Small-Scale Science Platform for Active Deep Learning in Large Spectra Archives

PETR ŠKODA[1], OLEXANDR BURAKOV[2] AND ALISHER LAIYK[2]

[1]Astronomical Institute of the Czech Academy of Sciences
Ondřejov, Czech Republic

[2]Faculty of Information Technology of the Czech Technical
University in Prague, Czech Republic

Abstract

The emerging technology of Data Science platforms addresses the need to analyze Big Data directly where it is stored, because moving large volumes of data is nearly impossible. Typical platforms, like Pangeo, ESA DataLab, and SciServer are complex cloud-based systems, running on large clusters, providing hundreds of users access to petabyte-scale astronomical data archives. They use flexible orchestration of multiple containers to allow script-driven data processing and complex database queries, as well as interactive exploratory analysis via a web GUI.

Similar principles can be applied to smaller, task-specific infrastructures, particularly for machine learning experiments that demand continuous user interaction. A representative example is the active learning–based classification of millions of stellar spectra from large surveys. In such cases, iterative training requires annotators to label candidate spectra across multiple cycles while simultaneously accessing complementary data such as catalog metadata, images, and spectra from other surveys.

To address this need, we introduce the `ml-job-manager`, a dedicated multi-tier cloud platform designed for orchestrating parallel experiments with human-in-the-loop active deep learning on several million spectra from the LAMOST archives. The system adopts modern DevOps strategies, leveraging a microservices backend with RESTful asynchronous job control inspired by the IVOA UWS protocol. The entire environment in several containers can be rapidly deployed using Docker Compose, ensuring reproducibility and ease of installation.

1 Active Deep Learning

The Active Learning method is a novel approach to overcome the need for a large and representative training set, as is required by classical deep learning. It is based on idea that the algorithm will perform better if it is allowed to choose data for its training.

A machine learning algorithm combined with active learning queries unlabelled data samples to be labelled by an *Oracle* (usually a human expert). The samples are selected in batches of a given size according to a certain informativeness measure. Commonly used is the *uncertainty sampling*, which selects data with the least certain labelling (based on information entropy)

A particular version of Active deep learning was introduced into astronomical spectra analysis and successfully applied to search for emission-line spectra in archive of the LAMOST telescope by (Škoda et al., 2020). The code used in this paper is available on GitHub (Podsztavek, 2019).

2 Active Deep Learning Cycle

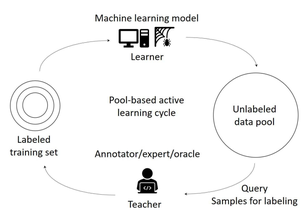
2.1 Labelling of New Training Samples by Oracle

The role of the Oracle (usually a human annotator) is to assign the correct label to the queried sample. The batch for Oracle labelling is selected from the sample where the network is least sure about the right classification. It is expected that the oracle can make the decision immediately just by looking at the visual properties of the sample.

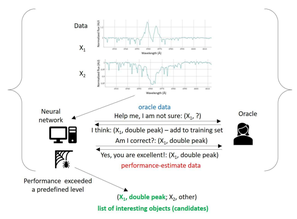
In real cases, other data and metadata about the sample may be helpful to estimate the correct label.

The batch samples from the labelled Oracle set are added to the training set and the network is retrained on them.

Below is an example of the active deep learning cycle as depicted in Mazel (2020).



Schema of active deep learning workflow



The role of oracle in assigning the correct label

2.2 Performance Estimation Sample

In addition to the Oracle data set, there is another set of samples shown to to the expert as well. When it is correcly labelled, it is used for calculating the classification accuracy of the network. It helps to identify the moment when the training iterations should stop, as the performance of the algorithm is not further improving. The samples in this set are selected randomly, and after computing the performance their label is discarded.

2.3 Network Iterations

After the network is trained on the new training set with added samples from Oracle set, it is asked to provide classification of all data. They are ordered by the cross entropy, and a some small number (a batch) of samples from the top is converted into a an oracle set shown to the expert. This repeats until the accuracy no longer improves.

