

DEEP-Hybrid-DataCloud

Project summary and current status

IBERGRID 2019

Santiago de Compostela, Spain

September 24, 2019

Álvaro López García

aloga@ifca.unican.es

CSIC



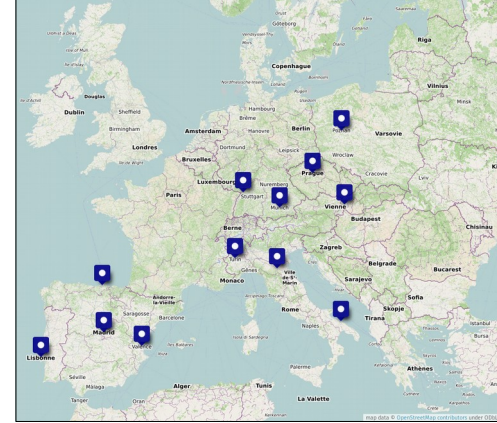
DEEP-HybridDataCloud has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777435.



Project introduction

DEEP project in 1 slide

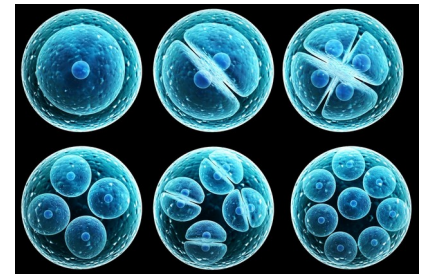
- **Designing and Enabling E-Infrastructures** for intensive data **P**rocessing in a **Hybrid DataCloud** (Grant agreement number 777435)
- **Global objective:** Promote the use of **intensive computing services** by different research communities and areas, an the **support by the corresponding e-Infrastructure** providers and open source projects
 - Focusing on **Machine learning, Deep learning, and Post processing**



- We need to build added value and advanced services on top of bare IaaS and PaaS infrastructures
- Key: **Service Oriented Architectures and platforms**
- Ease and lower the entry barrier for **non-skilled** scientists
 - Transparent execution on e-Infrastructures through specialized services and platforms → **lower entry barrier**
 - Build ready to use modules and offer them through a catalog or marketplace → **ease integration of services** into EOSC portal (ongoing)
 - Implement common software development techniques also for scientist's applications (DevOps) → **software quality strengthening**
- Build and promote the use of **intensive computing services** by different research communities and areas, and the support by the corresponding e-Infrastructure providers and open source projects

DEEP pilot use cases

- Three techniques of wide interest, involving
 - Large, heterogeneous data sets
 - Intensive computing demands that would benefit from using hardware accelerators (GPUs, low latency interconnects)

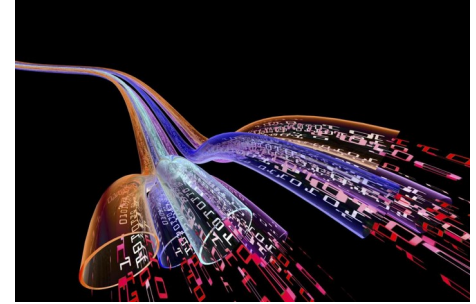


Deep learning applications

- Pilot cases: diabetic retinopathy detection, biodiversity applications.
- Objective: Provide a general, distributed architecture and pipeline to **train, retrain** and **use** deep learning (and other machine learning) models

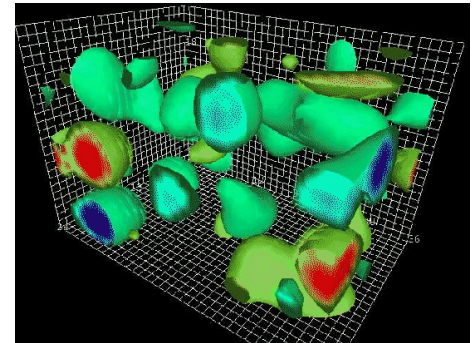
Online analysis of data streams

- Pilot case: intrusion detection systems, anomaly detection
- Provide an architecture able to analyze massive on-line data streams, also with historical records



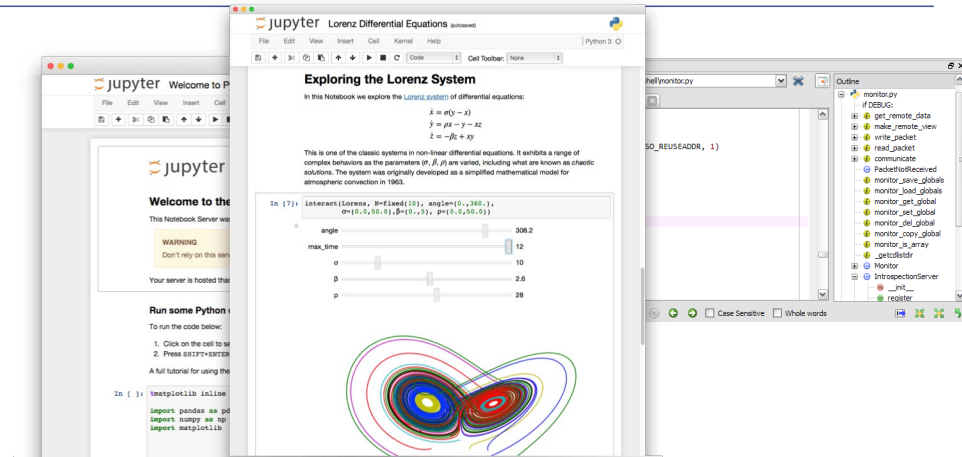
Post-processing

- Pilot cases: post-processing of HPC simulations
- Flexible pipeline for the analysis of simulation data generated at HPC resources



