Advanced AI applications in EO: fine-tuning foundation models

Dr. Lorenzo Papa, Dr. Ruben Cartuyvels, Dr. Valerio Marsocci

Internal Research Fellows, ESA ESRIN Frascati, Rome, Italy

Email: [lorenzo.papa, ruben.cartuyvels, valerio.marsocci]@esa.int

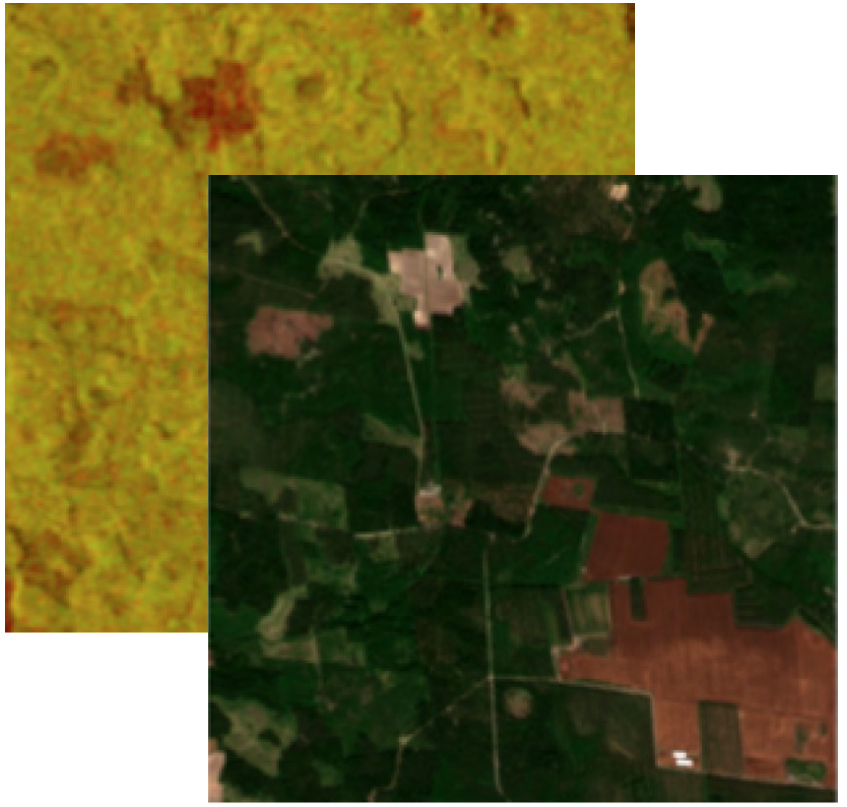
Outline

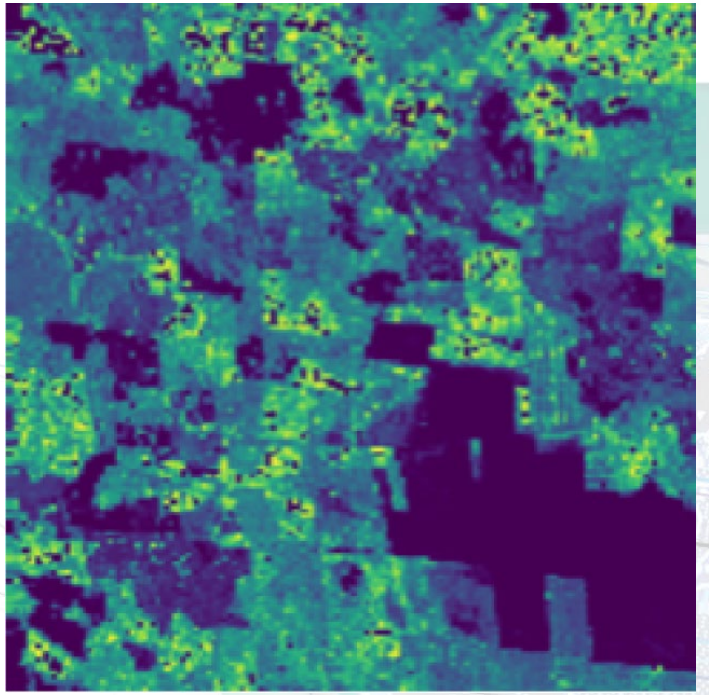
- What are Geospatial Foundation Models?
- How do we obtain a GFM?
- Benchmarking GFMs
- Conclusion

https://github.com/lorenzopapa5/BiDs_training_course_2025/tree/main

WHAT ARE GEOSPATIAL FOUNDATION MODELS?

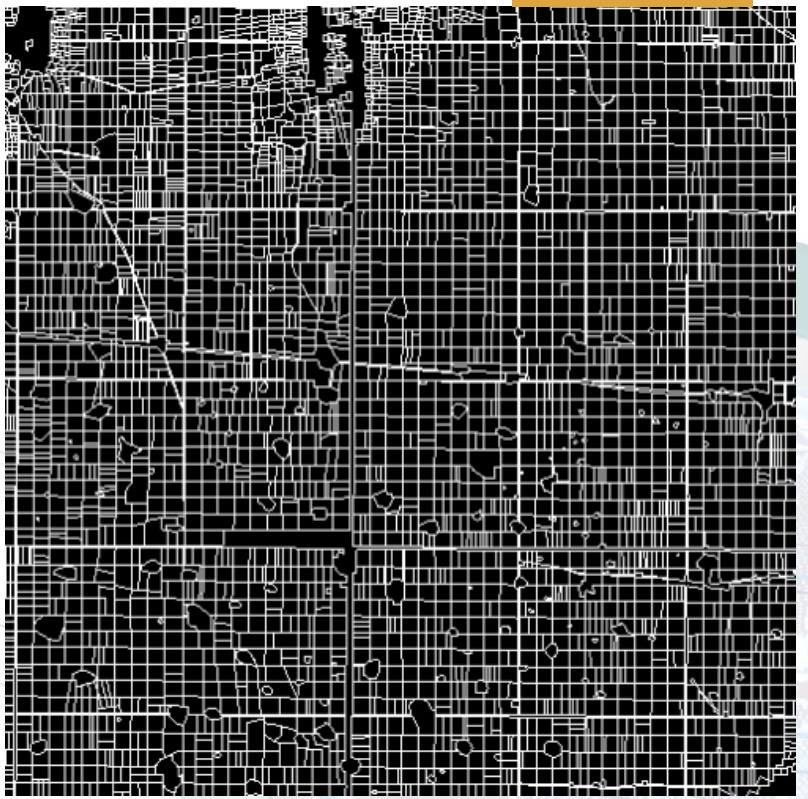
Think of a problem connected to an SDG





Think of another one

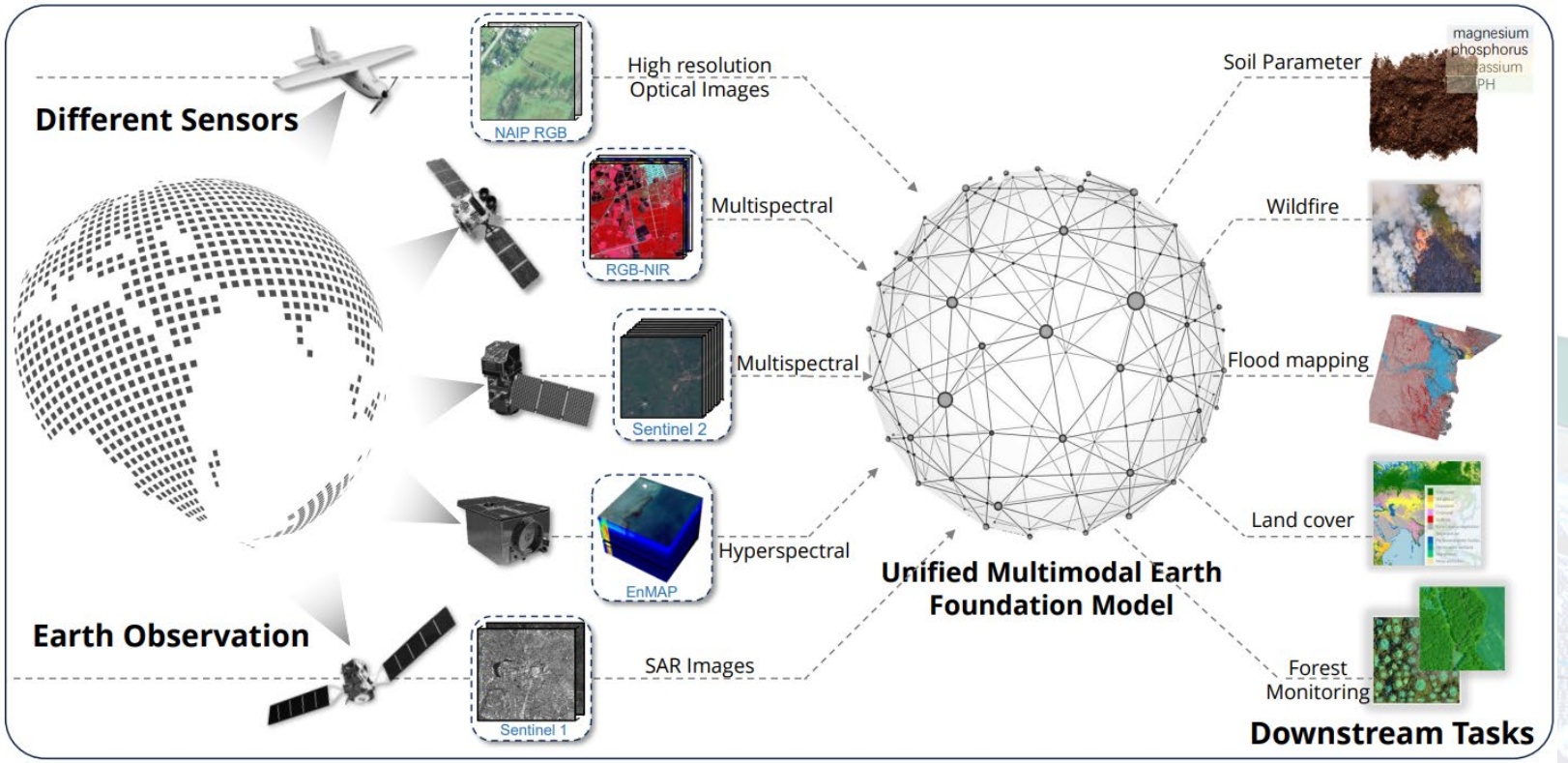
One more



How to solve all these tasks with the same model?

Let's use a **Geospatial Foundation Model**

- task agnosticism
- spatio-temporal awareness
- sensor agnosticism
- multimodality
- adaptability



Source: DOFA - 2403.15356

What Are Geospatial Foundation Models?

- Geospatial foundation models are **large-scale** AI models trained on vast amounts of diverse **geospatial data**

- GFMs learn **temporal** and **spatial** patterns, relationships, and **features** across different geographic scales

What Are Geospatial Foundation Models?

- **Not limited to a single task**
  Can generalize across multiple geospatial applications

- **Enhances scalability & efficiency**
  Reduces need for retraining and
  Supports diverse use cases

Task Agnosticism

What Are Geospatial Foundation Models?

- Model understands both spatial (**geographic**) and temporal (**time-related**) patterns

- Essential for **dynamic applications** like climate monitoring, disaster response, and urban development

Spatio-temporal awareness

What Are Geospatial Foundation Models?

- Model can process data from **various sensors** (e.g., satellite, UAV, LiDAR)

- Increases **versatility**
  Allows **integration** of **diverse** data sources

Sensor agnosticism

What Are Geospatial Foundation Models?

- Model integrates and processes **multiple types** of data (e.g., imagery, text, GIS, data point)

- Provides richer insights by leveraging **diverse information**

Multimodality

What Are Geospatial Foundation Models?

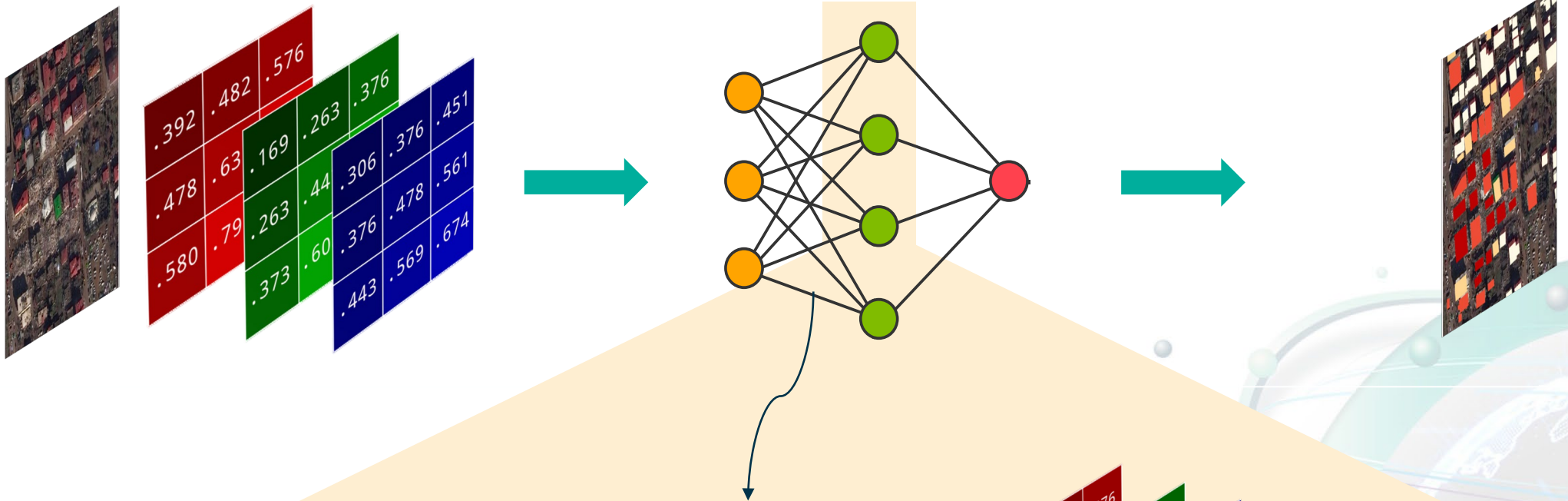
- Model can be **easily fine-tuned** or adjusted for different tasks or datasets

- **Reduces the need for retraining from scratch**, saving time and resources

Adaptability

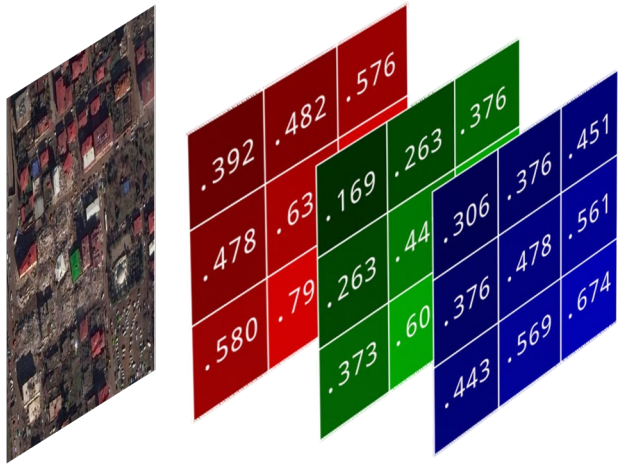
DEVELOPING GEOSPATIAL FOUNDATION MODELS

Neural networks (short recap)

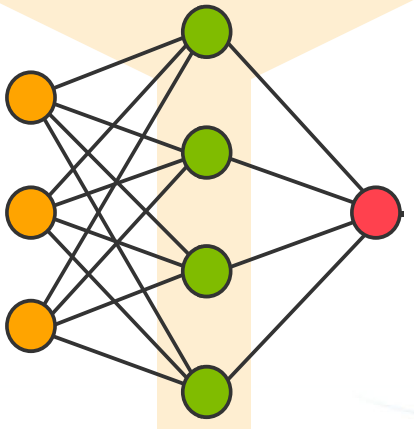


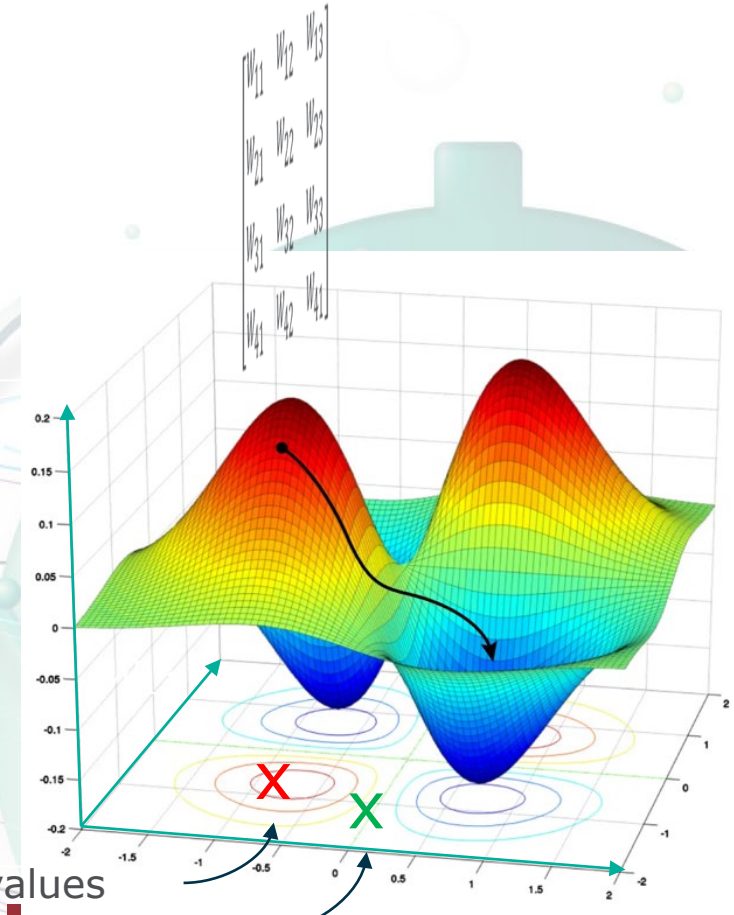
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \sigma\left(\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{41} \end{bmatrix} * \begin{bmatrix} .392 & .482 & .576 \\ .478 & .63 & .79 \\ .580 & .79 \end{bmatrix} \begin{bmatrix} .169 & .263 & .376 \\ .263 & .44 & .60 \\ .373 & .60 \end{bmatrix} \begin{bmatrix} .306 & .376 & .451 \\ .376 & .478 & .561 \\ .443 & .569 & .674 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}\right)$$

Neural networks (short recap)



