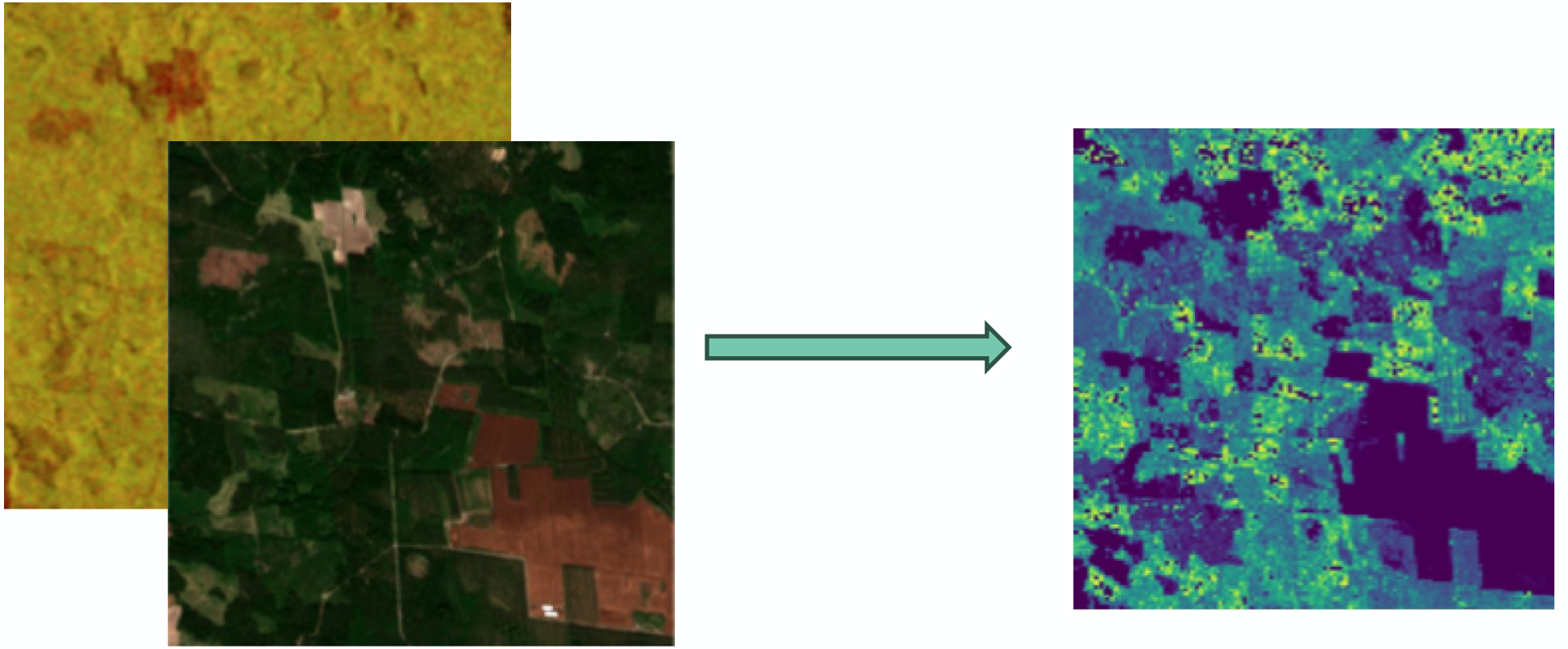
Benchmarking Geospatial Foundation Models

