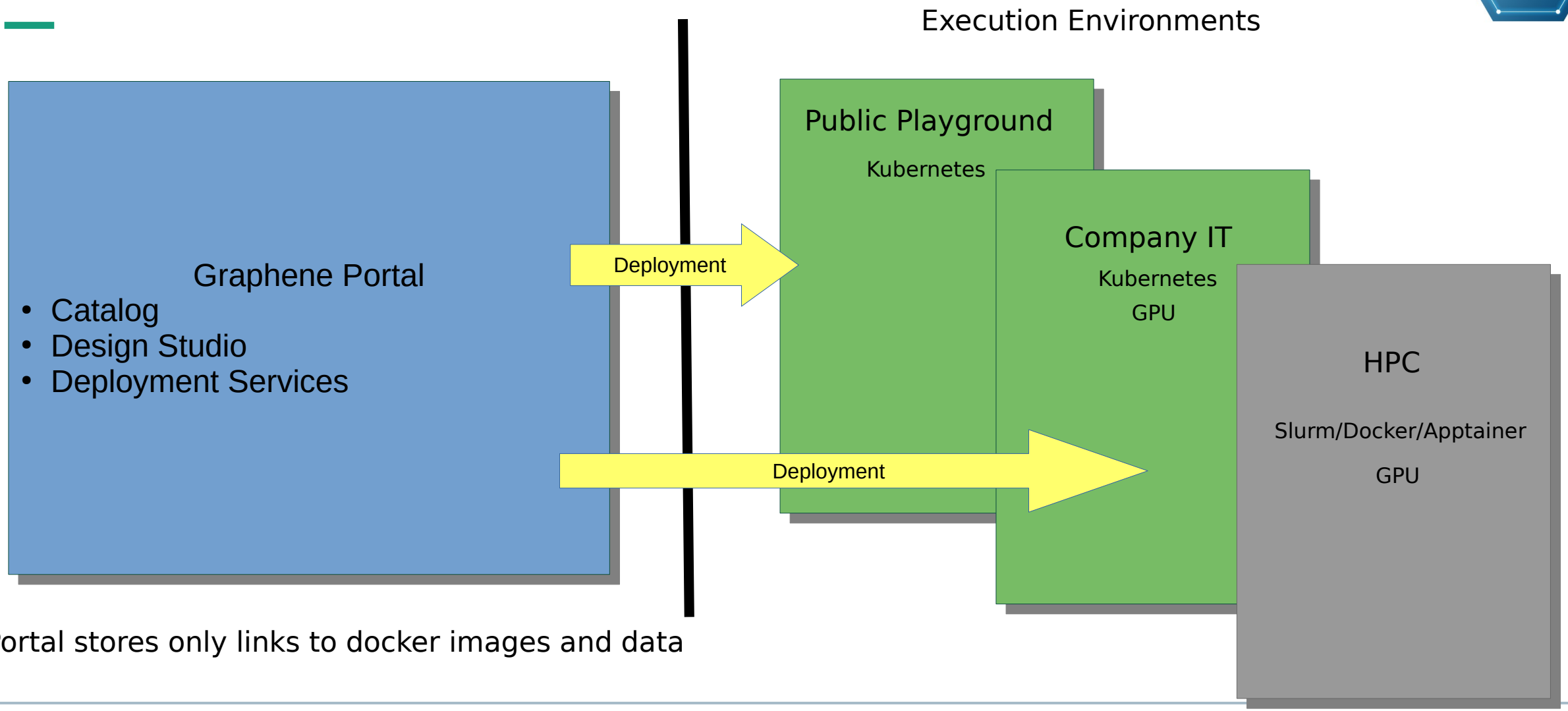
EclipseCon 2023

Applied AI with Eclipse Graphene

Martin Welß
Project Lead Graphene

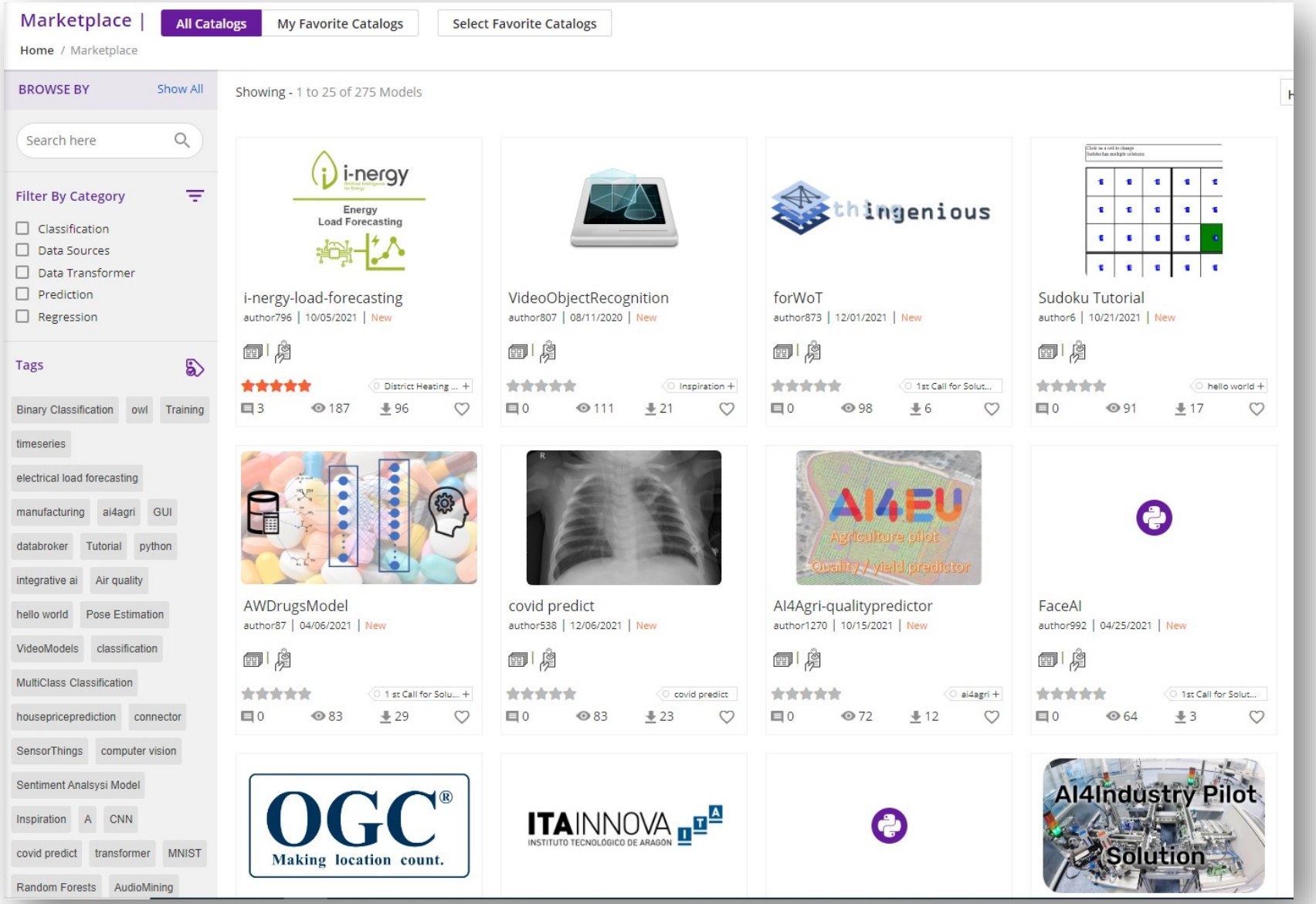
System Overview



Portal stores only links to docker images and data