$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{41} \end{bmatrix}$$



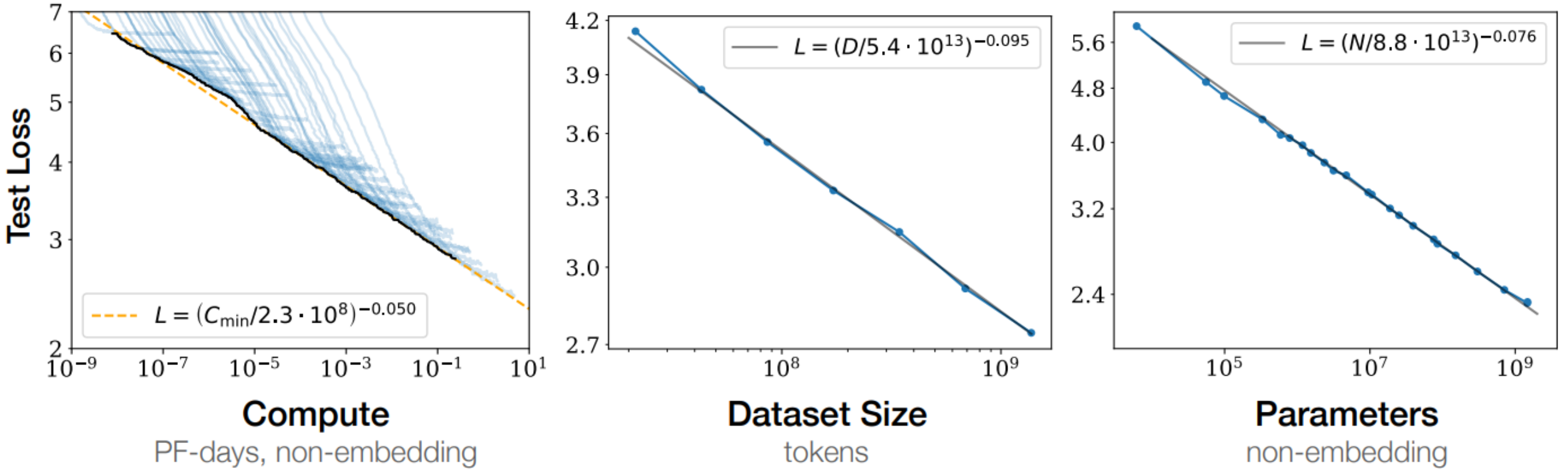


Objective function:

$$L\left(\quad , \quad \right) = 2.4$$

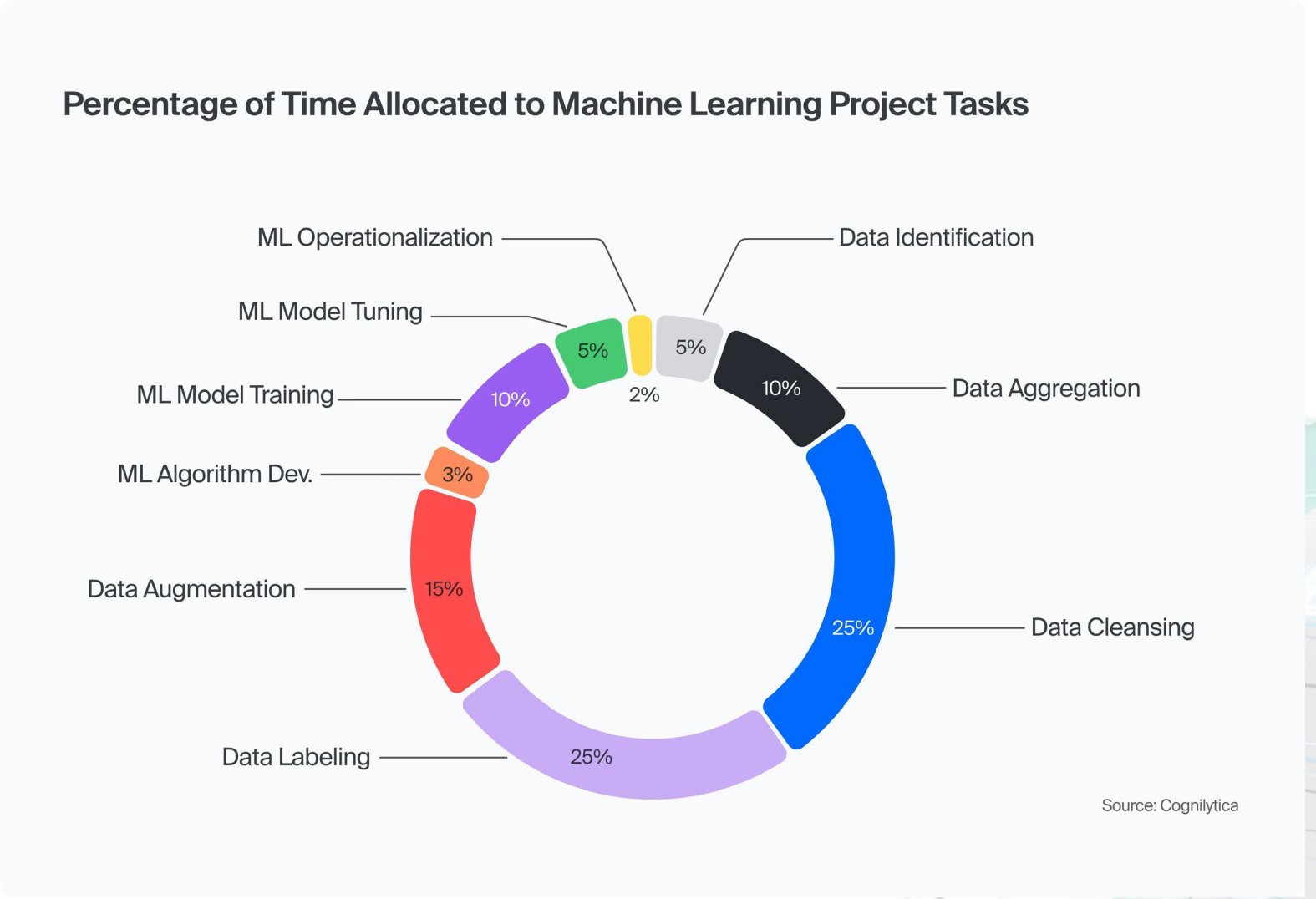
Neural networks need a lot of data

Scaling Laws for Neural Language Models



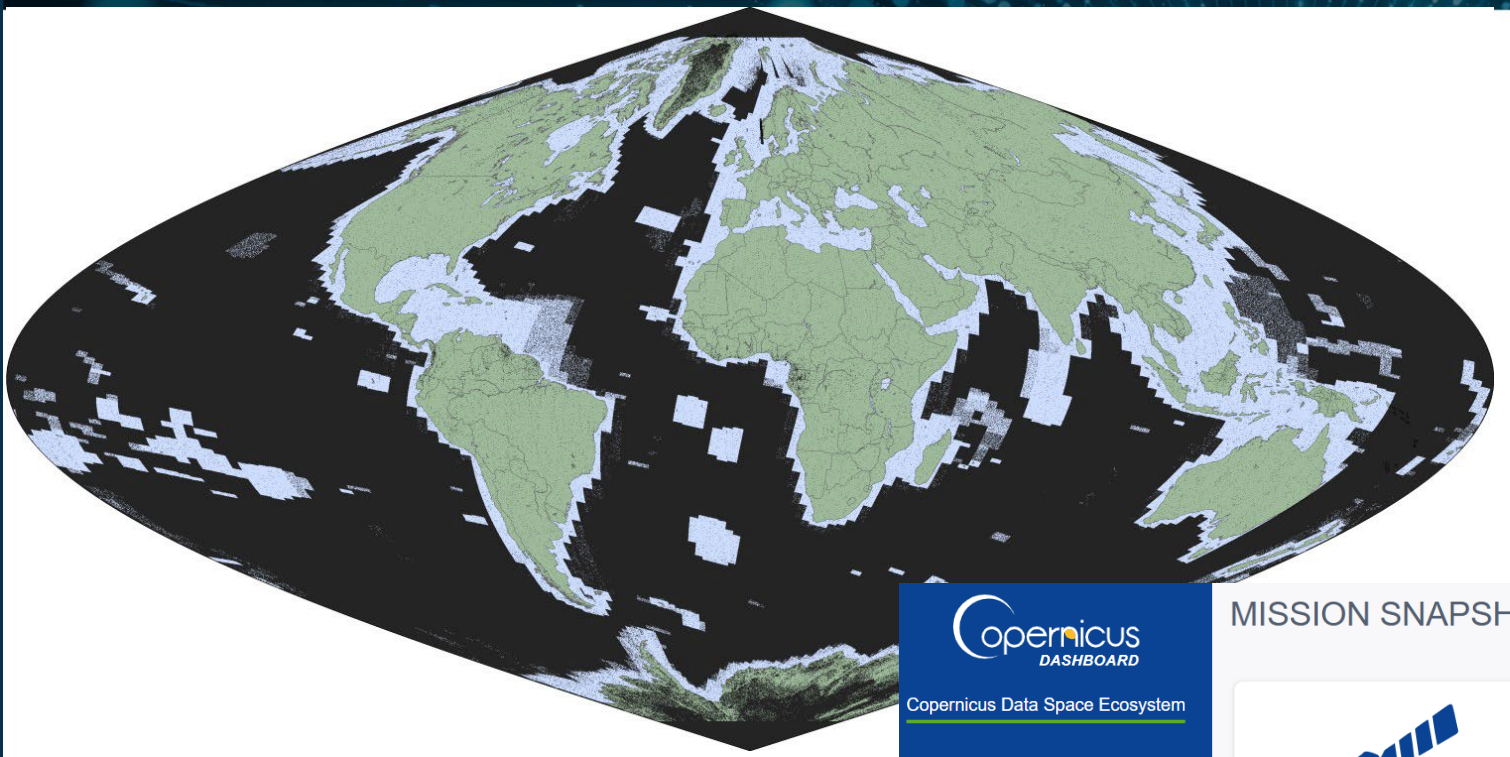
Kaplan, Jared, et al. "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361 (2020).

(Labelled) data is expensive



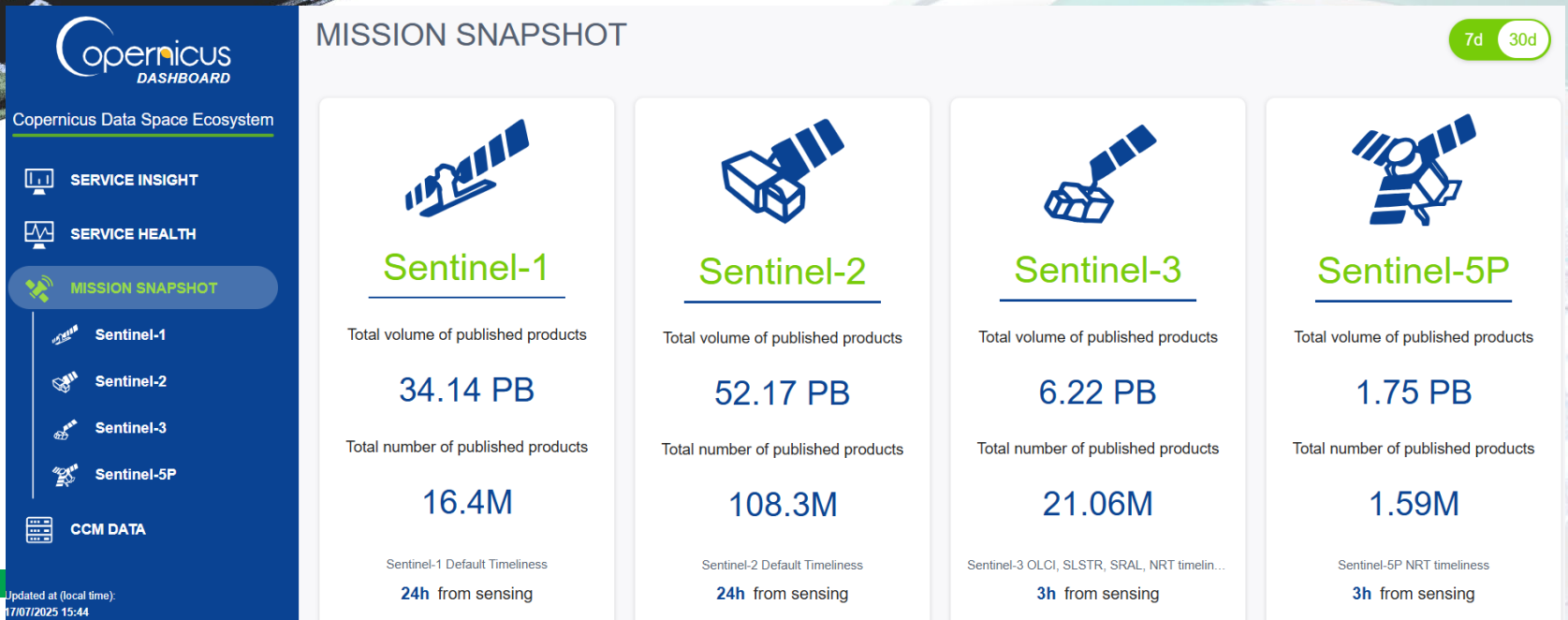
Cognilytica, A. I. "Data Engineering Lifecycle Checklist." (2020).

There is a lot of EO data

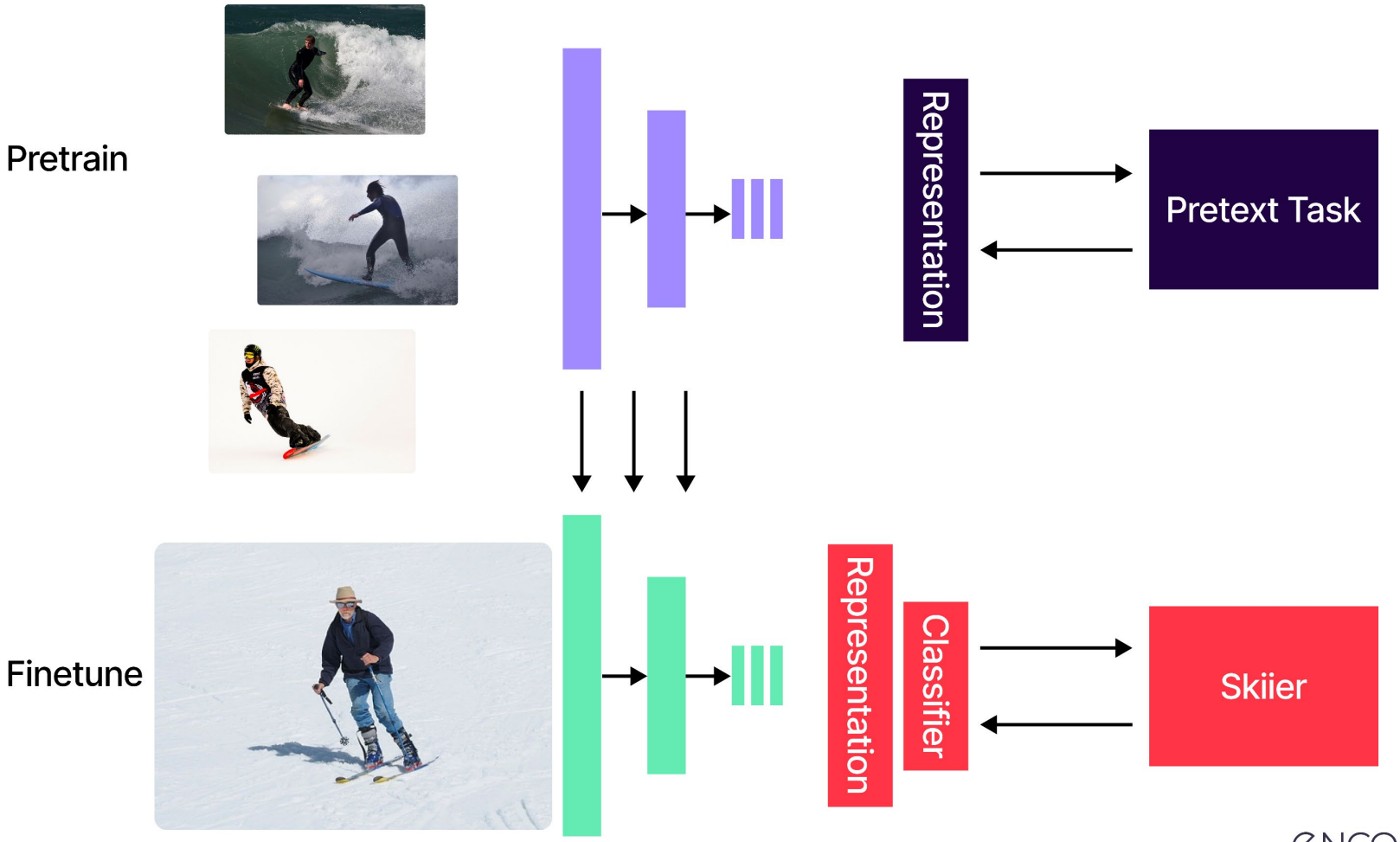


https://philab.esa.int/hello-major-tom-esa-%CF%86-lab-releases-largest-ml-ready-sentinel-2-dataset-ever-published/

https://dashboard.dataspace.copernicus.eu/#/mission-snapshot



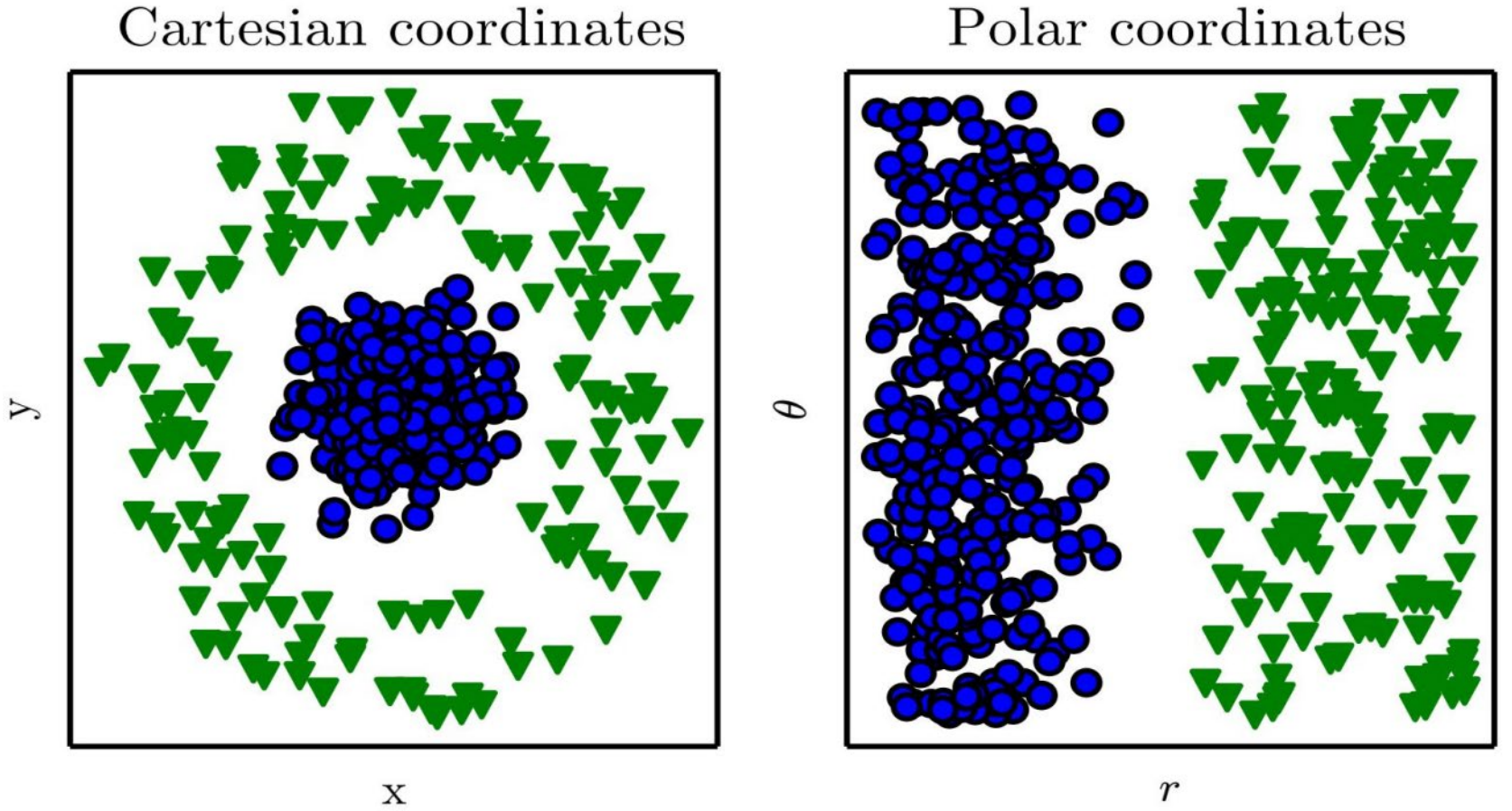
Self-supervised learning



https://encord.com/blog/self-supervised-learning/

Representations are important

- Rotation example



https://encord.com/blog/self-supervised-learning/

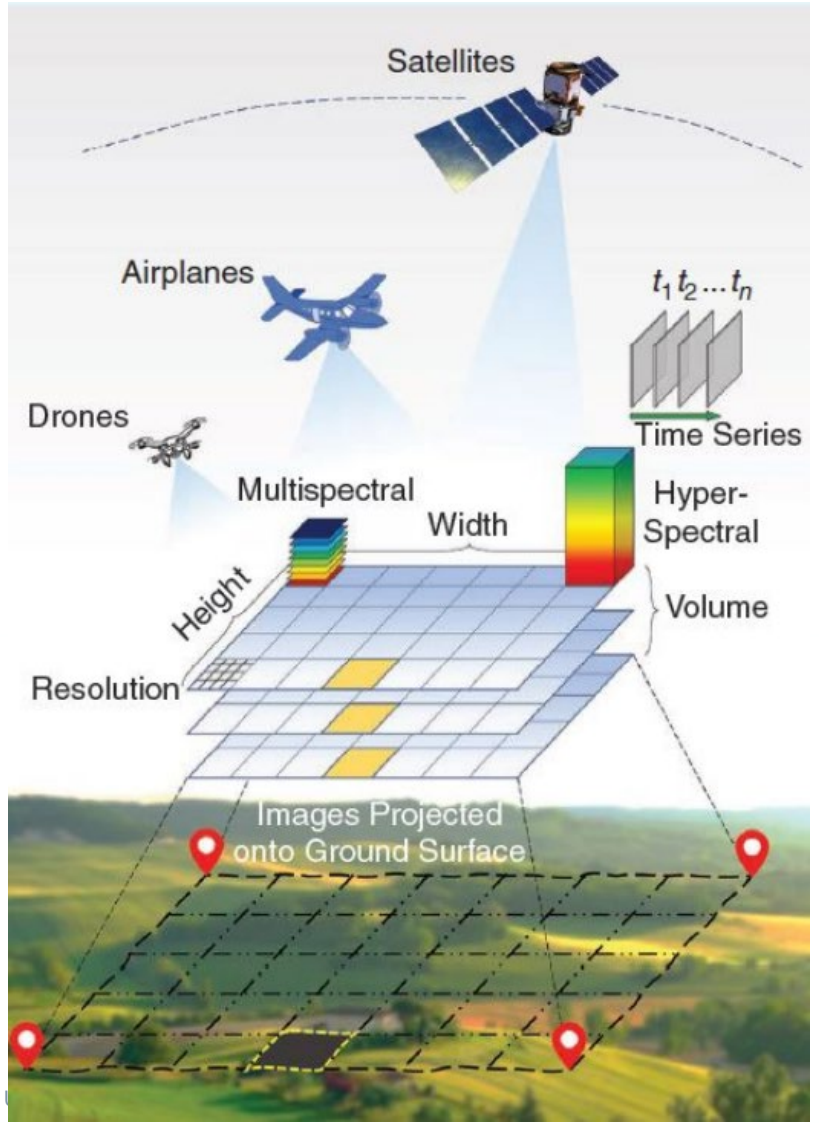
Self-supervised learning

- We have:
  - Large dataset of unlabelled inputs x
  - Small downstream dataset(s) of input, target pairs (x, y)

- Pretext task:
  - Learn model $f: x_{in} \rightarrow x_{out}$
  - For transformation g: $x \rightarrow (x_{in}, x_{out})$

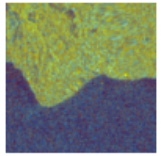
- Goal:
  - Obtain strong model (representations) f from large dataset
  - That improves supervised prediction by downstream model f_{down} on small dataset: $y = f_{down}(f(x))$

Input modalities



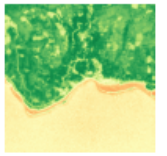
Sentinel-2 L2A
Sentinel-2 L1C
Sentinel-2 RGB

Sentinel-1 GRD
Sentinel-1 RTC



Digital Elevation Model (DEM)

Vegetation index (NDVI)



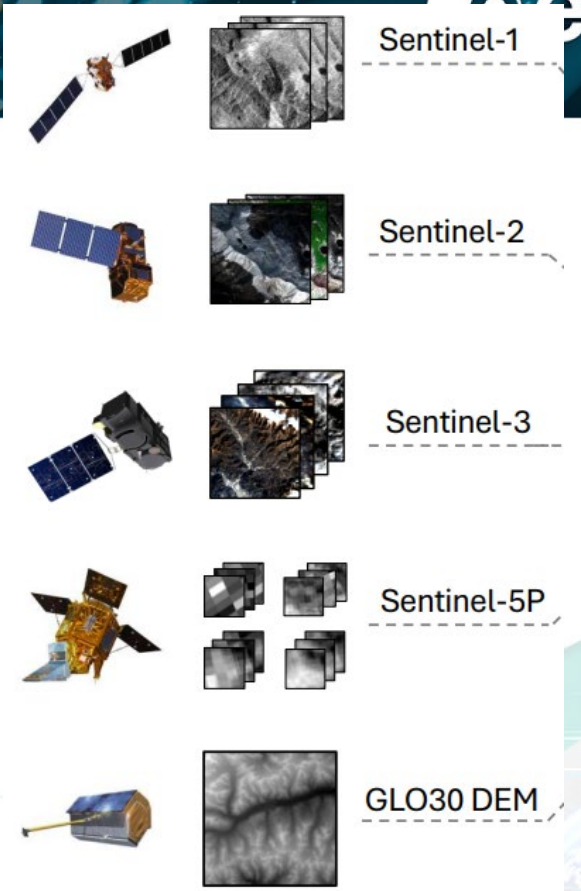
Land-use/ land-cover maps (LULC)

Image caption

A satellite image of a coastline with ...