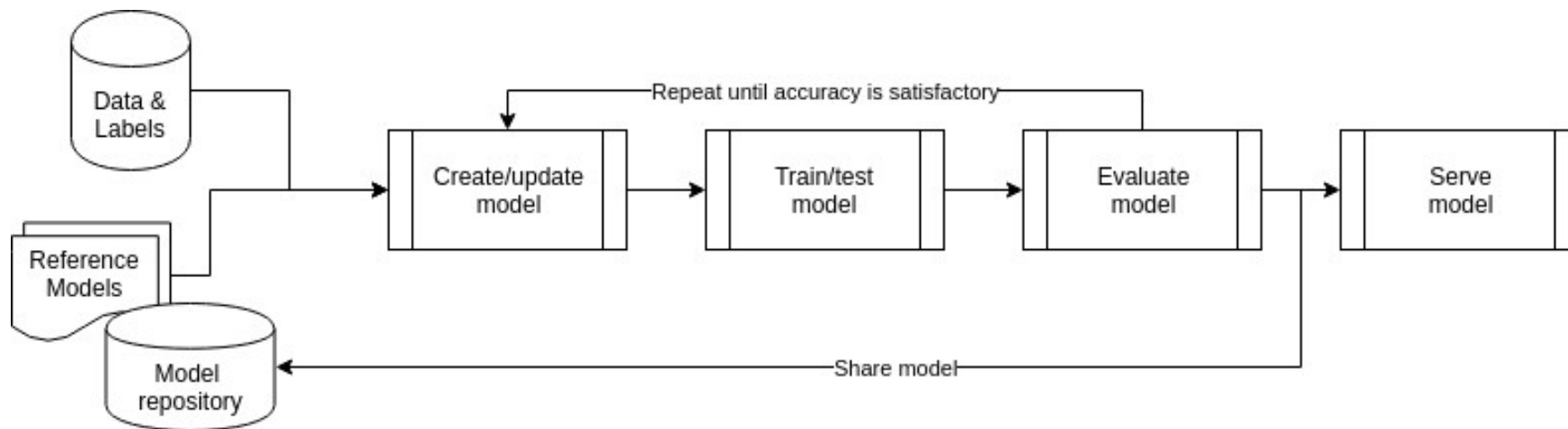
Previously...

- Scientists create a deep learning application on their personal computers, sometimes as a collection of scripts
- The deep learning model is trained in a GPU node (maybe also locally)
 - What happens if they do not have access to one?
- The work is published (or not)
 - Model architecture, configuration, dataset, scientific publication, etc.
- But:
 - How to ease sharing of models?
 - How to reuse an existing model?
 - How to offer the model to a broader audience? (other colleagues, citizens, etc.)
 - What about software development processes?



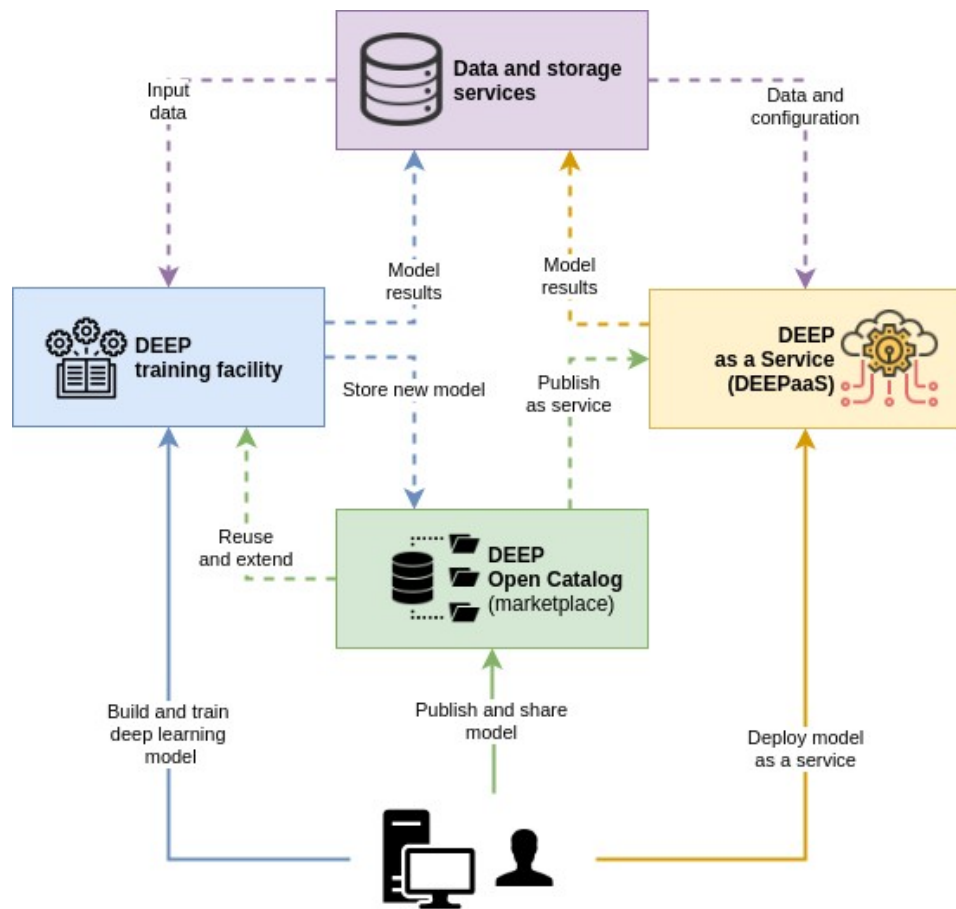
- **Category 1:** Deploy a readily trained network for somebody else to use it on his/her data set
 - Domain knowledge
- **Category 2:** Retrain (parts of) a trained network to make use of its inherent knowledge and to solve a new learning task
 - Domain + machine learning knowledge
- **Category 3:** Completely work through the deep learning cycle with data selection, model architecture, training and testing
 - Domain + machine + technological knowledge