AI Model Catalog

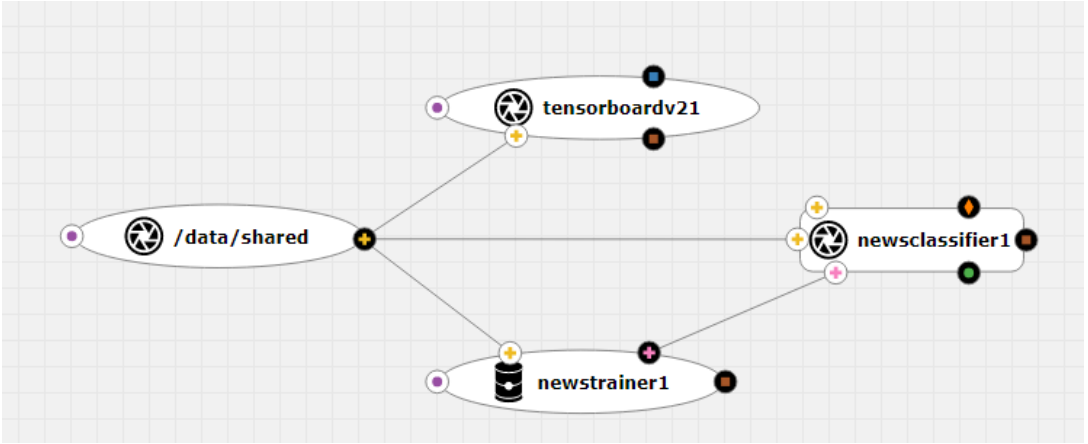
- Catalog of re-usable AI Modules
- The AI-Modules are Docker Containers
- Work in teams
- Support for commercial and open source licenses for models