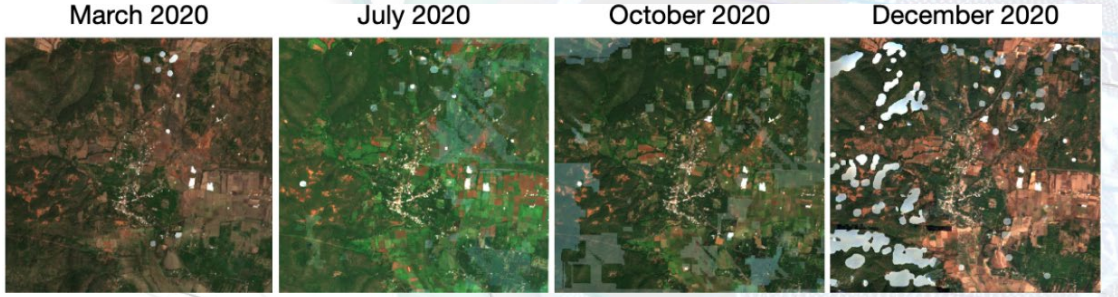
Coordinates at 0.25 x 0.25 deg

0°30′N 122°45′E



Wang, Yi, et al. "Towards a unified copernicus foundation model for earth vision." arXiv preprint arXiv:2503.11849 (2025).

Tseng, Gabriel, et al. "Lightweight, pre-trained transformers for remote sensi timeseries." arXiv preprint arXiv:2304.14065 (2023).

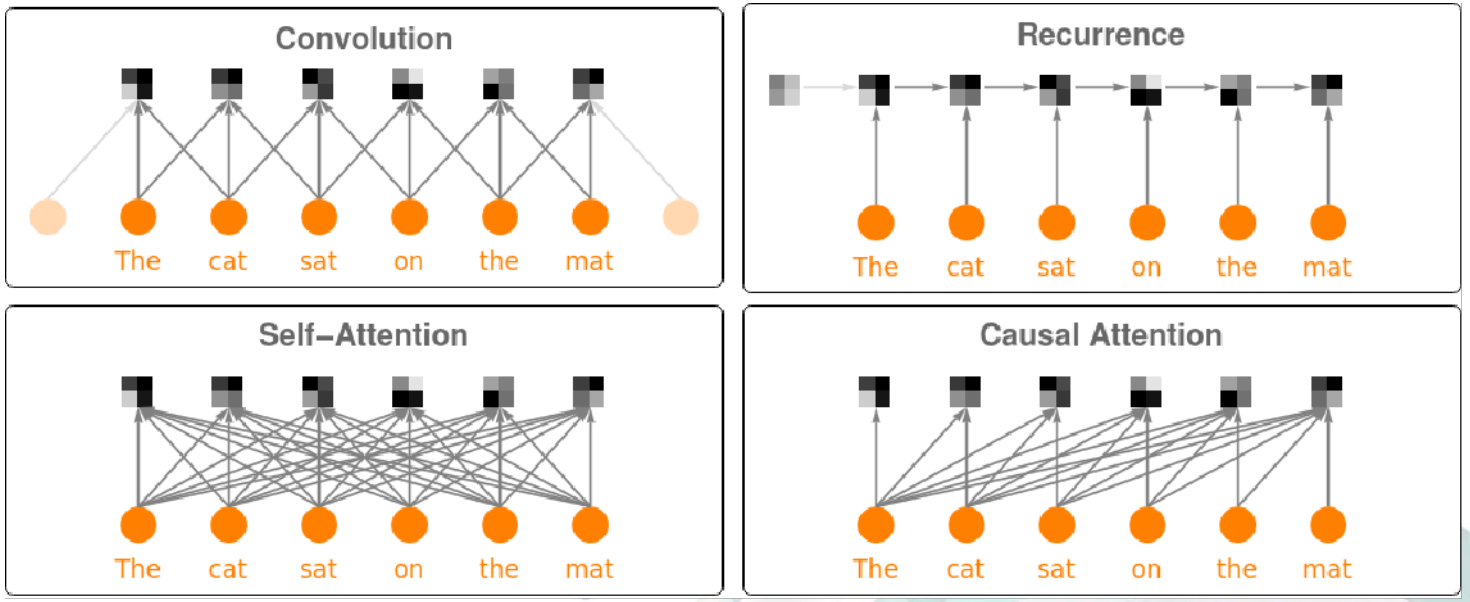


Jakubik, Johannes, et al. "Terramind: Large-scale generative multimodality for earth observation." arXiv preprint arXiv:2504.11171 (2025).

Schmitt, Michael, et al. "There are no data like more data: Datasets for deep learning in earth observation." IEEE Geoscience and Remote Sensing Magazine 11.3 (2023): 63-97.

Input modalities

Architectures



https://www.wolfram.com/language/12/neural-network-framework/use-transformer-neural-nets.html

Convolutional Neural Networks (CNNs):

Local, spatial feature extraction

Recurrent Neural Networks (RNNs):

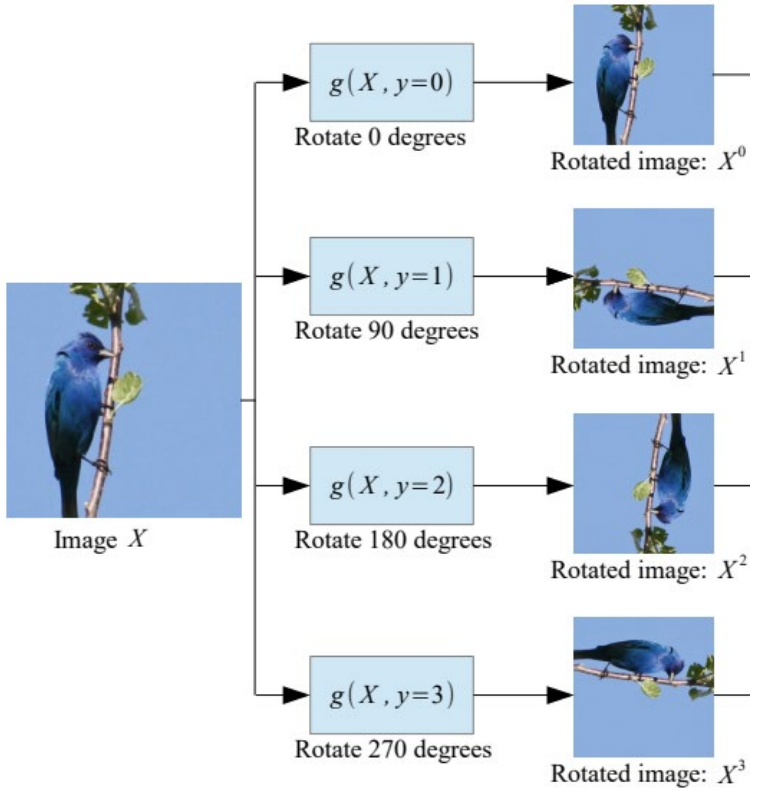
Process sequential data step by step, ideal for time-series analysis

Transformers:

(Self-)attention to capture global spatial, temporal, inter-modality relationships

Train in parallel so especially fit for large datasets

Learning metadata



SSL:

Learn model $f: x_{in} \rightarrow x_{out}$

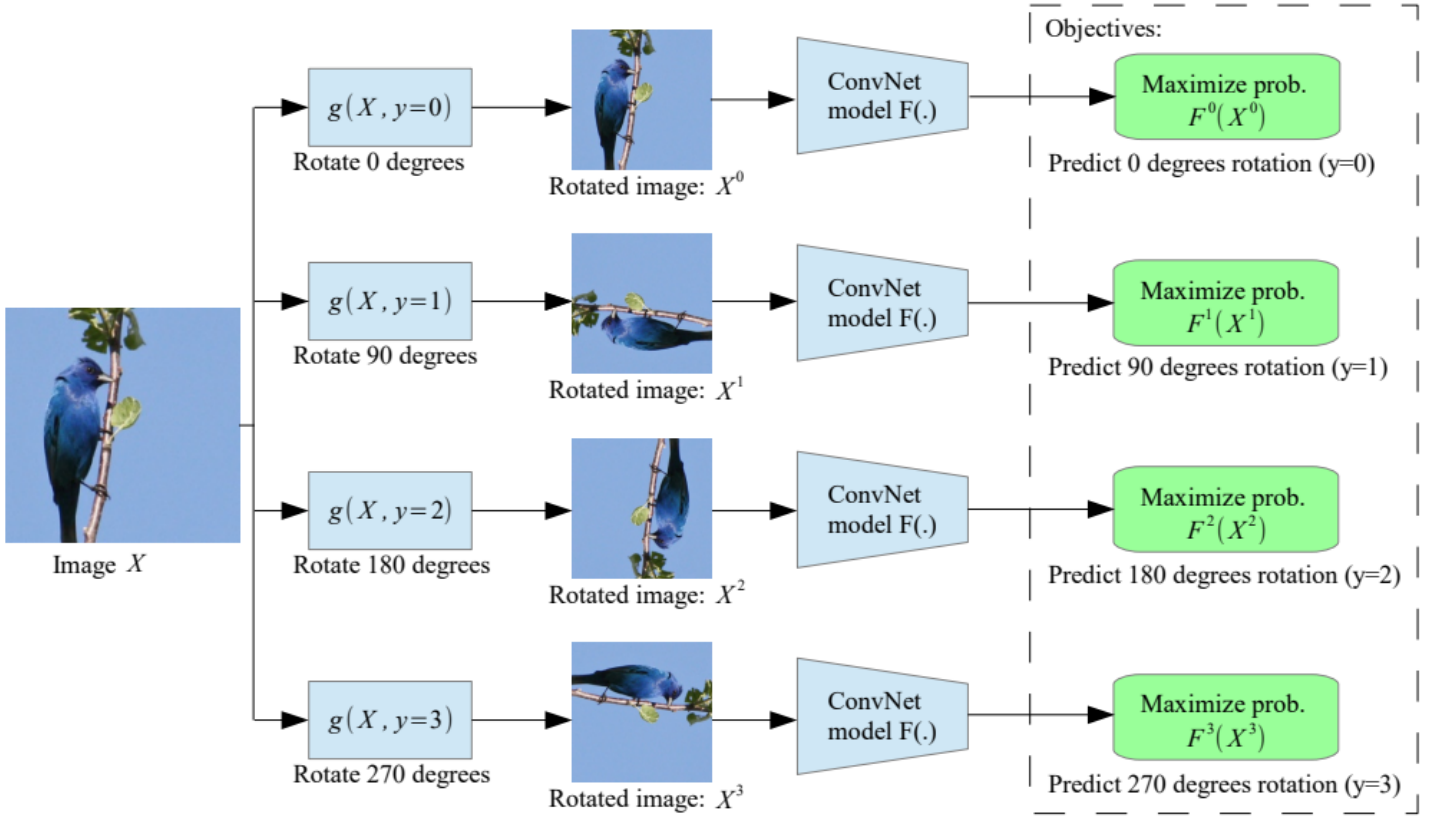
For transformation g: $x \rightarrow (x_{in}, x_{out})$

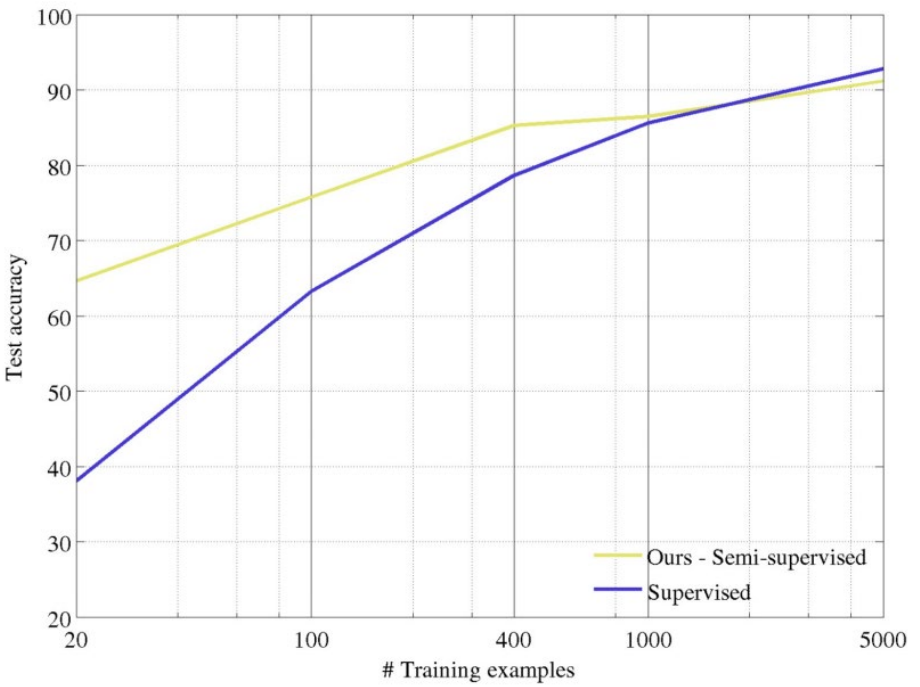
Learning rotations:

- $x_{out} = angle \in \{270, 180, 90, 0\}$
- g: $x \rightarrow (x_{rotated}, angle)$

Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." arXiv preprint arXiv:1803.07728 (2018).

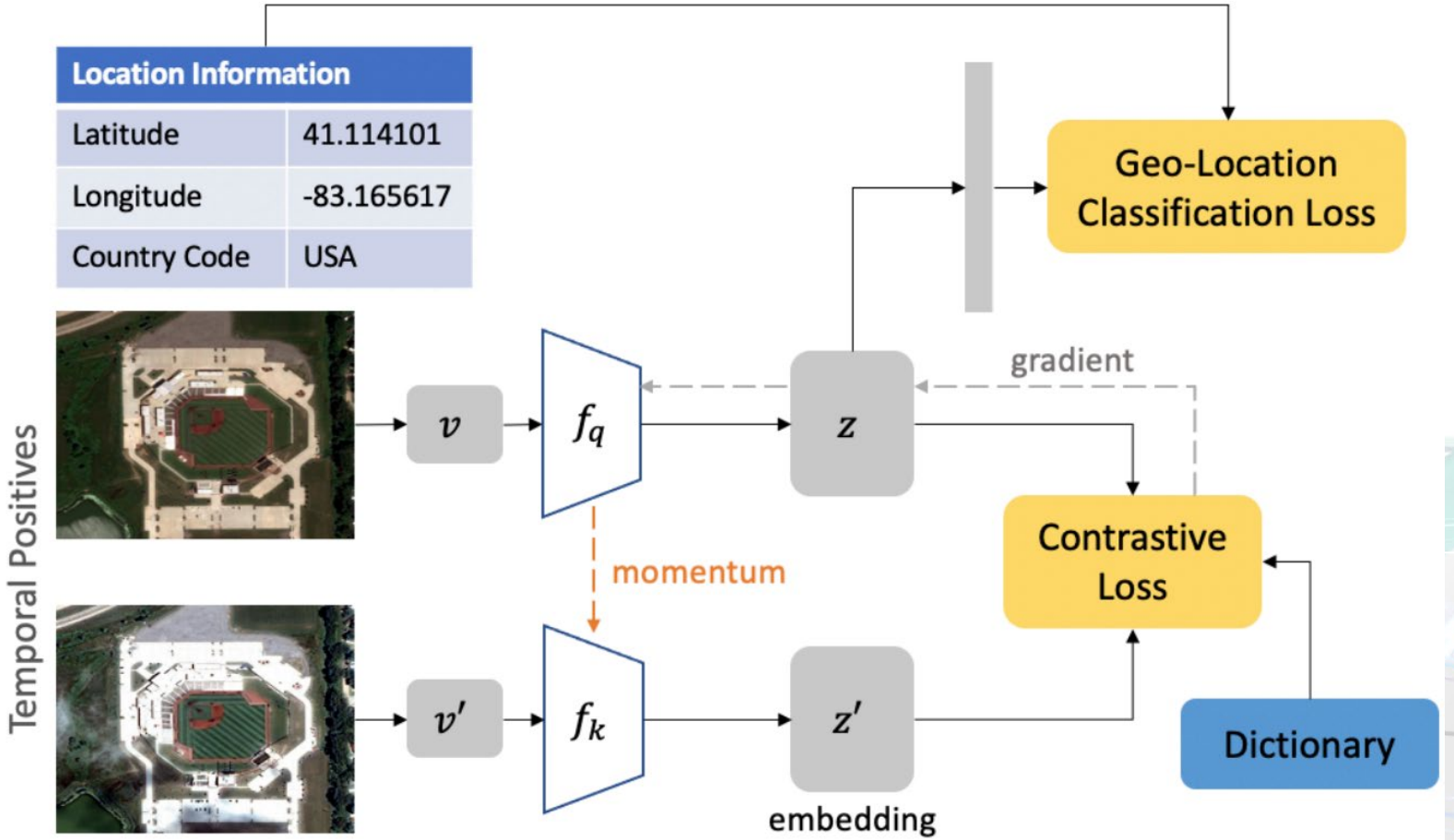
Learning metadata





Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." arXiv preprint arXiv:1803.07728 (2018).