Machine learning development cycle



- We are covering all the ML cycle phases
 - Create and update model
 - Train and testing
 - Evaluation
 - Serving and sharing

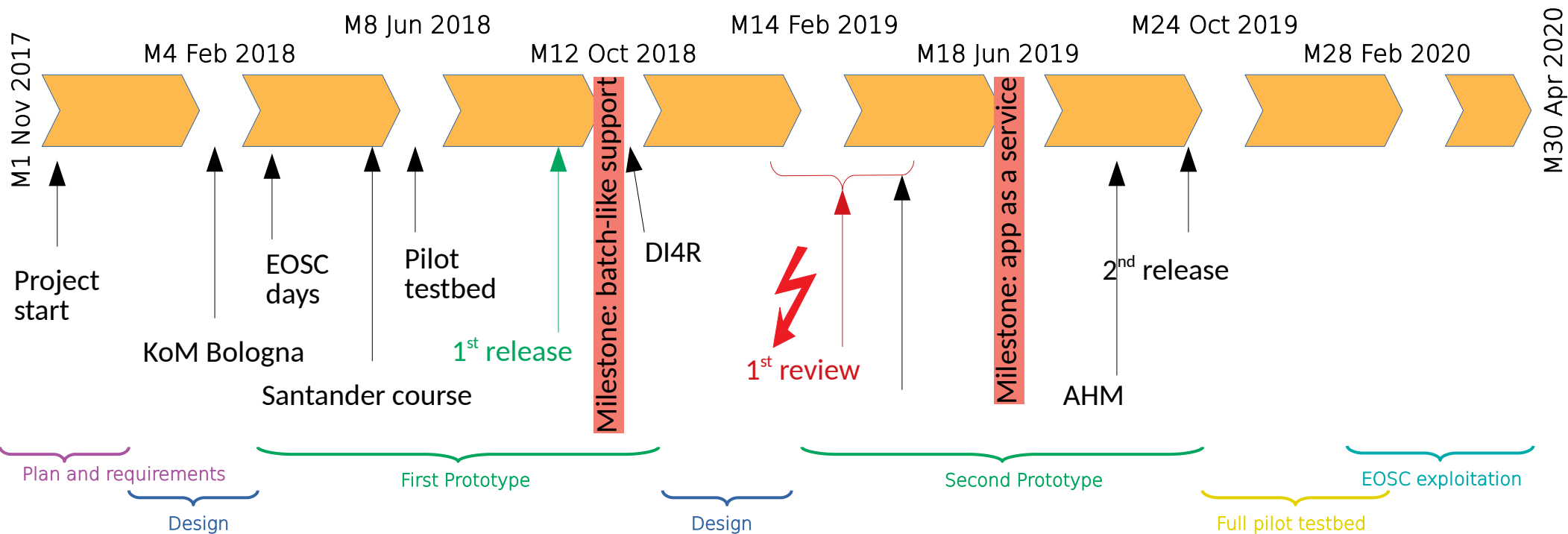
DEEP high level decomposition



- Position as technology providers to support DL/ML in the EOSC
- Generic building blocks (services) for exploitation through EOSC
 - Training facility
 - DEEPaaS facility
 - DEEP Open Catalog
- Integration with storage from external initiatives (eXtreme-DataCloud)