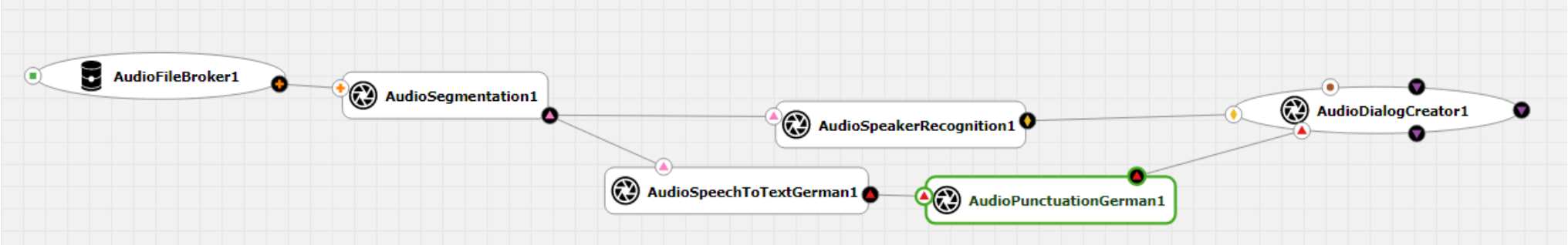
Fraunhofer IAIS

Compose AI Pipelines in the Design Studio

- Trainingpipeline with Tensorboard Integration



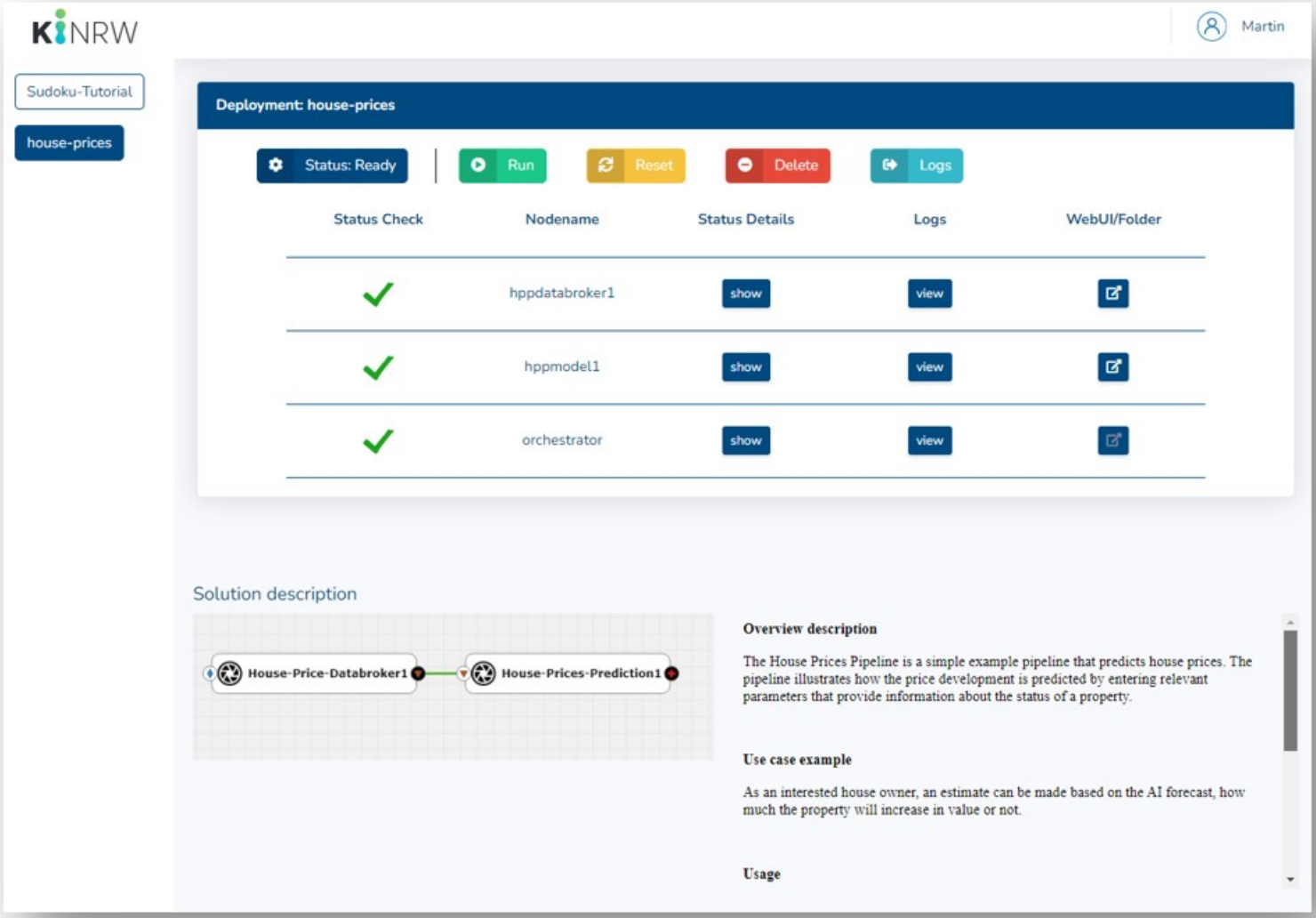
- Shared-folder concept
- Show matching models
- Application Pipeline



Fraunhofer IAIS

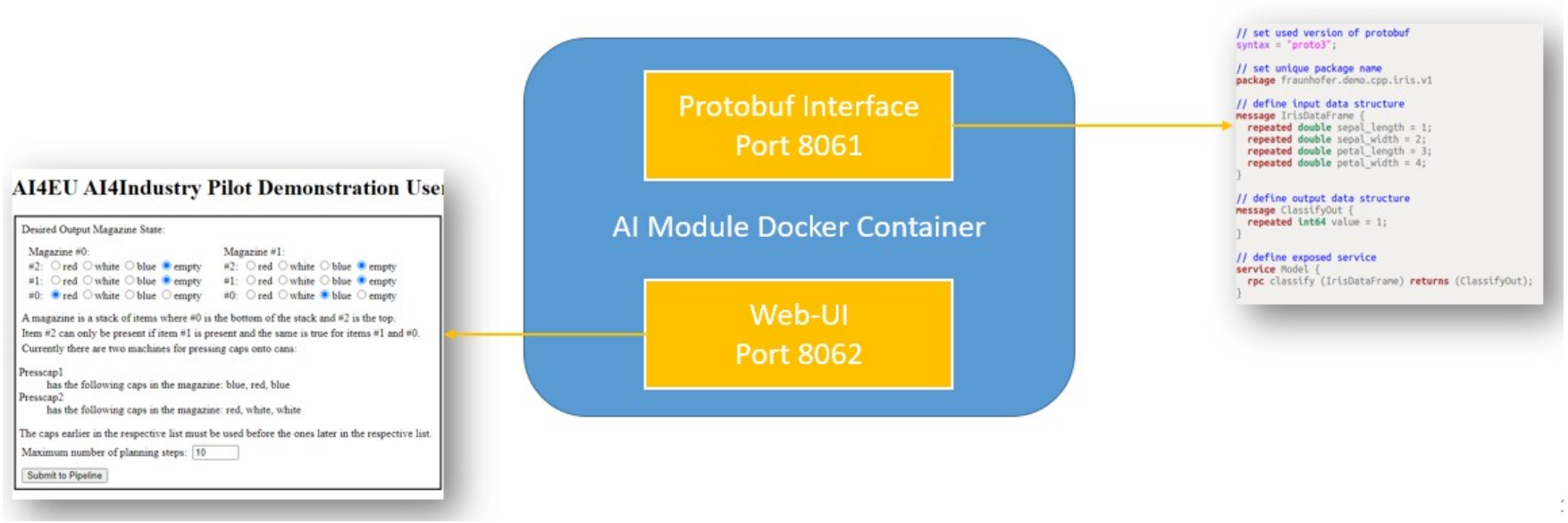
AI Playground

- Execution Environment
- One-Click-Deployment
- Based on Kubernetes
- Playground-App is a Graphene repo
- One-Click Update: Docke images are re-pulled on Reset