Learning metadata: Geography-aware self-supervised learning



Ayush, Kumar, et al. "Geography-aware self-supervised learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Contrastive learning

SSL:

Learn model $f: x_{in} \rightarrow x_{out}$

For transformation g: $x \rightarrow (x_{in}, x_{out})$

Contrastive:

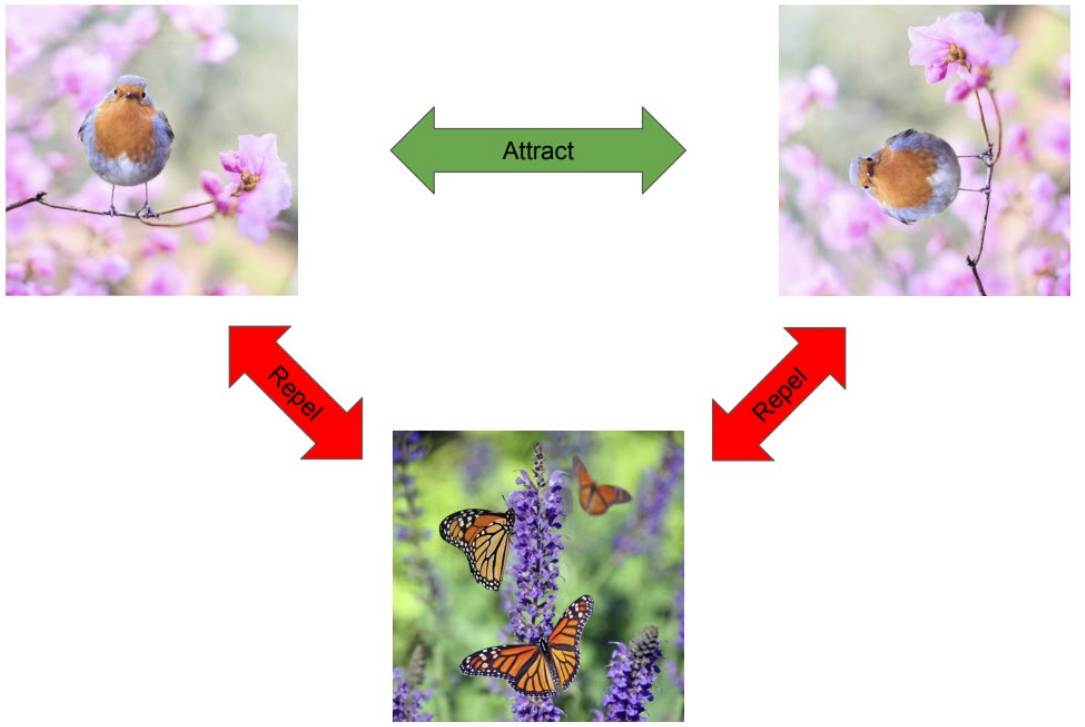
Learn model $f: x \rightarrow \mathbb{R}^D$

For positive and negative view x^+, x^-

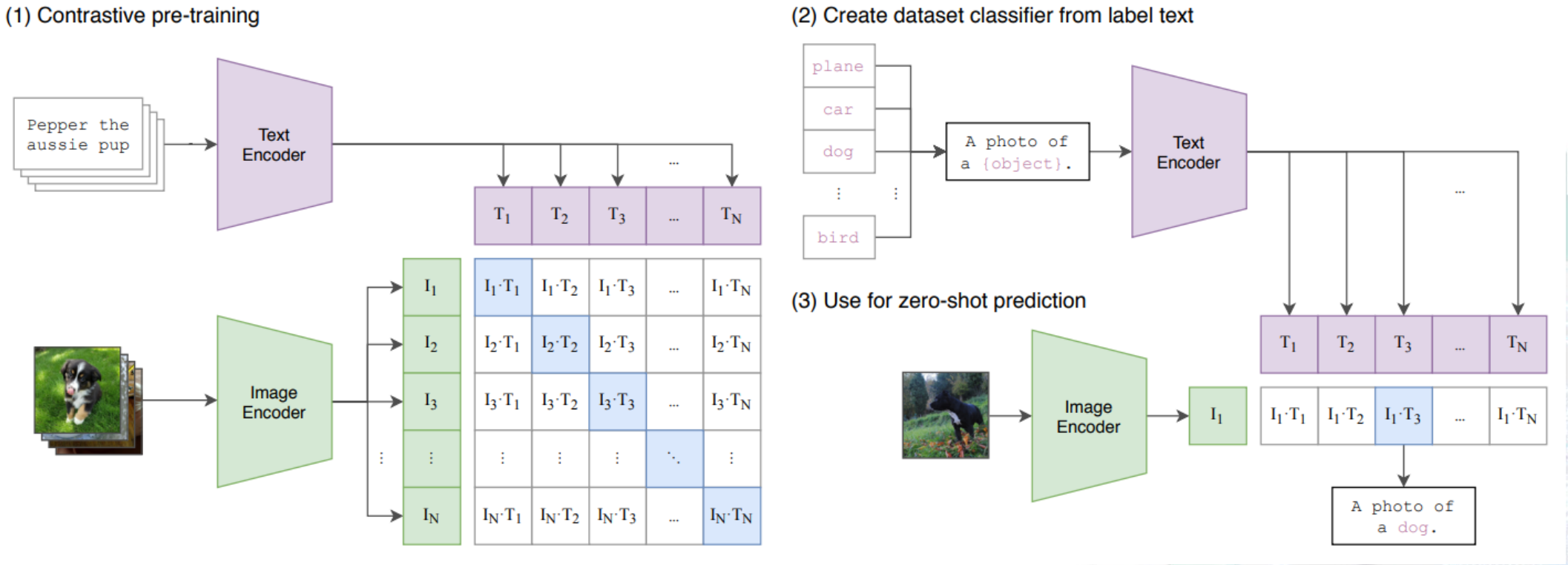
And similarity $s: x \rightarrow \mathbb{R}^1$

So that:

$$s\big(f(x), f(x^+)\big) > s\big(f(x), f(x^-)\big)$$

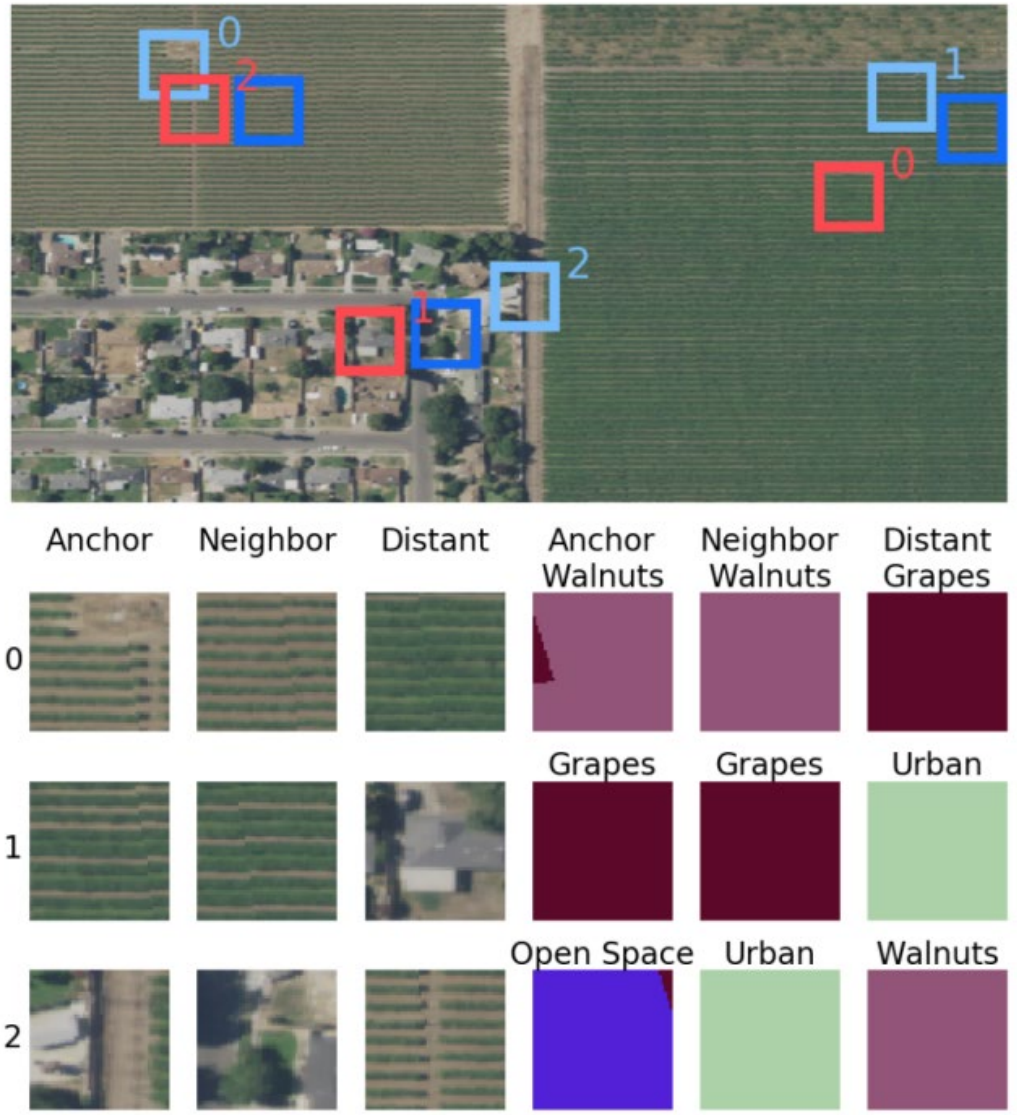


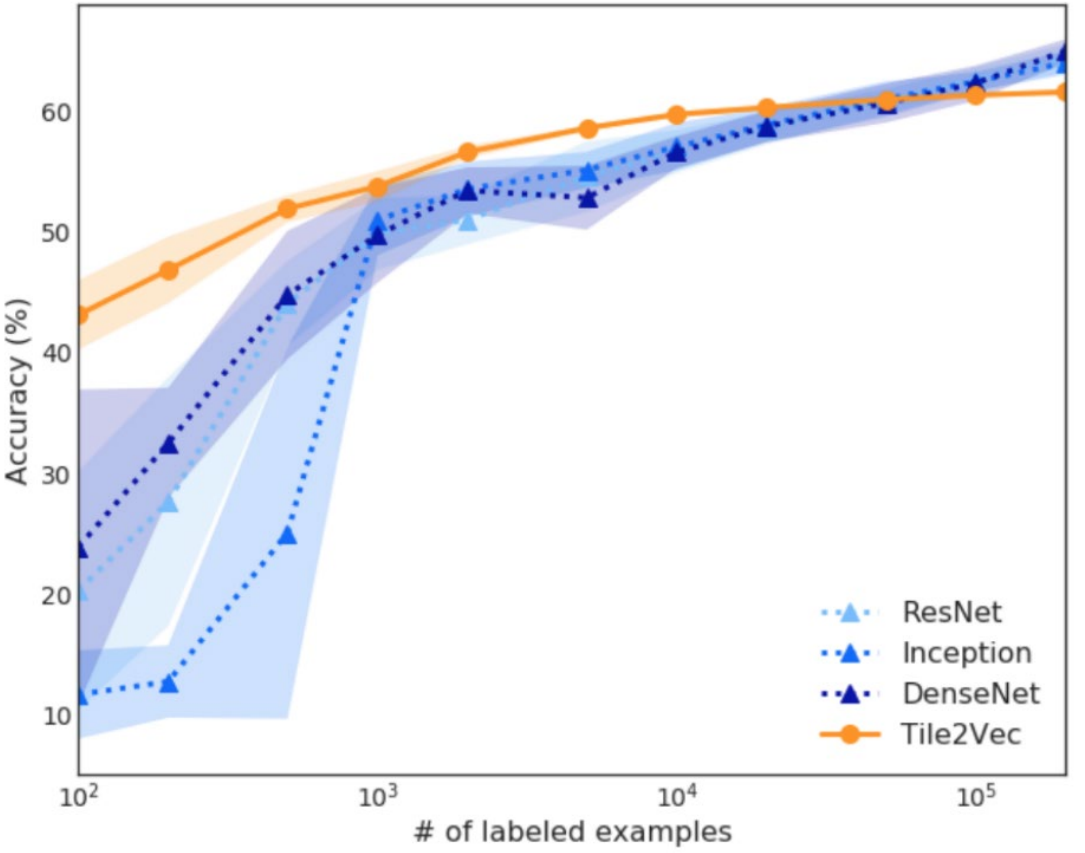
CLIP



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

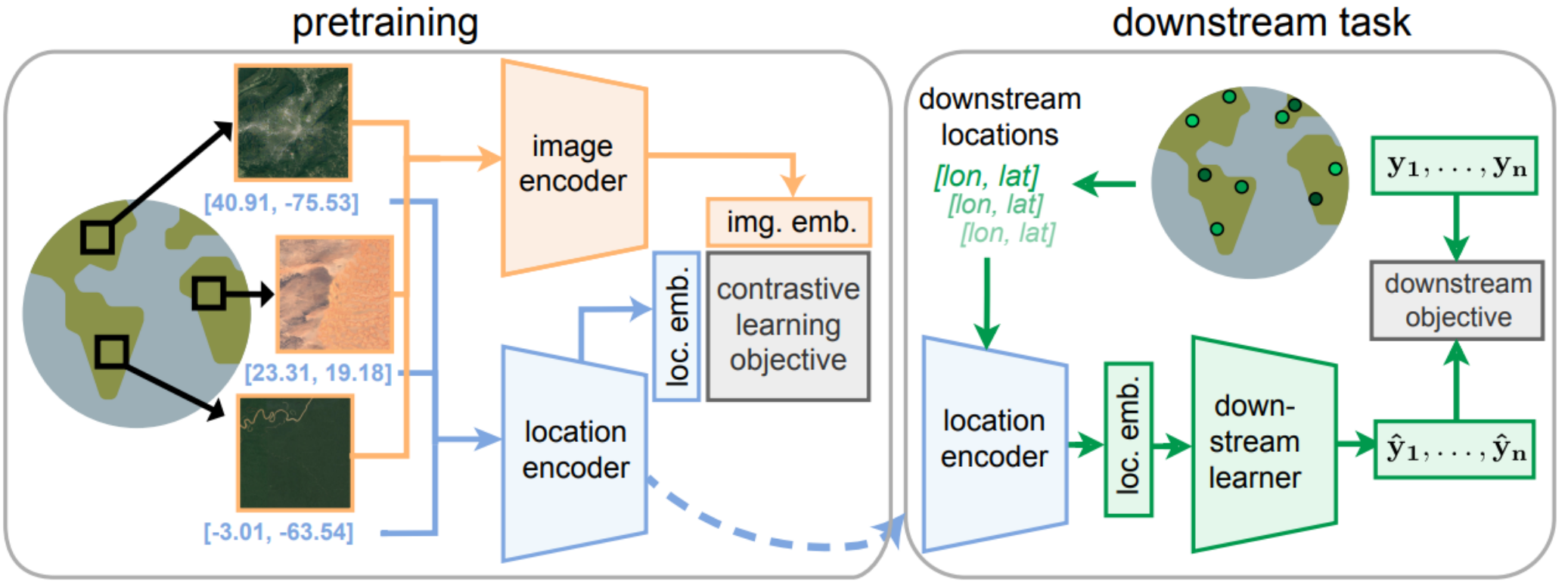
Tile2vec



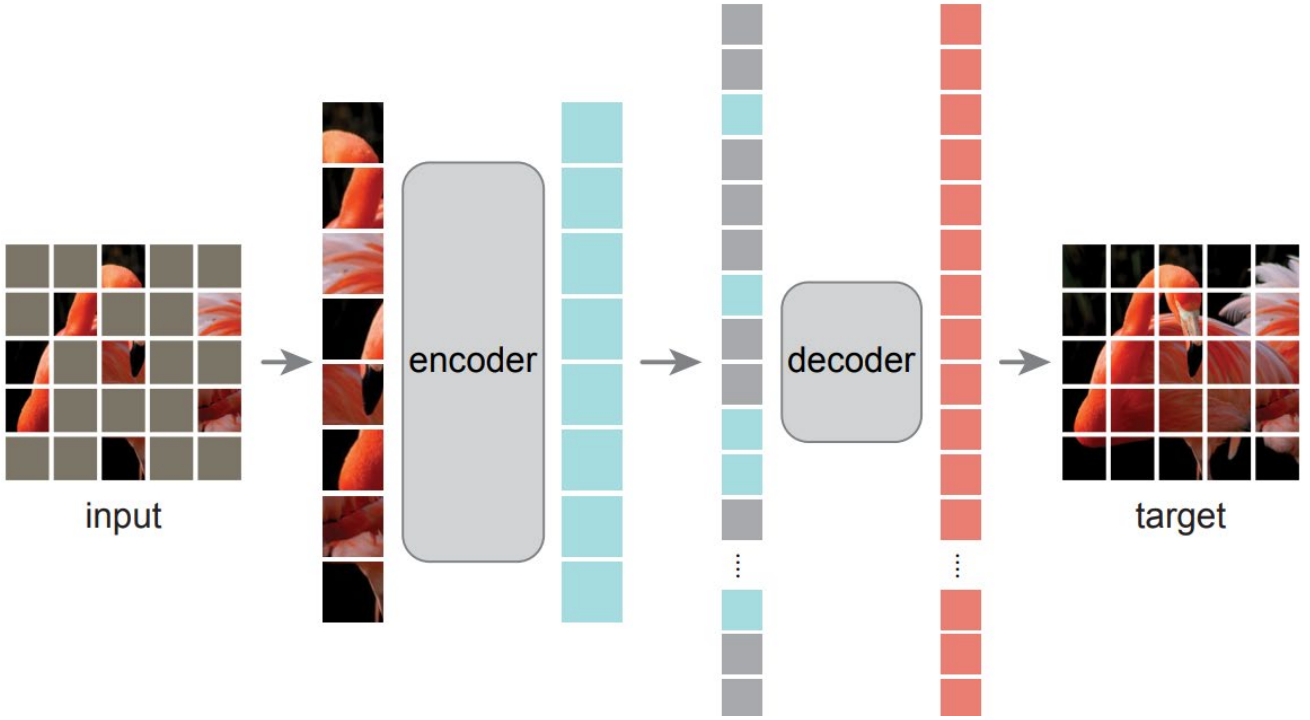


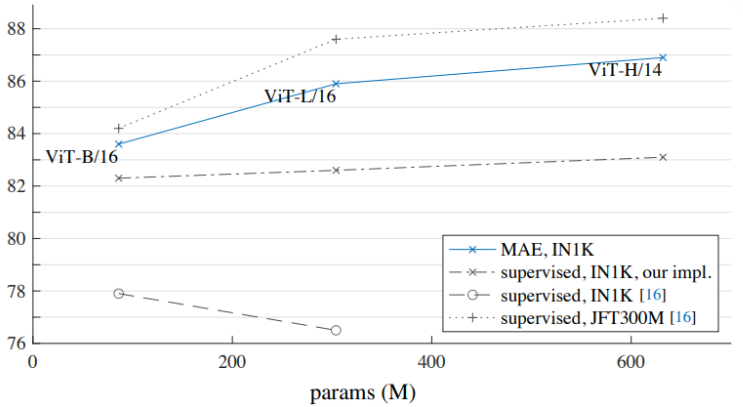
Jean, Neal, et al. "Tile2vec: Unsupervised representation learning for spatially distributed data." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.