Valerio Marsocci

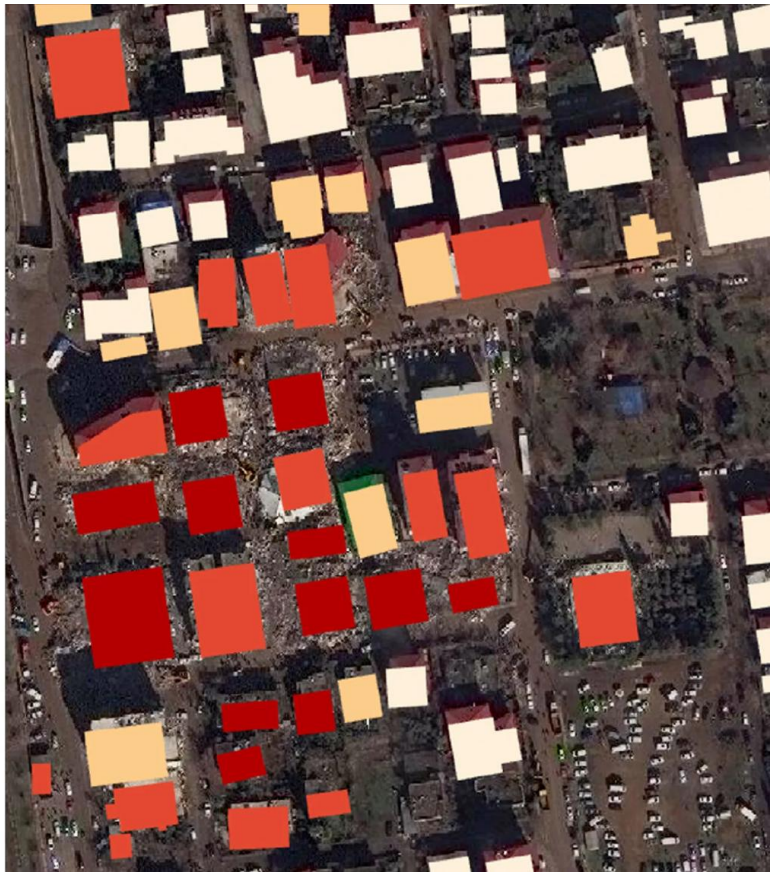
Think of a real-world SDG

Φ-Lab



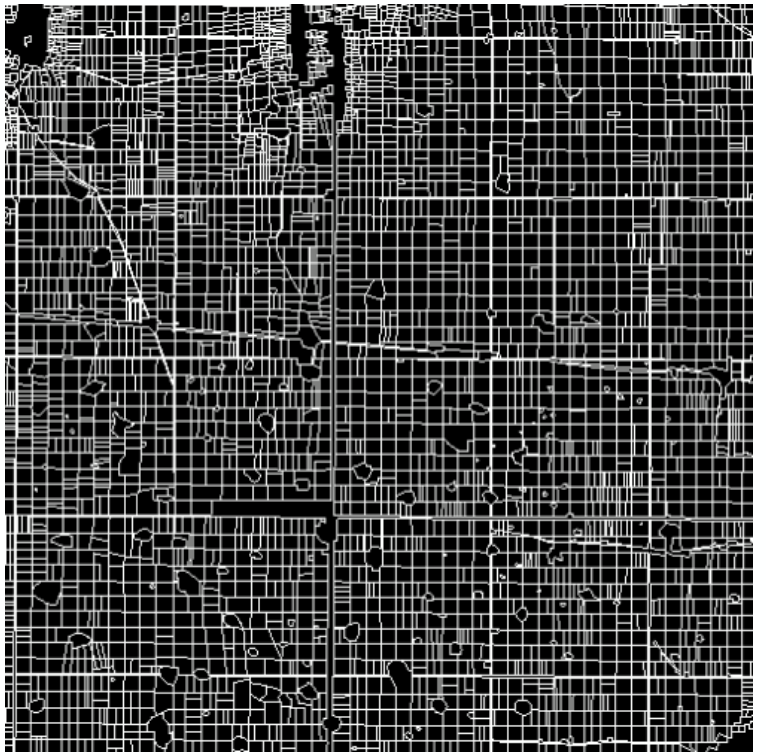
Think of another SDG

Φ-Lab



One more

Φ-Lab

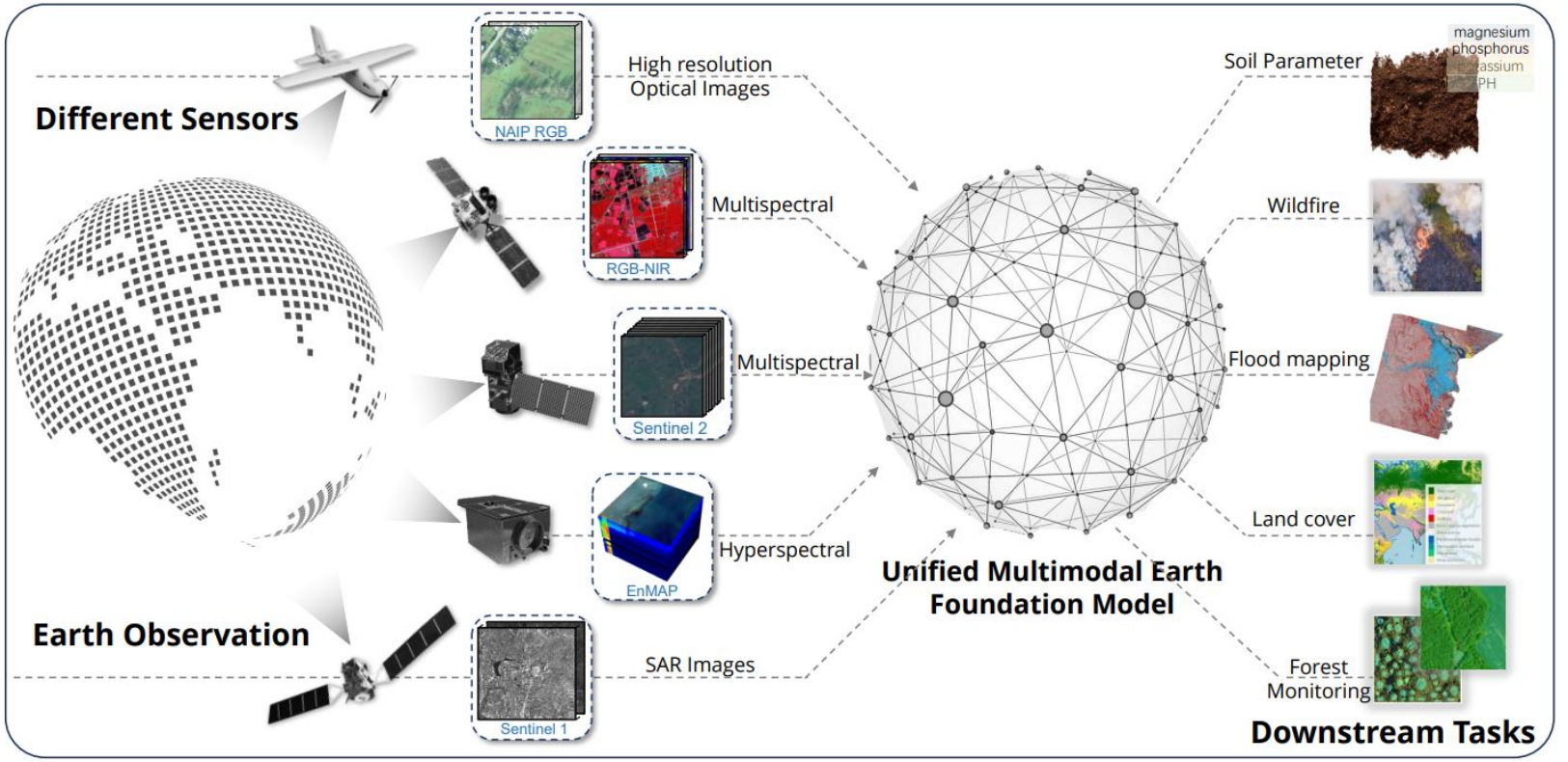


How to solve both tasks?

