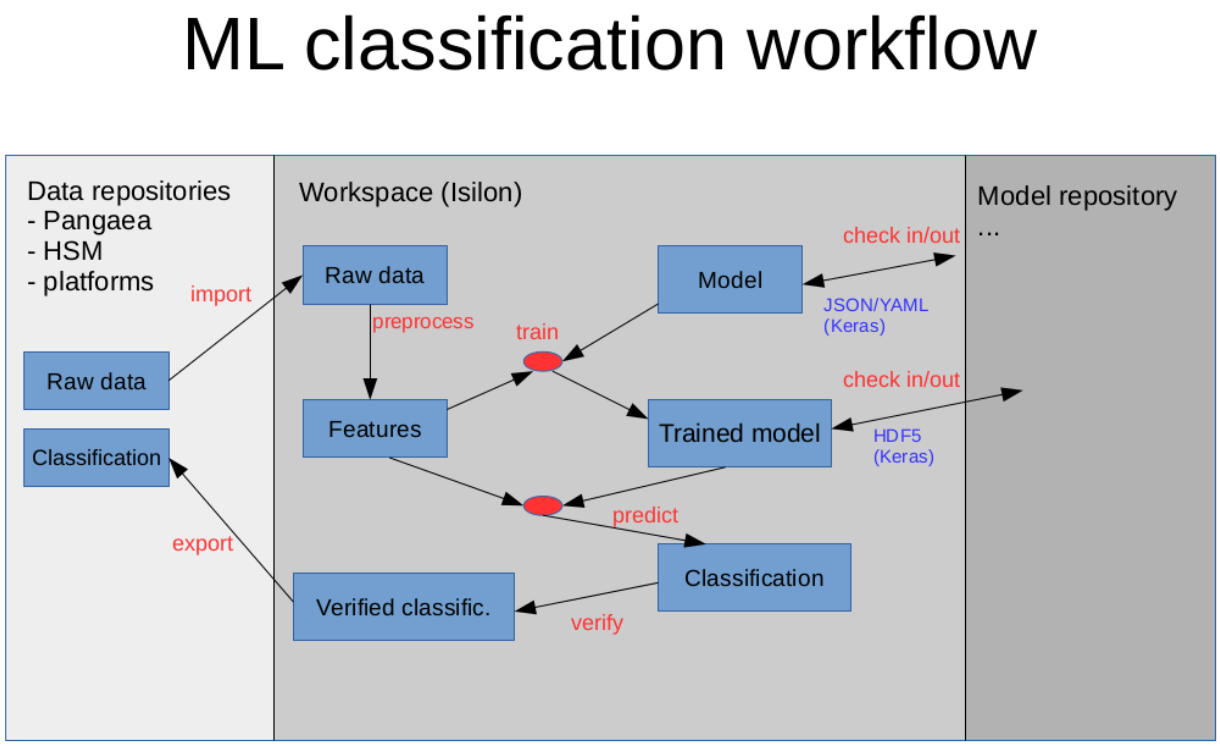
ML Workflow

Here we describe a generic workflow that can be followed when using machine learning within the AWI infrastructure. For now, we focus on classification tasks: other applications such as regression, clustering, etc can be considered in the future.

This document is organized as follows:

- In the Storage Section, we discuss where the data is stored at each phase of the workflow.
- In the "Model Repository" Section, we discuss how ML models should be organized and archived.
- In the "Computing resources" Section, we discuss the different computing resources that can be used for training and running ML models.

The general data flow is illustrated in the image below.



Storage

Data analytics is performed on some initial raw data which has been produced elsewhere, e.g. by sensors installed on a research vessel. We distinguish between two categories of data storage:

- **Data archive**: a read-only data storage in which the raw data is archived for access by different users.
- **Workspace**: a read-write data storage that is used locally by a project for storing custom software tools, ML models, input data, intermediate files, and final results.

The workspace

While it is possible to use the local disk of a personal computer as a workspace, it is recommended to use the file service (see the documentation) provided by the Computing and Data Centre at the AWI. In the documentation page you find information on how to mount / access files on the file service from your personal computer. The file service has the following advantages:

- All users working on the same project can share the workspace.
- All the data in the workspace is automatically backed up.
- Data in the workspace can be accessed both from a PC in the AWI intranet (e.g. a laptop) and from AWI computing facilities. In this way, each user can develop a model prototype on their own PC and then switch to more powerful computing resources that can access the same data.

Projects on the file server are found in the projects folder. A project can be created under cloud.awi.de (My projects).

Each project directory should contain at least:

- One read-only folder for the input raw data.
- One src folder for the source code needed to train the model.
- One lib folder for storing libraries that are common to the project.
- One model folder for storing the trained models.
- One or more subfolder in which preprocessed data and features are stored.

The data archives

The input raw data for an analytics task can come from different sources, called data archives in this document. Within the AWI, there are three main data sources:

- **Pangaea**: An information system archiving georeferenced data from earth system research. See the Pangaea homepage.
- **Hierarchical Storage Management (HSM)**: A high-capacity tape-archive. See the HSM documentation.
- **Platforms**: An online storage in which sensor data is archived.

Additionally, external data archives can also be used by a project, e.g. Ecotaxa.

Currently, users have to transfer the raw data to the workspace manually. We plan to develop a library for transferring data between data archives and a project workspace using a unified API.

Model Repository

ML libraries and frameworks such as Keras normally support saving models to file for later use. Models are made of two different components:

- **Model architecture**: this specifies the structure (e.g. network architecture) and possibly the hyperparameters of a model.
- **Model parameters / weights**: these parameters are learned during training and are used afterwards to run the model in a production environment.

It is important to have archive both model architectures and parameters in order to provide reproducibility. Additionally, one should store metadata such as:

- References to the data sets used for training (for parameters only).
- Other metadata such as computing resources used for training, running time, and so on.
- Model versioning: using symbolic names and a version control system like git it is possible to recall an existing model efficiently.

We are currently still investigating the requirements for archiving ML models and evaluating appropriate software tools.

Computing Resources

Going from the raw data to an ML model involves several data processing steps:

- Data cleaning and normalization: data may come in different files, possibly containing errors or slightly different formats (different column labels, different units, and so on).
- Feature extraction: the features used in the machine learning method may not be present directly in the raw data. E.g. acoustic data analysis may be performed on a spectrogram: in this case, the spectrogram represents the features that must be extracted from the raw acoustic data.
- Training of the model: features from each data sample are fed into the model and used to learn model parameters.

All these data processing steps require computing resources. Since the workspace can be mounted as a network folder, users working inside the AWI-intranet can run their computations on their PCs. An alternative is to use computing resources of the data centre, which can be provided on request.