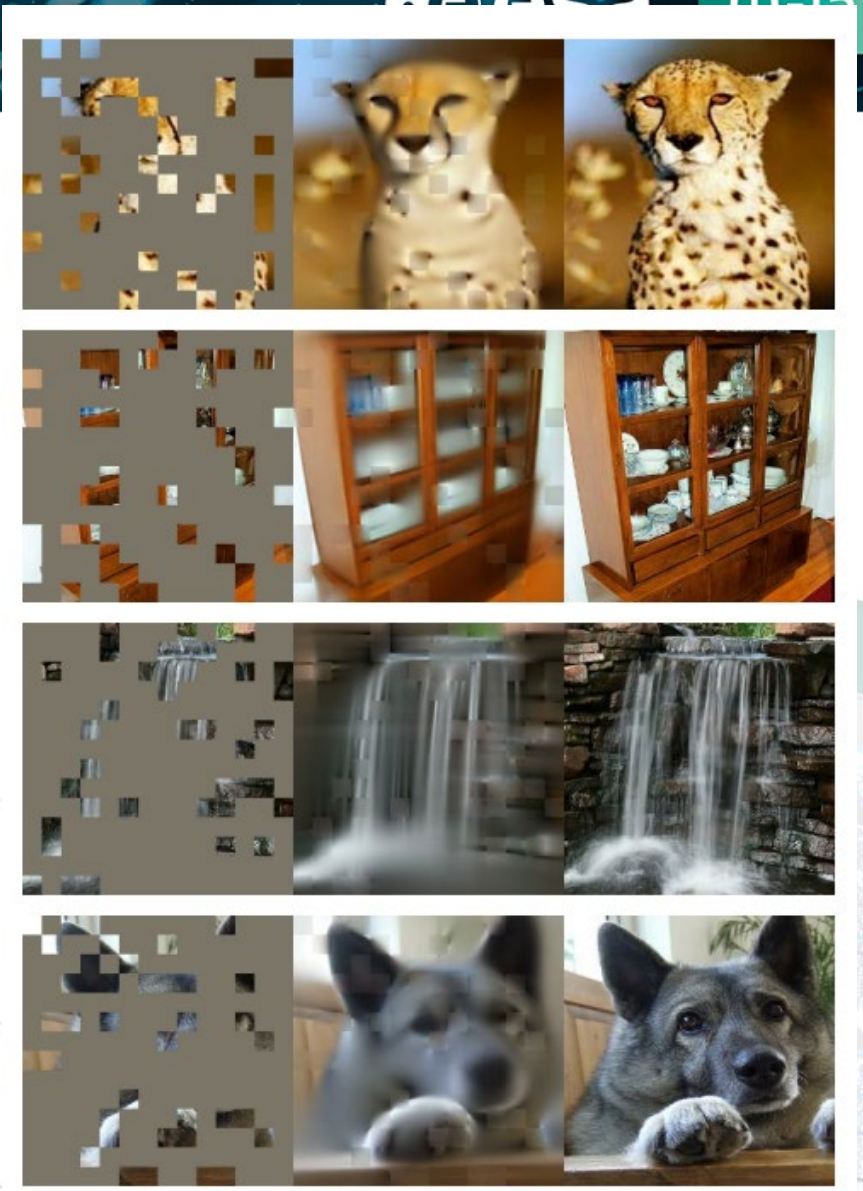
SatCLIP



Reconstruction

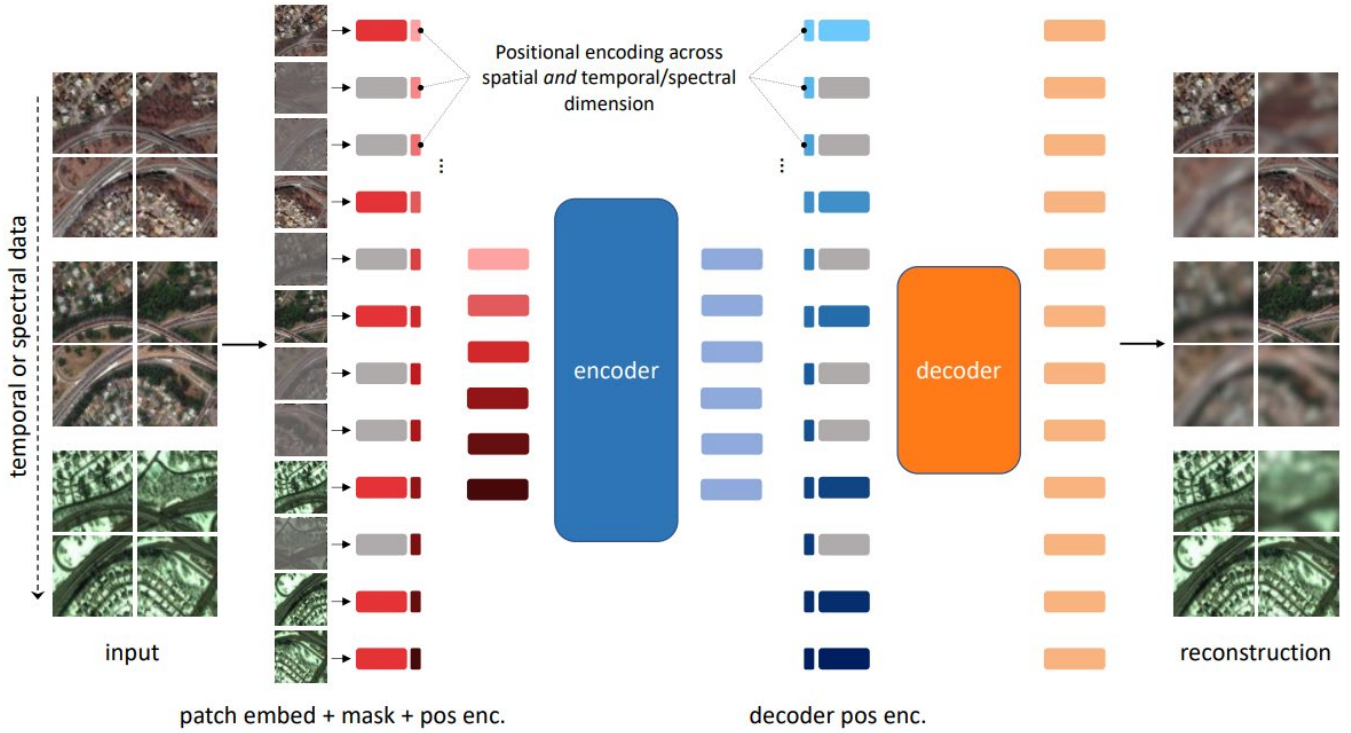


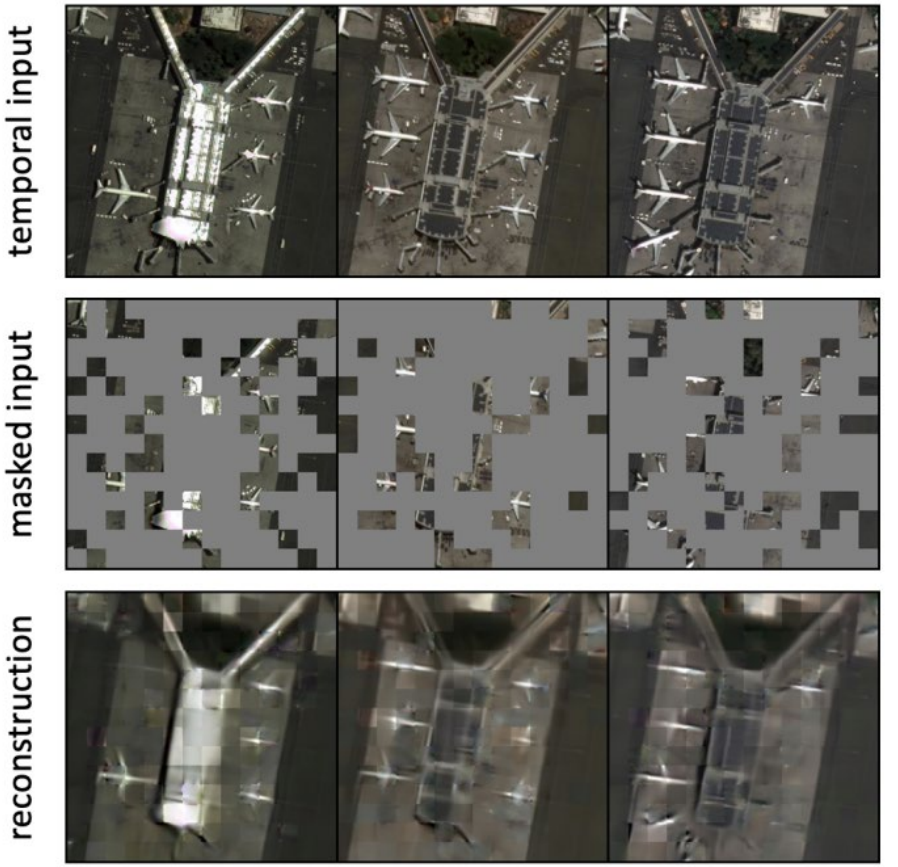




He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

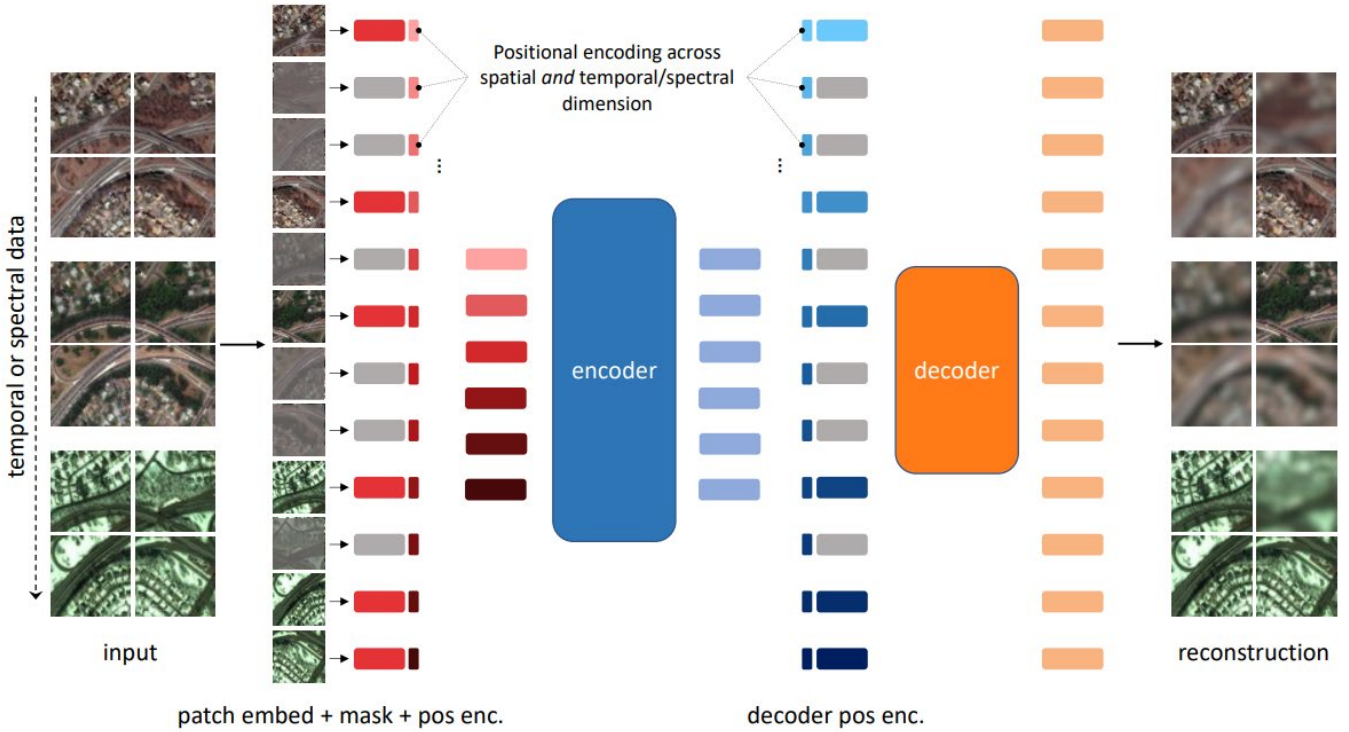
SatMAE



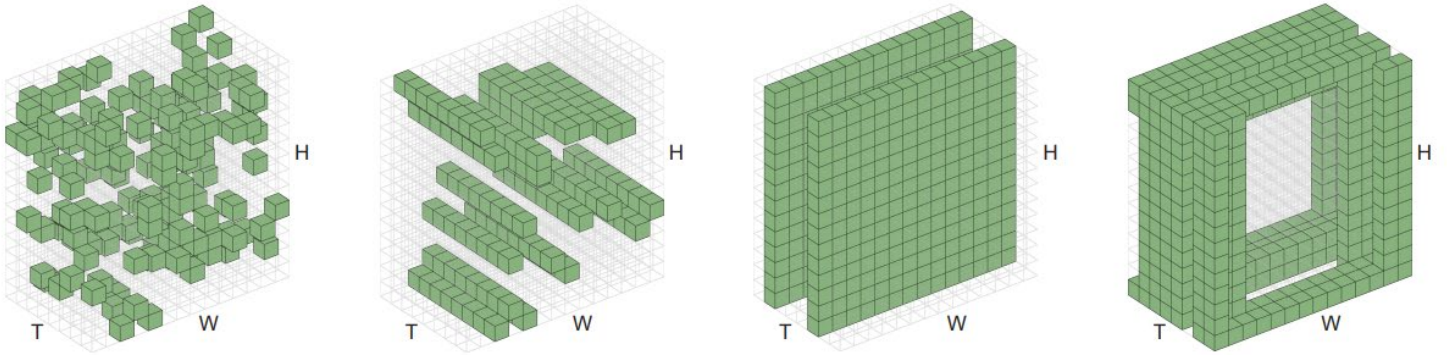


Cong, Yezhen, et al. "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery." Advances in Neural Information Processing Systems 35 (2022): 197-211.

SatMAE

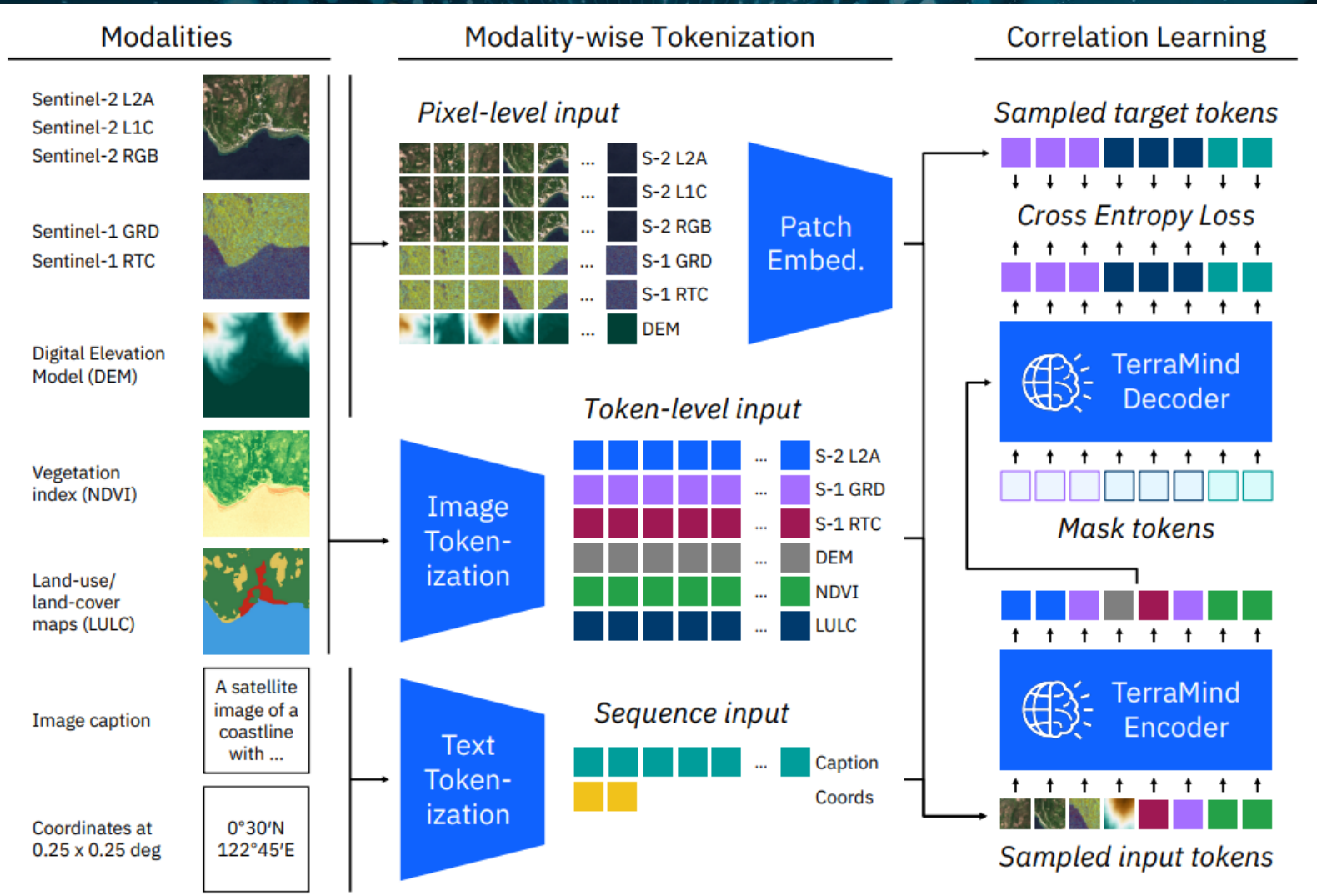


Feichtenhofer, Christoph, Yanghao Li, and Kaiming He. "Masked autoencoders as spatiotemporal learners." Advances in neural information processing systems 35 (2022): 35946-35958.



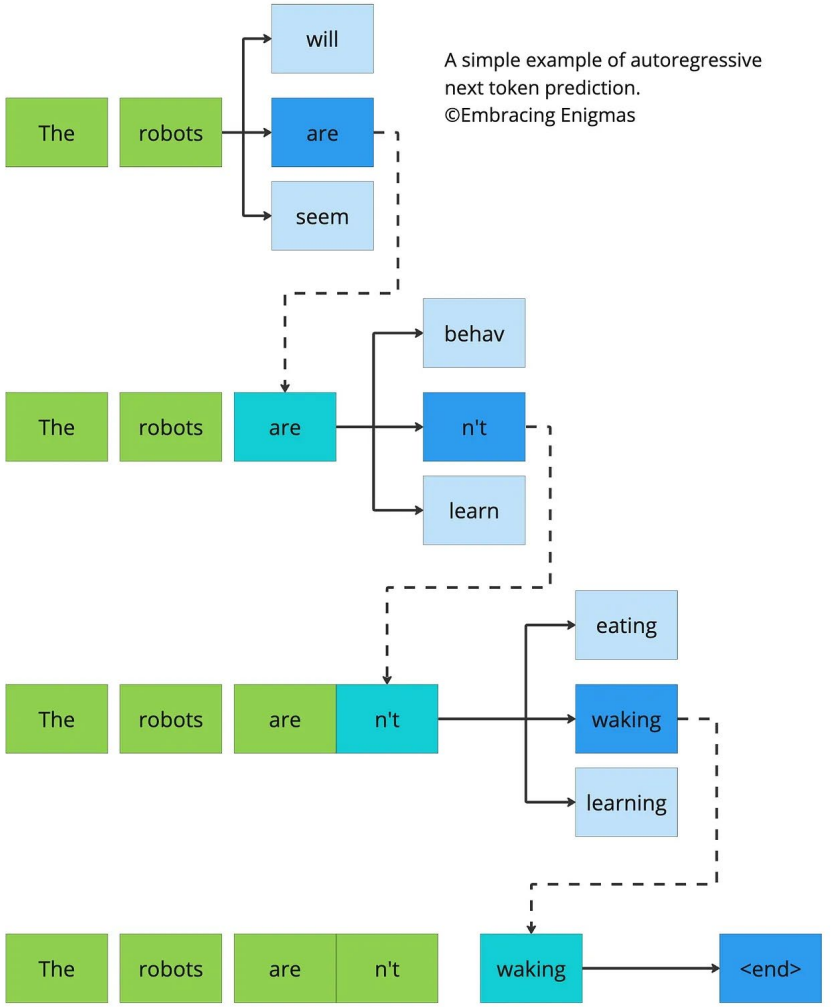
(a) agnostic, 90% (b) space-only, 90% (c) time-only, 75% (d) block, 75%

TerraMind



Jakubik, Johannes, et al. "Terramind: Large-scale generative multimodality for earth observation." arXiv preprint arXiv:2504.11171 (2025).

Generative models



Learn a model $p_\theta(x)$ that approximates the true underlying probability distribution $p_{data}(x)$

Generate new data by sampling from the learned model: $x_{new} \sim p_\theta(x)$

Foundation model: use learned internal representations as inputs for various tasks

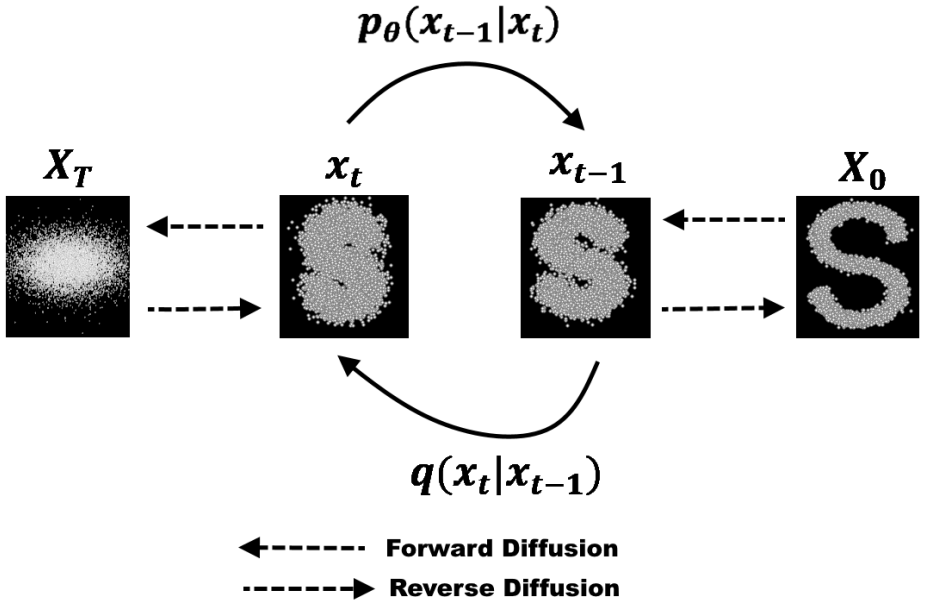
E.g. autoregressive language models:

$$p_\theta(x) = \prod_i p_\theta(x_i | x_0, \ldots, x_{i-1})$$

$$p_\theta(the, robots, are) = p_\theta(the) \cdot p_\theta(robots \mid the) \cdot p_\theta(are \mid the, robots)$$

https://embracingenigmas.substack.com/p/next-token-prediction-is-a-fundamental

Generative models



https://towardsdatascience.com/diffusion-models-made-easy-8414298ce4da/

Learn to reverse a process of adding noise

Forward (Fixed):

Start with a real image x_0

Gradually add Gaussian noise until image equals pure noise x_T

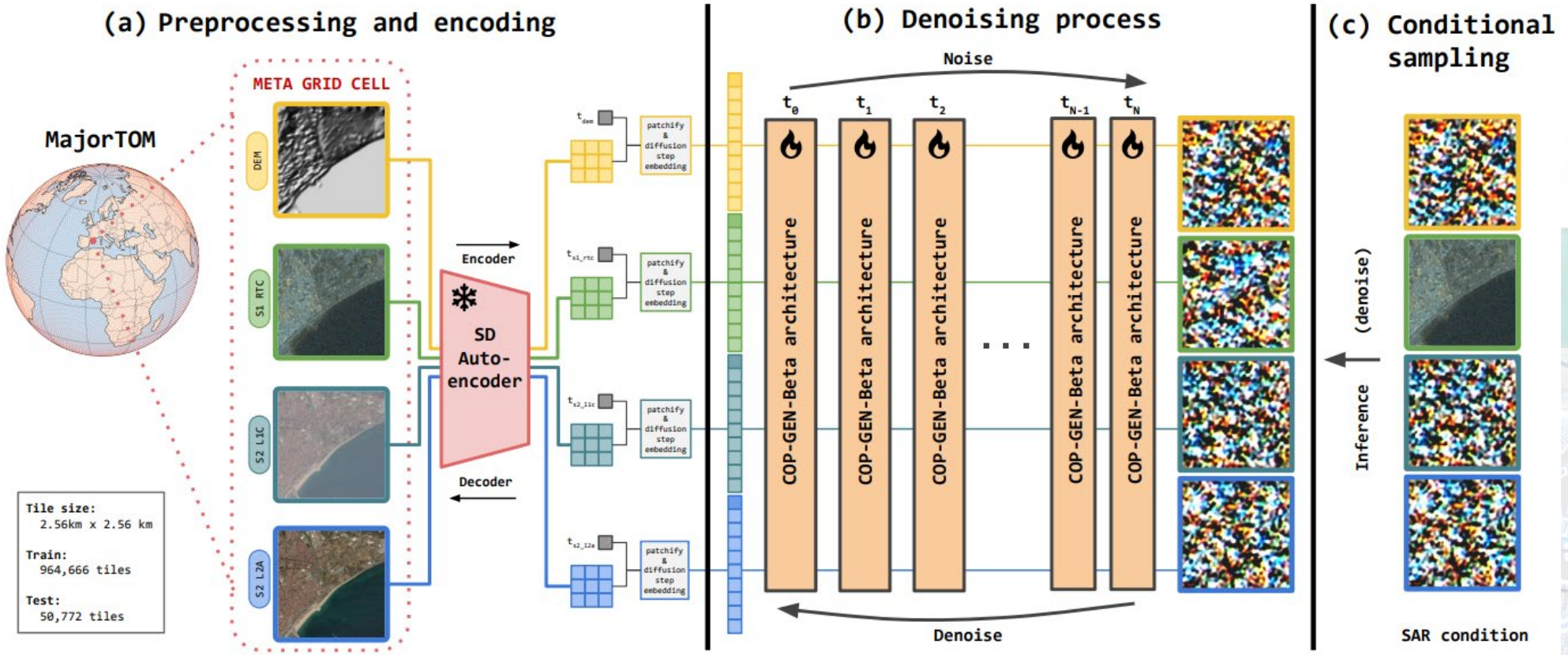
Reverse (Learned):

Train a neural network $p_\theta(x)(x_{t-1}|x_t)$ to denoise the image

Model learns to predict a slightly less noisy image x_{t-1} from x_t .

