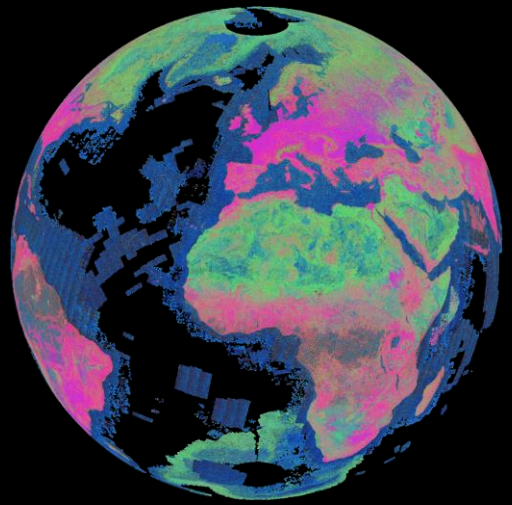
CloudFerro

Redefining Earth Observation: embeddings as the next data layer for AI-ready EO ecosystems



Artificial intelligence and Earth Observation: From innovation to services
Brussels 9-10 March 2026

Jędrzej Bojanowski

01

Why embeddings?

CloudFerro

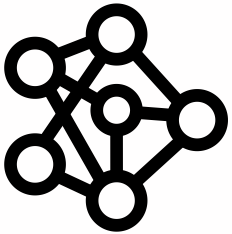
Decades-old remote sensing challenges

- **Clouds, gaps, irregular time series** – require corrections, gap-filling, compositing, etc.
- **SAR data is not intuitive** and processing is more demanding than for optical data
- No single sensor has **ideal spatial/temporal/spectral properties** – needs tradeoffs of one dimension against another
- **Data fusion** of optical, SAR, etc. still requires a lot of manual engineering
- Supervised models (classifiers, regressors) **need many labelled scenes**
- **Too much data** – petabytes of raw pixels are hard to turn into analysis-ready
- **Limited scalability** of algorithms (models) from local to regional/global data
- Task-specific pipelines (algorithms) with **poor reuse across projects**

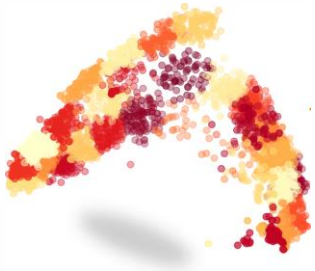
Turning complexity into simple format

Spatial data

AI model



Vector embedding



```
[ -0.369384765625, -0.1776123046875,
0.3779296875, -0.2294921875, 0.262939453125,
0.1444091796875, 0.249755859375, …]
```


CloudFerro

Different kinds of EO embeddings

Input granularity

- Pixels, patches, scenes

Modalities & semantics

- Single modality (optical, SAR, meteo, thematic etc.) or multi-modal fusion
- Vision-only or vision+text

Spatial & temporal dimensions

- Single to multi-resolution (e.g. 10–1000 m)
- Single date to all-year stacks (time series)

CloudFerro

Embeddings advantage

- Less data, more signal – compressed vectors instead of huge multi-dimentional stacks (time, space, bands, sensors, etc.)
- Semantic meaning – vectors encode state/change/temporal evolution, not just radiance
- Unified view – one representation fusing multiple sensors and time
- Gap-free – embeddings provide downstream models clean and regular feature vectors
- Faster & cheaper training – smaller ML retrieval models on embeddings than on big multi-dimensional data
- Better search & analytics – „easy” image-to-image and text-to-image similarity search, clustering and anomaly detection

02

Infrastructure & Data needs for GeoAI

CloudFerro

Requirements for training GeoAI models

- Consolidation of heterogeneous spatial data
- Fast and autonomous access to data in optimised formats
- Guarantee of data authenticity
- Sovereignty and security of cloud infrastructure
- Use of artificial intelligence models in a secure environment

CloudFerro computing resources for AI models & data

High-performance compute cloud (**CPU & GPU**)

Low-latency **SSD** storage

High-speed network for data access

100+ PB EO DATA (CDSE and more)

