

AI4

eoosc

AI4Compose: low-code composition of AI inference pipelines

Amanda Calatrava

Universitat Politècnica de València (UPV)



Co-funded by
the European Union

Objectives

Objective 1

Feature rich services and platform to build and deploy custom AI applications in the EOSC

Objective 2

Support for building AI systems on distributed datasets, with a particular focus on federated learning

Objective 3

Services to compose AI tools, enabling the development of complex data-driven applications

Objective 4

AI exchange/hub in the context of the EOSC, enhancing and increasing the application offer currently available

Objective 5

Extend the service offer and the capabilities being offered through the EOSC portal, with focus on AI

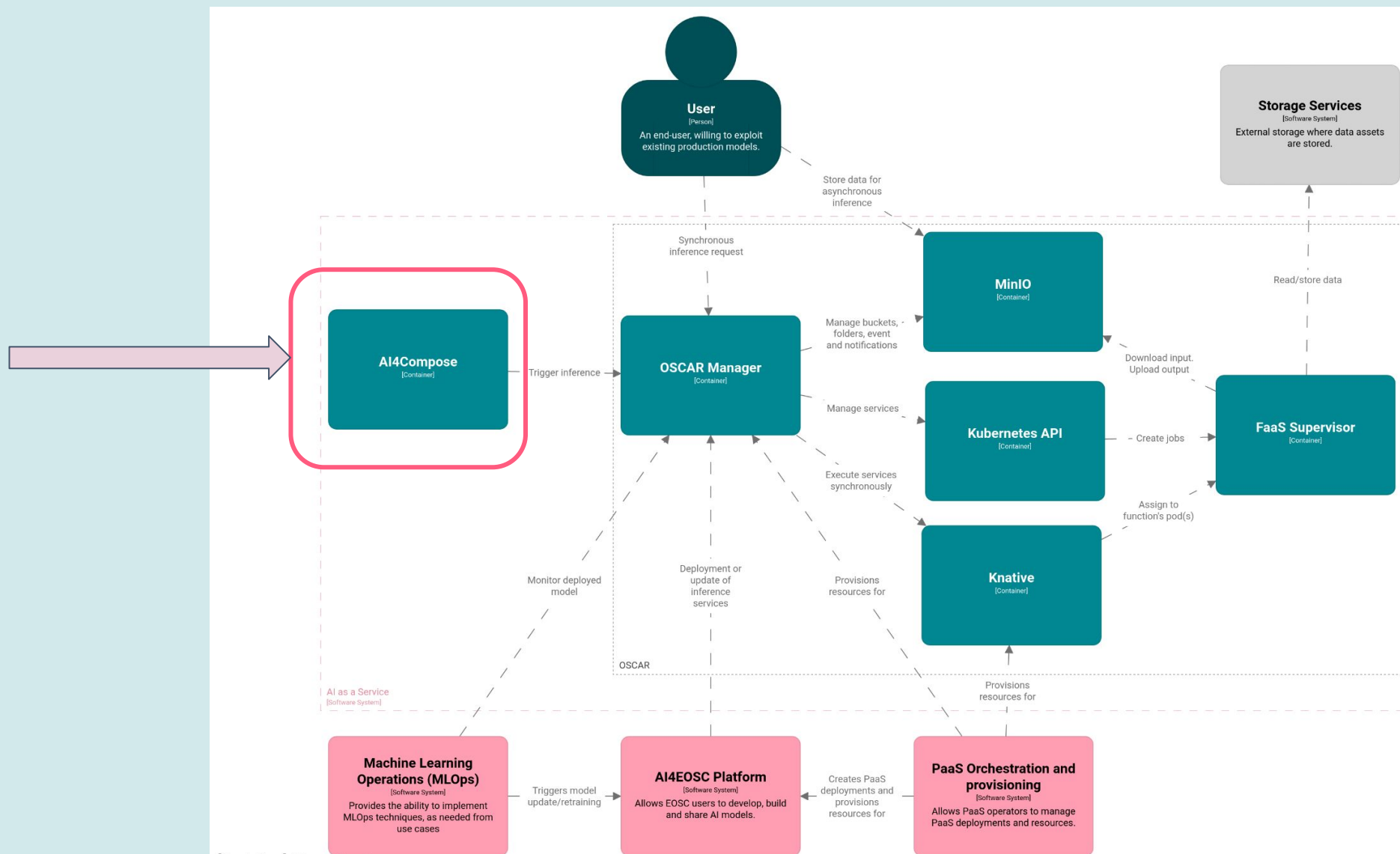
Goal

Foster an AI exchange in the EOSC context transforming the development of AI applications in the EOSC



Graphical Composition of **AI Inference Pipelines**

AI4EOSC Platform architecture



OSCAR: an overview

AI4

eosc



Co-funded by
the European Union



OSCAR is an open-source platform that:

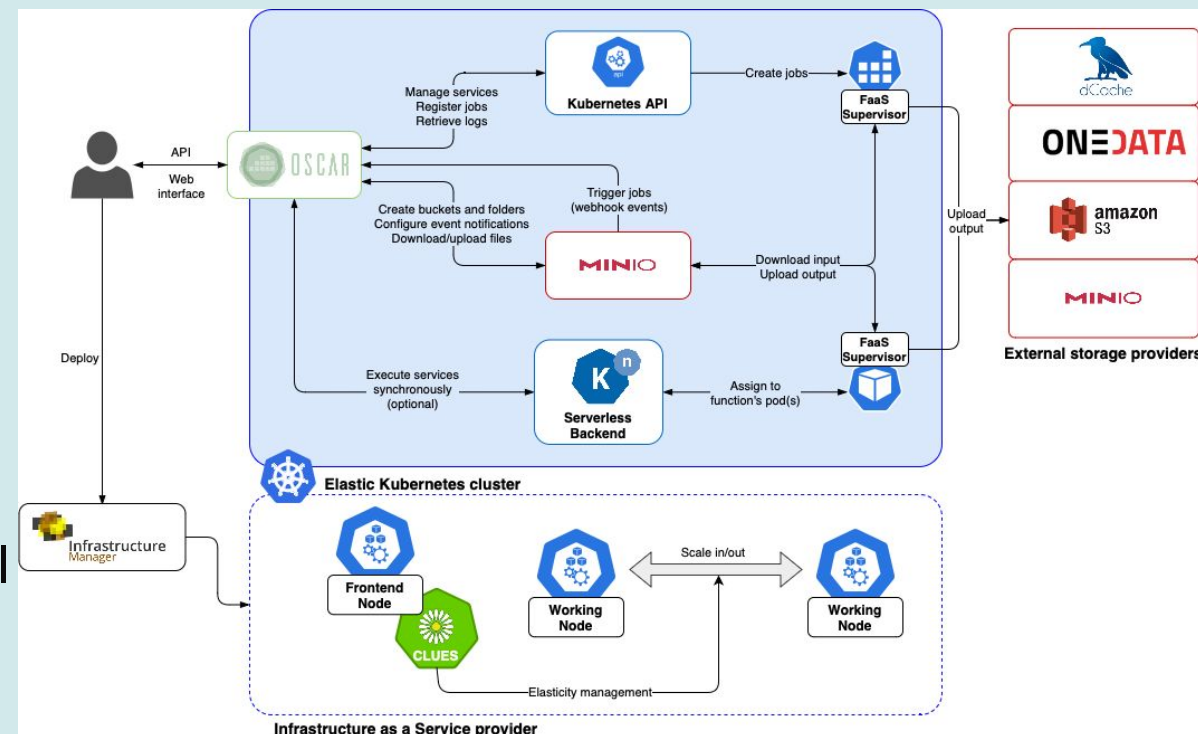
- Support the serverless (event-driven + automated elasticity) computing model for data-processing applications on Kubernetes.
- Is easily deployed on multi-Clouds via the Infrastructure Manager (IM).
- Supports the execution of highly-parallel event-driven data-processing serverless applications.
- Integrated with IM Dashboard, EGI DataHub, EGI Check-In and dCache (via WebDav).

OSCAR is used for:

- Scalable AI inference of pre-trained AI models.

OSCAR instance for AI4EOSC:

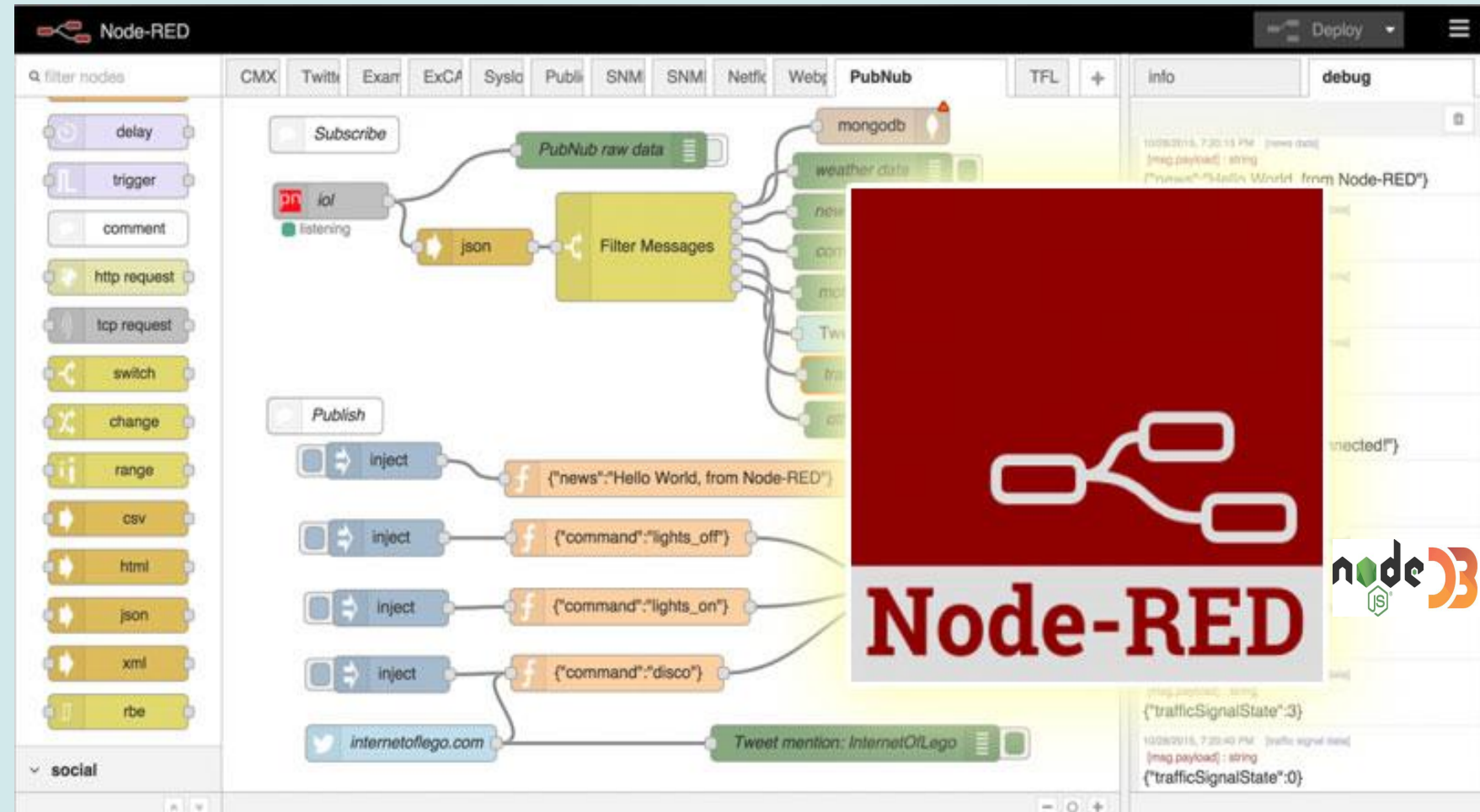
<https://inference.cloud.ai4eosc.eu#/login>



