

·earthpulse!

EOX



# THE EARTH OBSERVATION TRAINING DATA LAB EOTDL

---

24/10/2023



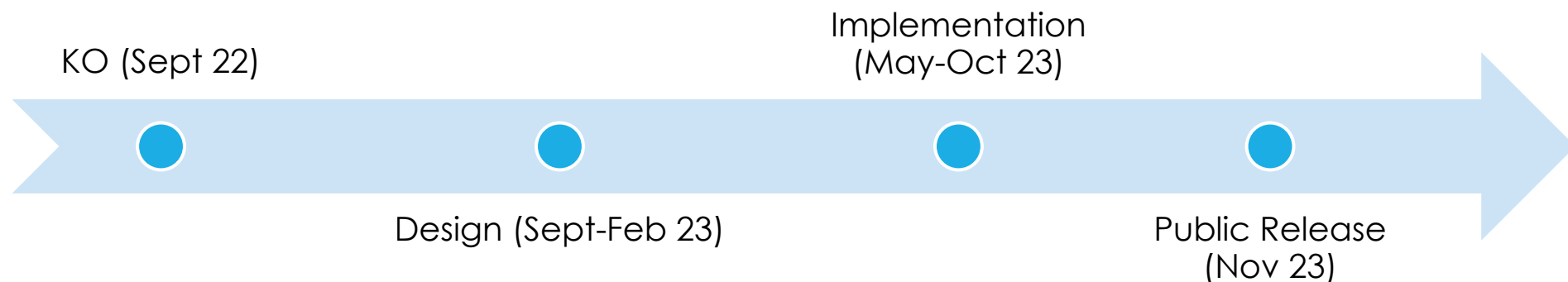
European Space Agency

# Background

- ▶ One of the most limiting factors of AI for EO applications is the scarcity of suitable and accessible **Training Datasets** (TDS). As the name suggests, TDS are used to train an AI model to perform a specific task. Currently, the main barrier is that gathering and labelling EO data is a convoluted process. Some techniques exist that can help alleviate this issue, for example transfer learning or unsupervised learning, but annotated data is always required for fine-tuning and final validation of AI models.
- ▶ Generating TDS is time consuming and expensive. Data access is usually limited and costly, especially for Very High Resolution (VHR) images that allow objects like trees to be clearly identified. In some cases, domain experts or even in-person (in-situ) trips are required to manually confirm the objects in a satellite image are correctly annotated with a high degree of quality. This results in the field of AI for EO applications lagging when compared to other fields, impeding the development of new applications and limiting the full potential of AI in EO.

# The Earth Observation Training Data Lab

- ▶ The European Space Agency (ESA) **Earth Observation Training Data Lab (EOTDL)** address key limitations and capability gaps for working with Machine Learning (ML) training data in EO by providing, on the one hand, a set of open-source tools to create, share, and improve datasets as well as training ML algorithms in the cloud. On the other hand, EOTDL offers an online repository where datasets and models can be explored and downloaded.
- ▶ The curation mechanism and repository will cover a wide range of dataset types: training, validation, test, benchmark and reference datasets (in-situ data, product validation datasets).
- ▶ The EO-TDL is expected to launch in September 2023, with over 100 available datasets for a wide range of applications and data sources.