Φ-Lab

Let's use a
Geospatial Foundation Model

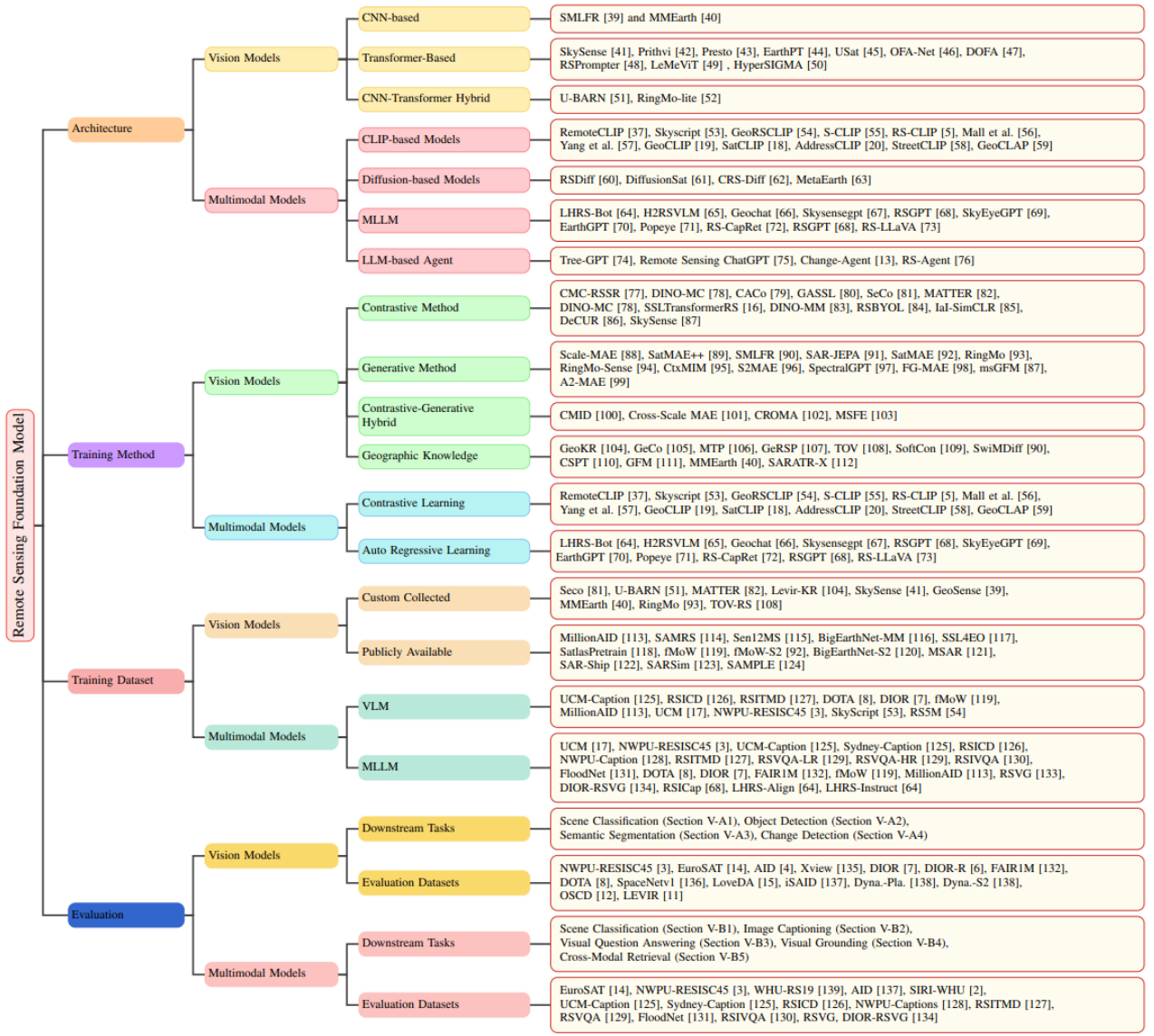
- task agnosticism
- spatio-temporal awareness
- sensor agnosticism
- multimodality
- adaptability



Source: DOFA - [2403.15356](2403.15356)

Which is the best among all available GFMs?

Φ-Lab



Let's benchmark them and find out

Source: [2503.22081]

Why Benchmarking GFMs

Φ-Lab

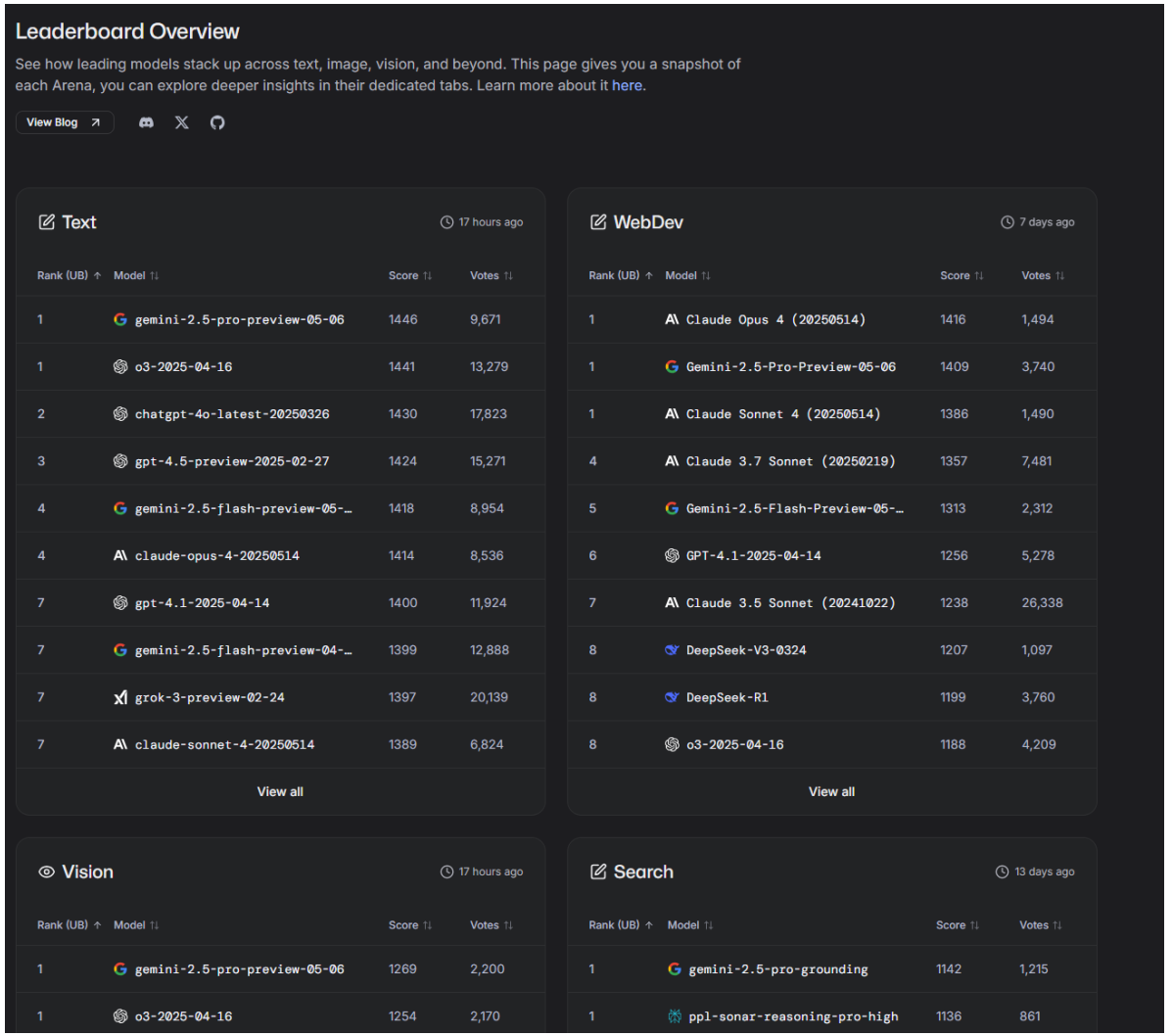
1. Performance Evaluation
2. Fairness & Robustness
3. Guiding Improvements