Project status and current achievements

DEEP timeline



DEEP-Genesis: 1st platform and release



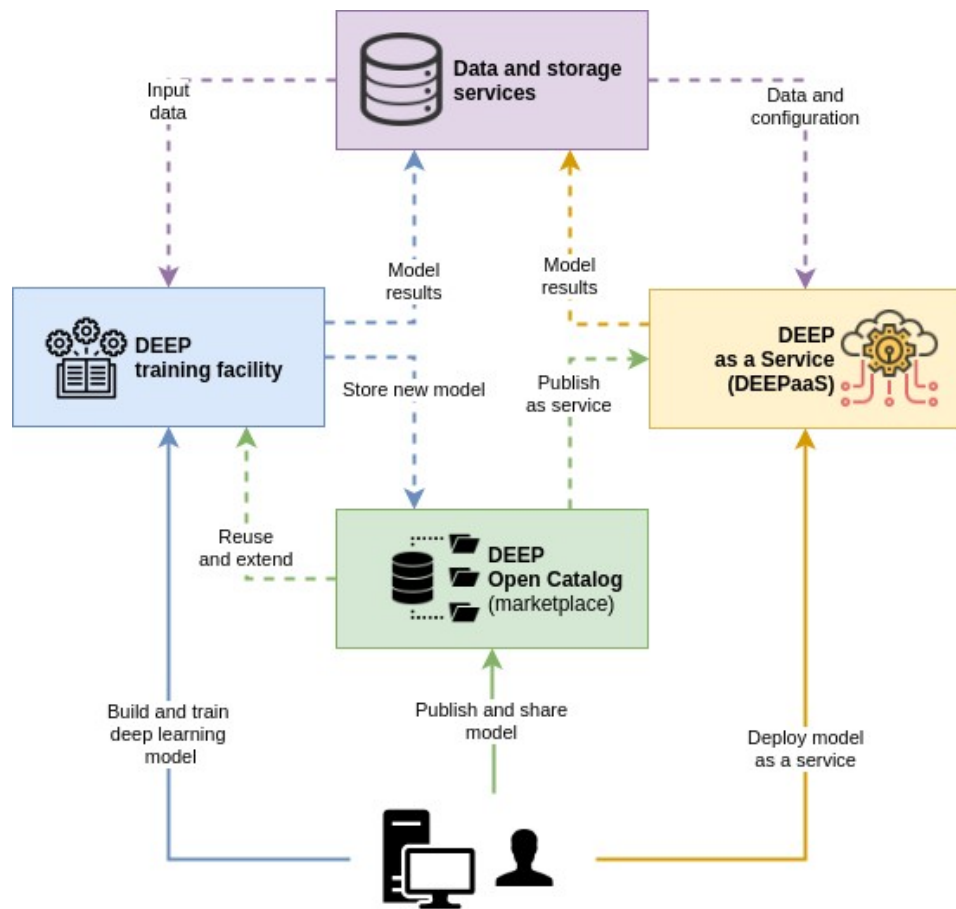
- First software release and prototype platform released January 2018
- More than 12 software components, 4 different services, several upstream contributions, more than 10 models in marketplace

DEEP 1st release: services

Service	Functionalities	Preview endpoint
Visual application topology composition and deployment	<ul style="list-style-type: none">• Graphical composition of complex application topologies• Deployment through PaaS orchestrator	https://a4c.ncg.ingrid.pt
ML/DL training facility as a service	<ul style="list-style-type: none">• Provide continuous training and retraining of developed models	https://train.deep-hybrid-datacloud.eu/
DEEP as a Service	<ul style="list-style-type: none">• Deployment of DEEP Open Catalog components as server-less functions	https://deepaas.deep-hybrid-datacloud.eu/
DEEP Open Catalog	<ul style="list-style-type: none">• Ready-to-use machine learning and deep learning applications, including:<ul style="list-style-type: none">➤ Machine learning frameworks + JupyterLab➤ ML/DL ready to use models➤ BigData analytic tools	https://marketplace.deep-hybrid-datacloud.eu

- All services are OIDC-ready, following AARC/AARC2 blueprint recommendations
- Also work on:
 - TOSCA templates and TOSCA types
 - Documentation and configuration recipes for GPU support
 - Patches to upstream projects (Apache Libcloud, Apache OpenWhisk, OpenStack)

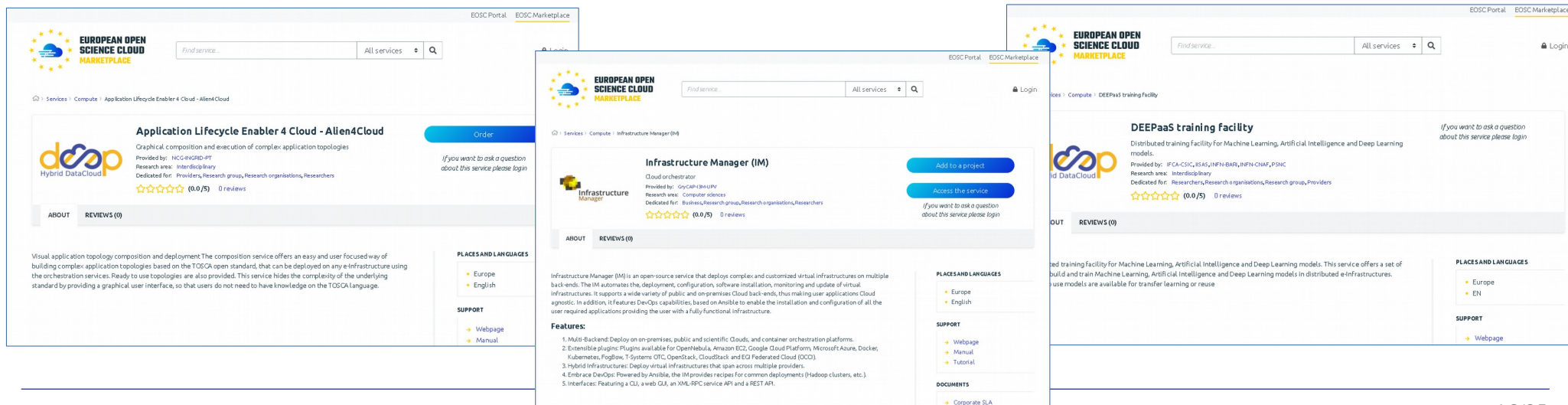
DEEP high level decomposition



- Position as technology providers to support DL/ML in the EOSC
- Generic building blocks (services) for EOSC exploitation
 - Training facility
 - DEEPaaS facility
 - DEEP Open Catalog

Services in EOSC-Hub catalog

- We have succeeded in integrating some of these services into the EOSC portal
 - DEEPaaS training facility: <https://marketplace.eosc-portal.eu/services/deepaas-training-facility>
 - Alien4Cloud: <https://marketplace.eosc-portal.eu/services/application-lifecycle-enabler-4-cloud-alien4cloud>
 - Infrastructure Manager: <https://marketplace.eosc-portal.eu/services/infrastructure-manager-im>
- Difficult and cumbersome process, feedback provided to EOSC-Hub



The image displays three overlapping screenshots of the EOSC Marketplace interface, showcasing different services available on the platform.