Own tech stack/Open Source

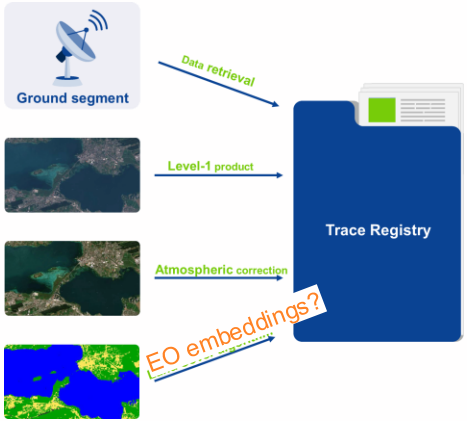
Own IaaS + PaaS built on own and open-software + Specialized EO tools and apps

Natural place for operational embeddings generation: fresh data availability, highest timeliness, data already in processing pipelines

CloudFerro

Traceability for data authenticity

- It allows you to precisely reconstruct the process of acquiring and processing each product (by checksums of every file).



Traceability

Ensuring product traceability across locations.

Search About

Product Overview

Product Name: S1A_IW_SLC__1SDV_20230418T032847_20230418T032914_048145_05C9D2_1BBD.SAFE.zip
Size: 8019.36 MB
Checksum: d1afd1d39385876f49d9ee6aed7a2b8c08066ab65ece3a3a0e68ae81cb319d17
Origin: archive@cdse.copernicus.eu

Contents

S1A_IW_SLC__1SDV_20230418T032847_20230418T032914_048145_05C9D2_1BBD.SAFE

annotation	
manifest.safe	76d752b4855ddf388d494ba924be08b231bf2cad7c893ebaa551535451bbdb74
measurement	
preview	
S1A_IW_SLC__1SDV_20230418T032847_20230418T032914_048145_05C9D2_1BBD.SAFE-report-20230418T062516.pdf	8066040f832f20b452d6231d00876c0f93a4d41fe064684e13b38bcfefee69dc
support	

History

Date	Event	Origin	Product Name	Checksum	Signed
18/04/2023	CREATE	archive@cdse.copernicus.eu	S1A_IW_SLC__1SDV_20230418T032847_20230418T032914_048145_05C9D2_1BBD.SAFE.zip	d1afd1d39385876f49d9ee6aed7a2b8...	
19/07/2023	CREATE	cdse_csc@dataspace.copernic...	S1A_IW_SLC__1SDV_20230418T032847_20230418T032914_048145_05C9D2_1BBD.SAFE.zip	8e5627fc88bfa649c371f573ae46208...	✓

CloudFerro

Integrating Large Language Models

Sherlock

Generative AI, Zero Logs, Total Privacy

- A privacy-first generative AI platform
- Access to leading language and vision models via unified API
- As a response to limited trust in companies training own models
- Zero logs policy with no data retention
- Natural language catalogue query

```
from openai import OpenAI
import os

# Set your API key and base URL
client = OpenAI(
  api_key=os.environ['SHERLOCK_API_KEY'],
  base_url="https://api-sherlock.cloudferro.com/openai/v1"
)

# Specify your model
model = "Llama-3.3-70B-Instruct"
```