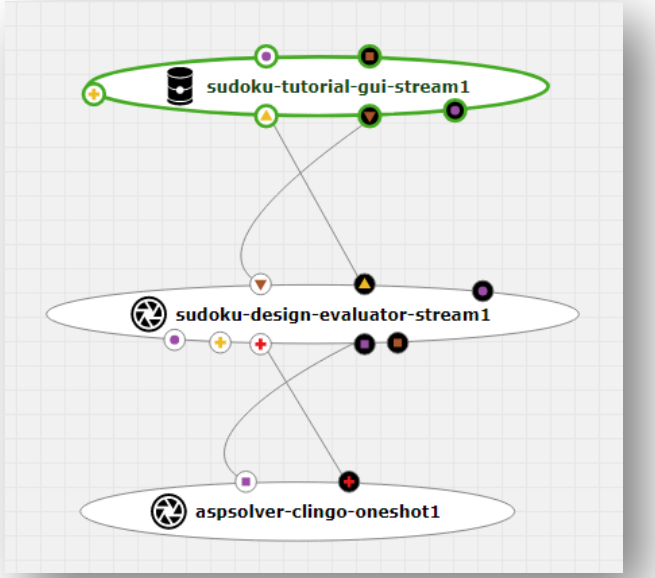
Container Specification for composable AI-Modules

- Docker Container with additional properties
- public interface in protobuf / grpc
- optional Web-UI (highly recommended)
- No Lock-In: it continues to be a normal Docker container
- If done right, the same Docker container can be used across different execution environements (e.g. GPU or no GPU)



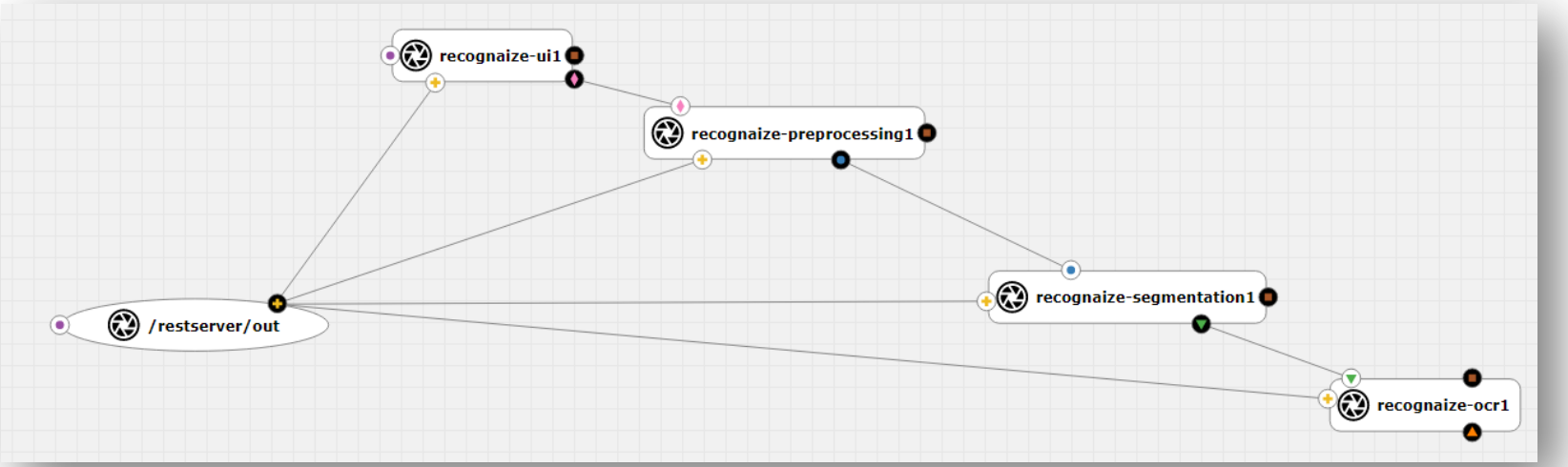
Generic Parallel Orchestrator

- The orchestrator executes the pipeline according to the topology file (generated by the design studio)
- The orchestrator dispatches the messages among the nodes following the edges
- Thanks to protobuf/grpc the communication stubs can be generated on the fly
- Support for grpc streaming:
  - Sensor-Input
  - UI-Input
  - Media-Streams
  - The edges can be served in parallel
- Different orchestrator implementations possible



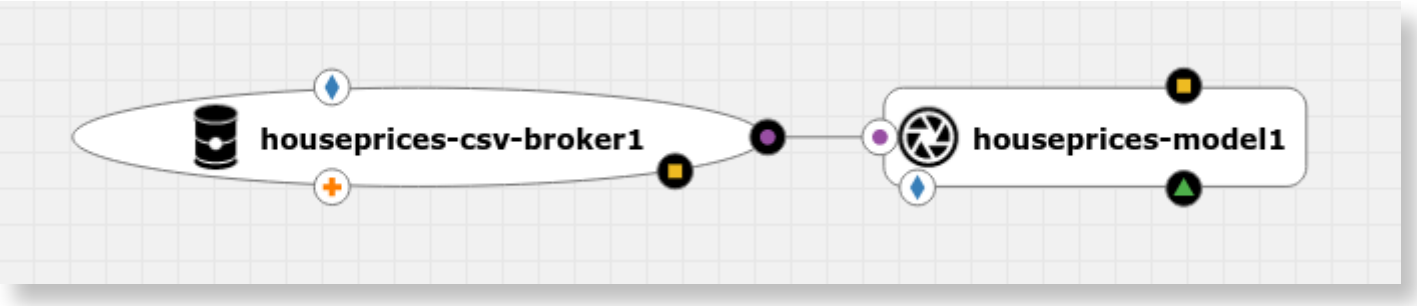
Pipeline Definition (Topology)