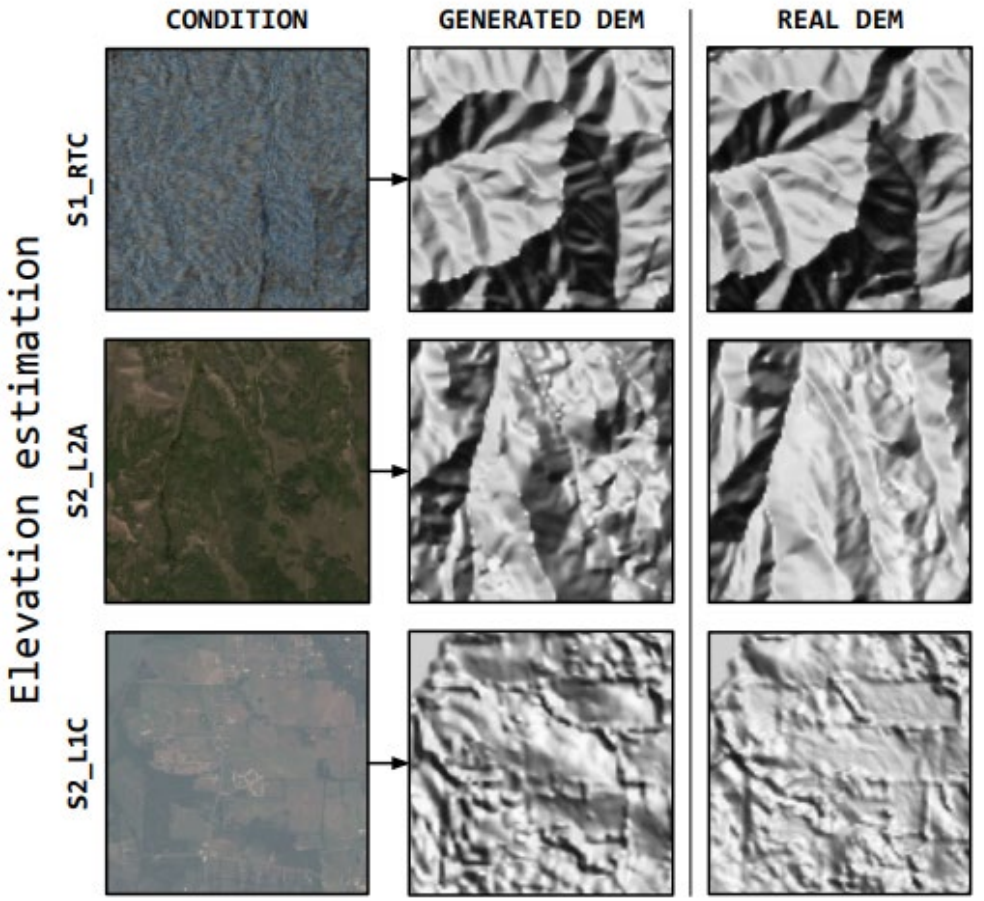
After T steps, it generates a clean image.

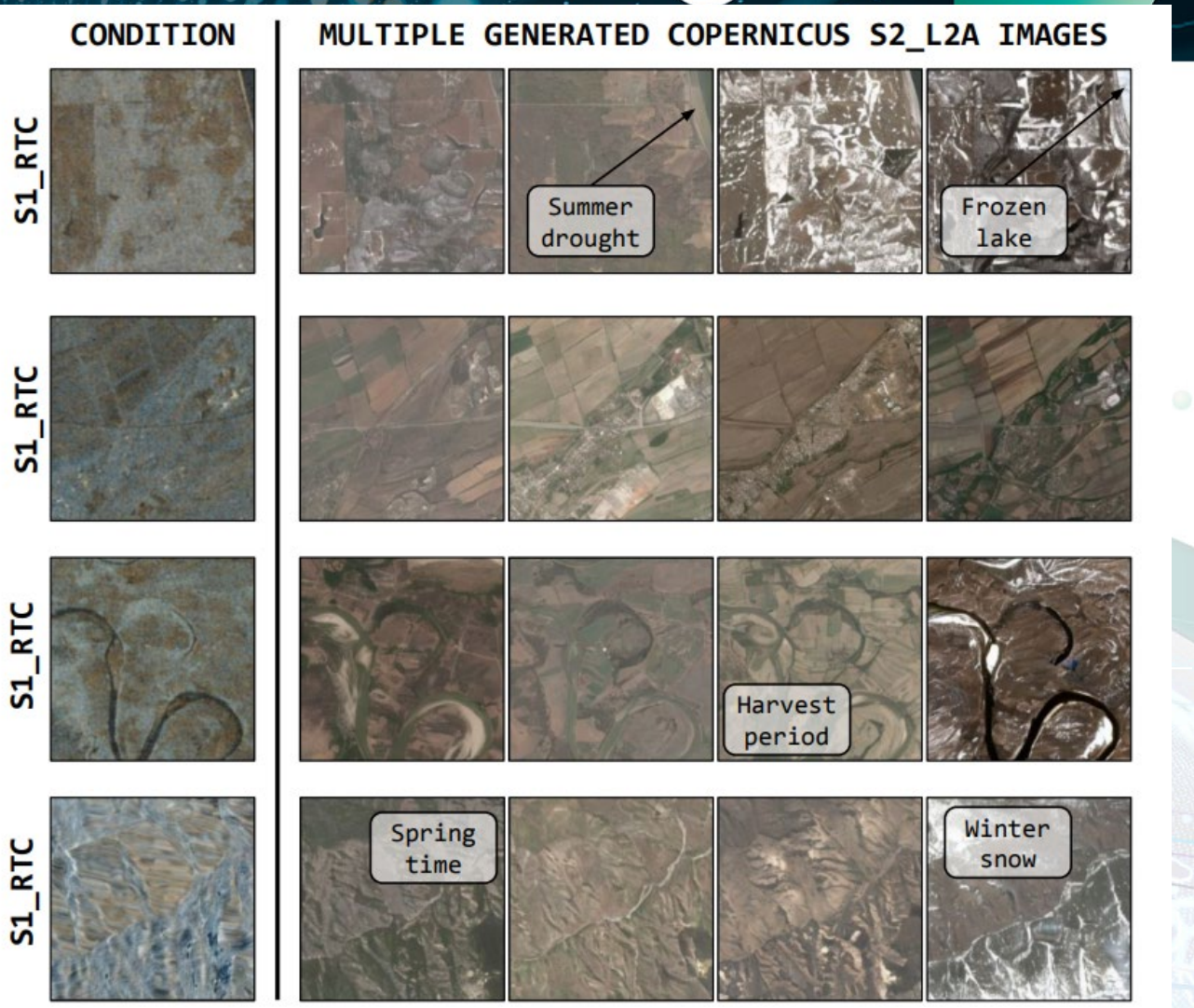
Generative models



Espinosa, Miguel, et al. "COP-GEN-Beta: Unified Generative Modelling of COPernicus Imagery Thumbnails." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Generative models



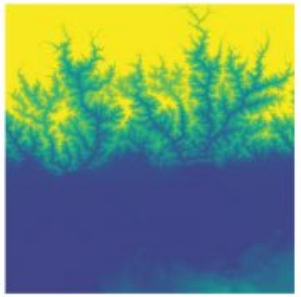


Espinosa, Miguel, et al. "COP-GEN-Beta: Unified Generative Modelling of COPernicus Imagery Thumbnails." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

What is next?

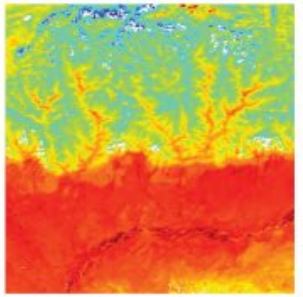
Mission Critical – Satellite Data is a Distinct Modality in Machine Learning

Products

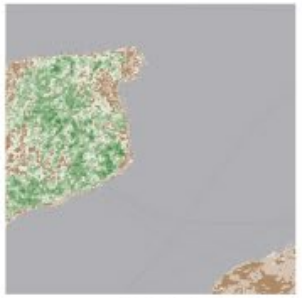


ALOS DEM

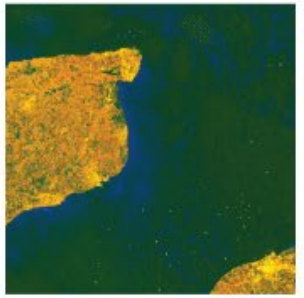
Eastern Himalayas, English Channel



MODIS day temp.



NDVI



SAR (Sentinel 1)

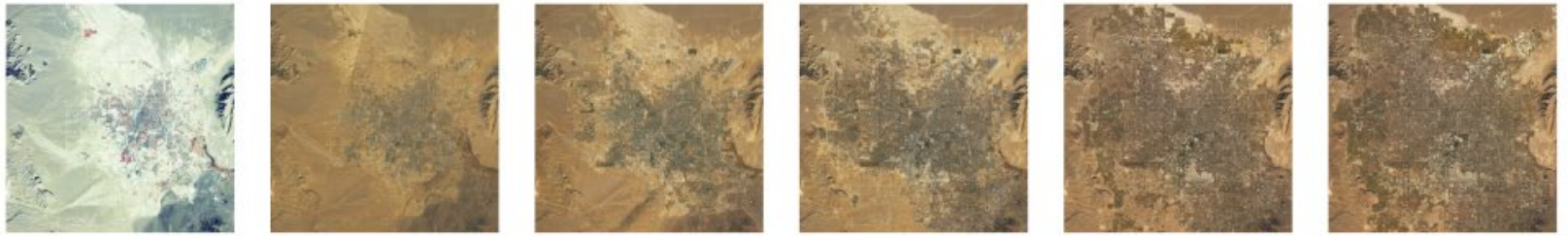
Spatial resolutions

Las Vegas, Nevada, USA

GOES-18 at 2000m/px | MODIS at 250m/px | Landsat 9 at 30m/px | Sentinel 2 at 10m/px | NAIP at 0.6m/px

Time steps

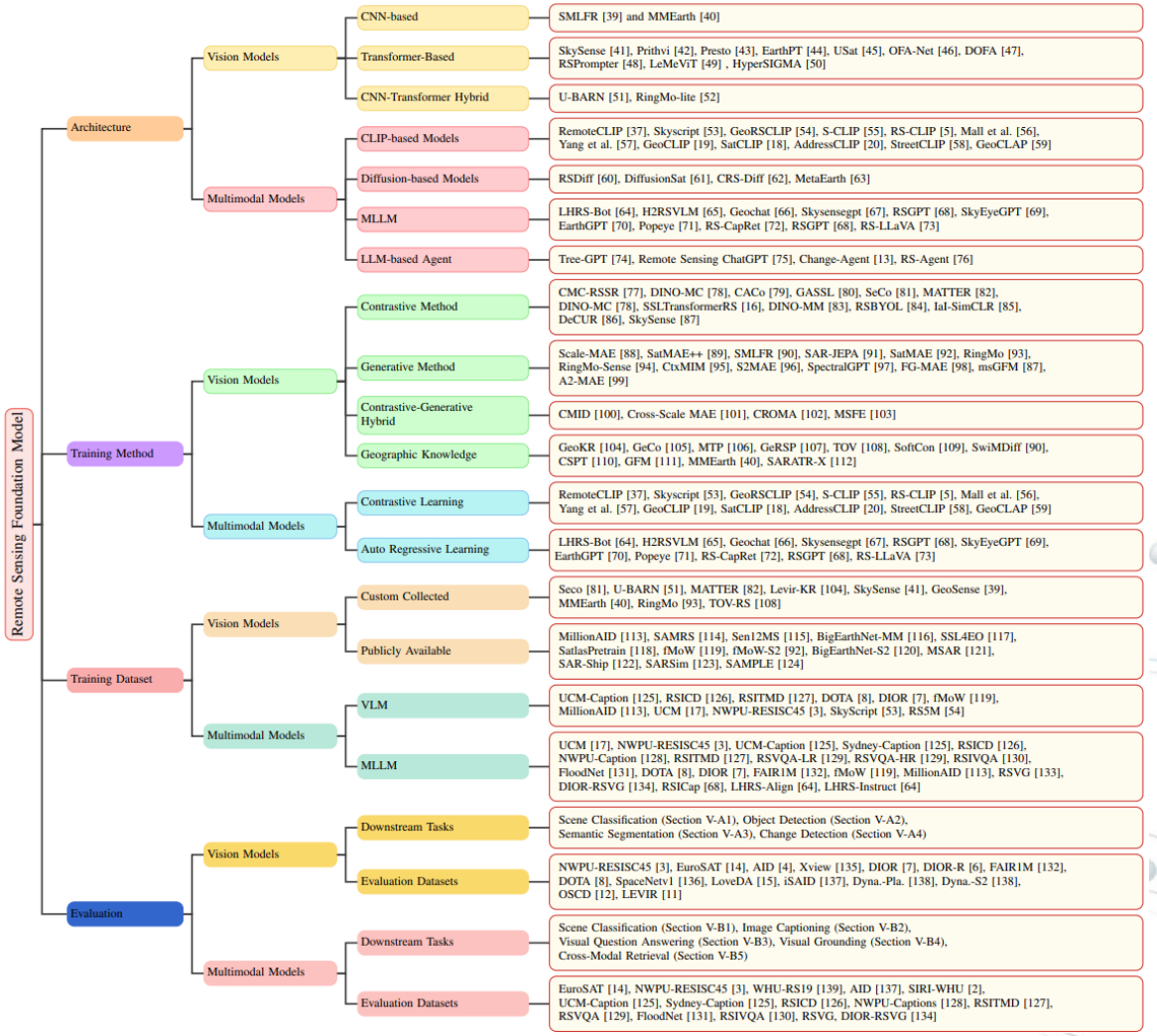
Las Vegas, Nevada, USA



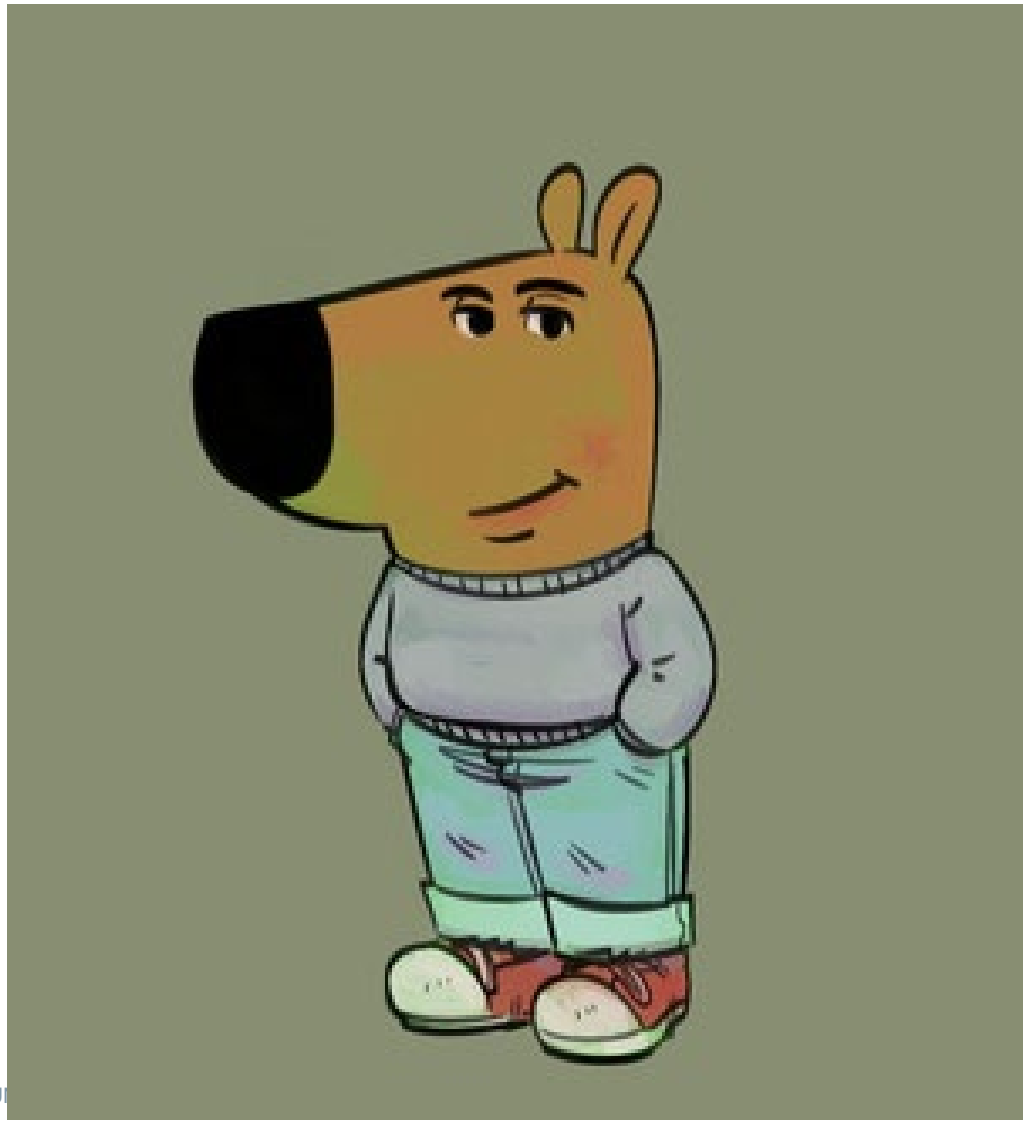
Dec. 25, 1973 | Dec. 3, 1982 | Dec. 9, 1993 | Dec. 24, 2001 | Dec. 23, 2013 | Dec. 28, 2023

BENCHMARKING GEOSPATIAL FOUNDATION MODELS

Which is the best among all available GFMs?



Which is the best among all available GFMs?