<https://oscar.grycap.net>

What is Node-RED?

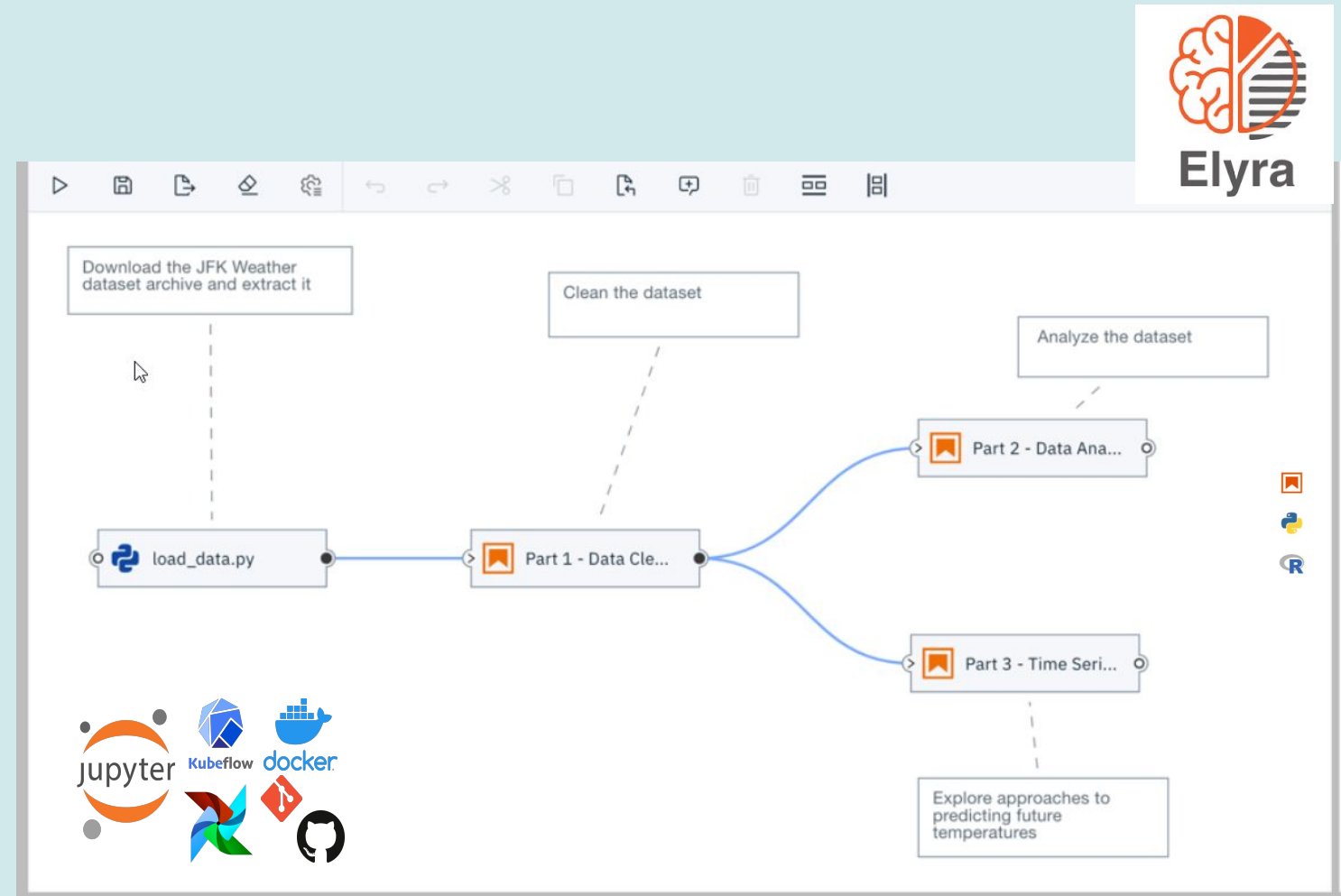
- Node-RED is a flow-based programming tool for connecting hardware devices, APIs, and services.
- Built on NodeJS and the D3.js library.
- The minimal structure are the nodes.
- Nodes are organized in flows that connect nodes.