BIELIK PLLuM Meta deepseek MISTRAL AI CloudFerro

03

GeoAI data availability

CloudFerro

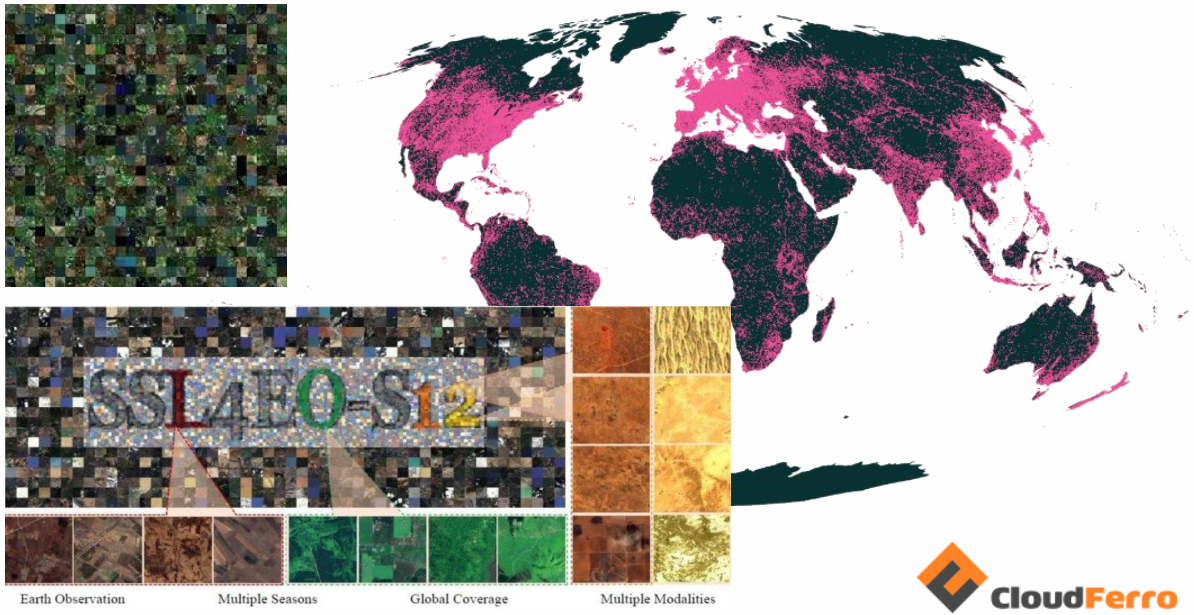
Training datasets

supporting pretraining, self-supervised learning, classification and multimodal learning

Making geoAI datasets available within CDSE/CREODIAS significantly lowers the barrier to AI development by enabling faster access, easier testing, and tight integration with Copernicus data.

Datasets:

- MajorTOM
- SSL4EO-S12
- OlmoEarth
- MMEarth
- BigEarthNet-S1/S2
- EuroSAT MS/RGB
- CloudSEN12
- SatlasPretrain
- TerraMesh

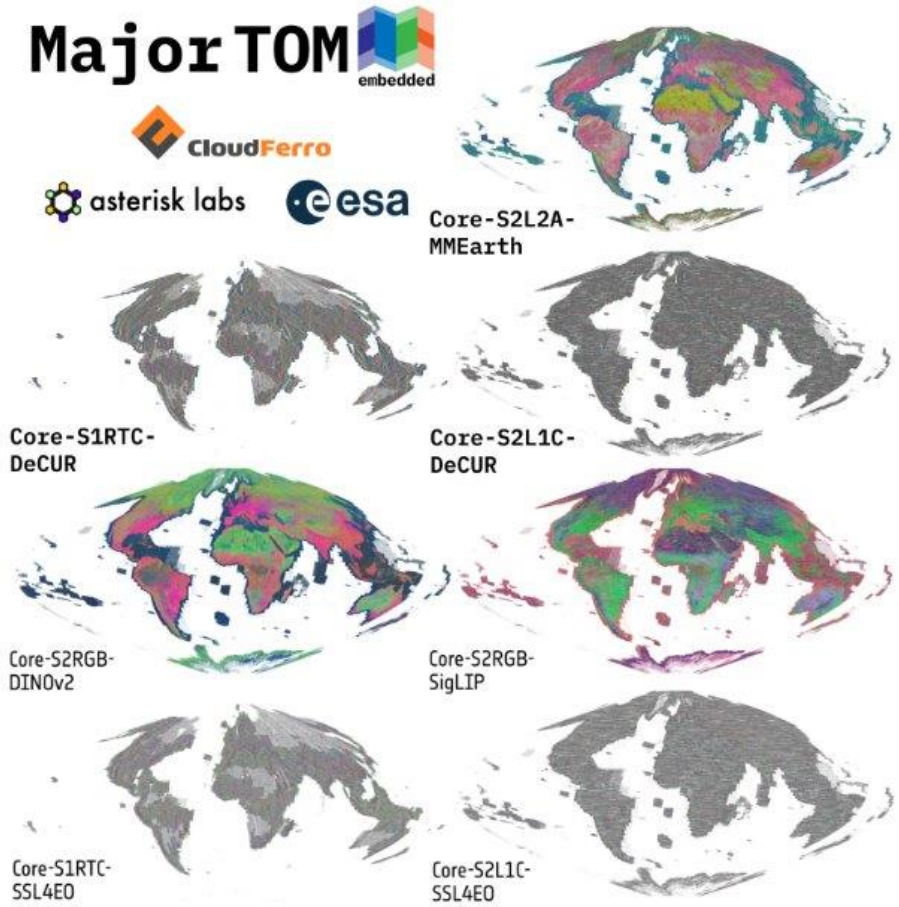


CloudFerro

Open access satellite data embeddings for EO-AI applications

Embeddings datasets summary:

- **51 TB** of AI-embeddings generated from processed Sentinel data
- over **40 billion** embedding vectors
- processing of **147 TB** of raw satellite data
- analysis covering more than **15 million** Sentinel-1 and Sentinel-2 scenes and more than **16 trillion** pixels



Paper

Global and Dense Embeddings of Earth: Major TOM Floating in the Latent Space

arXiv 10.48550/arXiv.2412.05600

Available open source via: CREODIAS Hugging Face

CloudFerro

04

Anomaly detection

CloudFerro

Inference text-to-image

satellite image of city

satellite image of sea coast

satellite image of deforestation

CloudFerro

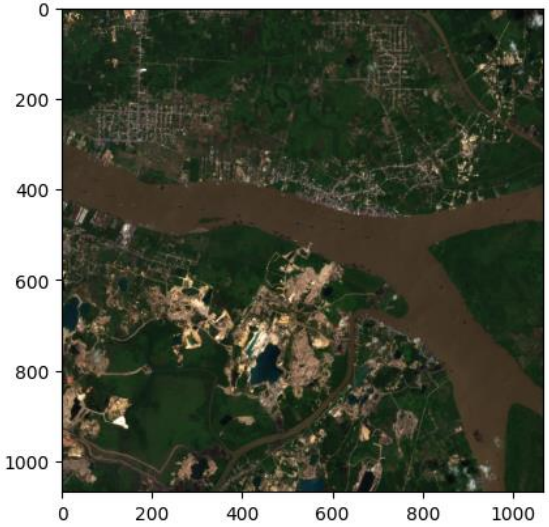
Inference image-to-image

Reference image

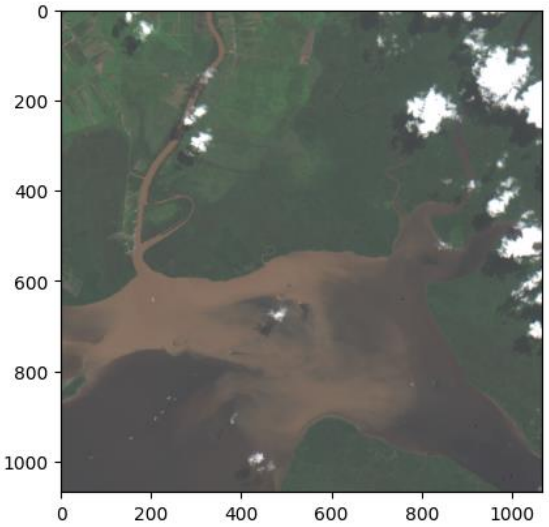
Image for grid cell '7D_1308R'