Research themes:

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

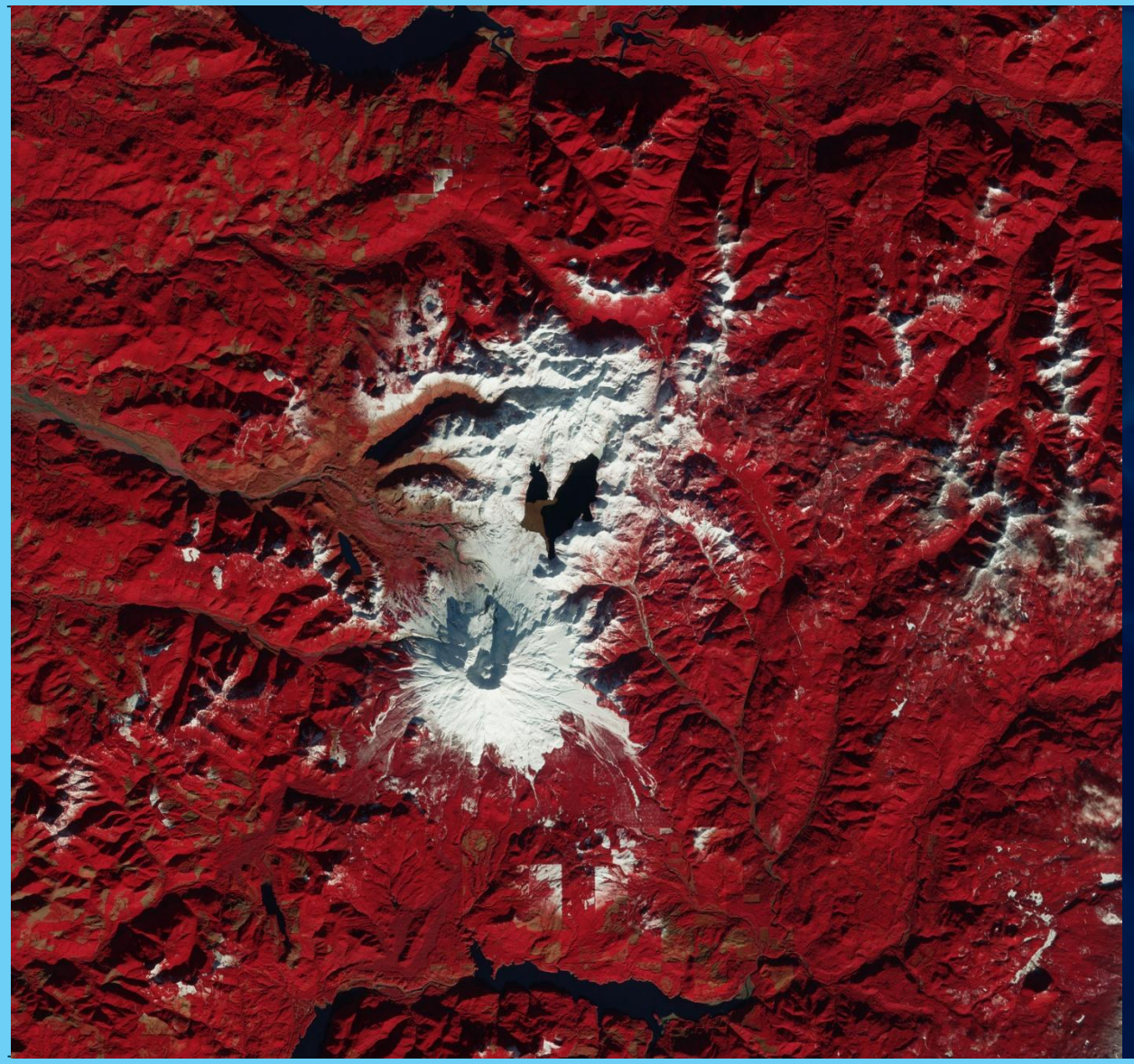
Think of LLMs

Φ-Lab



PANGAEA

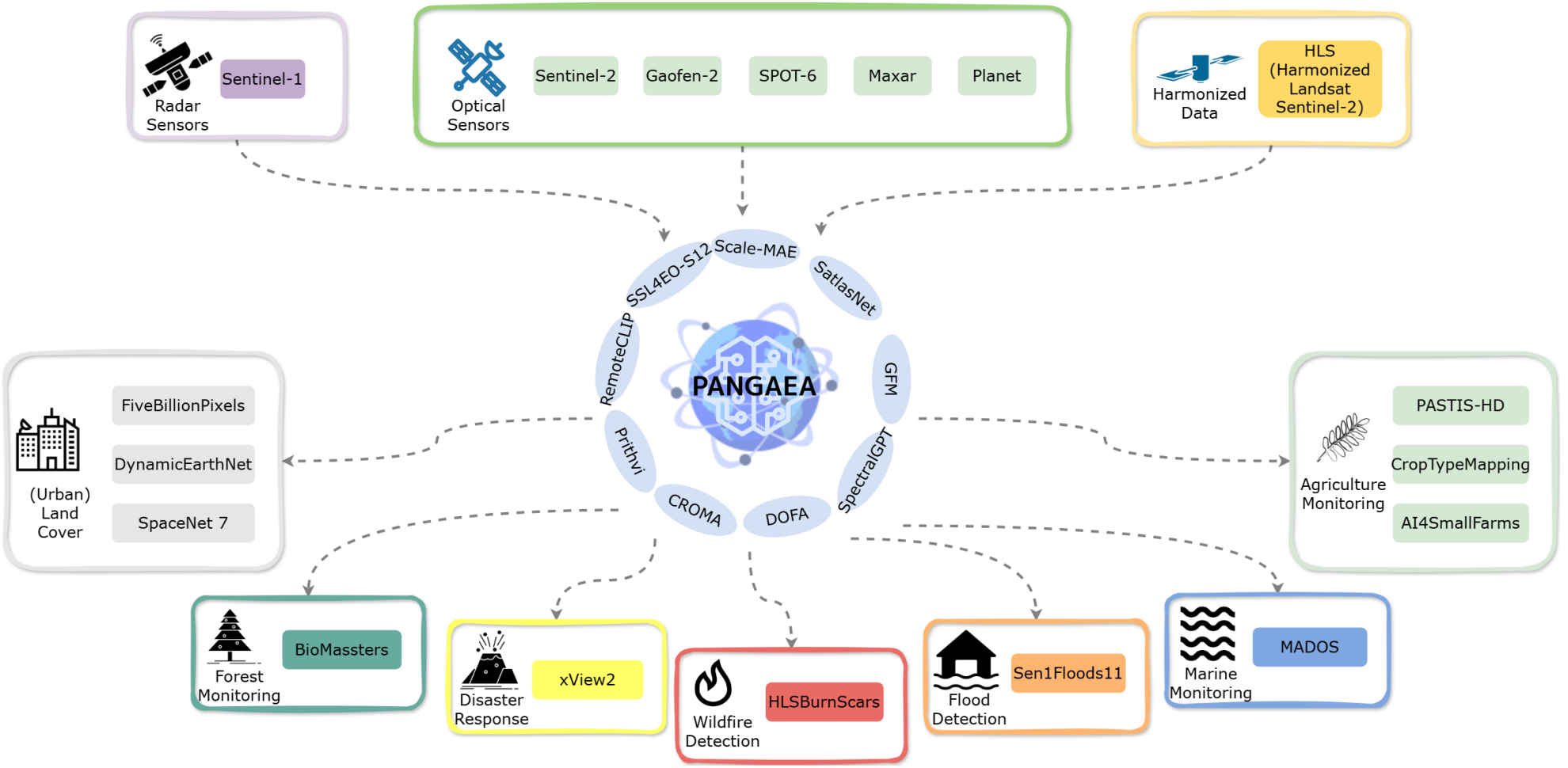
esa Φ-lab



→ THE EUROPEAN SPACE AGENCY

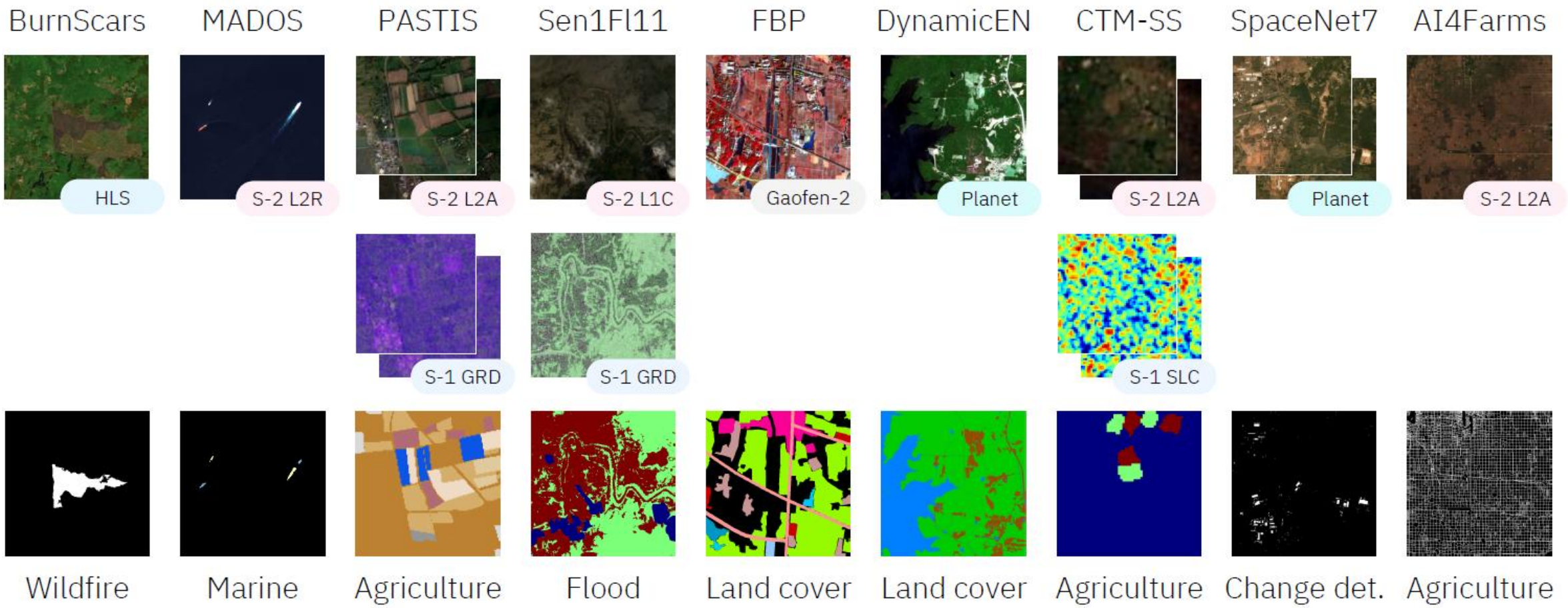
PANGAEA

Φ-Lab



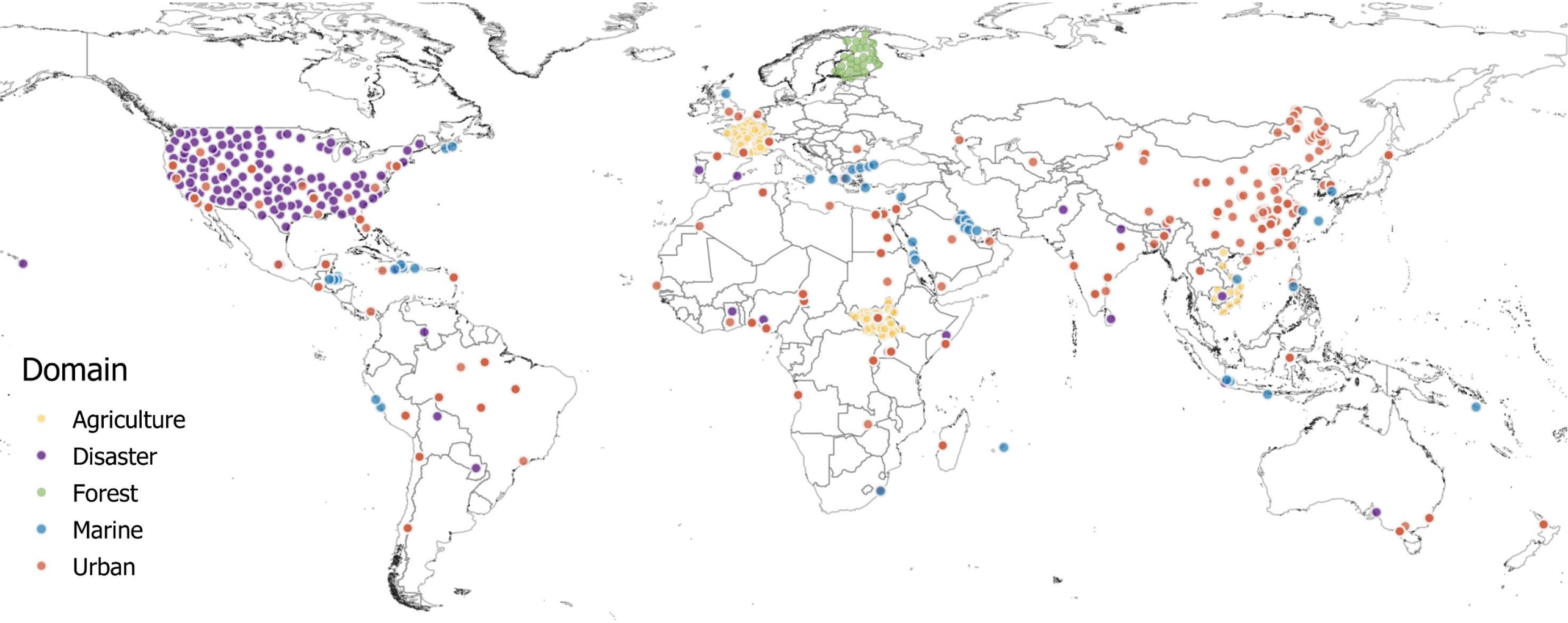
PANGAEA: the datasets

Φ-Lab



PANGAEA: the datasets

Φ-Lab