- A JSON-file which contains nodes and edges (technology neutral)
- Is generated by the Design Studio, but can be generated elsewhere (e.g. for AutoML)
- During deployment, the specific configurations for an execution environment are generated (MLOps)
- Different orchestrators are possible
- Shared Folder support



Databroker

- A Databroker is a shallow component that makes data available to the pipeline
- Usually the first node of a pipeline
- Specific for an execution environment (e.g. company IT)
- Can access the data source (credentials)
- Does not contain the dataset!
- We have already a prototype of an Eclipse Dataspace Connector



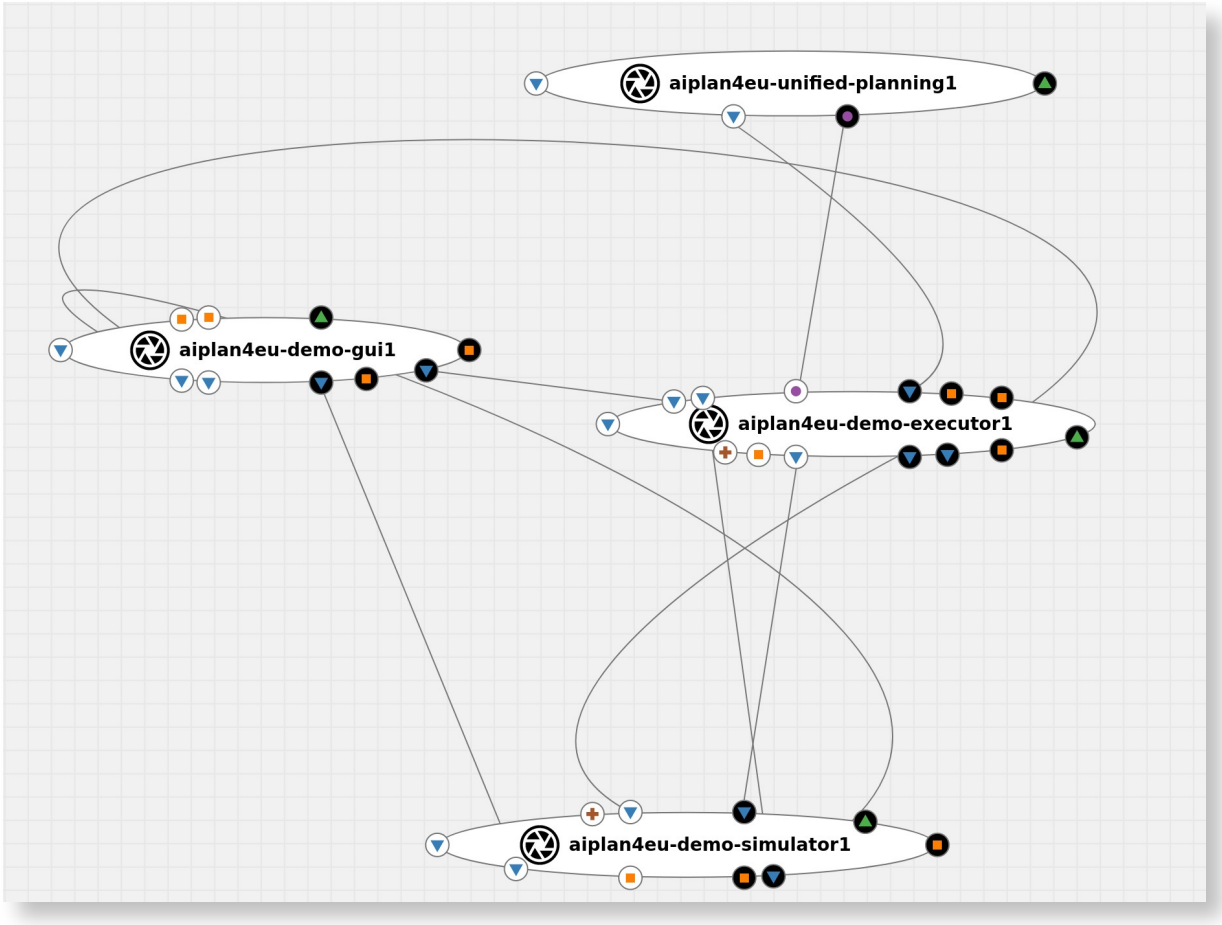
What’s New in Graphene Release 1.1?

- Installation uses Kubernetes 1.26, Calico 3.26, Helm 3.12.3, Ingress-Nginx 4.7.1
- Playground-App is faster
- Support for more protobuf features:
  - enums
  - oneof
  - recursive message definitions
  - nested definitions
- Better support for directed cyclic topologies
- Jupyter-Connect
- First-Steps for Reproducibility
- And Benchmarking (Metrics)

```
2   syntax = "proto3";
3
4   // As in s-expression, an Expression is either an atom or list repres
5   message Expression {
6       // If non-empty, the expression is a single atom.
7       // For instance `3`, `+`, `kitchen`, `at-robot`, ...
8       Atom atom = 1;
9       // If the `atom` field is empty, then the expression is a list of
10      // typically representing the application of some arguments to a
11      // For instance `(+ 1 3)`, (at-robot l1)`, `(>= (battery_level) 2
12      repeated Expression list = 2;
13
14      // Type of the expression. For instance "int", "location", ...
15      string type = 3;
16      // Kind of the expression, specifying the content of the expressi
17      // This is intended to facilitate parsing of the expression.
18      ExpressionKind kind = 4;
19  }
20
```