Found images

similarity: 0.93

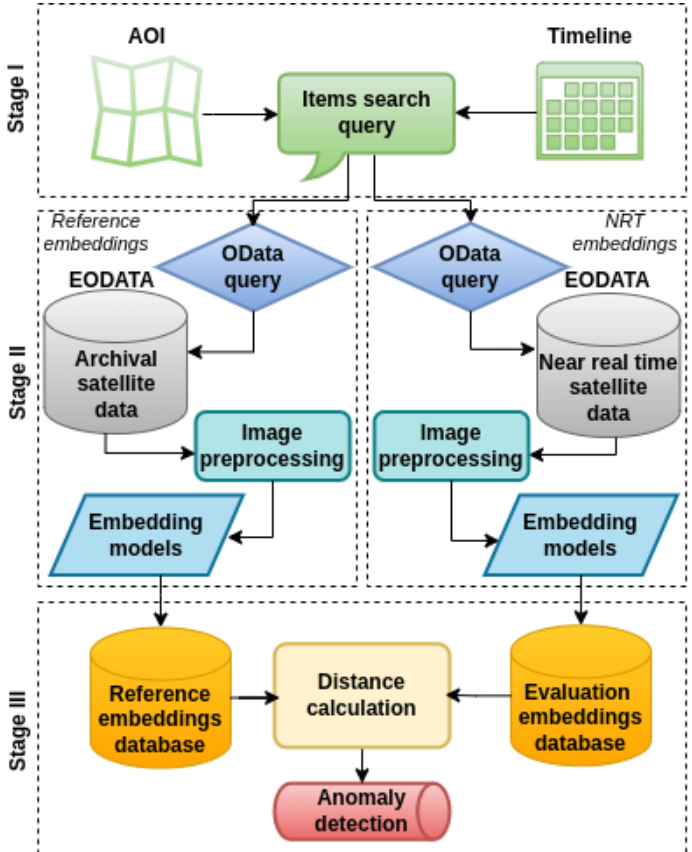


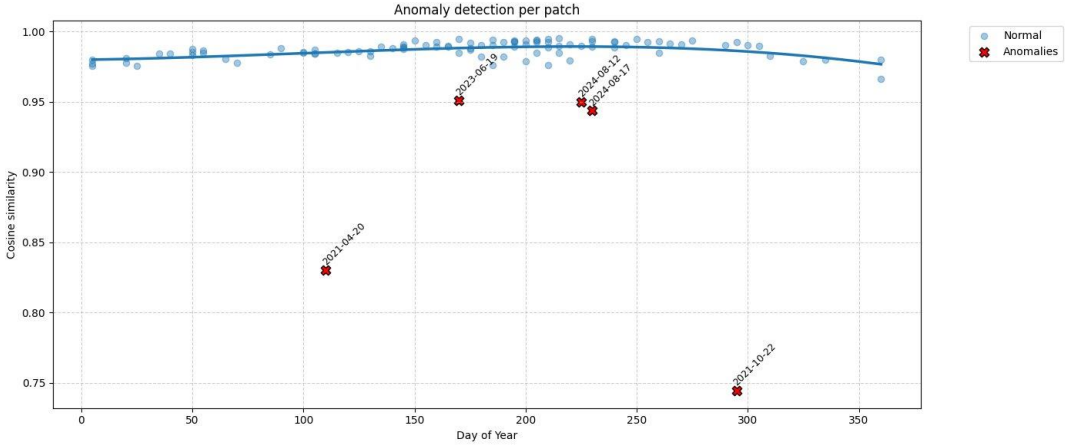
similarity: 0.86



CloudFerro

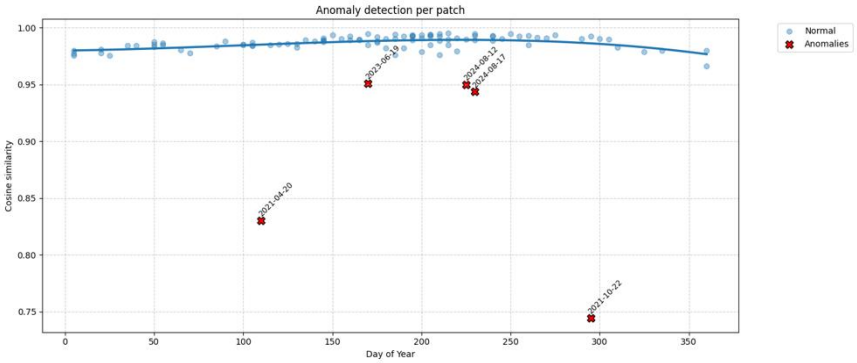
Anomaly detection using Copernicus data

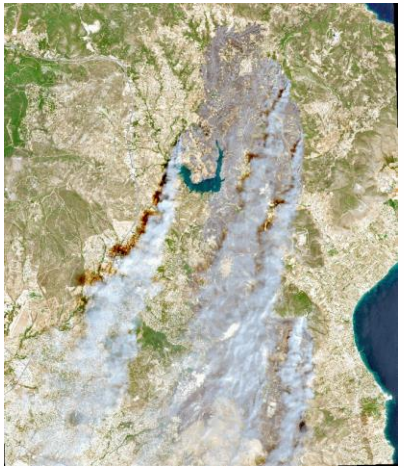