Let’s benchmark them and find out

Why Benchmarking GFMs

1. Performance Evaluation

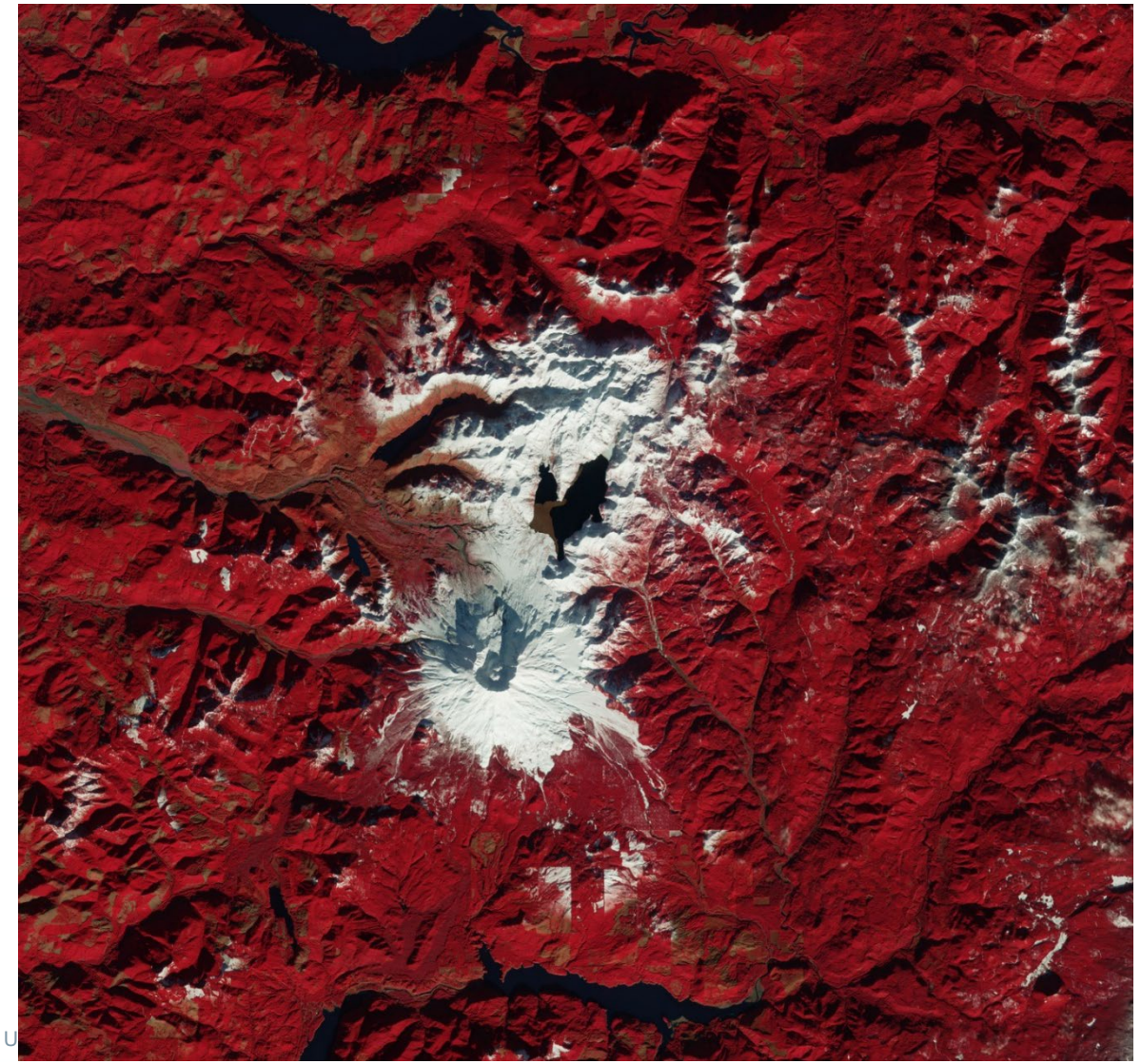
2. Fairness & Robustness

3. Guiding Improvements

Research themes:

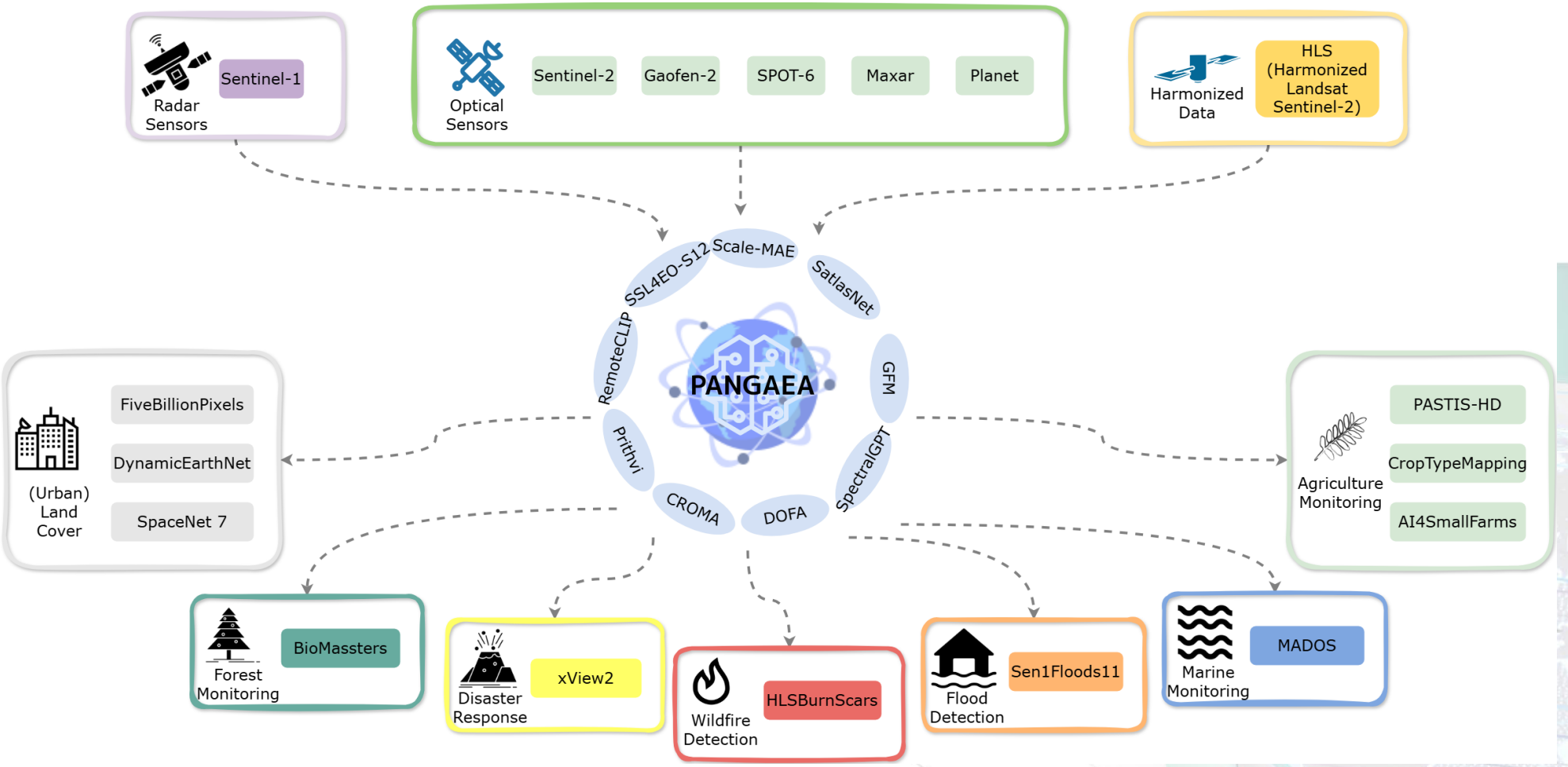
Effective benchmarking can enhance the capabilities of GFMs themselves, improving the easiness of adoption and finding the limitations to be considered

PANGAEA

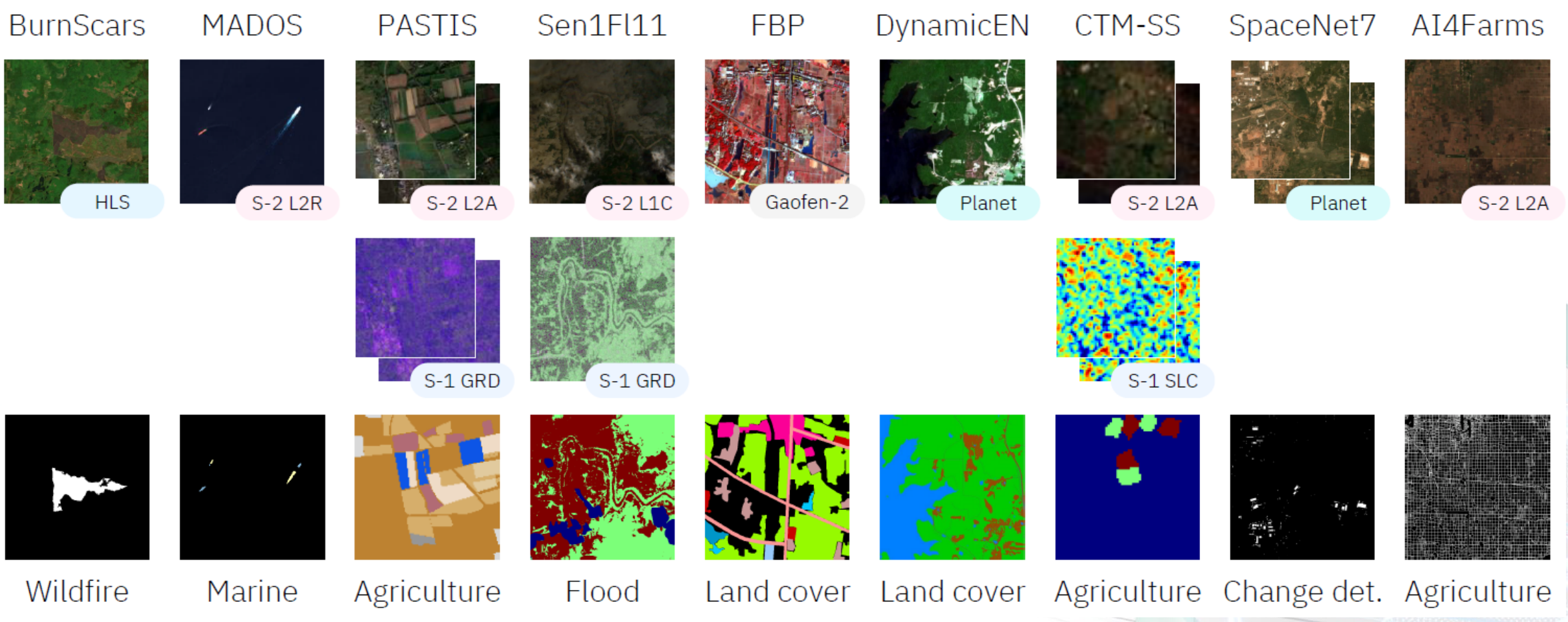


ESA U

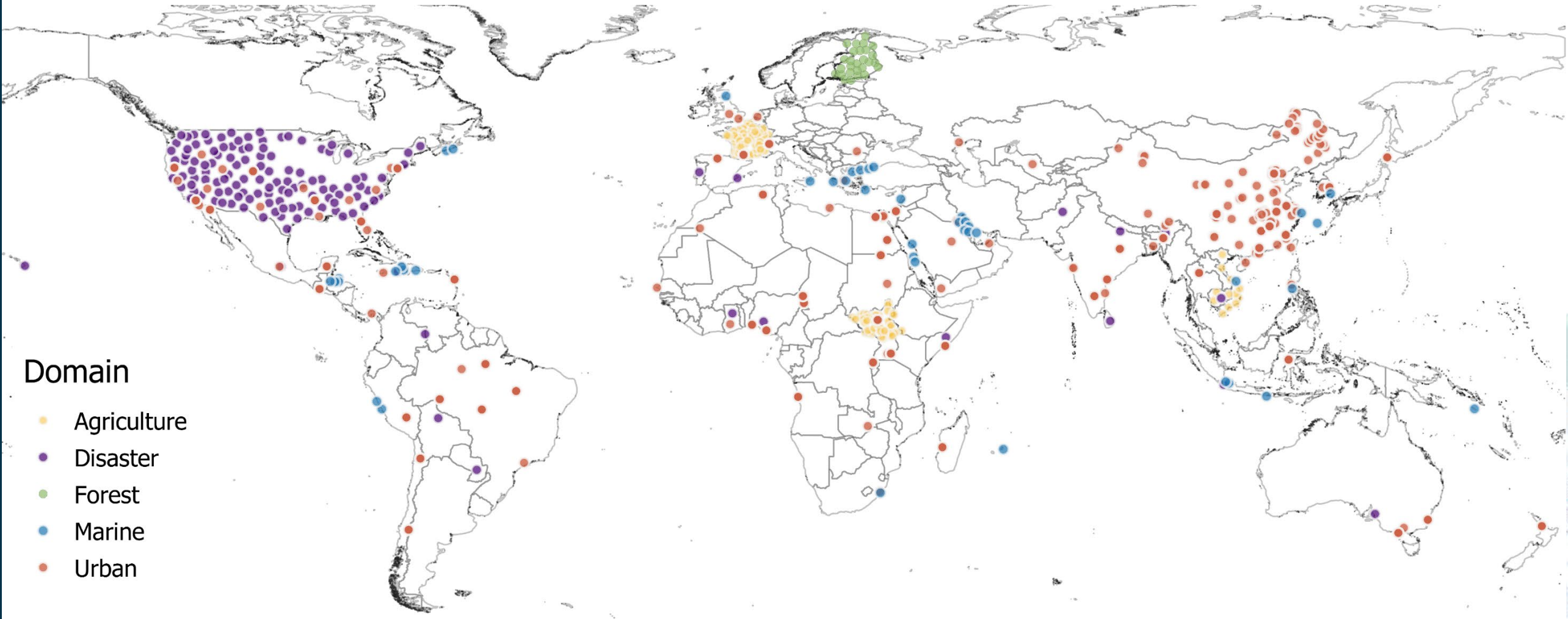
PANGAEA



PANGAEA: the datasets



PANGAEA: the datasets



PANGAEA: the models

Table 3: Overview of the pretraining datasets and the number of patches used by the selected GFMs. For Prithvi, the data volume is reported.

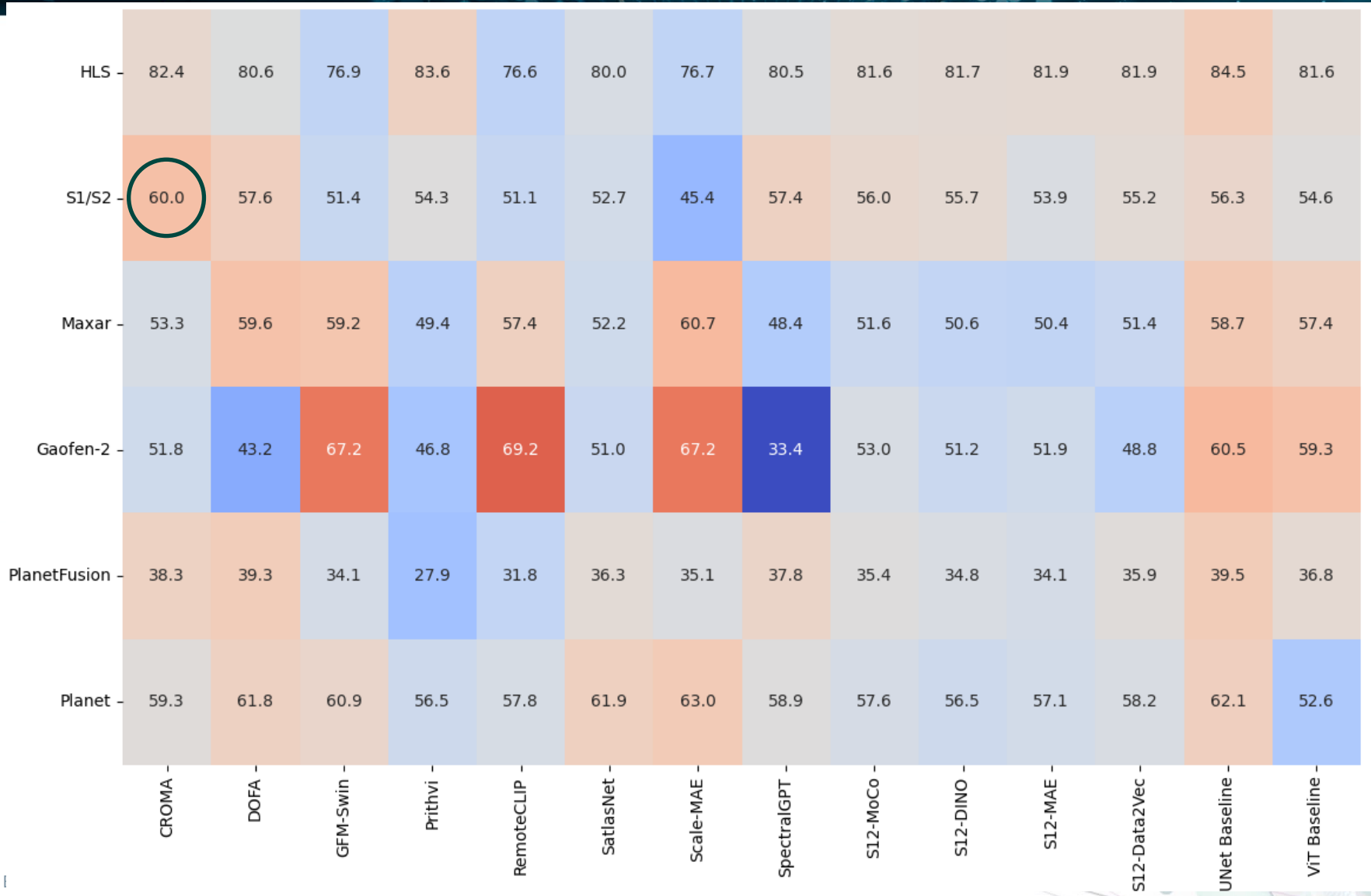
Model	Pretraining Images	Patches/Volume
CROMA	Sentinel-1, Sentinel-2	3M
DOFA	Sentinel-1, Sentinel2, Gaofen-2, NAIP, EnMAP	8.08M
GFM-Swin	NAIP, RSD46-WHU, MLRSNet, RESISC45, PatternNet	600K
Prithvi	Harmonized Landsat Sentinel-2 (HLS)	1TB
RemoteCLIP	SEG-4, DET-10, RET-3	165K
SatlasNet	Sentinel-2, NAIP	856K
Scale-MAE	FMoW-RGB	363.6K
SpectralGPT	fMoW-S2, BigEarthNet	1.47M
SSL4EO-S12	Sentinel-1, Sentinel-2	3M

Some results

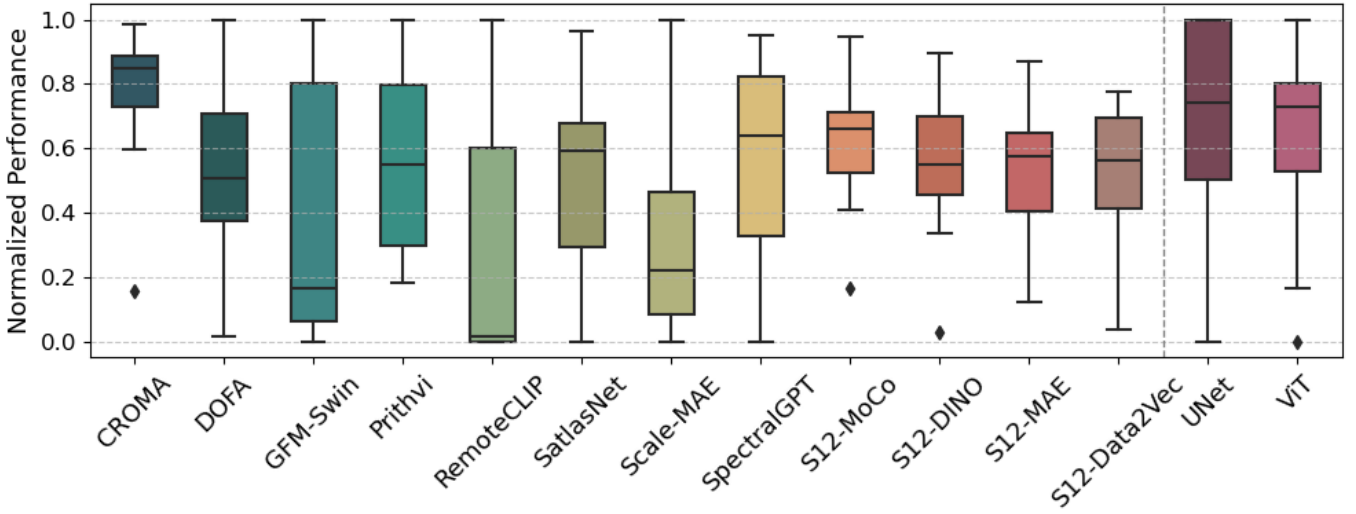
Table 5: Performance evaluation of Geospatial Foundation Models across 11 benchmark datasets using 100% of the data. For semantic segmentation and change detection tasks, the mIoU ↑ is reported. For regression task, RMSE ↓ is reported. #Top2 indicates the number of datasets where the models achieve top-2 performance across all evaluated datasets.