Cognitive Architectures

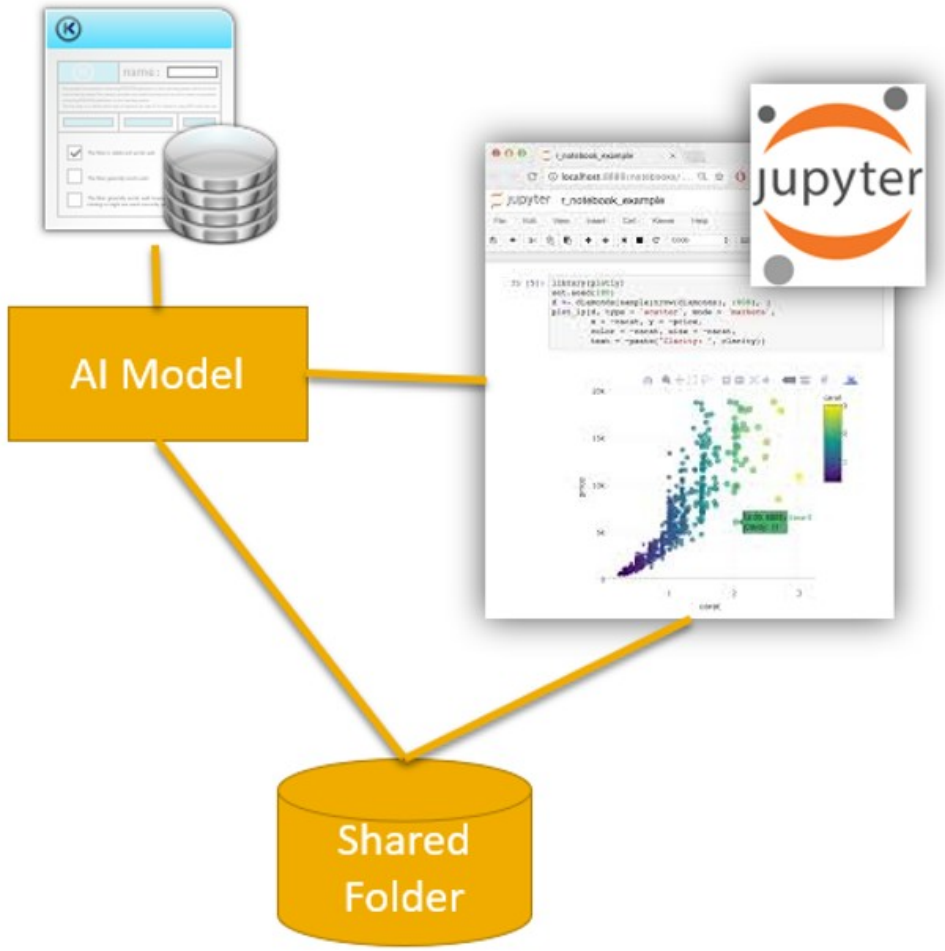
- To fully leverage AI power, we need to combine LLM with other AI technologies:
  - Planning, Logic, Knowledge Graphs, Filter, Image Object Detection, OCR, Audio
  - Constant Learning, Feedback loops
  - Cognitive Architectures require directed cyclic graphs (supported by Graphene 1.1)
  - And recursive data structures (supported by Graphene 1.1)



Maze-Planner

Jupyter Connect: Interactive Model Exploration

- If the user deploys a single model
- Graphene automatically generates a suitable Juypter-Lab container
- and adds the protobuf interface definition
- and a shared folder
- for comfortable interaction with a model / AI-Module



Metadata / Benchmarking

Capture Metadata about:

- System (CPU/GPU/RAM)
- Datasets
- Docker Image Checksums
- Different Metrics
- See Container-Spec

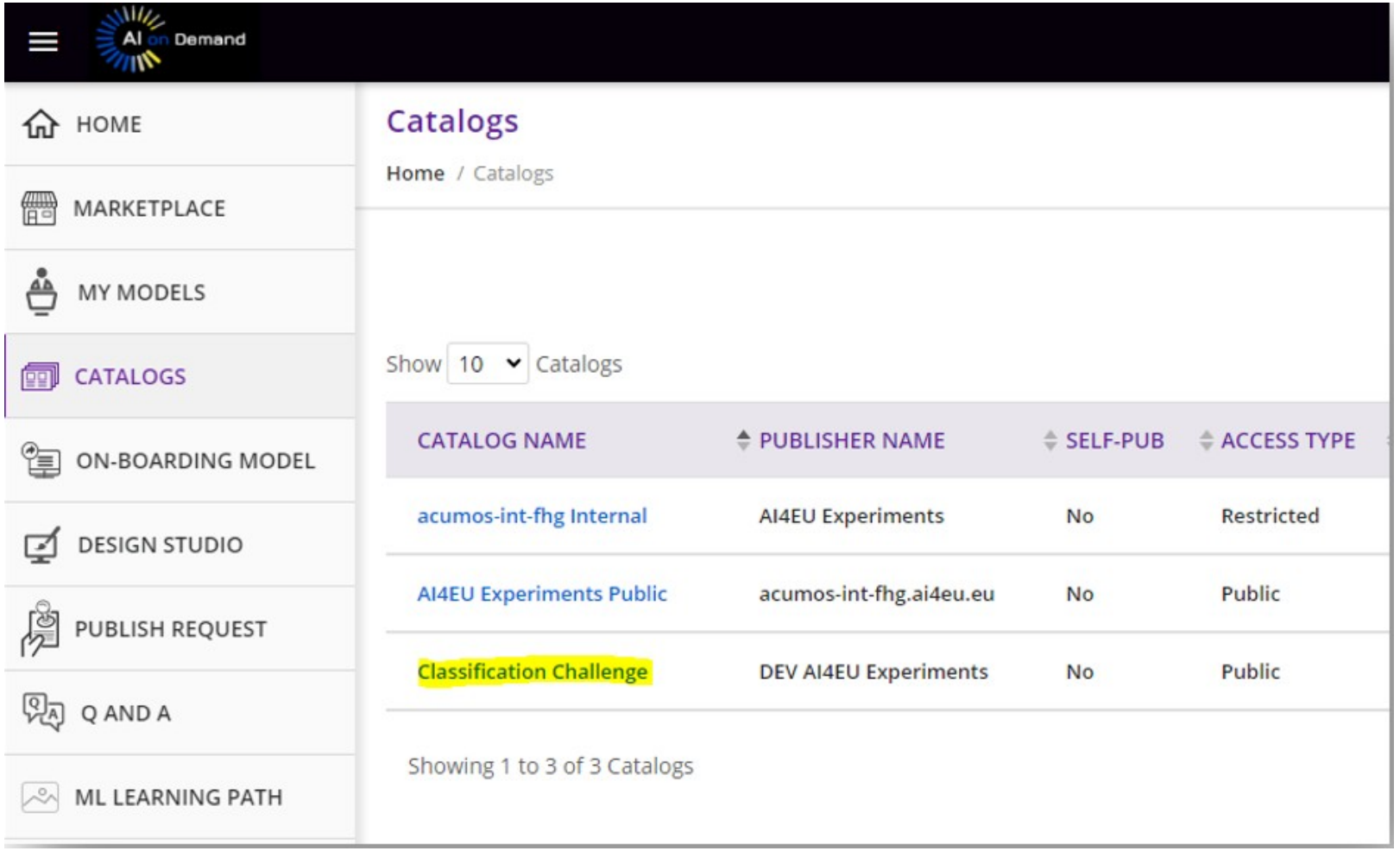

```
"checksum": "docker-pullable://cicd.ai4eu-dev.eu:7444/tutorials/house_price_prediction/hpp_databroker@sha256:6d8c5f203a4bcc8a794b54d82a7428b36ffeb9f53b164968f576442e2bdde89f",
"dataset_features": {
  "type": "aiod-dataset/v1",
  "datasetname": "The Reuters Dataset",
  "description": "http://kdd.ics.uci.edu/databases/reuters21578/README.txt",
  "size": "4MB",
  "DOI_ID": "Not available"
}
```

```
"metrics": [
    {
        "type": "classification-training-metrics/v1",
        "date_time": "2023-09-07 07:28:51",
        "accuracy": 0.9216,
        "validation_loss": 0.9244,
        "status_text": "success"
    },
    {
        "type": "classification-testing-metrics/v1",
        "date_time": "2023-09-07 07:28:51",
        "F1 Score": 0.85,
        "Specificity": 0.90,
        "ROC-AUC": 0.92,
        "status_text": "success"
    }
]
```

```
"running_time": "345s",
"system_info": {
  "system_name": "mwtest",
  "fqdn": "test.playground.org",
  "cpu": "10",
  "gpu": "",
  "memory": "65851412Ki"
}
```