Left Screenshot: Application Lifecycle Enabler 4 Cloud - Alien4Cloud

- Header:** EUROPEAN OPEN SCIENCE CLOUD MARKETPLACE, Find service..., All services, Search icon.
- Breadcrumbs:** Services > Compute > Application Lifecycle Enabler 4 Cloud - Alien4Cloud
- Service Card:** Application Lifecycle Enabler 4 Cloud - Alien4Cloud. Provided by: HECQUAD-PI. Research area: Interdisciplinary. Dedicated for: Providers, Research groups, Research organisations, Researchers. (0.0/5) 0 reviews.
- Buttons:** Order, Add to project, Access this service.
- Text:** Graphical composition and execution of complex application topologies. Visual application topology composition and deployment. The composition service offers an easy and user focused way of building complex application topologies based on the TOSCA open standard, that can be deployed on any e-Infrastructure using the orchestration services. Ready to use topologies are also provided. This service hides the complexity of the underlying standard by providing a graphical user interface, so that users do not need to have knowledge on the TOSCA language.
- PLACES AND LANGUAGES:** Europe, English.
- SUPPORT:** Webpage, Manual.

Middle Screenshot: Infrastructure Manager (IM)

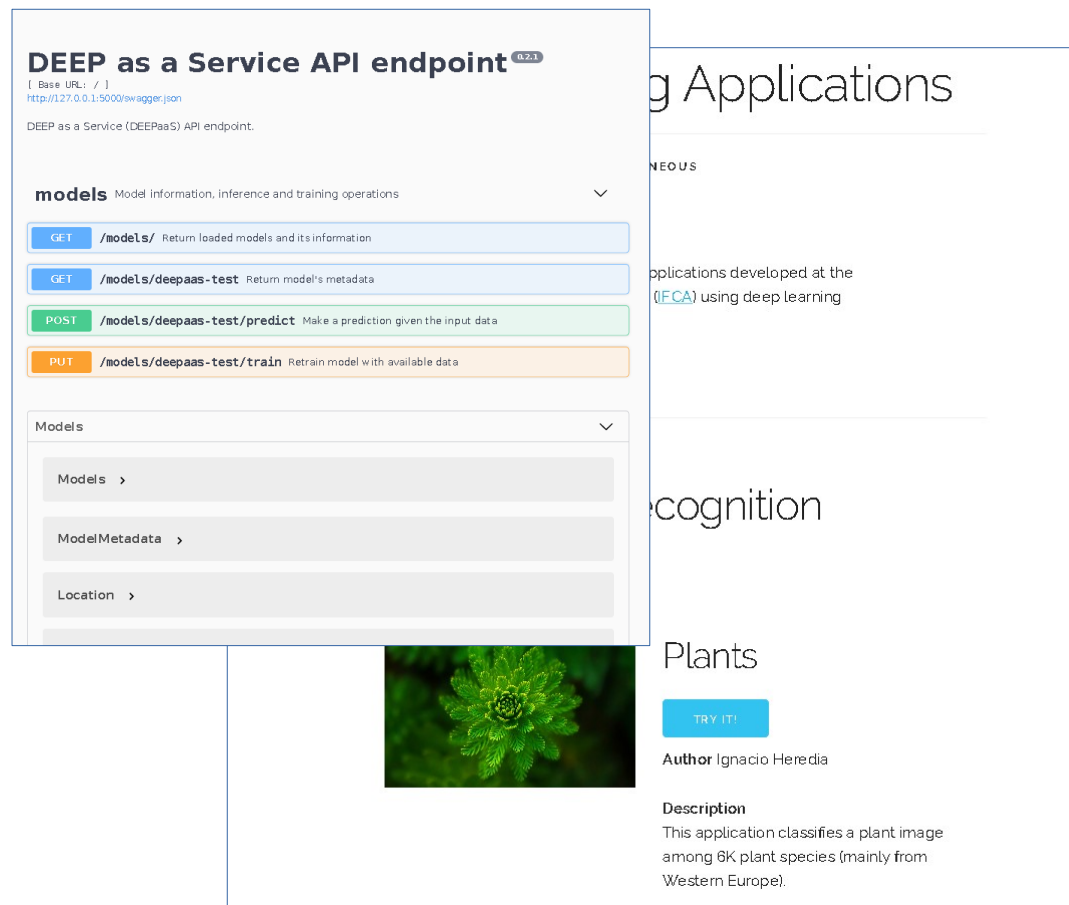
- Header:** EUROPEAN OPEN SCIENCE CLOUD MARKETPLACE, Find service..., All services, Search icon.
- Breadcrumbs:** Services > Compute > Infrastructure Manager (IM)
- Service Card:** Infrastructure Manager (IM). Cloud orchestrator. Provided by: GYCAH-BIAUIN. Research area: Computer sciences. Dedicated for: Business, Research groups, Research organisations, Researchers. (0.0/5) 0 reviews.
- Buttons:** Add to project, Access this service.
- Text:** Infrastructure Manager (IM) is an open-source service that deploys complex and customized virtual infrastructures on multiple back-ends. The IM automates the deployment, configuration, software installation, monitoring and update of virtual infrastructures. It supports a wide variety of public and on-premises Cloud back-ends, thus making user applications Cloud agnostic. In addition, it features DevOps capabilities, based on Ansible to enable the installation and configuration of all the user required applications providing the user with a fully functional infrastructure.
- PLACES AND LANGUAGES:** Europe, English.
- SUPPORT:** Webpage, Manual, Tutorial.
- DOCUMENTS:** Corporate SLA.