CloudFerro

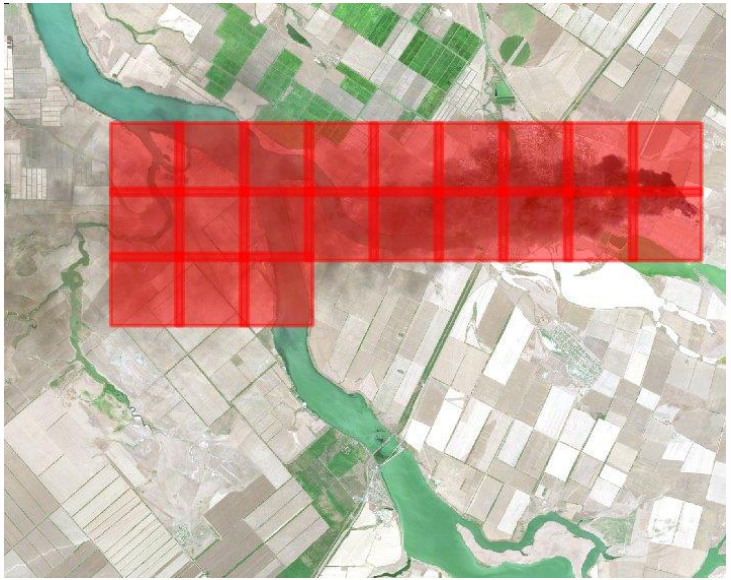
Abrupt changes (wildfires in Greece, 2024)



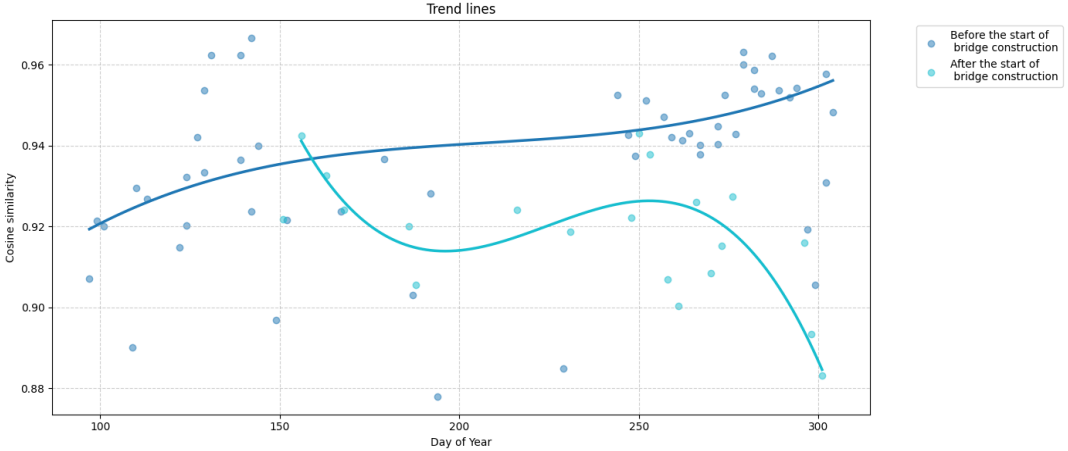


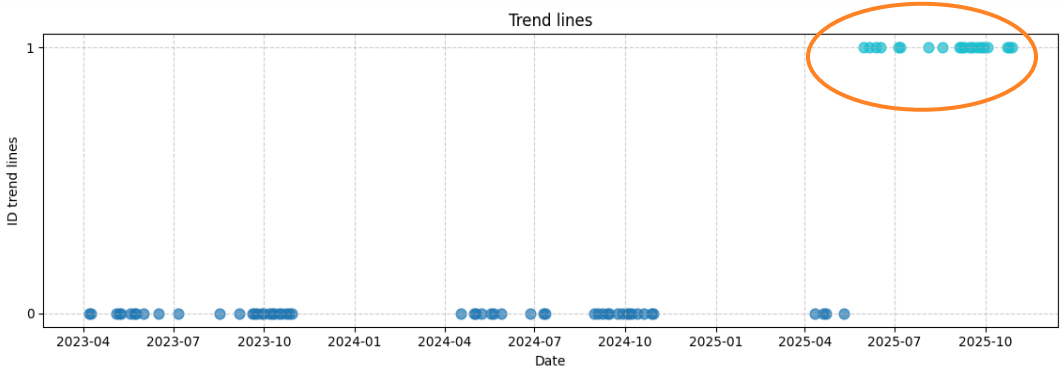
CloudFerro

Abrupt changes (fire in Proletarsk, 2024)