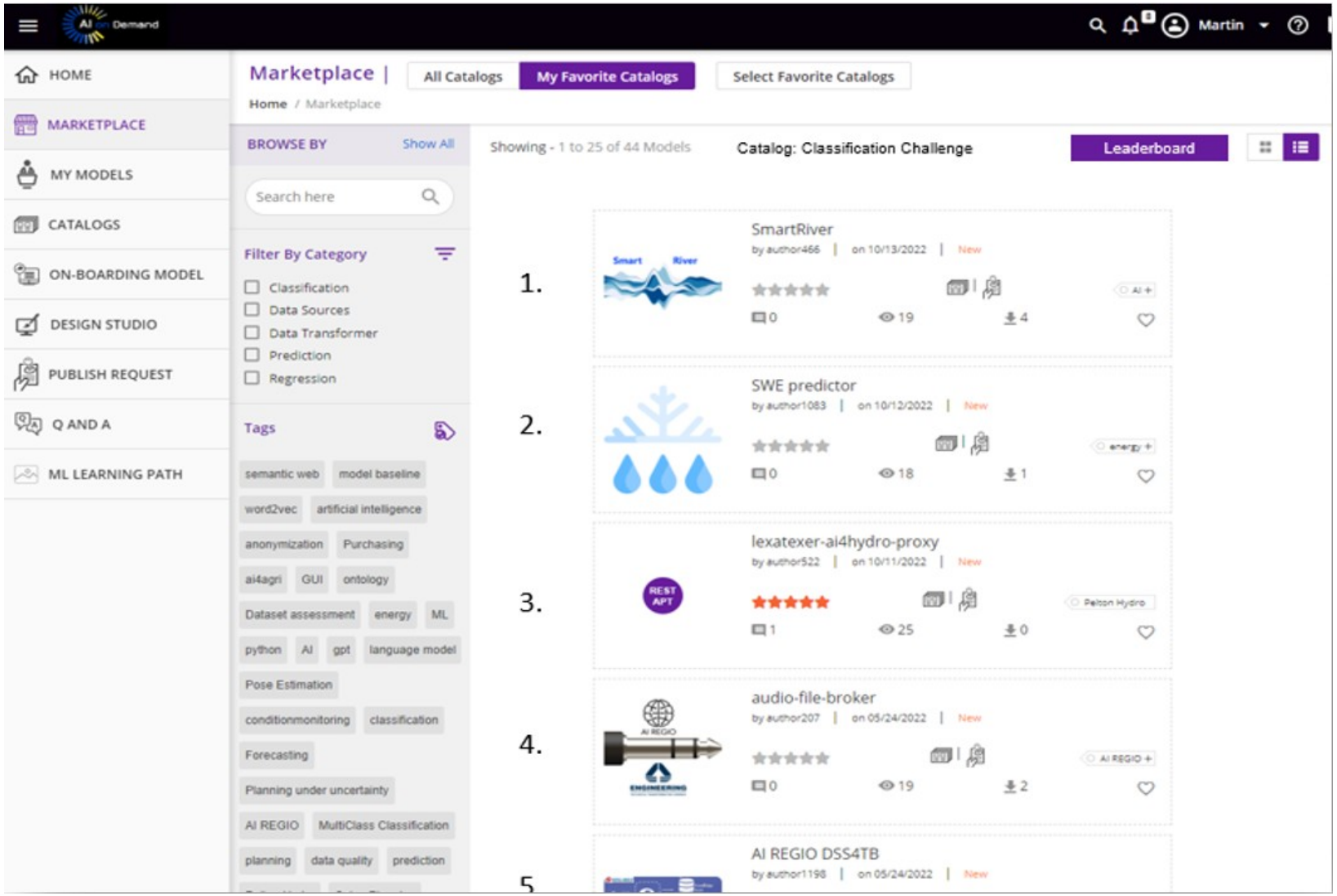

Challenges (Planned)