Right Screenshot: DEEPaaS training facility

- Header:** EUROPEAN OPEN SCIENCE CLOUD MARKETPLACE, Find service..., All services, Search icon.
- Breadcrumbs:** Services > Compute > DEEPaaS training facility
- Service Card:** DEEPaaS training facility. Distributed training facility for Machine Learning, Artificial Intelligence and Deep Learning models. Provided by: IFCA-CSIC, ISAS, INFN-BARI, INFN-CNAF, PSNC. Research area: Interdisciplinary. Dedicated for: Researchers, Research organisations, Research groups, Providers. (0.0/5) 0 reviews.
- Buttons:** Add to project, Access this service.
- Text:** Distributed training facility for Machine Learning, Artificial Intelligence and Deep Learning models. This service offers a set of build and train Machine Learning, Artificial Intelligence and Deep Learning models in distributed e-Infrastructures. User models are available for transfer learning or reuse.
- PLACES AND LANGUAGES:** Europe, EN.
- SUPPORT:** Webpage.

Offering models as a service

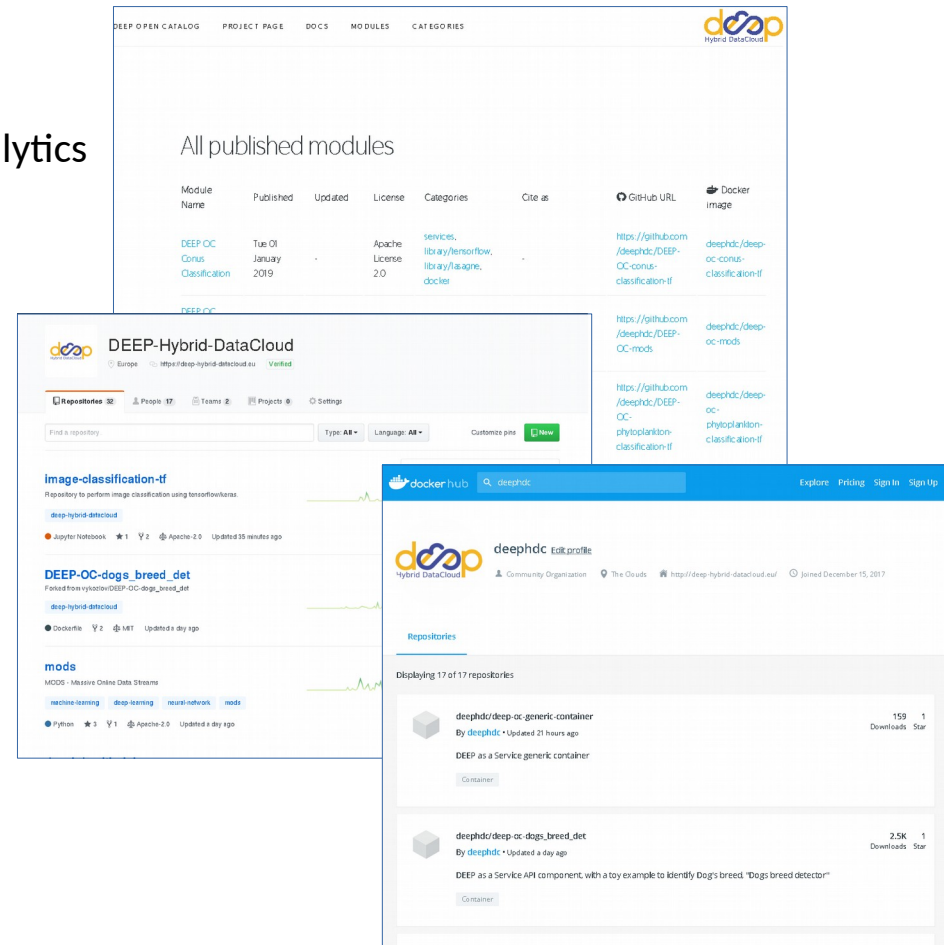
- Knowledge of API development and web applications
- Scientists need to know what an API is: REST, GET, POST, PUT...
- Lack of API consistency (different versions) → hard for external developers to consume them
- **DEEPaaS API:** Provide users with a generic API (based on OpenAPI) component where they application can be easily plugged



The image shows a composite of two screenshots. The left screenshot displays the 'DEEP as a Service API endpoint' documentation, version 0.2.1, with a Swagger-style interface. It lists four API endpoints: GET /models/ (Return loaded models and its information), GET /models/deepaas-test (Return model's metadata), POST /models/deepaas-test/predict (Make a prediction given the input data), and PUT /models/deepaas-test/train (Retrain model with available data). Below the endpoints is a 'Models' section with expandable items for Models, ModelMetadata, and Location. The right screenshot shows a web application titled 'g Applications' (partially visible) with a 'TRY IT!' button. Below the button, it identifies the author as 'Ignacio Heredia' and provides a description: 'This application classifies a plant image among 6K plant species (mainly from Western Europe)'. A small image of a green plant is visible above the description.