<https://nodered.org>

What is Elyra?

- Elyra is an open-source project for developing and running machine learning workflows in JupyterLab.
- Provides tools for users to create visual pipelines for machine learning workflows.
- The minimum structure for a workflow are the nodes (Python script, R script, and Jupyter notebook).
- Elyra is programmed in Python and allows the use of many libraries focused on data analysis.



<https://elyra.readthedocs.io/en/stable/>

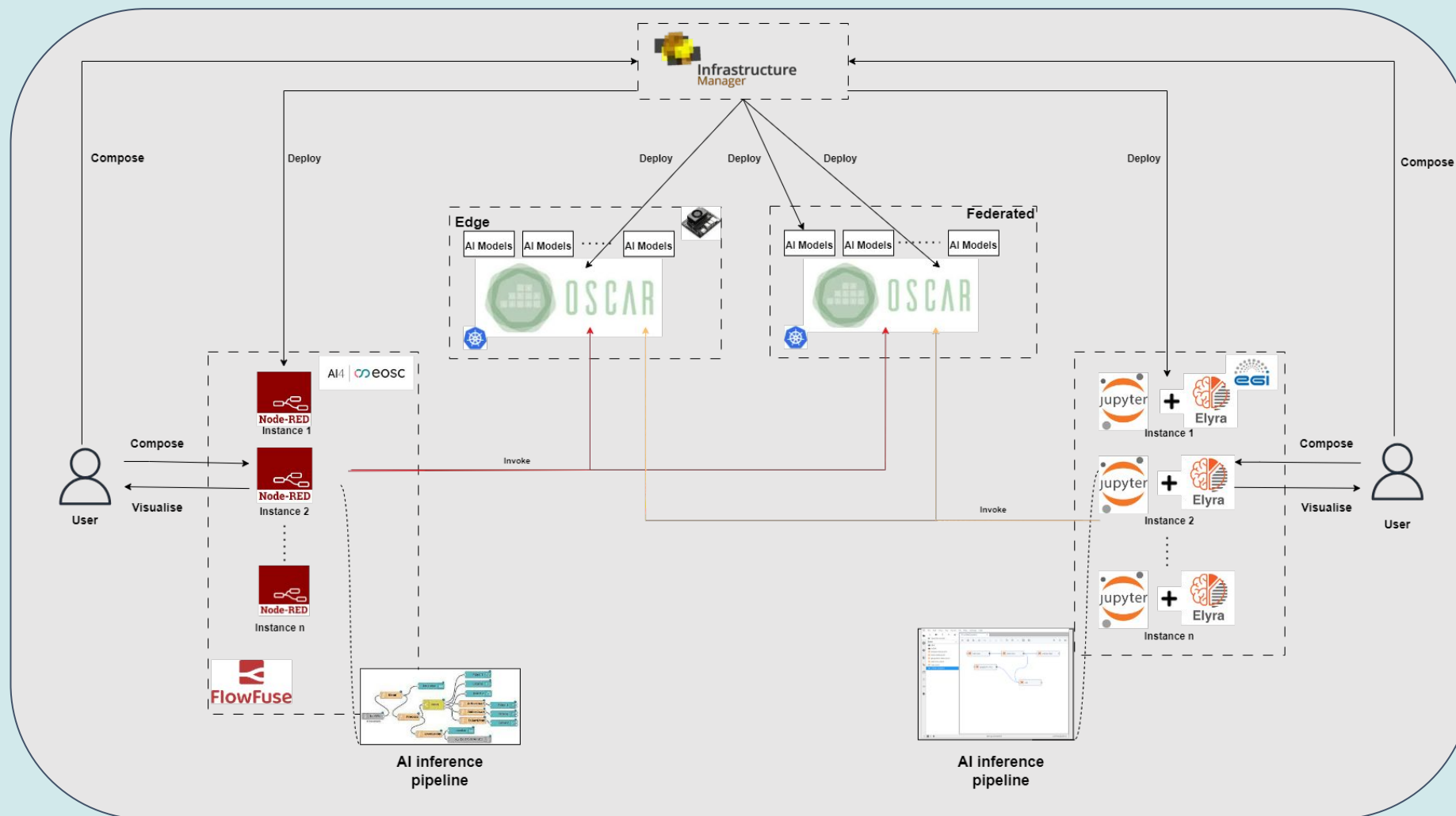
AI4Compose: Low-code composition of AI inference pipelines.