Model	HLS Burns	MADOS	PASTIS	Sen1Floods11	xView2	FBP	DynEarthNet	CropMap	SN7	AI4Farms	BioMassters	#Top2
CROMA	82.42	**67.55**	32.32	90.89	53.27	51.83	38.29	49.38	59.28	25.65	36.81	2
DOFA	80.63	59.58	30.02	89.37	59.64	43.18	39.29	51.33	61.84	27.07	42.81	2
GFM-Swin	76.90	64.71	21.24	72.60	59.15	67.18	34.09	46.98	60.89	27.19	46.83	1
Prithvi	83.62	49.98	33.93	90.37	49.35	46.81	27.86	43.07	56.54	26.86	39.99	1
RemoteCLIP	76.59	60.00	18.23	74.26	57.41	**69.19**	31.78	52.05	57.76	25.12	49.79	2
SatlasNet	79.96	55.86	17.51	90.30	52.23	50.97	36.31	46.97	61.88	25.13	41.67	0
Scale-MAE	76.68	57.32	24.55	74.13	**60.72**	67.19	35.11	25.42	**62.96**	21.47	47.15	3
SpectralGPT	80.47	57.99	35.44	89.07	48.40	33.42	37.85	46.95	58.86	26.75	36.11	1
S12-MoCo	81.58	51.76	34.49	89.26	51.59	53.02	35.44	48.58	57.64	25.38	40.21	0
S12-DINO	81.72	49.37	36.18	88.61	50.56	51.15	34.81	48.66	56.47	25.62	41.23	1
S12-MAE	81.91	49.90	32.03	87.79	50.44	51.92	34.08	45.8	57.13	24.69	41.07	0
S12-Data2Vec	81.91	44.36	34.32	88.15	51.36	48.82	35.90	**54.03**	58.23	24.23	41.91	1
UNet Baseline	**84.51**	54.79	31.60	**91.42**	58.68	60.47	**39.46**	47.57	62.09	**46.34**	**35.67**	6
ViT Baseline	81.58	48.19	**38.53**	87.66	57.43	59.32	36.83	44.08	52.57	38.37	38.55	2

Some results

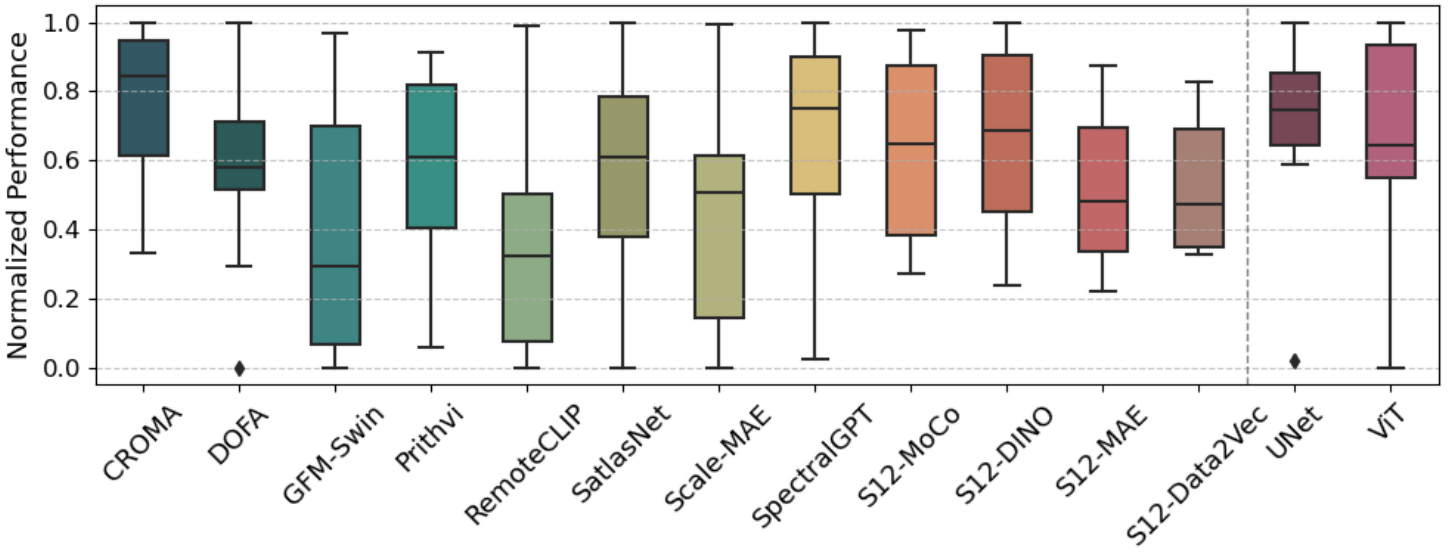


Some results: about data scarcity



50% labels

10% labels



The codebase

setup.py — Minor update of the team name — 7 months ago

README | GPL-3.0 license

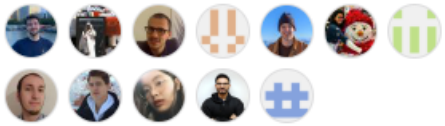
Tests passing

PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models

News

- [23/04/2025] we pushed a new version of the code, fixing different bugs (e.g. commands are working for all the datasets now, metric computation with ignore_index is fixed, etc...). In the next month, we will provide: all downloadable datasets and models, downloadable stratified subsamples for all the datasets, classification. Stay tuned!
- [22/04/2025] on EarthDay, PANGAEA was officialy adopted to benchmark TerraMind. Read the news and the pre-print. We will release the benchmarking code in PANGAEA very soon!
- [05/12/2024] the pre-print is out!

Contributors 12



Languages

- Python 100.0%

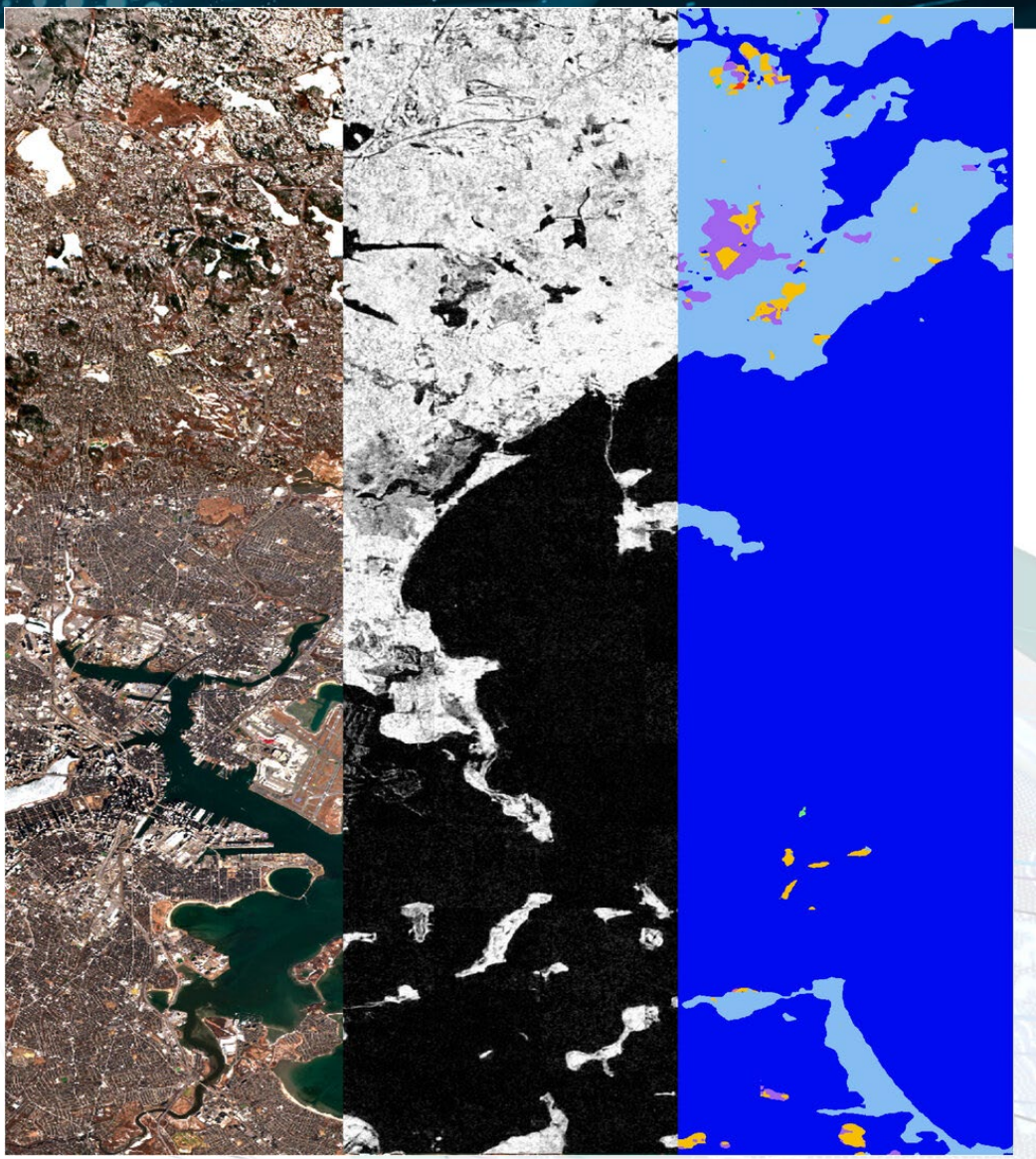
>200 stars

Some nice achievements

- **AnySat** [CVPR 2025 Highlight]:
  One Earth Observation Model for Many Resolutions, Scales, and Modalities

- **TerraMind** [ICCV 2025]:
  Large-Scale Generative Multimodality for Earth Observation

- Several newly added features
  (e.g. kNN, classification, thematic benchmarks, challenges)



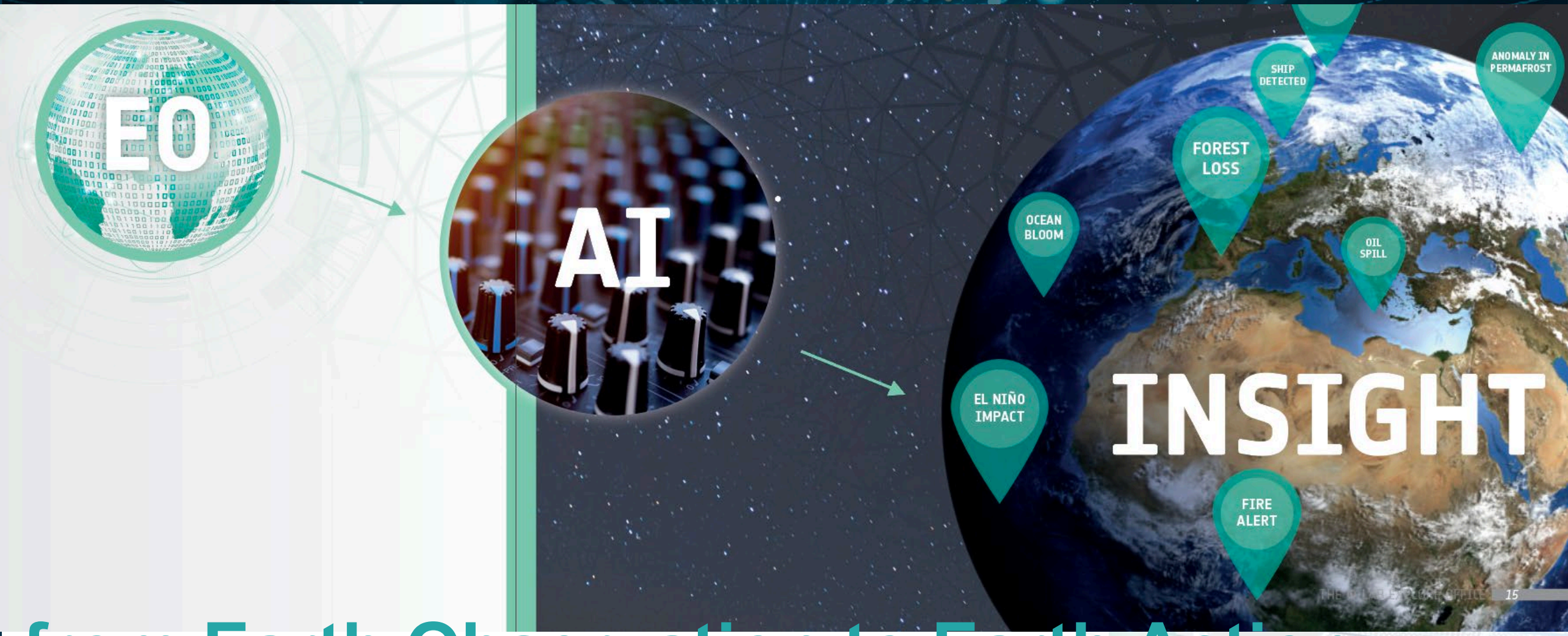
Final Remarks

- Geospatial Foundation Models: hype or the new big thing?
- Importance of benchmarking
- Importance of an open-source project

ESA Φ-lab

The ESA Φ-lab – Why ?

esa Φ-lab



from Earth Observation to Earth Action

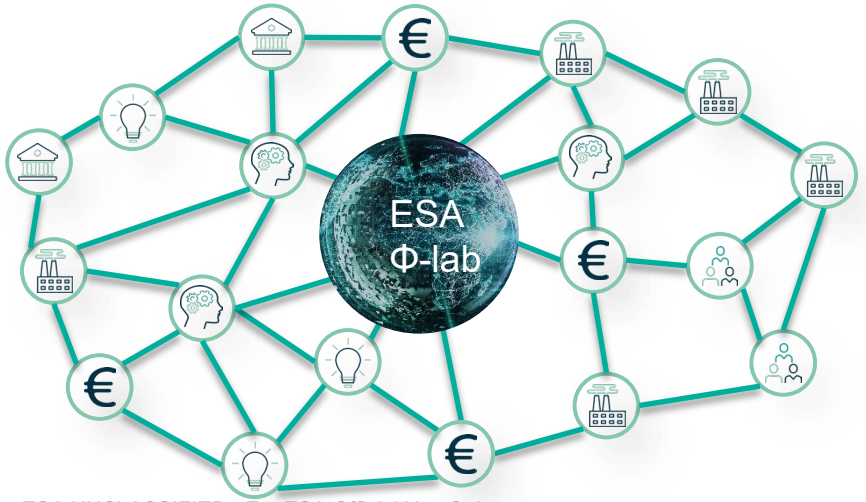
From data to actionable information

The ESA Φ-lab – How?

Φ-lab aims to become "the reference" for the transformational innovation and a key influencer (by reputation and authority) in the Earth Observation ecosystem

Catalyst

- **Attract EO academic and industrial researchers to generate transformative ideas**
- **Exploit fail fast ethos, rapidly prototyping concepts**
- **An informal but rigorous, multi-disciplinary, collaborative environment**
- **Act as facilitator to foster competitiveness growth and entrepreneurial initiatives**
- **Implement investment actions from ESA MSs or in the investors industry**

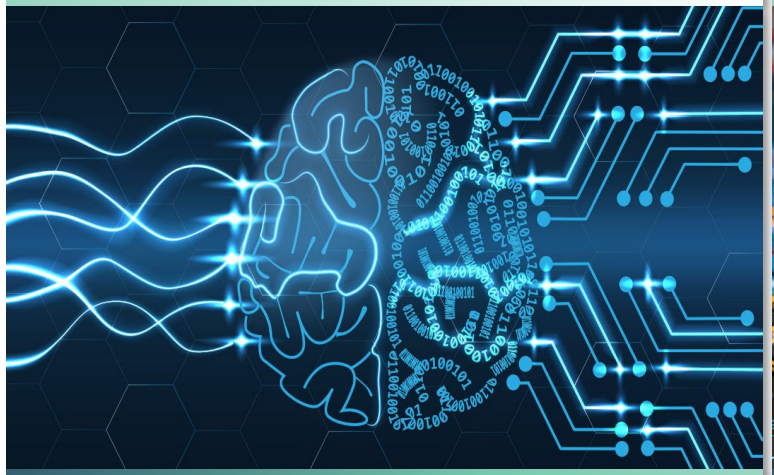


Bridge

- **Be the bridge between the European start-ups, academic and industrial researchers, New Space operators, Investors, ICT players, EO world leaders, and ESA**
- **Act as hub stimulating, connecting, and developing a growing ecosystem of talents and capabilities across Europe**

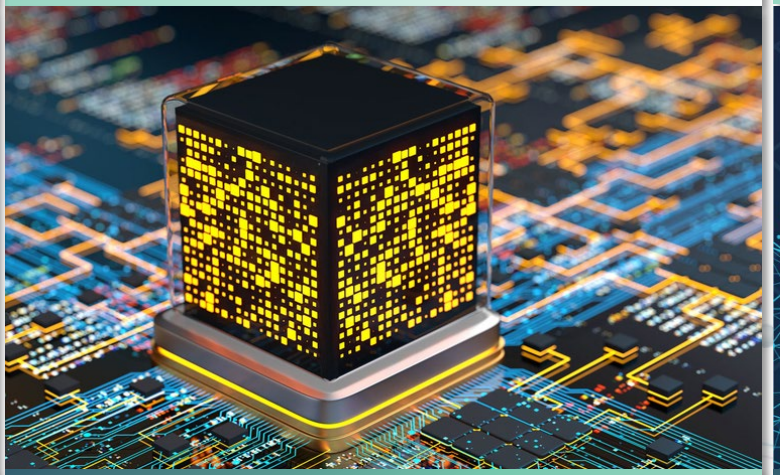
Innovation Technologies axis and Applications

AXIS I
Augmented Intelligence



Foundation Models
Digital Assistant and Twins
Generative AI
Decision Intelligence, Agentic AI
Explainability (xAI)
Physics-Informed ML
AI4EO for Climate, Health, and Human

AXIS II
Innovative Computing Paradigms



On-Board AI
Quantum Computing
Hybrid HPC Computing
Neuromorphic
Biocomputing and others

AXIS III
Innovative Computing Paradigms

Cognitive Space
VR/AR Immersive Visualisation
Web 3.0
IOT
Distributed Ledgers/Blockchain

Collaboration opportunities at Φ-lab

Shared mutual interest

Join the Φ-lab to explore disruptive ideas

as a Visiting Researcher (industry, academia),
Visiting Professor, Research Fellow, PhD, YGT, etc.

Funded

1. Φ-lab's Invitation To Tender on ESA-STARS
   - Foundation Models, Generative AI, QC4EO, Edge computing, Web 3.0, etc..
2. InCubed : partnership development of commercial products or services
3. Open Space Innovation Platform : co-funded research or researchers
4. EO Science4Society : no SOW, 100/200K, 6/18 months
5. ESA Technology Programmes like GSTP and TDE

Working with us it is easy

- Visiting Professor
- Visiting Researchers (Industrial and Scientific)
- ESA Research Fellowships
- ESA Co-funded PhD
- ESA Early Graduate Traineeships (EGT), Internships, National trainee

Join ESA Φ-lab through CIN

The **Collaborative Innovation Network (CIN) by ESA Φ-lab**, provides to leading researchers and University Professors the **opportunity to** join ESA Φ-lab and be actively involved in **accelerating the future of Earth Observation with ESA**.

The ESA Φ-lab CIN aims to:

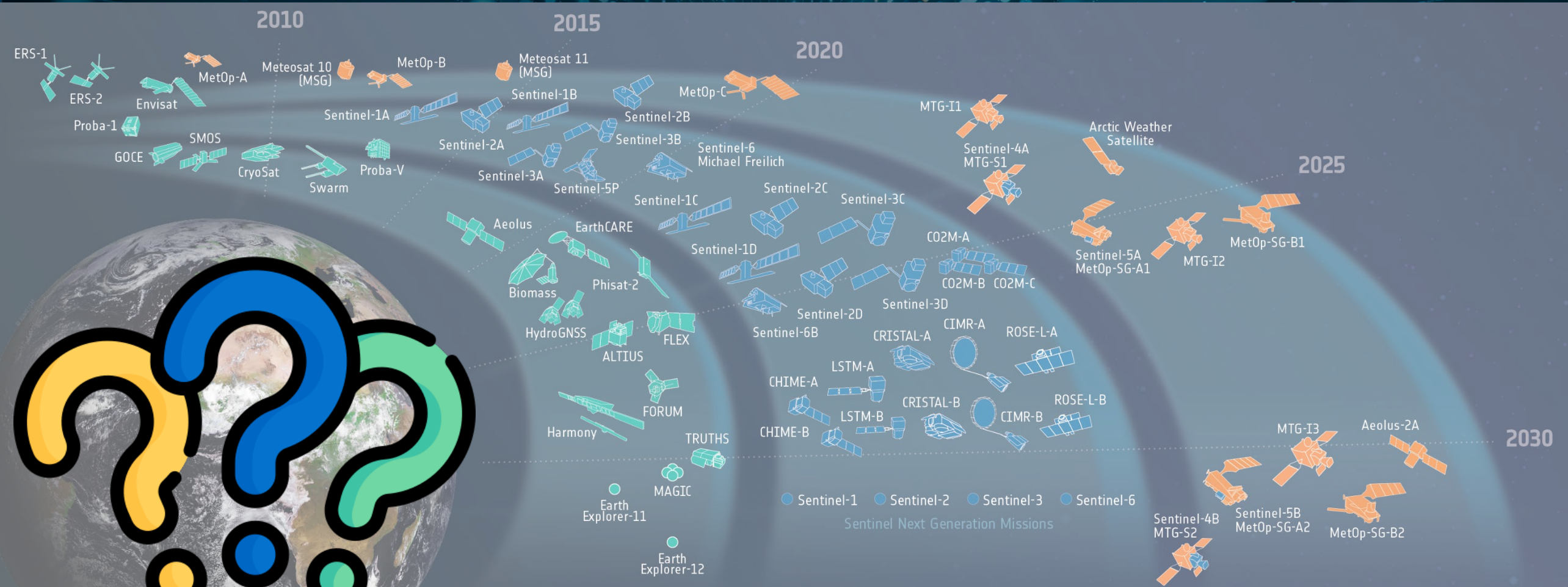
- Establish a global network through which researchers and innovators can **JOIN ESA Φ-LAB**
- **Promote knowledge** sharing and develop groundbreaking EO solutions

Check the open calls

Follow CIN on Linkedin

What to go deeper into Φ-lab disruptive innovation?

Don’t miss our spotlight session at Big Data from Space
(Day 2 – October 2nd at 17:50)



ESA – European Space Agency
Internal Research Fellow, Φ-lab Division
Earth Observation Climate Action, Sustainability and Science Department (EOP-S)
Directorate of Earth Observation Programme
lorenzo.papa@esa.int | www.esa.int | https://philab.esa.int