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

and counting…

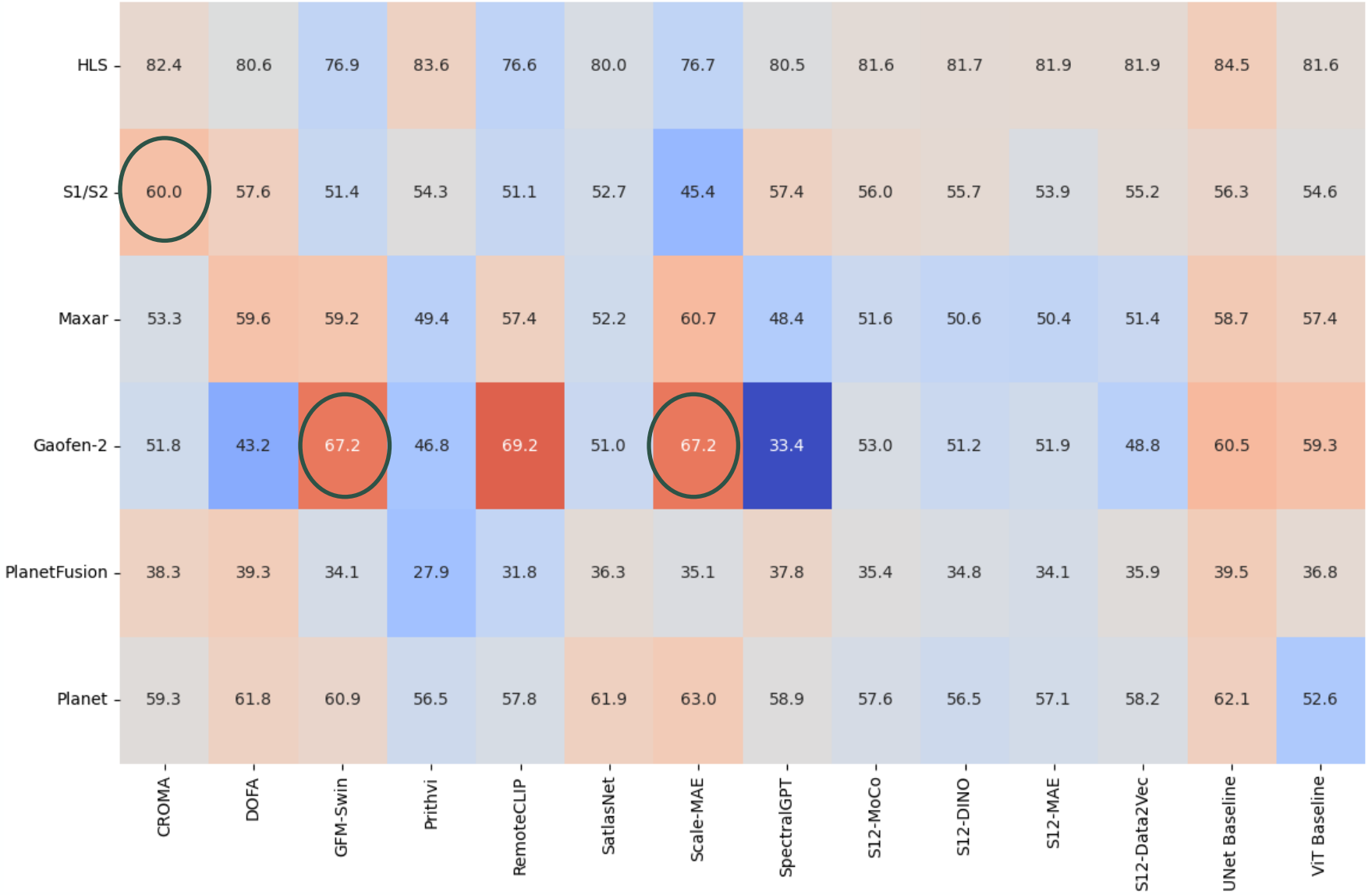
Some results

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

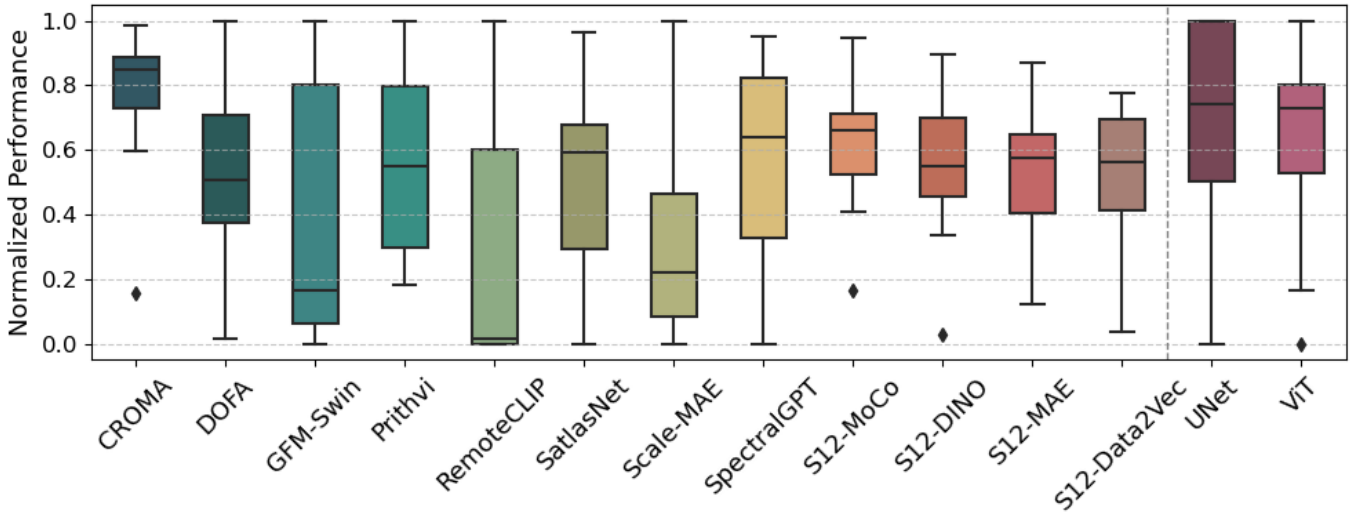
Some results: about resolution

Φ-Lab



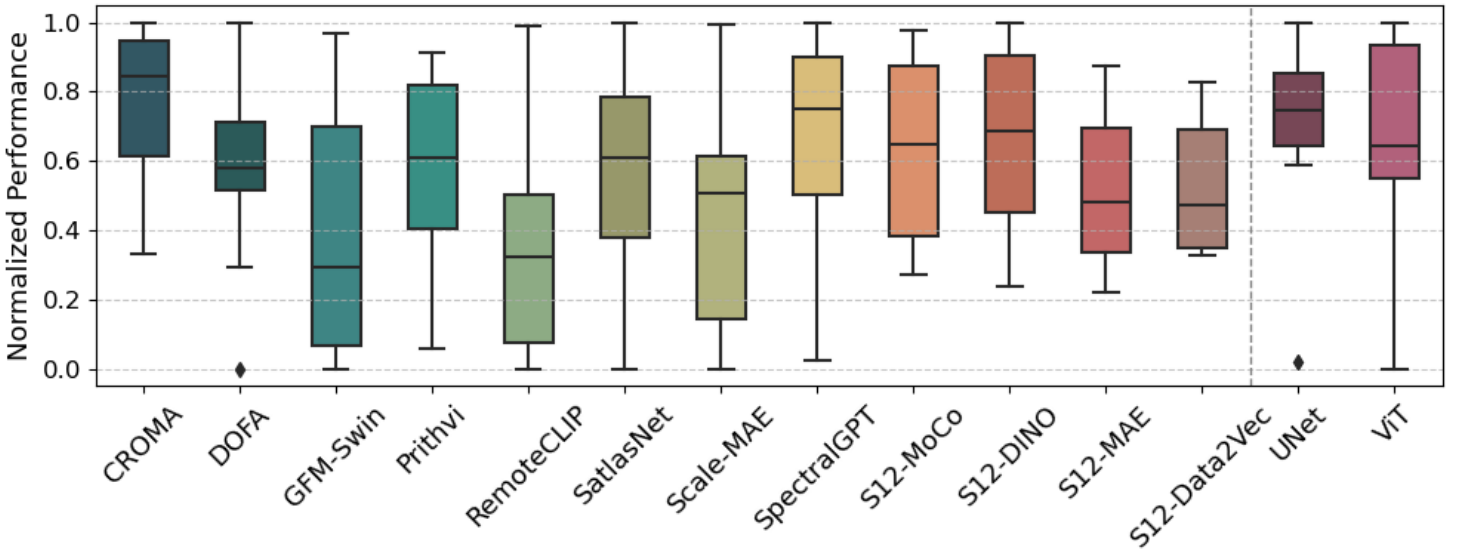
Some results: about data scarcity

Φ-Lab



50% labels

10% labels



The codebase

Φ-Lab

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

