS: these users could achieve computational resources from the EOSC Portal Catalogue and Marketplace[4], deploy these services in a reasonably simple way, and start developing and experimenting their own applications for scientific research, contributing to the overall open science movement.

Giuseppe Vizzari[a], Thomas Cecconello[a],
Eva Sciacca[b], Cristobal Bordiu[b],
Gabor Kertesz[c], Jozsef Kovacs[c]
[a]Università degli studi di Milano-Bicocca (UNIMIB)
[b]Istituto Nazionale di Astrofisica (INAF)
[c]Institute for Computer Science and Control (SZTAKI),
Eötvös Loránd Research Network (ELKH)

[4] https://marketplace.eosc-portal.eu/