Permanent changes (Russia-North Korea bridge)





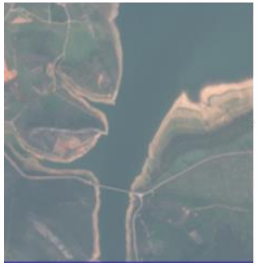
SSL4EO-S12

2024

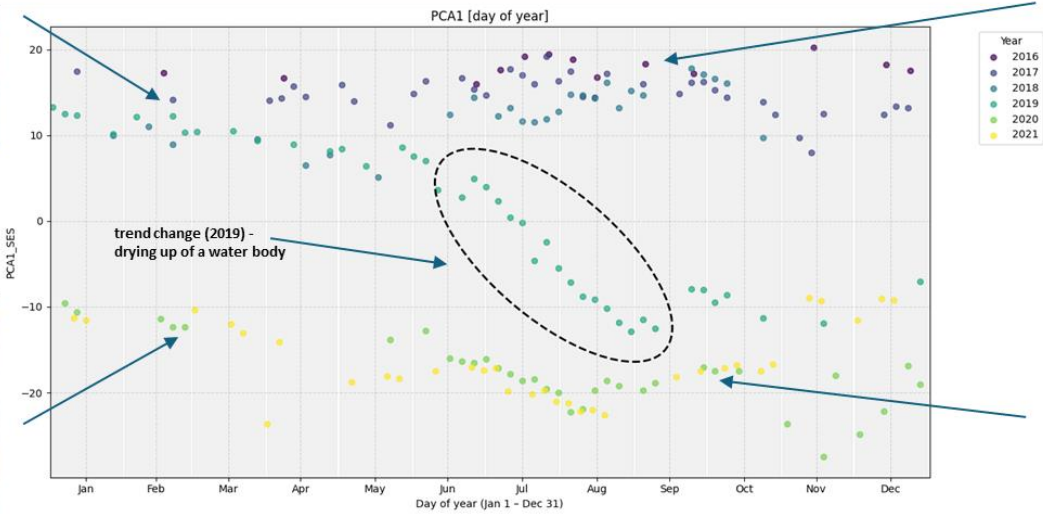
2025

Ongoing environmental changes (drying Embalse de la Serena)

2016



2017



2020

2021

CloudFerro

Key takeaways

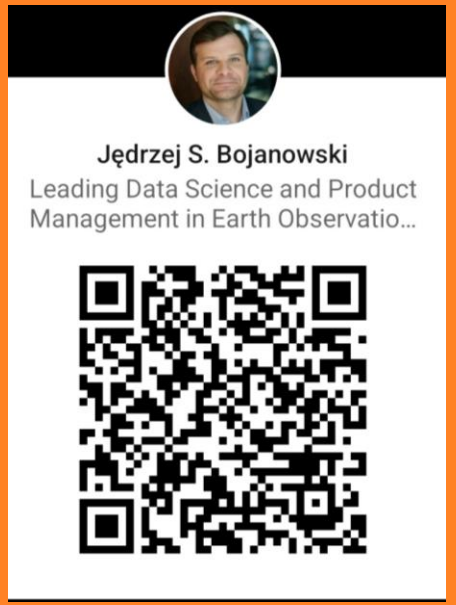
- Embeddings turn petabytes of raw pixels into analysis-ready AI features
- Sovereign infrastructure is essential for model training and operational embedding generation
- Pre-computed embeddings make global-scale monitoring fast and affordable
- Embeddings enable generic anomaly detection critical for civil security and dual-use
- A few embedding layers power many applications — reuse across tasks and projects

CloudFerro

www.cloudferro.com

Jędrzej Bojanowski

jbojanowski@cloudferro.com



facebook.com/cloudferro | linkedin.com/company/clfr/ | twitter.com/CloudFerro