AI4

eosc



Co-funded by
the European Union



<https://github.com/ai4os/ai4-compose/tree/main>

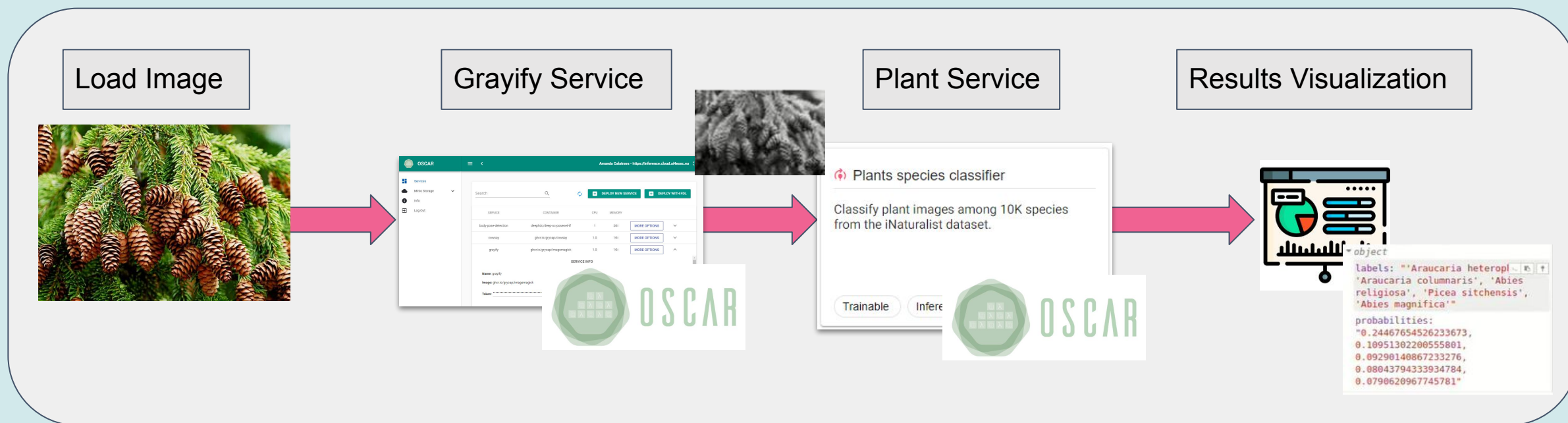
Composing AI pipelines: an example

AI4

eosc



Co-funded by
the European Union



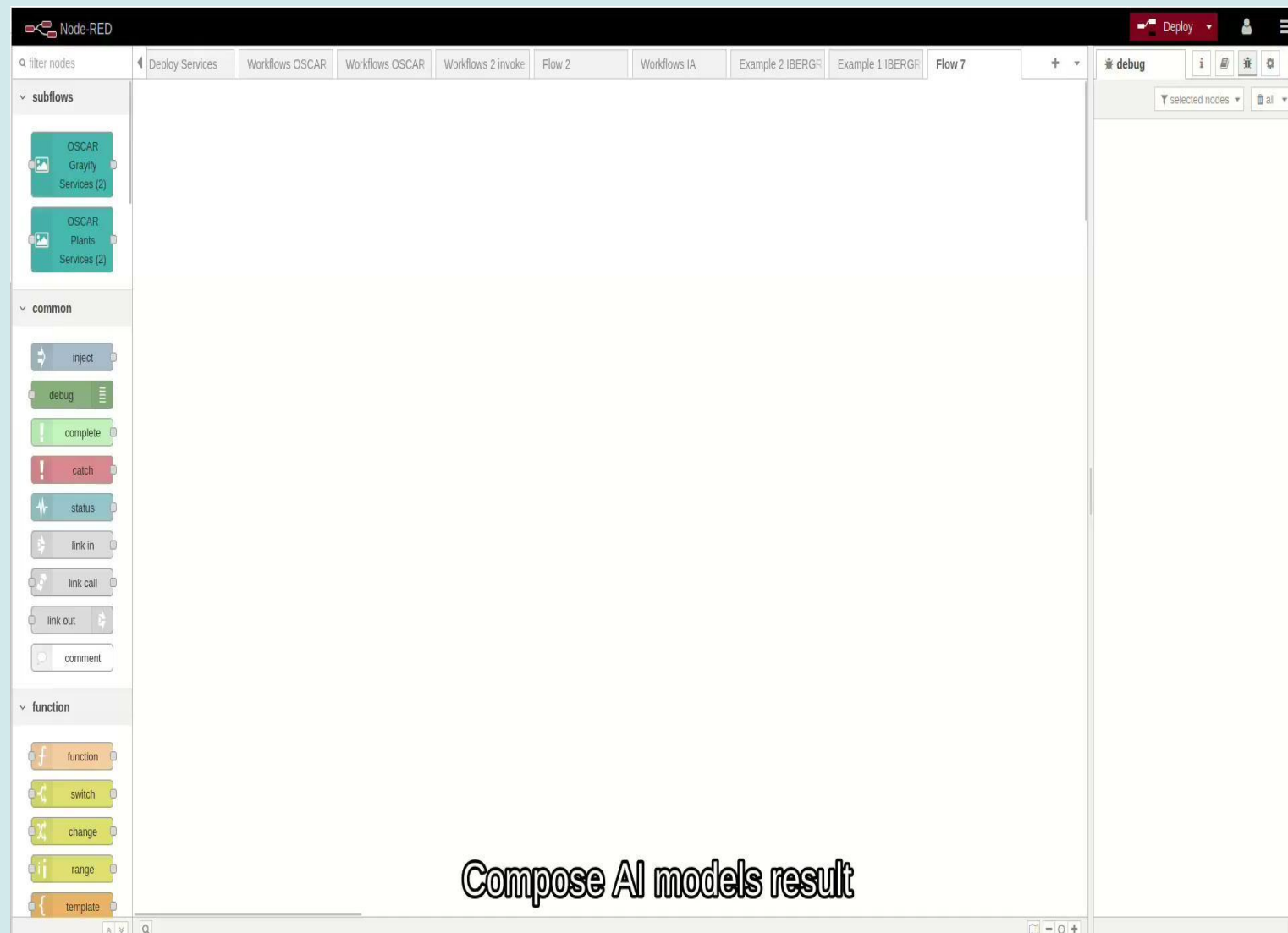
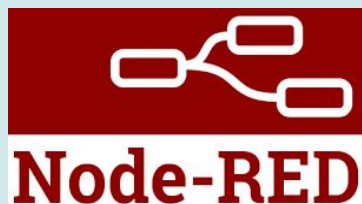
Composing AI pipelines: an example