DEEP Open Catalog

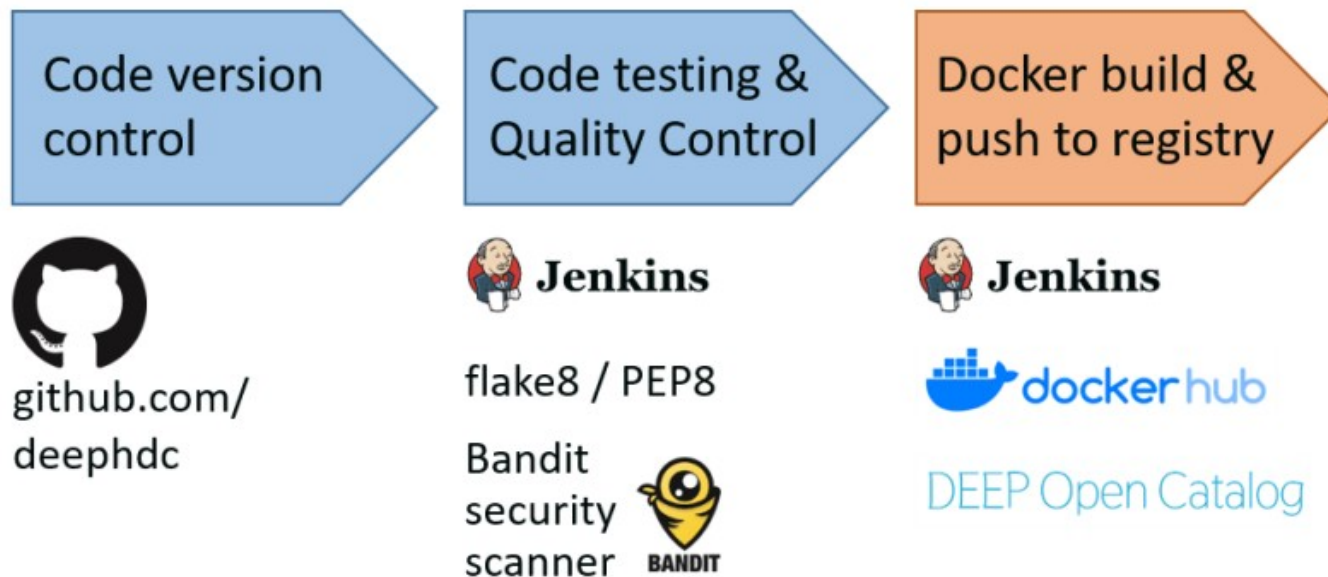
- **Crowdsourced** collection of ready-to-use modules (for inference, training, retraining, etc.)
 - Comprising machine learning, deep learning, big data analytics tools + corresponding TOSCA templates
 - ML/DL Marketplace:
<https://marketplace.deep-hybrid-datacloud.eu>
 - GitHub: <https://github.com/deephdc/>
 - DockerHub: <https://hub.docker.com/u/deephdc/>
- Based on **DEEPaaS API** component
 - Expose underlying model functionality with a common API
 - Follows OpenAPI specifications
 - Minimal modifications to user applications.
- **Goal: execute the same module on any platform and infrastructure:**
 - Laptop, workstation, HPC, Kubernetes, Mesos, DEEPaaS, other FaaS frameworks etc.



The image displays three screenshots related to the DEEP ecosystem:

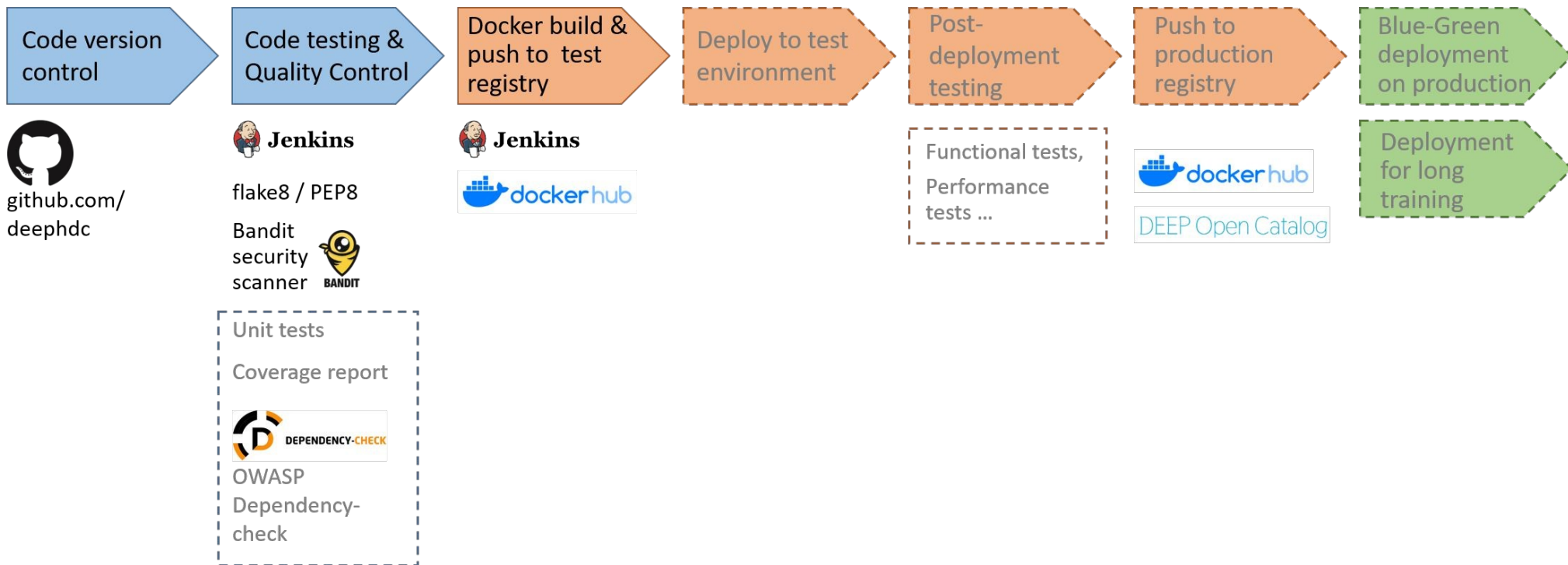
- Top Screenshot:** A view of the 'All published modules' page in the DEEP Open Catalog. It features a table with columns for Module Name, Published, Updated, License, Categories, Cite as, GitHub URL, and Docker image. One module listed is 'DEEP-OC-Consus Classification'.
- Bottom Left Screenshot:** A GitHub repository page for 'DEEP-Hybrid-DataCloud'. It shows repository statistics (30 repositories, 17 people, 2 teams, 2 projects, 0 settings) and lists several repositories including 'image-classification-tf', 'DEEP-OC-dogs_breed_det', and 'mods'.
- Bottom Right Screenshot:** A DockerHub profile page for 'deephdc'. It shows the profile information (Community Organization, The Clouds, joined December 15, 2017) and a list of repositories. Two repositories are highlighted: 'deephdc/deep-oc-generic-container' and 'deephdc/deep-oc-dogs_breed_det'.