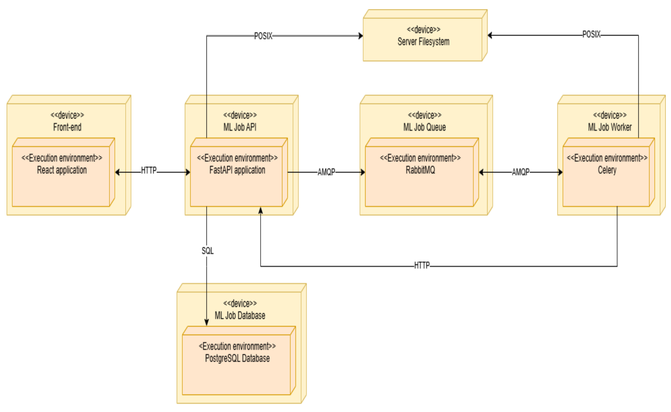
3 Cloud-based Science Platform `ml-job-manager`

A modern, cloud-based scientific platform `ml-job-manager` represents a comprehensive redevelopment of the now outdated system `VO-CLOUD` (Koza, 2015, 2017) and the its Active Deep Learning client (Mazel, 2020).

This redevelopment was undertaken in two Bachelor's theses (Burakov, 2025; Layik, 2025) within the software and web engineering program at the Czech Technical University in Prague.

The current implementation features two main types of processing units. The primary—and most sophisticated—component is the Active Deep Learning worker, offering an interactive interface for labeling and visualization. The secondary unit, the Pre-processing worker, efficiently handles large-scale data tasks such as converting, rescaling, trimming, and padding millions of spectral samples.

The system is composed of a backend server and a web-based graphical interface. Communication between these components utilizes a RESTful API, drawing inspiration from the IVOA Universal Worker Service (UWS) protocol (Harrison and Rixon, 2016).

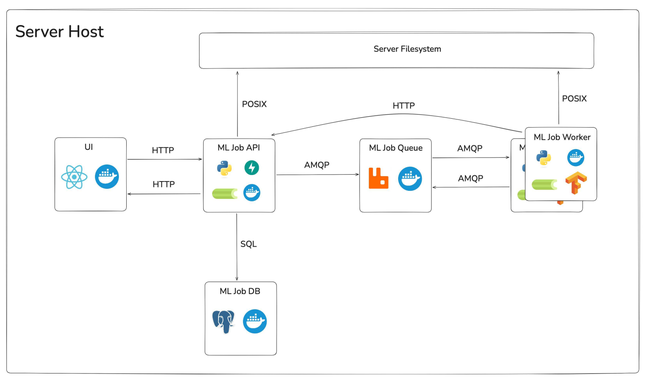


Schema of `ml-job-manager` science platform

4 Technology Stack of `ml-job-manager`

The platform is based on microservice architecture combining a number of modern technologies, frameworks and libraries. Its development was following the current DevOps best practices. The whole system is easily deployed in several containers using `docker-compose` (or compatible `podman-compose`).

- **Implementation language:** Python 3.13
- **Package management:** Poetry
- **Relational database:** PostgreSQL 17
- **Object relational mapping and migrations:** SQLAlchemy, Alembic
- **Message broker:** RabbitMQ
- **Web framework:** FastAPI
- **Data validation & serialization:** Pydantic
- **Job orchestration:** Celery
- **Async I/O:** aiofiles
- **Configuration format:** JSON
- **Scientific data handling:** Astropy (FITS), h5py (HDF5), NumPy
- **Machine learning:** Scikit-learn, TensorFlow , Keras
- **Containerization:** Docker (or Podman) composer
- **Web Frontend:** React, TypeScript, Tailwind CSS, HTTPX, Plotly



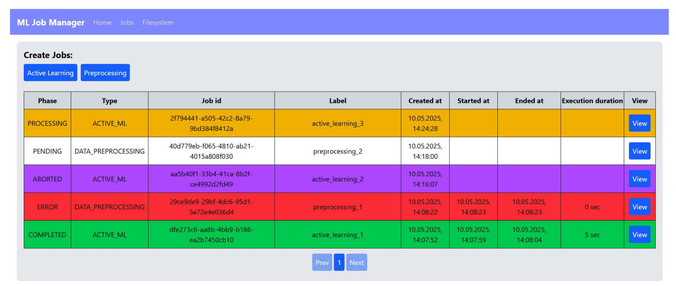
Software Architecture Diagram with logos of all frameworks used.

