Motivation

Training datasets (TDS) are crucial for ML and AI applications but they are a major **bottleneck due to:**

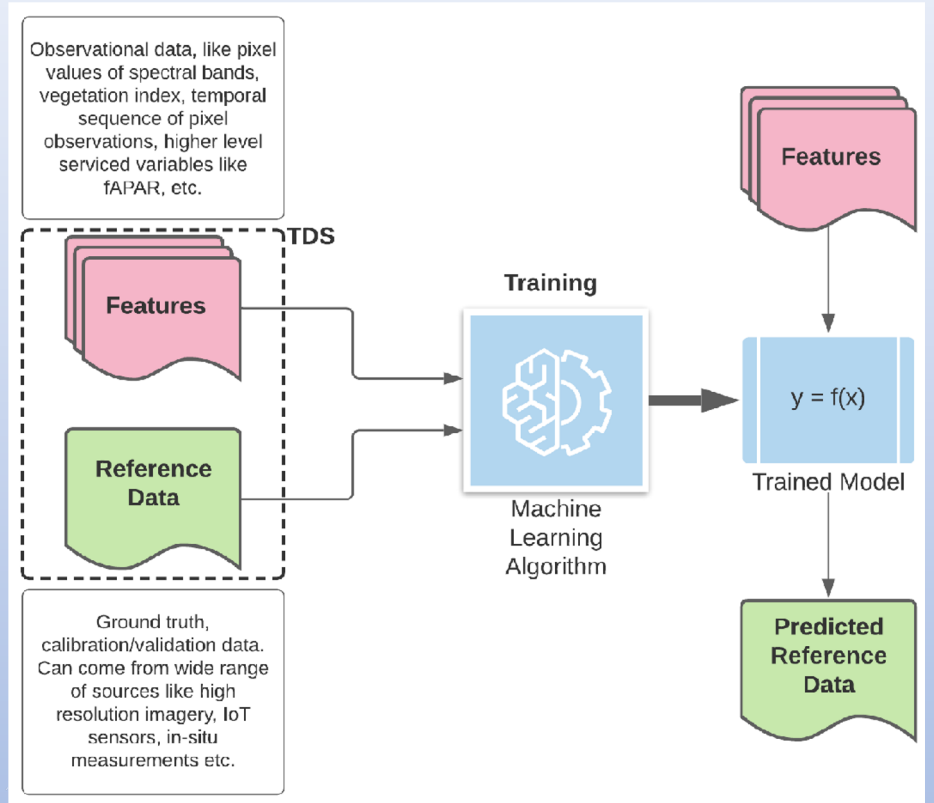
- Limited availability and willingness to share
- Lack of TDS and limited interoperability
- Lack of best-practices/standards to generate and describe TDS

Outputs:

- **AIREO Network** - to capture Community requirements
- **Specification and Best Practice guidelines**
  - V0 for release June 2021
- **Pilot Datasets**
  Benchmark
- **Library and Notebooks**
  - Easy access to Training Datasets
    - Xarray-, Pandas-compatible

Pilot Datasets

- **Biomass retrieval (BRIX):** canopy and biomass from 260 forest plots: regression, correlation task
- **AI4Arctic Sea Ice**: S1 and AMSR: boundary detection, segmentation, identification
- **Common Agricultural Policy (CAP) Austria**: information about crop types and field boundaries for segmentation, classification and boundary detection
- **SpaceNet7** Multi-temporal Urban development: One image per month of 100 locations with building footprints for segmentation, detection



Innovations:

- Quality Assurance Automation in Library: Consistency, completeness
- Provenance: Provide full data traceability, FAIR standards
- AIREO STAC Extensions for Cloud-native metadata and datasets
- Embedded Feature engineering recipe: Automated recreation of TDS