DevOps for user apps



- using Jenkins enables a *pipeline-as-code* approach
- tools being used are widely used in community
- methodology being developed for development of DEEP core components can also be used for development of applications

DevOps for users apps: what next



- **Consolidation of services and components**
 - Promotion into production of prototype services (i.e. TRL8 and above).
- **Preparation of 2nd release**
- **Exploitation and on-boarding of new communities** (EOSC and beyond)
 - Early-stage researchers, collaboration with Master programs
 - ML/DL model developers and research groups
 - Cloud providers and research e-Infrastructures
 - Exploitation through industrial partner and SMEs
- Integration of services into EOSC portal (ongoing)
- Promotion of open “DEEP Open Catalog” to external users → crowd-source of applications

- Main webpage
 - <https://deep-hybrid-datacloud.eu/>
- DEEP Open Catalog
 - <https://marketplace.deep-hybrid-datacloud.eu/>
- DEEP documentation
 - <http://docs.deep-hybrid-datacloud.eu/>
- DEEP YouTube channel
 - https://www.youtube.com/playlist?list=PLJ9x9Zk1O-J_UZfNO2uWp2pFMmbwLvzXa
- Social media:
 - https://twitter.com/DEEP_eu

https://twitter.com/DEEP_eu



Thank you
Any Questions?



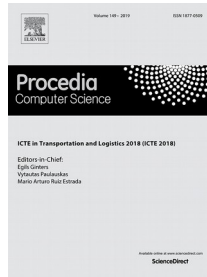
<https://deep-hybrid-datacloud.eu>



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777435.

Selected DEEP early results

- G. Nguyen, S. Dlugolinsky, M. Bobák, V. Tran, Á. López García, I. Heredia, P. Malík, and L. Hluchý. “Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey”. In: Artificial Intelligence Review (Jan. 2019). ISSN: 1573-7462. DOI: 10.1007/s10462-018-09679-z
- User communities publications:
 - N. Tran, T. Nguyen, B.M. Nguyen, G. Nguyen. “A multivariate fuzzy time series resource forecast model for clouds using LSTM and data correlation analysis”. Procedia Computer Science, Elsevier, 2018, Volume 126, pp. 636-645, ISSN 1877-0509. DOI 10.1016/j.procs.2018.07.298
 - G. Nguyen, B.M. Nguyen, D. Tran, L. Hluchý. “A heuristics approach to mine behavioural data logs in mobile malware detection system”. Data & Knowledge Engineering, Elsevier, 2018, Volume 115, pp. 129-151, ISSN 0169-023X, DOI 10.1016/j.datak.2018.03.002
 - B. M. Nguyen, H. Phan, D. Q. Ha, G. Nguyen. “An Information-centric Approach for Slice Monitoring from Edge Devices to Clouds”, Procedia Computer Science Volume 130, 2018, Pages 326-335. DOI: 10.1016/j.procs.2018.04.046
- Published articles by user communities not in the project, exploiting DEEP-HybridDataCloud software components:
 - I. Heredia Cacha. Application of a Convolutional Neural Network for image classification to the analysis of collisions in High Energy Physics. CHEP 2018 Conference, Sofia, Bulgaria. Oral Contribution
 - L. Lloret; I. Heredia; F. Aguilar; E. Debusschere; K. Deneudt; F. Hernández. Convolutional Neural Networks for Phytoplankton identification and classification. Biodiversity Information Science and Standards. 2018. Oral Contribution
 - F. Pando; I. Heredia; C. Aedo; M. Velayos; L. Lloret; J. Calvo. Deep learning for weed identification based on seed images. Biodiversity Information Science and Standards. 2018. Oral Contribution



DEEP high level Architecture