AI4

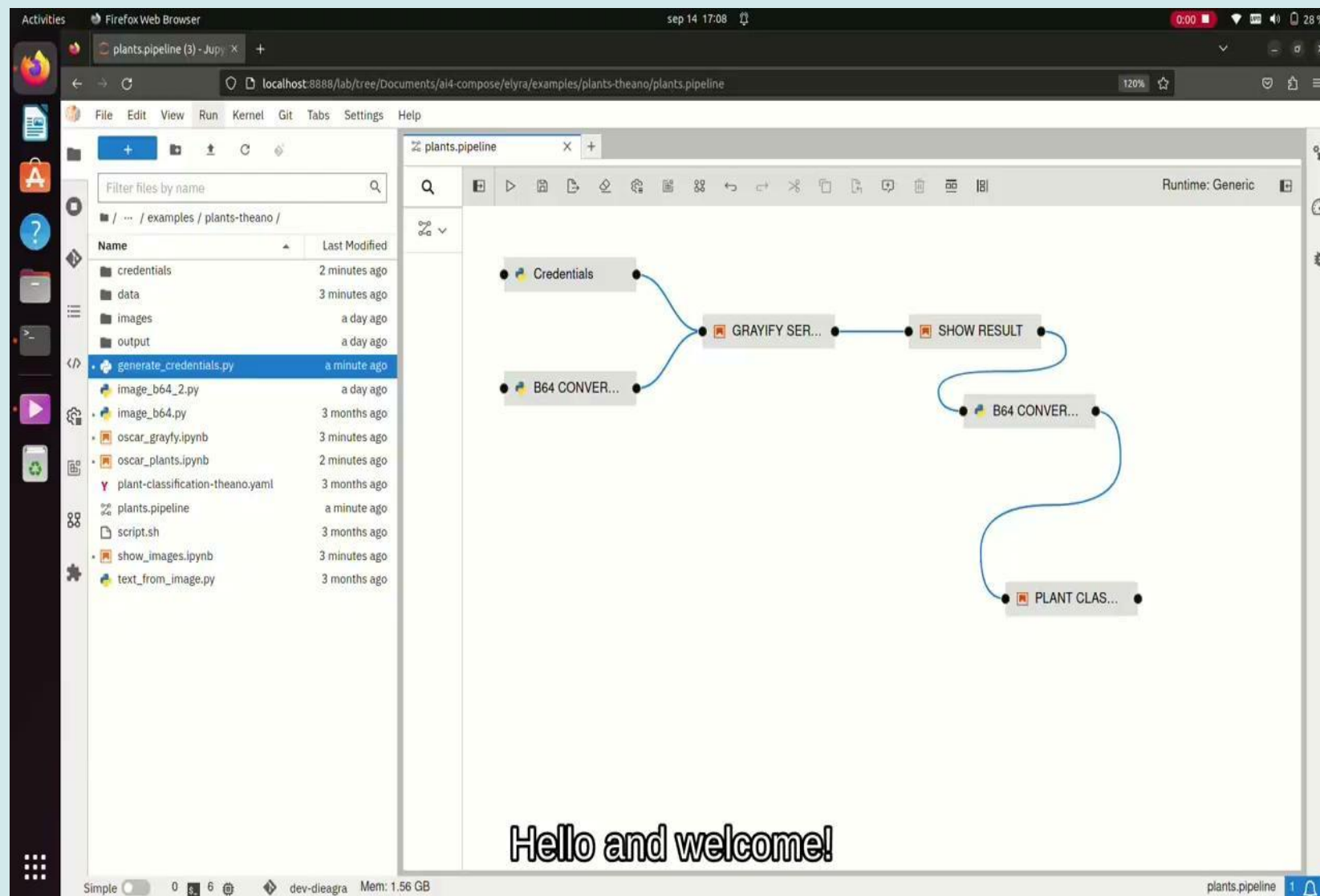
eosc



Co-funded by
the European Union



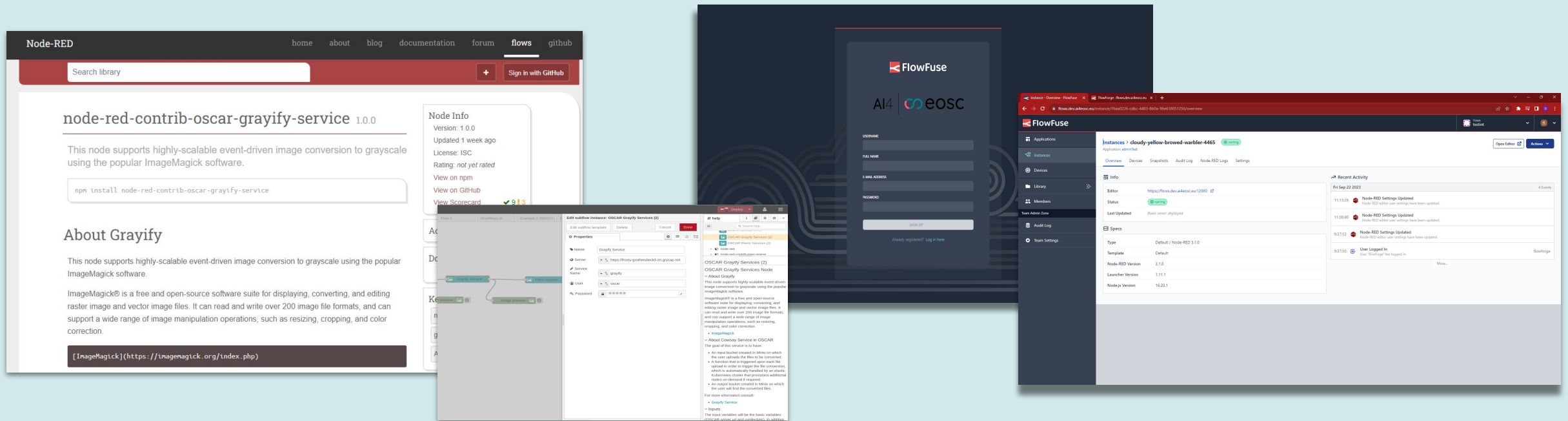
Composing AI pipelines: an example



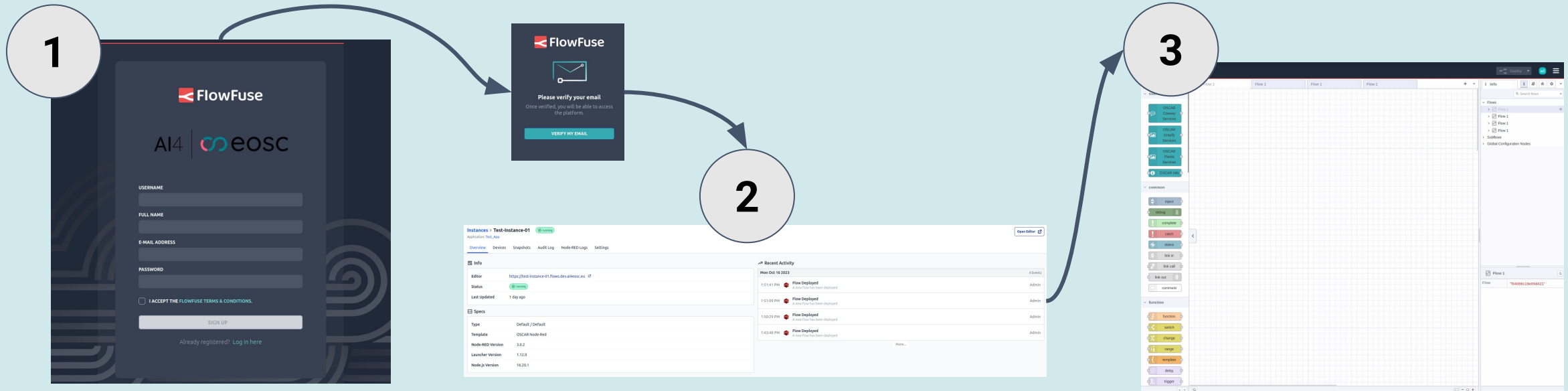
The screenshot displays the Elyra JupyterLab interface. On the left, a file browser shows a directory structure with files like 'credentials', 'data', 'images', 'output', and 'generate_credentials.py'. The main workspace shows a pipeline workflow with the following steps: 'Credentials' and 'B64 CONVER...' (likely B64 encoding) feed into 'GRAYIFY SER...', which then feeds into 'SHOW RESULT'. 'SHOW RESULT' feeds into 'B64 CONVER...' (likely B64 decoding), which finally feeds into 'PLANT CLAS...'. The interface includes a top menu bar with options like File, Edit, View, Run, Kernel, Git, Tabs, Settings, and Help. The bottom status bar shows 'Simple' mode, 0 errors, 6 warnings, and a memory usage of 1.56 GB. The text 'Hello and welcome!' is visible at the bottom of the workspace.

AI4Compose: Innovation in AI4EOSC

- New modules to interact with OSCAR, published in the official Node-RED library (*more coming soon*):
 - <https://flows.nodered.org/search?term=AI4EOSC>
- [Flowfuse](https://flowfuse.com): to manage users and multiple instances of Node-RED. Official instance of the project:
 - <https://forge.flows.dev.ai4eosc.eu>
 - OSCAR templates ready to offer a pre-configured Node-RED instance with OSCAR support.