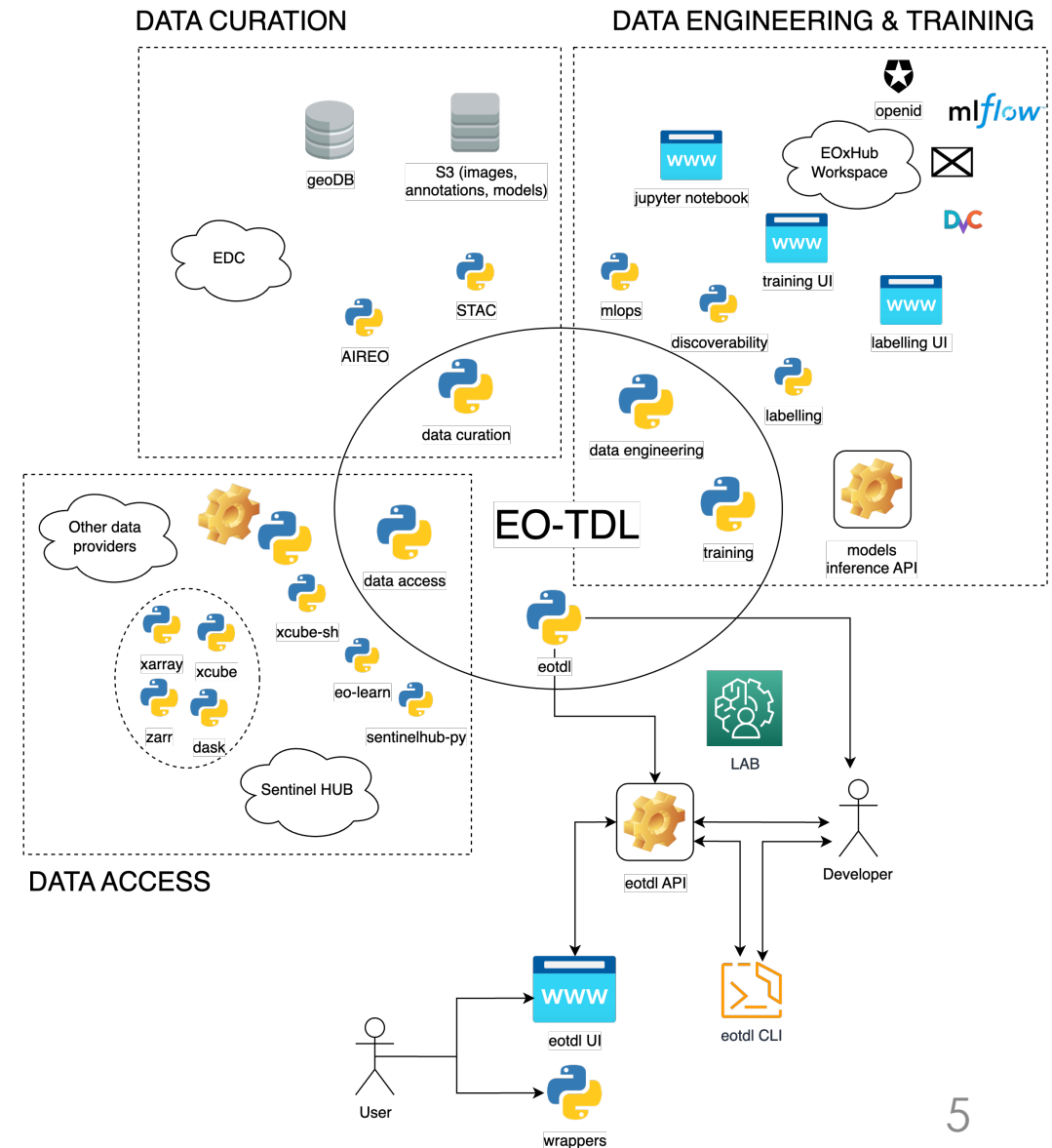
# Objectives

- ▶ To contrast the lack of suitable TDS, the EOTDL has the **objective of providing a set of open-source tools to generate, curate, analyse, and use AI-ready training datasets**. This environment will act as a cloud repository, where TDS can be created, imported, maintained, and improved by everyone.
- ▶ EOTDL is connected to effective data access mechanisms and feature engineering tools, allowing (among other things) training models with the hosted datasets directly on the cloud with multi-GPU machines.
- ▶ The EOTDL is a lab for ML training data, with a focus on EO applications. There is a particular emphasis on datasets to overcome existing barriers such as Unsupervised learning, Data Fusion, multitask Learning and the development of custom architectures.
- ▶ EOTDL provides interoperability with third party platforms, such as Radiant Earth MLHub, and provides a diverse set of AI ready datasets.
- ▶ Community engagement is incentivised to stimulate collaboration in dataset creation, enhancement and quality assurance.



# Ecosystem

- ▶ The EOTDL ecosystem is built on top of open source projects, and is open source as well.
- ▶ Users can access data for dataset creation, selecting data source, time range and areas of interest.
- ▶ Metadata for data curation and quality assurance is generated following the STAC specification. Automatic QA mechanisms and versioning is applied during the process of dataset ingestion.
- ▶ Engineering tools enable reproducible feature engineering, labelling, bias discoverability, etc. Training ML models in the cloud with multi-GPU machines is transparently enabled.
- ▶ The EOTDL is accessible at multiple levels thanks to user interfaces, web APIs, CLIs and Python libraries.



# Who is this for?

- ▶ The EOTDL is available to everyone, with special tiers for prime commercial users.
- ▶ Many areas can benefit from the EOTDL. Having a repository of AI-ready training datasets will strengthen industry capabilities for exploiting EO data as a whole and help accelerate EO market penetration. Furthermore, to enable Digital Twin Earth simulations, access to these quality datasets is necessary for researchers and engineers as they build and apply quality models.
- ▶ The initial data population will consist of over 100 datasets especially selected to cover a wide range of applications and ML techniques, ranging from Computer Vision tasks such as classification or object detection to parameter estimation or 3D applications on different data sources such as Sentinel 1 and 2, Airbus SPOT and PLEIADES, UAV imagery or vector data. But, with the provided tools, it is our objective to increase this quantity over time.

# Community

- ▶ The flagship characteristic of the EOTDL is its community and open-source nature. All the code is hosted on Github (<https://github.com/earthpulse/eotdl>), where the community can contribute with improvements and new features. A public Discord server enables discussion and community engagement (<https://discord.gg/hYxc5AJB92>).
- ▶ Not only will users be able to train their AI models on the cloud with the available datasets, but they will also be encouraged and incentivised to contribute to the enrichment of the platform. Users making significant contributions, such as the addition of new datasets or the enhancing of existing ones, will be rewarded in the form of credits. A larger community will thus correspond to a better and more versatile platform for everyone.
- ▶ The EOTDL facilitates integration by popular Machine Learning libraries, like Scikit-Learn, Pytorch and Tensorflow. Furthermore, the platform provides tools to ingest datasets (including legacy data) with automated quality assurance procedures.

·earthpulse!

EOX



# THE EARTH OBSERVATION TRAINING DATA LAB EOTDL

---

24/10/2023



European Space Agency