README | GPL-3.0 license

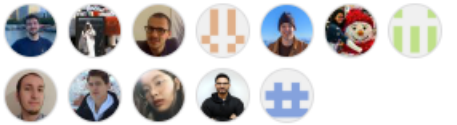
Tests passing

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%

Starred 178

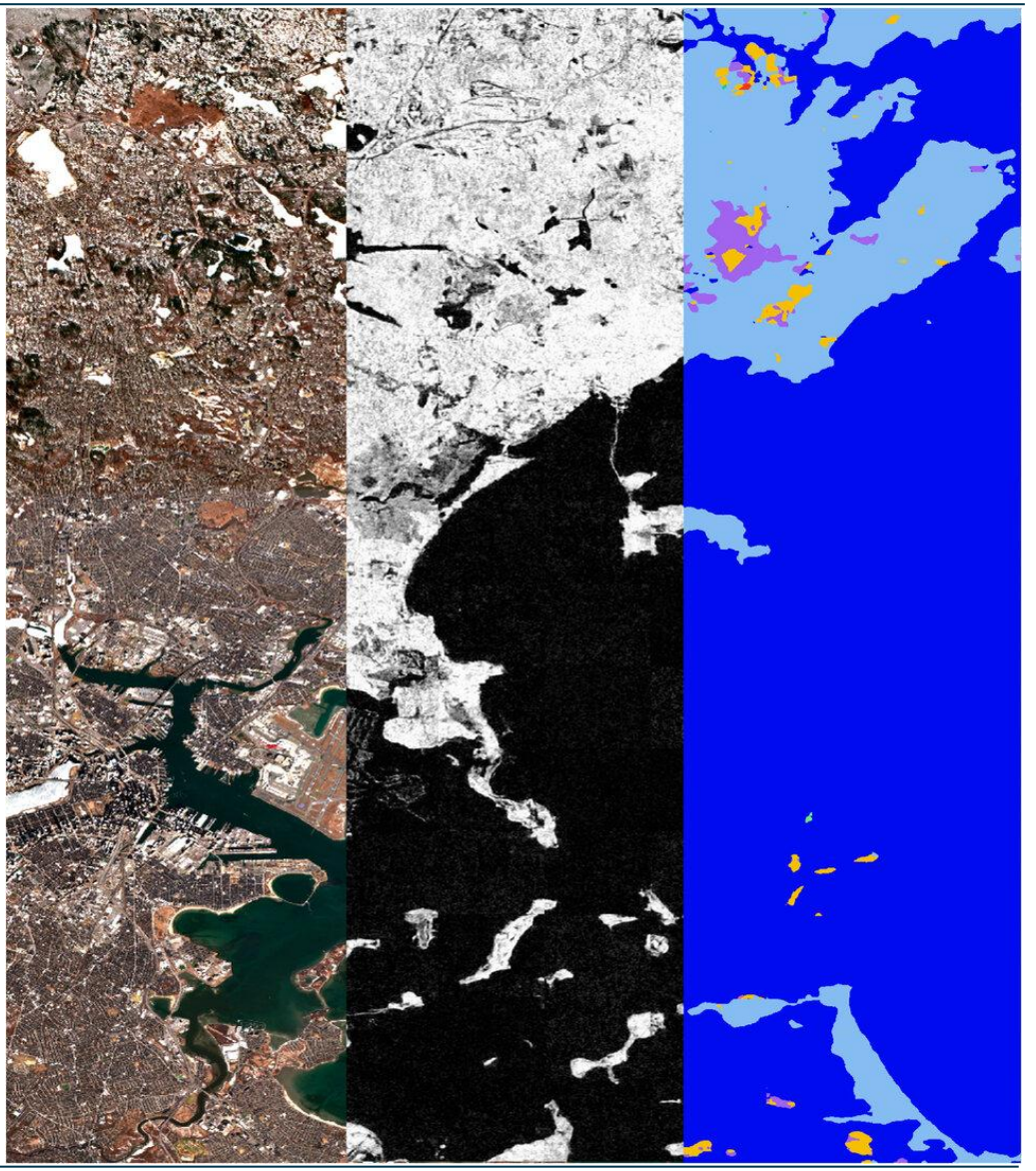
Some nice achievements

Φ-Lab

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

- **TerraMind:**
  Large-Scale Generative Multimodality for Earth Observation

- **Embed2Scale** challenge winner:
  KTH&Friends used PANGAEA for it



Some recent news/next steps

- GEO-Bench datasets included (with **kNN** and **classification**)
- Planning to add a lot of new **models** and **thematic benchmarks**
- Launch cool **challenges** with ITU

Final Remarks

Φ-Lab

- Importance of benchmarking for SDGs
- Importance of an open-source project
- Stay tuned for many exciting next steps

Acknowledgments

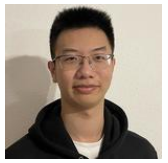
Φ-Lab

Thanks to all the amazing team!

Yuru Jia

Georges Le Bellier

David Kerekes



Liang Zeng

Sebastian Hafner

Sebastian Gerard

Eric Brune

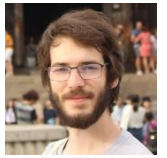
Ritu Yadav

Ali Shibli

Heng Fang

Yifang Ban

Maarten Vergauwen



Nicolas Audebert

Andrea Nascetti

Question time

Happy to chat with you about this!

SCAN ME

valerio.marsocci@esa.int

Me at my wedding

Me benchmarking your GFM