AI4Compose: Flowfuse User Journey



Sing up [in the portal](#)

- Wait for admin approval
- OIDC integration desired, currently not possible

Create and/or connect to a Node-RED instance

- Inside "Applications"

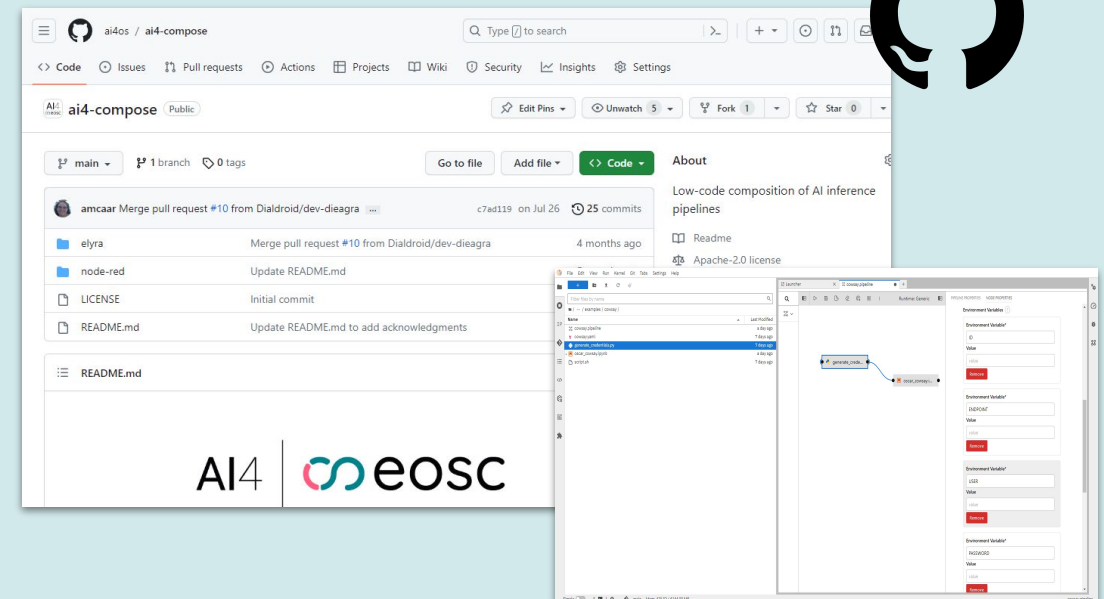
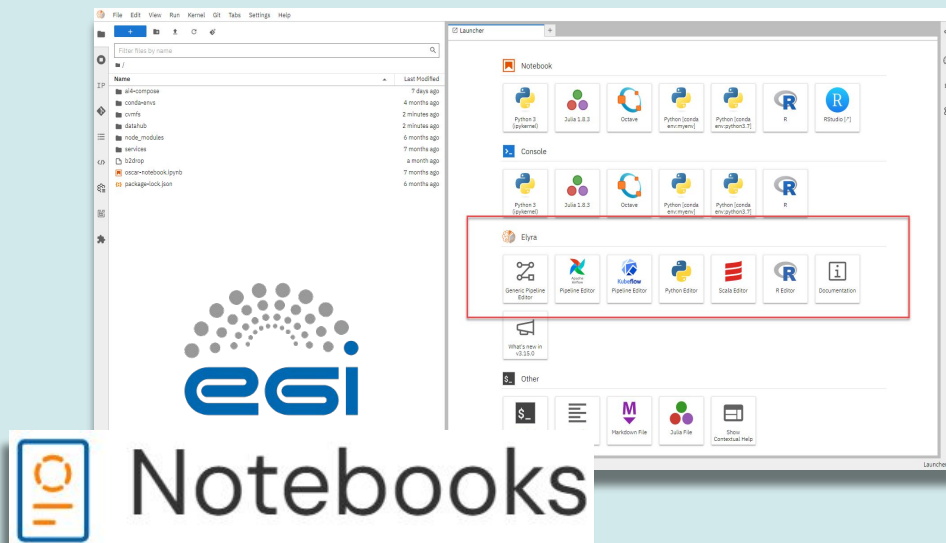
Use the Node-RED instance:

- Import Flows & Subflows
- Design your pipelines
- Execute them

➔ [Official documentation available in Confluence.](#)

AI4Compose: Innovation in AI4EOSC

- Elyra in [EGI Notebooks](#): we have contributed to have official support of Elyra inside the service, so users can easily access the tool.
- Notebooks to be used in Elyra for AI models of the DEEP Open Catalog:
 - <https://github.com/ai4os/ai4-compose/tree/main/elyra/examples>
- Documentation in [Confluence](#): soon we will contribute to the official documentation of the project: <https://docs.ai4eosc.eu/en/latest/index.html>



Conclusions



- AI4Compose offers a visual support (drag & drop + customization) to compose AI inference pipelines.
- It minimizes the orchestration effort:
 - Multiple AI models can be triggered for inference and later aggregate the results for enhance accuracy.
- Reusable functions:
 - Pre-defined workflows can be created to facilitate interaction among the AI models in the DEEP Open Catalog.
 - Specific nodes can be created for the different AI models for a simpler definition of workflows.
- Workflows along the computing continuum can be supported (e.g. OSCAR clusters in disparate computing infrastructures).
- Each node can be configured to invoke an OSCAR service within a specific OSCAR clusters.
- Deployment of pre-defined Node-Red instances to facilitate AI composition of workflows (multi-tenant support thanks to FlowFuse).

AI4

 eosc



Co-funded by
the European Union



AI4EOSC



amcaar@i3m.upv.es



ai4eosc.eu

Thank you! Any questions?

Amanda Calatrava