- Challenges are special types of catalogs

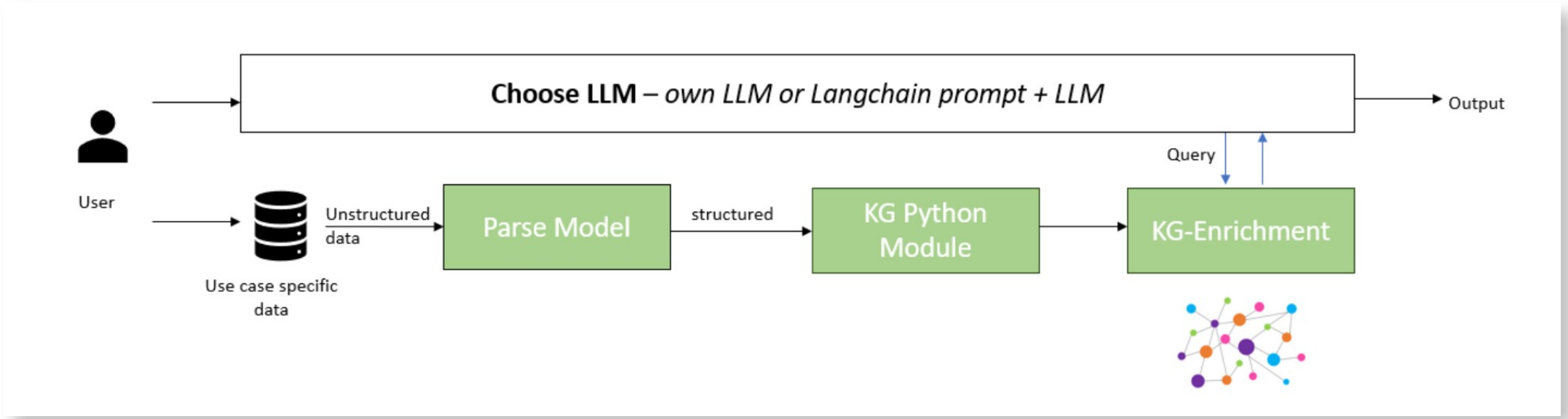


Leaderboard (Planned)

- Order based on Metrics



LLM Pipelines: Grounding LLM with Knowledge Graphs

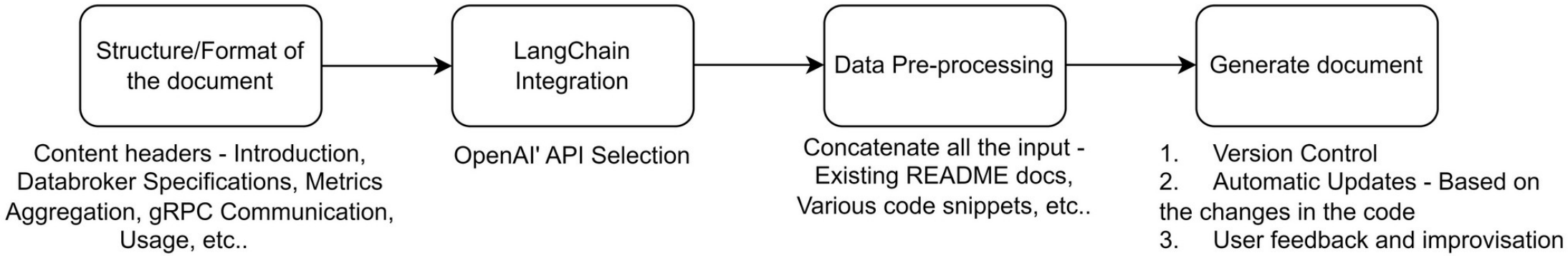


Work in progress

LLM Pipelines: Interactive Graphene Documentation

The issue aims at automatically generating documents using large language models and LangChain. By doing so, we can effectively streamline the creation of comprehensive and informative documentation.

The initial overview of the process is shown below,



Work in progress

Graphene Use Cases

- General: Building AI-pipelines from re-usable AI-modules
- AI Research: model training, benchmarking, Jupyter-Connect
- Education: practical live demonstrations and student excercises in the playground
- MLOps: simplify deployment, training pipelines, data cleaning pipelines, can support many tools
- Model showcase, catalog (internal, external)
- Application: let non ai-experts compose and deploy pipelines for thier business domains
- Collaboration: work in mixed teams on PoCs
- Automated certification

Supporting Projects: Graphene Funding at least until 2028

- KI.NRW
- AI4Europe
- HumanE-AI Net
- West-AI
- Deploy-AI (starting 2024)

Useful Links

- AI4EU Experiments: https://aiexp.ai4europe.eu/#/home
- KI.NRW AI-Lab: https://www.ai-lab.nrw/
- Eclipse Graphene Project: https://projects.eclipse.org/projects/technology.graphene
- Eclispe Graphene Gitlab: https://gitlab.eclipse.org/eclipse/graphene
- Container Spec: https://gitlab.eclipse.org/eclipse/graphene/tutorials/-/tree/main/Container_Specification
- Tutorials: https://gitlab.eclipse.org/eclipse/graphene/tutorials

Contact

Martin Welß
Senior AI Solution Architect
martin.welss@iais.fraunhofer.de

Fraunhofer-Institut für intelligente Analyse- und Informationssysteme IAIS
Schloss Birlinghoven 1
53757 Sankt Augustin
Germany

www.iais.fraunhofer.de