Source Code

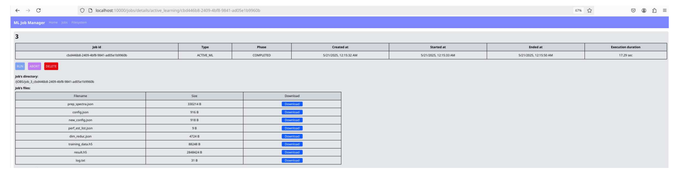
`https://github.com/bursasha/ml-job-manager`

`https://github.com/sparki0/mj-job-manager-client-modules`

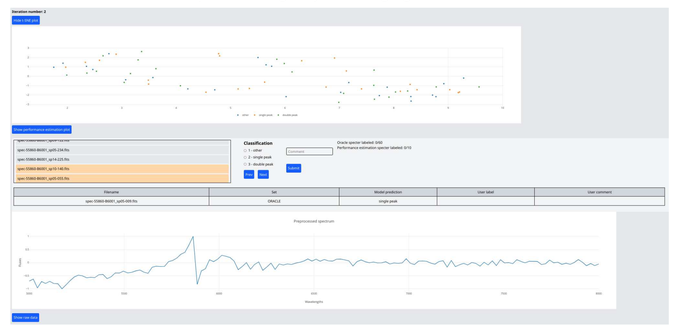
5 Front End Graphics Screen Shots



Window with jobs listing. Various types of jobs in different stages of processing are shown



Details about the individual job including the files belonging to it



Labelling window showing list of files selected for extending the training set after correct labelling by Oracle (grey) or performance estimation files (yellow). The previwed spectrum may be zoomed or panned interactively thanks to embedded Plotly JavaScript library. The dark blue buttons call various actions, as assigning the label to the spectrum, opening the tSNE plot of current batch (as shown), opening the performance estimation graph of all iterations, or showing the original view of spectrum in the archive

6 Conclusions

The active deep learning of large spectral archives requires all the training iterations to be supported by the interactive labelling performed by the domain expert in role of the Oracle. The `ml-job-manager` cloud-based science platform provides the expert with the flexible and user-friendly visual environment allowing to run a number of experiment in parallel.

Acknowledgements

Astronomical Institute of the Czech Academy of Sciences is supported by the project RVO 67985815. For testing the platform the spectra from the public LAMOST DR2 archive were used extensively.

References

O. Burakov. Science platform for machine learning of big astronomical data - cloud infrastructure, 2025. URL `https://doi.org/10.5281/zenodo.17420073`. Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology.

P. Harrison and G. Rixon. IVOA universal worker service pattern, 2016. URL `https://www.ivoa.net/documents/UWS`. IVOA Recommendation.

J. Koza. Design and implementation of a distributed platform for data mining of big astronomical spectra archives, 2015. URL `https://doi.org/10.5281/zenodo.17537102`. Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology.

J. Koza. Interactive cloud-based platform for parallelized machine learning of astronomical big data, 2017. URL `https://doi.org/10.5281/zenodo.17537158`. Master's thesis. Czech Technical University in Prague, Faculty of Information Technology.

A. Layik. Science platform for machine learning of big astronomical data - data analysis modules, 2025. URL `https://doi.org/10.5281/zenodo.17420100`. Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology.

T. Mazel. Cloud-based platform for active learning of astronomical spectra, 2020. URL `https://doi.org/10.5281/zenodo.17420002`. Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology.

O. Podsztavek. active-cnn source code for active deep learning on github.com. online, 2019. URL `https://github.com/podondra/active-cnn`. [Accessed 2025-11-05].

P. Škoda, O. Podsztavek, and P. Tvrdík. Active deep learning method for the discovery of objects of interest in large spectroscopic surveys. *Astronomy & Astrophysics*, 643:A122, 2020. doi: 10.1051/0004-6361/201936090.

Poster presented at ADASS 2025, 9 – 13th November 2025, Goerlitz, Germany

Astronomical Institute
of the Czech Academy of Sciences

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE