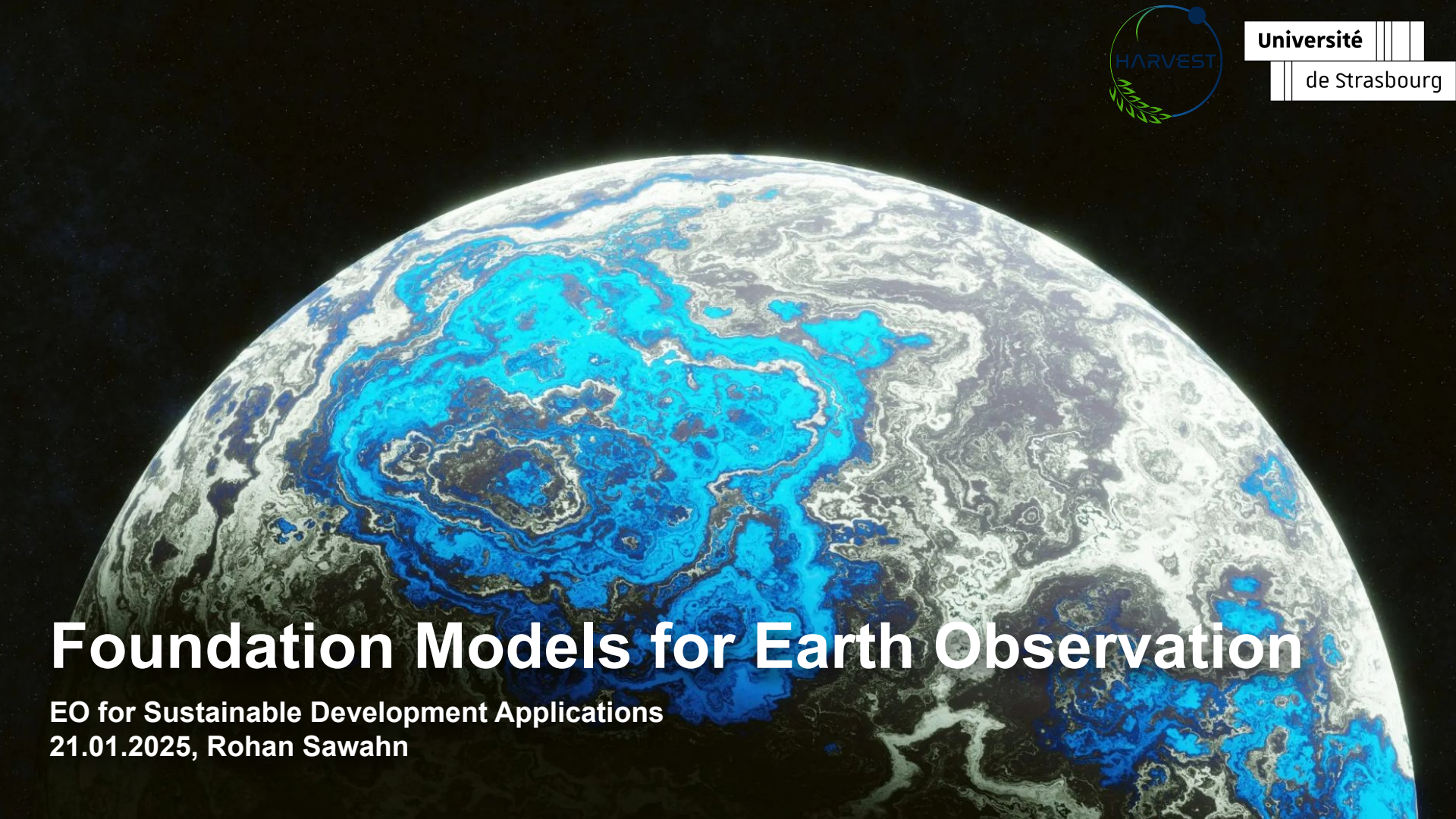


Foundation Models for Earth Observation

EO for Sustainable Development Applications
21.01.2025, Rohan Sawahn

Lecture Overview

Part 1 (50min)

1. Introduction to Foundation Models
2. Foundations Models in EO
3. Understanding Embeddings
4. Training Paradigms for Foundation Models for EO
5. Using EO Foundation Models in Practice

Part 2 (30min)

Hands-On Session: Applying Foundation Models for Croptype classification & Deforestation Detection

Part 3 (20min)

Continue to Work on your Presentations

What is a Foundation Model?

What is a Foundation Model?

*“Foundation Models serve as a **foundation** that has some **general understanding** of the domain and can be applied to **broad range of specialized tasks** without requiring to retrain the model from scratch.”*

What is a Foundation Model?

*“Foundation Models serve as a **foundation** that has some **general understanding** of the domain and can be applied to **broad range of specialized tasks** without requiring to retrain the model from scratch.” ~ a foundation model*

Foundation Models Characteristics

Large-Scale Pretraining

Pretrained on a vast corpus of (unlabelled) data. Typically millions of samples requiring 100s of GPU hours.

Unified Representation

Shared, general purpose representation feature space of the domain (ie. earth) that can be reused for different tasks.

Task Agnostic

Not designed for one specific task, but can easily be adopted to a variety of downstream tasks by the end users.

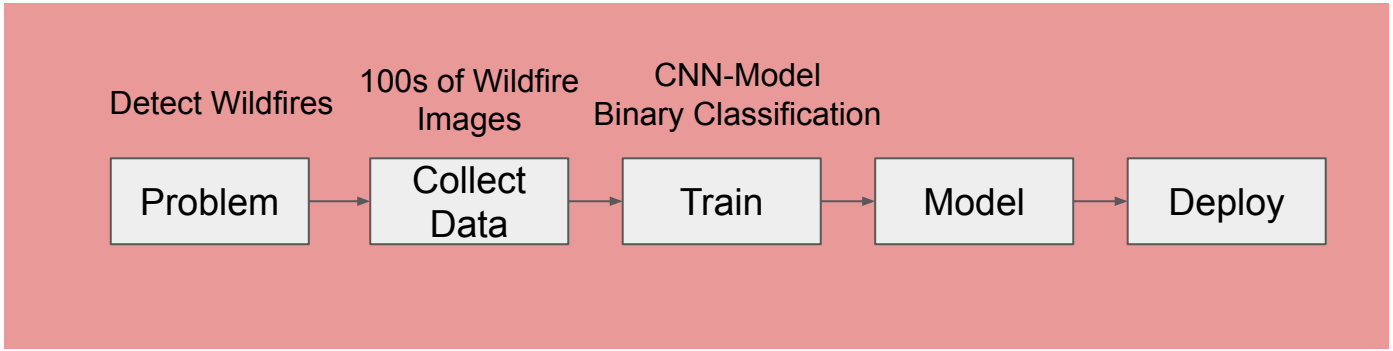
Transfer Learning

Can be fine-tuned for specific downstream tasks using limited data, thus enabling domain adoption.

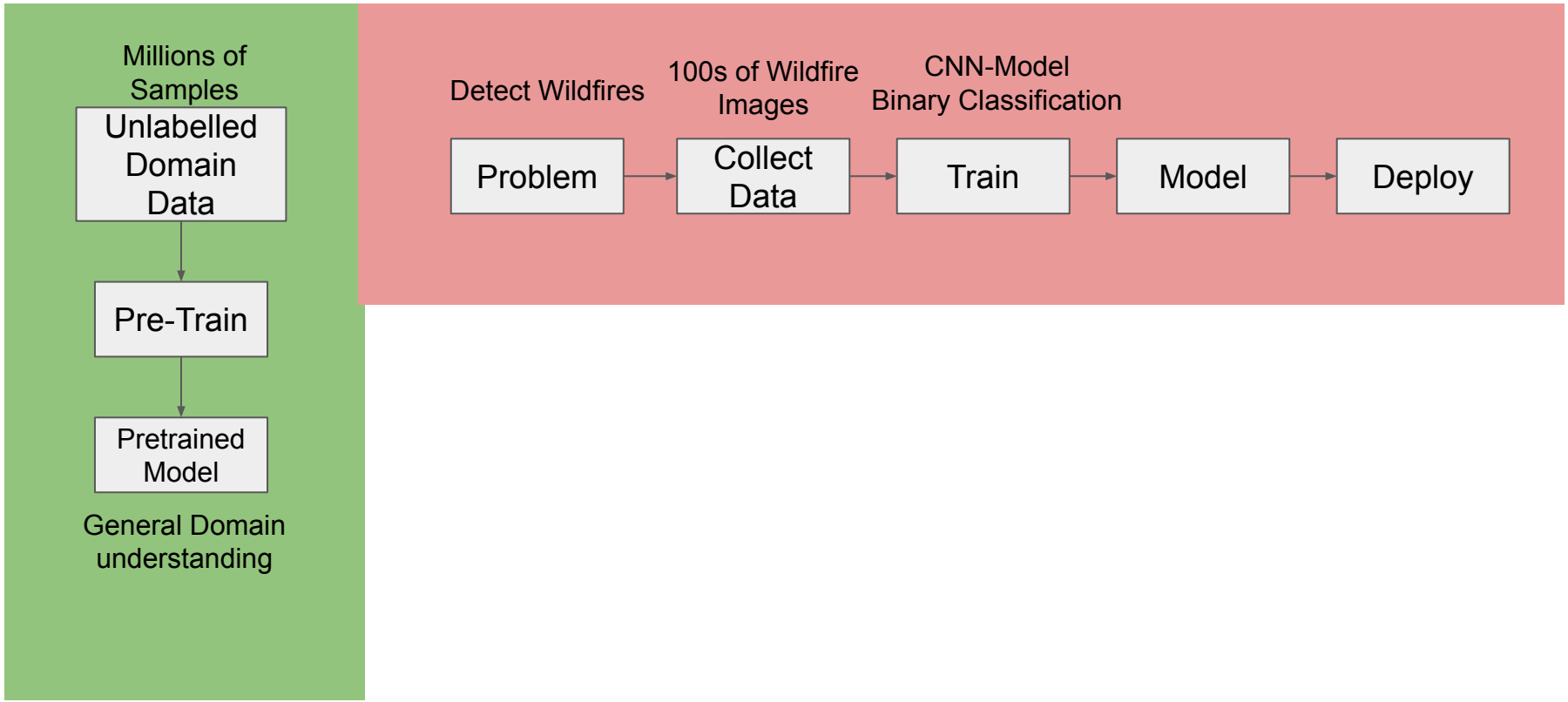
Zero-Shot Capabilities

Can be used without labelled instances with unsupervised methods like clustering etc.

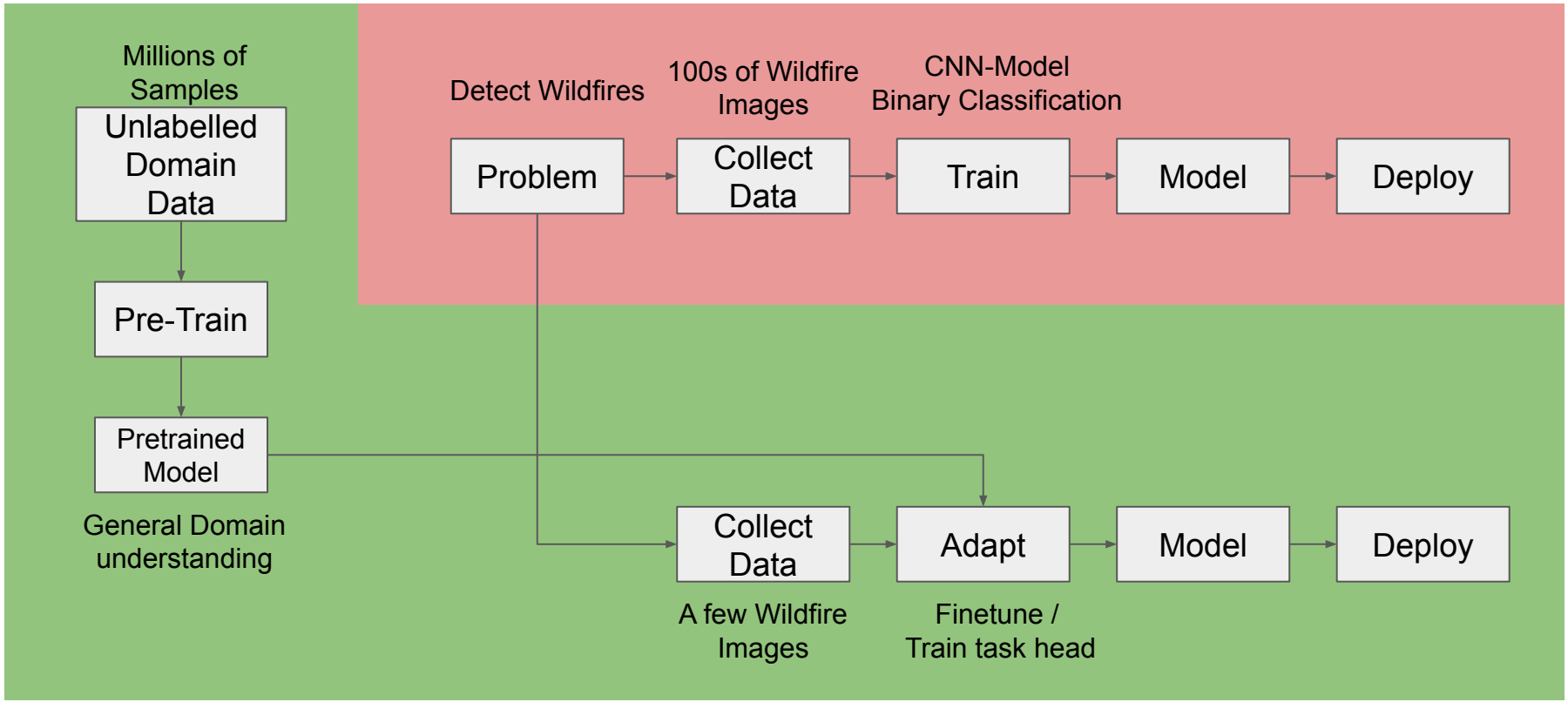
Traditional ML vs Foundation Models



Traditional ML vs Foundation Models



Traditional ML vs Foundation Models



Why Foundation Models in EO?

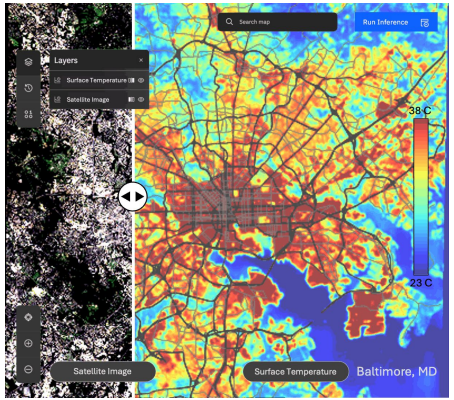
Data Availability
- Petabytes of unlabelled data
- Collection task-specific data is expensive & requires expertise

Common Visual Patterns
- Similar visual patterns for different processes - exploit shared understanding
- Similar environments often exhibit similar patterns

Multimodal Nature of Data
- Optical, SAR, Hyperspectral…
- Climate, Location, Elevation, Descriptions
- Variety of spatial and temporal resolution

Democratization of EO & ML
- Reduces requirements for Compute & Data
- Faster development of solutions for e.g disaster response

Foundation Models in EO Use Cases

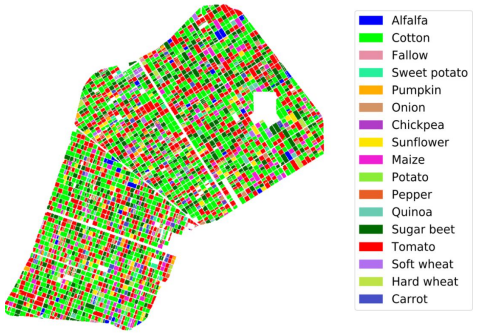


https://research.ibm.com/blog/prithvi2-geospatial

https://www.nytimes.com/2024/09/15/world/africa/floods-africa.html

https://earth.org/amazon-rainforest-deforestation-facts/

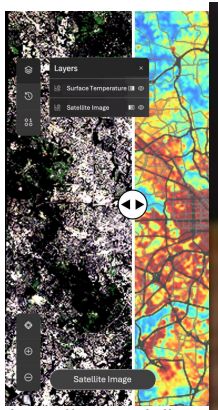
https://www.bbc.com/news/science-environment-64920236



https://www.esa.int/ESA_Multimedia/Images/2021/02/Crop-type_map

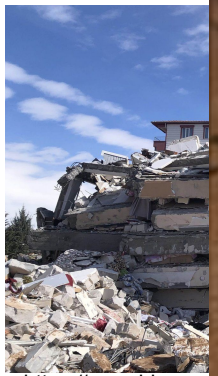
https://airqoon.com/resources/articles/urban-air-pollution-sources-and-pollutants/

Foundation Models in EO Use Cases



https://research.ibm.c

eforestation-facts/



https://www.bbc.com

s/urban-air-pollution-sour

Overview of Foundation Models in EO

Model Name	Architecture	Pre-training Dataset	Resolution (m)	Geographic Coverage	Image Analysis Levels	Pretrain methods	# of Params
CMC-RSSR [84]	ResNet-50	SWPU-DOTA [114], BigEarthNet [85], ImageNet [18]	0.2 to 60	Global	Image-level	Contrastive Multiview Coding	23M
SeCo [66]	ResNet-50	Sentinel-2 Imagery	10, 20, 60	200k Locations Worldwide	Image-level, Spacial-temporal	CL	23.5M
GeoKR [58]	ResNet-50	Levir-KR [56]	0.8 to 16	Global	Image-level, Pixel-level, Region-level	Geographical Knowledge Supervision	23.5M/138M
MATTER [2]	ResNet-34	Sentinel-2 Imagery	-	Rural and Remote Regions with Little Changes	Image-level, Pixel-level	SSL	21.3M
GASSL [6]	ResNet-50	fMoW [13], GeoImageNet [18]	-	7 Continents	Image-level, Pixel-level Region-level	CL	23.5M
RSP [96]	ViTAE-2-S	MillionAID [63], [64]	0.5 to 153	Global	Image-level, Pixel-level, Region-level, Spacial-temporal	Supervised Learning	24.8M/23.5M/29M
DINO-MM [105]	ViT-S/8	BigEarthNet-MM [96]	10	Global	Image-level	SSL	22M
Scheibenreif, et al. [79]	Swin Transformer	SEN12MS [80]	10	Global	Image-level, Pixel-level	CL	-
RingMo [87]	ViT/Swin Transformer	2 million RS images	0.3 to 30	6 Continents	Image-level, Pixel-level, Region-level, Spacial-temporal	MIM	-
GeCo [57]	ResNet-50	Levir-KR [56]	0.8 to 16	Global	Image-level, Pixel-level, Region-level	SSL	23.5M
RS-BYOL [45]	BYOL	Sen12MS [80]	10 to 20	Global	Image-level, Pixel-level	SSL	23.5M
CSPT [124]	ViT-B	ImageNet-1K [18]	-	Global	Image-level, Region-level	SSL	86M
RVSA [100]	ViT	MillionAID [63], [64]	0.5 to 153	Global	Image-level, Pixel-level, Region-level	MAE	100M
SatMAE [16]	MAE-based Framework	fMoW Sentinel-2 [13]	10, 20, 60	Global	Image-level, Pixel-level	MAE	307M
TOV [89]	TOV	TOV-NL TOV-RS	-	Global	Image-level, Pixel-level, Region-level	SSL	-
CMID [70]	Teacher-student Self-distillation	MillionAID [63], [64]	Varied	Global	Image-level, Pixel-level, Region-level, Spacial-temporal	SSL	25.6M/87.8M
CACo [67]	ResNet-18/50	Sentinel-2 Imagery	10	Global	Image-level, Pixel-level, Spacial-temporal	SSL	11.7M/23.5M
Ial-SimCLR [77]	ResNet-18	SEN12MS ImageNet [18],	-	Global	Image-level	CL	11.7M
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Model Name	Architecture	Pre-training Dataset	Resolution (m)	Geographic Coverage	Image Analysis Levels	Pretrain methods	# of Params
GFM [69]	Teacher-Student	GeoPile [69]	-	Global	Image-level, Pixel-level	Continual Pretraining	-
SatlasPretrain [7]	SatlasNet	GeoPile [69]	1, 10	Global	Image-level, Pixel-level	Multi-task Learning	88M
RingMo-Sense [119]	Multi-Branch	RS Spatiotemporal Dataset	-	Global	Pixel-level	SSL	-
Scale-MAE [76]	ViT-Large	fMoW [13]	-	Global	Image-level, Pixel-level	MAE	322.9M
RingMo-lite [109]	CNN-Transformer	AID [115]	0.3 to 30	Global	Image-level, Pixel-level, Region-level, Spacial-temporal	FD-MIM	60% less than RingMo
DeCUR [102]	Multimodel SSL	SSL4EO-S12 [107], RGB-DEM/depth	Varied	Global	Image-level, Pixel-level	SSL	23.5M
Feng et al. [27]	MSFEs+MMFH	Multi-modal Dataset	Varied	Global	Image-level, Pixel-level, Region-level, Spacial-temporal	SSL	-
FG-MAE [108]	ViT	SSL4EO-S12 [107]	10	Global	Image-level, Pixel-level	MAE	-
Prithvi [40]	ViT	Harmonized Landsat Sentinel 2	30	Contiguous U.S.	Pixel-level	MAE	100M
CROMA [28]	Multimodal Encoder	SSL4EO [107]	10	Area Surrounding Human Settlements	Image-level, Pixel-level	CL, MAE	86M
USat [40]	ViT	Satlas [7]	Varied	Global	Pixel-level	MAE	-
Cross-Scale MAE [88]	ViT-B	fMoW [13]	-	Global	Image-level, Pixel-level	MAE	86M
U-BARN [26]	Unet+Transformer	Sentinel-2 Imagery	Varied	France	Pixel-level	SSL	-
EarthPT [82]	Transformer	Sentinel-2 Imagery	10	UK	Image-level	Autoregressive SSL	700M
GeRSP [42]	Teacher-Student Network	ImageNet [18], MillionAID [63], [64]	0.5 to 153	Global	Image-level, Pixel-level, Region-level	SSL, SL	-
SwiMDiff [91]	Dual-Branch	Sen12MS [80]	Varied	Global	Image-level, Spacial-temporal	SSL	11.7M
SMLFR [22]	Generative ConvNet	GeoSense [23]	0.05 to 150	Multiple Continents	Pixel-level, Region-level	SSL	88M/197M
SpectralGPT [40]	3D GPT	Sentinel-2 Imagery	Varied	Global	Image-level, Pixel-level, Spacial-temporal	MAE	100M/300M/600M
Presto [92]	MAE-based Framework	Presto-21.5M [92]	10	Global	Crop Type Segmentation	MAE	402K
SatMAE++ [72]	SatMAE	fMoW [13]	Varied	Global	Image-level	Multi-Scale Pre-training	-

Model Name	Architecture	Pre-training Dataset	Resolution (m)	Geographic Coverage	Image Analysis Levels	Pretrain methods	# of Params
FoMo-Bench [8]	ViT	Multiple	Varied	Global	Image-level, Pixel-level, Region-level	MAE	101M/110M
SkySense [32]	Factorized Multi-Modal Spatiotemporal Encoder	Multiple	Varied	Global	Image-level, Pixel-level, Region-level, Spacial-temporal	CL	2.06B
UPetu [14]	Multi-Modules	GeoSense [23]	-	Global	Image-level, Pixel-level, Spacial-temporal	SSL	0.65M
msGFM [33]	Swin Transformer	GeoPile-2 [69]	0.1 to 153	Global	Image-level, Pixel-level	MIM	89M
DINO-MC [111]	DINO	SoCo-100K [66]	10 to 60	Global	Image-level, Spacial-temporal	SSL	-
OFA-Net [118]	OFA-Net	Multi-modal Dataset	Varied	Global	Image-level, Pixel-level	MIM	-
MTP [99]	Shared Encoder Task-Specific Decoders	SAMRS [86]	Varied	Global	Image-level, Pixel-level, Region-level, Spacial-temporal	Multi-Task Pretraining	over 300M
BFM [11]	ViT	MillionAID [63], [64]	0.5 to 153	Global	Pixel-level, Region-level	MAE	86M/605.26M/1.36B/2.42B
MMEarth [72]	MP-MAE	Multi-modal, Geospatial Data	-	Global	Image-level, Pixel-level	MP-MAE	3.7M to 650M
CtxMIM [123]	ViT	WorldView-3 Imagery	Varied	Asia	Image-level, Pixel-level, Region-level	MIM	88M
SARATR-X [54]	HiViT	SAR Datasets	0.1 to 3	Global	Image-level, Region-level	MIM	66M
SoftCon [106]	Siamese Network with ResNet and ViT Backbones	SSL4EO-S12-ML [107]	-	Global	Image-level, Pixel-level, Spacial-temporal	Multi-label Soft Contrastive Learning	23M, 25M, 86M
LuMeViT [69]	Hierarchical ViT	MillionAID [63], [64]	-	-	Image-level, Pixel-level, Region-level , Spacial-temporal	Dual Cross-Attention with Learnable Meta Tokens Adaptation	15M/[illegible]
S2MAE [59]	3D Transformer-based MAE	fMoW-Sentinel [13], BigEarthNet [85]	-	Global	Image-level, Spacial-temporal	3D MAE	-
RS-DFM [110]	Multi-platform Inference Framework	AirCo-MultiTasks [110]	-	-	3D Region-level, Pixel-level	Generalized Feature Mapping with Relative Depth Estimation	-
A2-MAE [122]	ViT-Large	STSSD (Spatial-Temporal-Spectral Structured Dataset)	0.8 - 30m	Global	Image-level, Pixel-level, Spacial-temporal	Anchor-aware Masking Strategy and Geographic Encoding Module	-
HyperSIGMA [93]	ViT-based	HyperGlobal-450K [93]	30m	Global	Image-level, Region-level, Anomaly Detection, Spacial-temporal	MAE	over 1B

Lu et al, 2025, Vision Foundation Models in Remote Sensing: A survey

AlphaEarth

Presto/Galileo

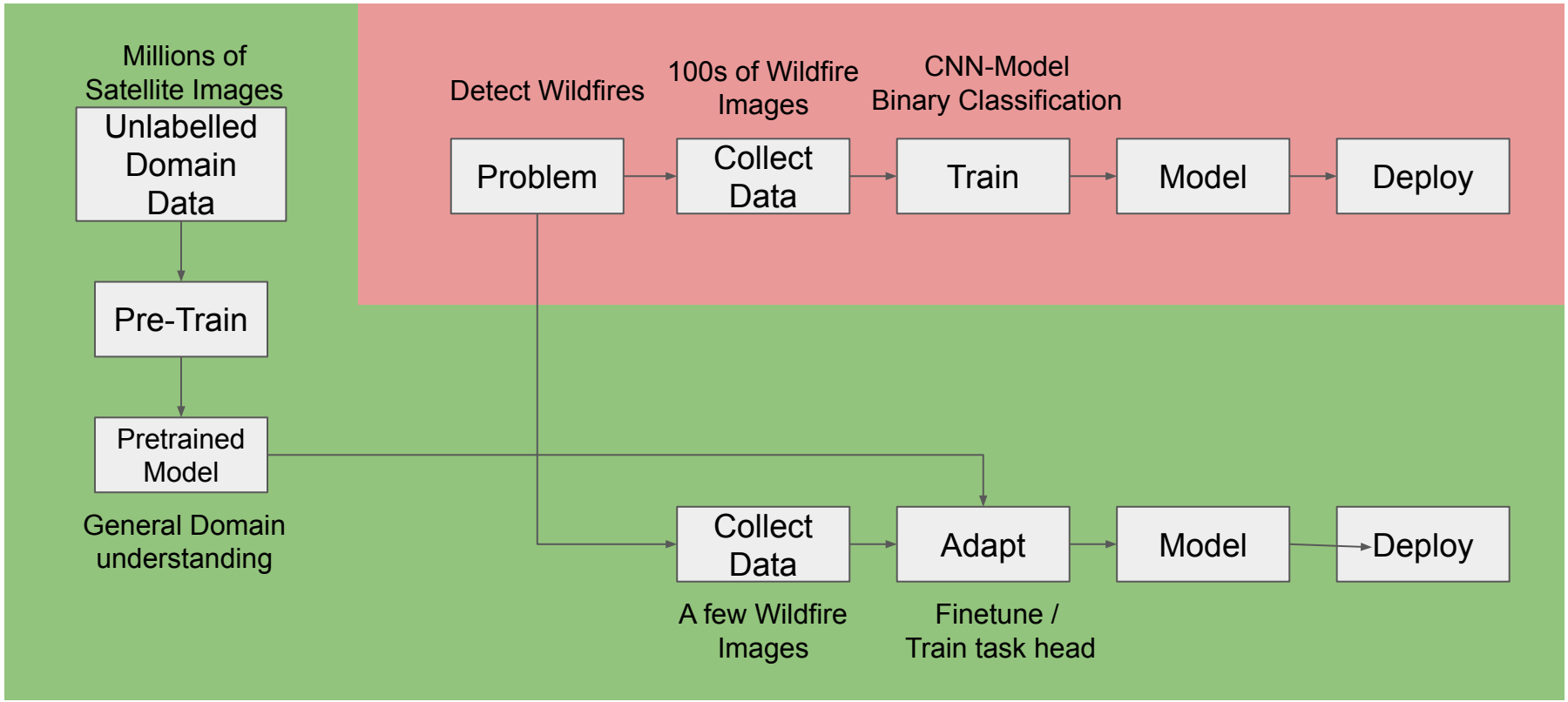
Prithvi

MajorTom

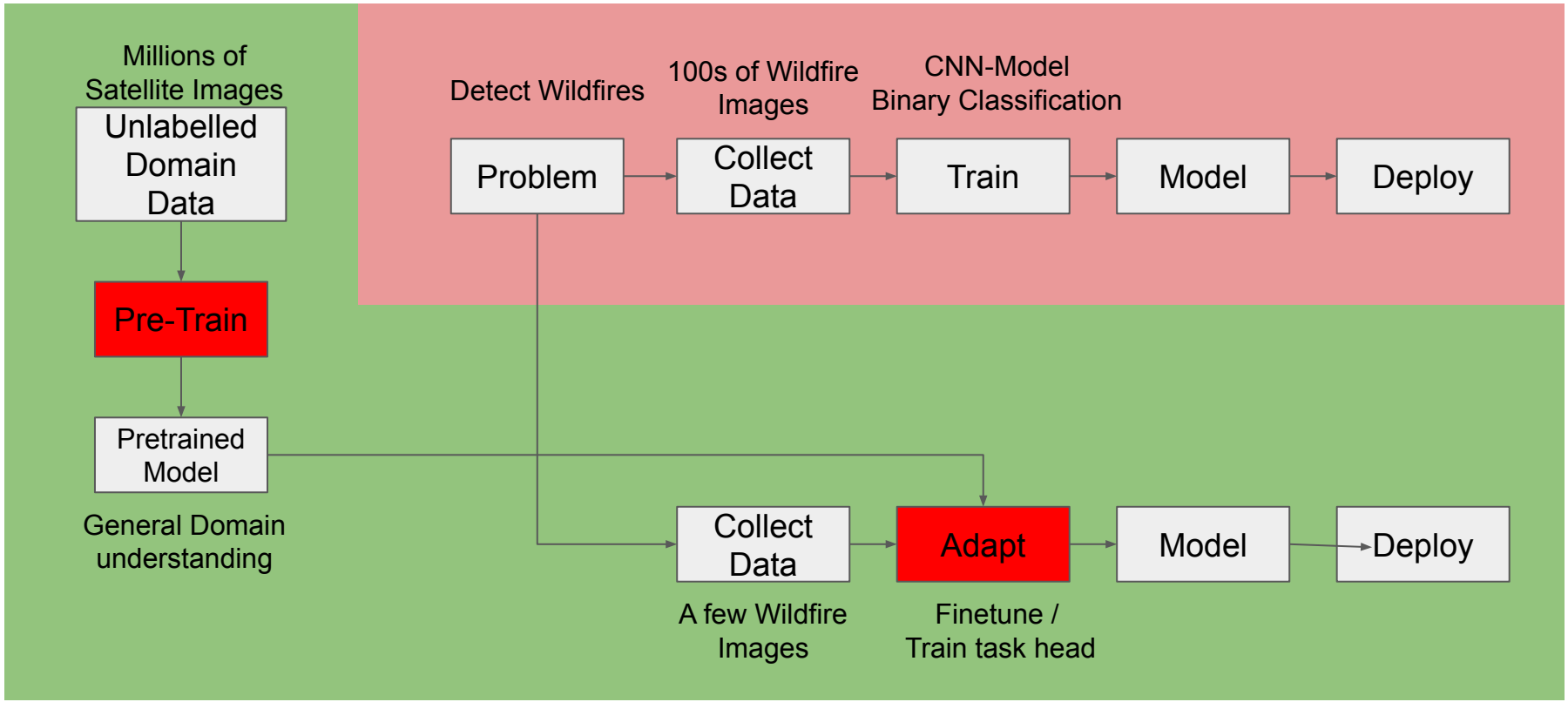
Tessera

OlmoEarth

Traditional ML vs Foundation Models

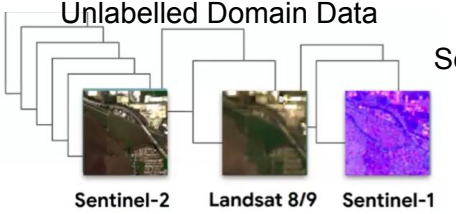


Traditional ML vs Foundation Models



Foundation Models in EO

Global Dataset
Unlabelled Domain Data



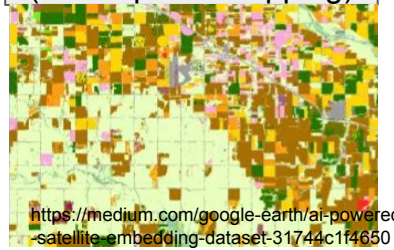
Optical, Hyperspectral, SAR, Precipitation, Climate, Elevation, Location, Text

Self Supervised Learning

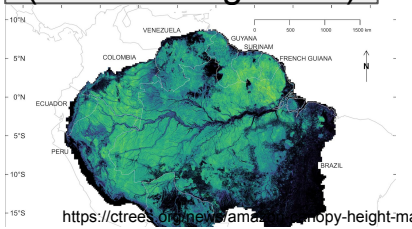
Pretrained Foundation Model

Embedding
(Latent Representation)

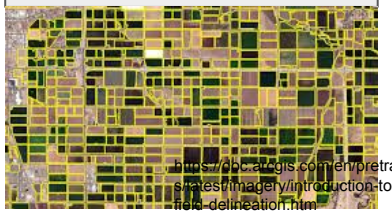
Classifier
(ie. Cropland mapping)



Regressor
(i.e Tree height est.)



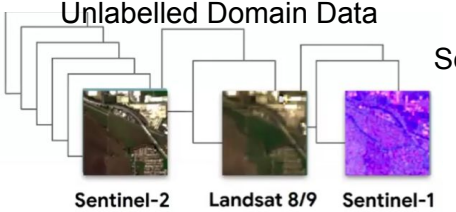
Segmentation



Images from Valerie Pasquarella,
AlphaEarth Foundations and the Satellite Embedding Dataset: A New Foundation for Geospatial Analysis, IEEE GRSS

Foundation Models in EO

Global Dataset
Unlabelled Domain Data



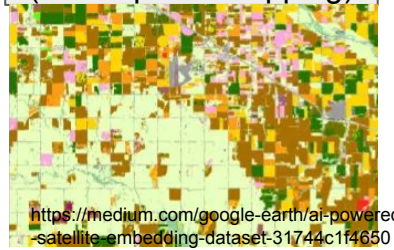
Optical, Hyperspectral, SAR, Precipitation, Climate, Elevation, Location, Text

Self Supervised Learning

Pretrained Foundation Model

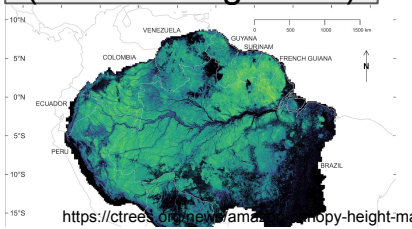
Embedding
(Latent Representation)

Classifier
(ie. Cropland mapping)



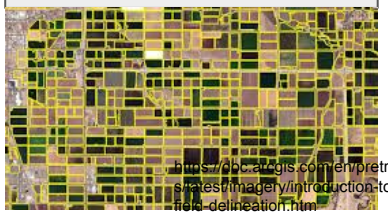
https://medium.com/google-earth/ai-powered-pixels-i
-satellite-embedding-dataset-31744c1f4650

Regressor
(i.e Tree height est.)



https://ctrees.org/new/amaz opy-height-map

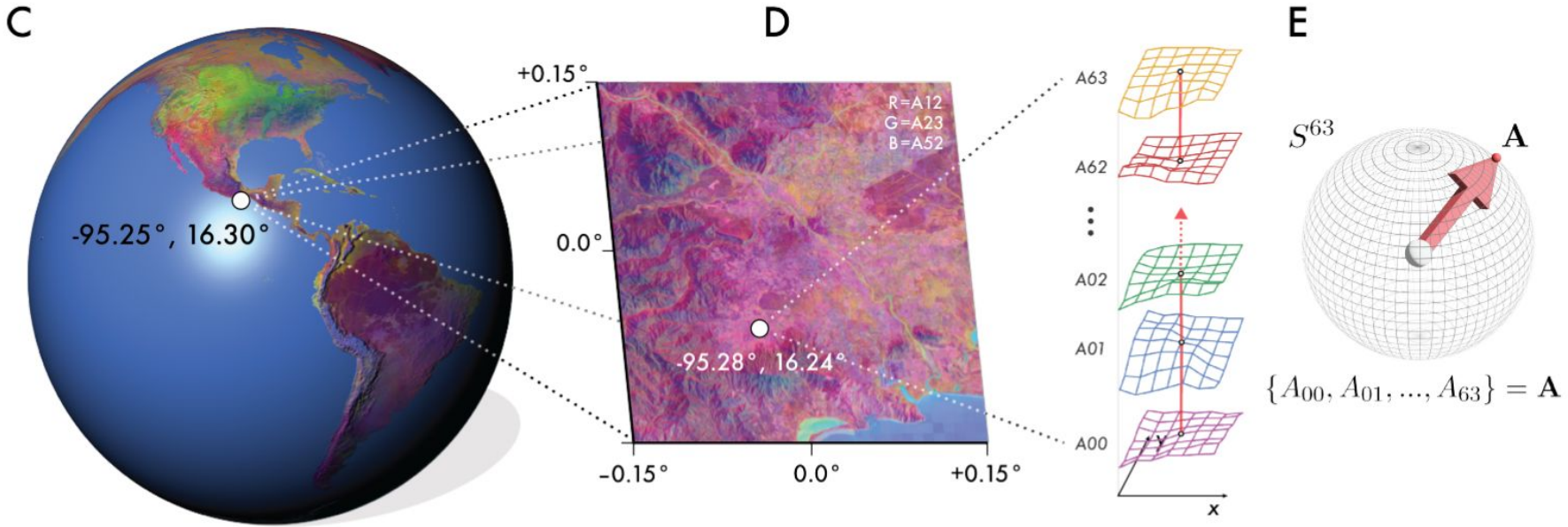
Segmentation



https://doc.arcgis.com/en/pretrained-mod
sources/imagery/introduction-to-agricultura
-field-delineation.htm

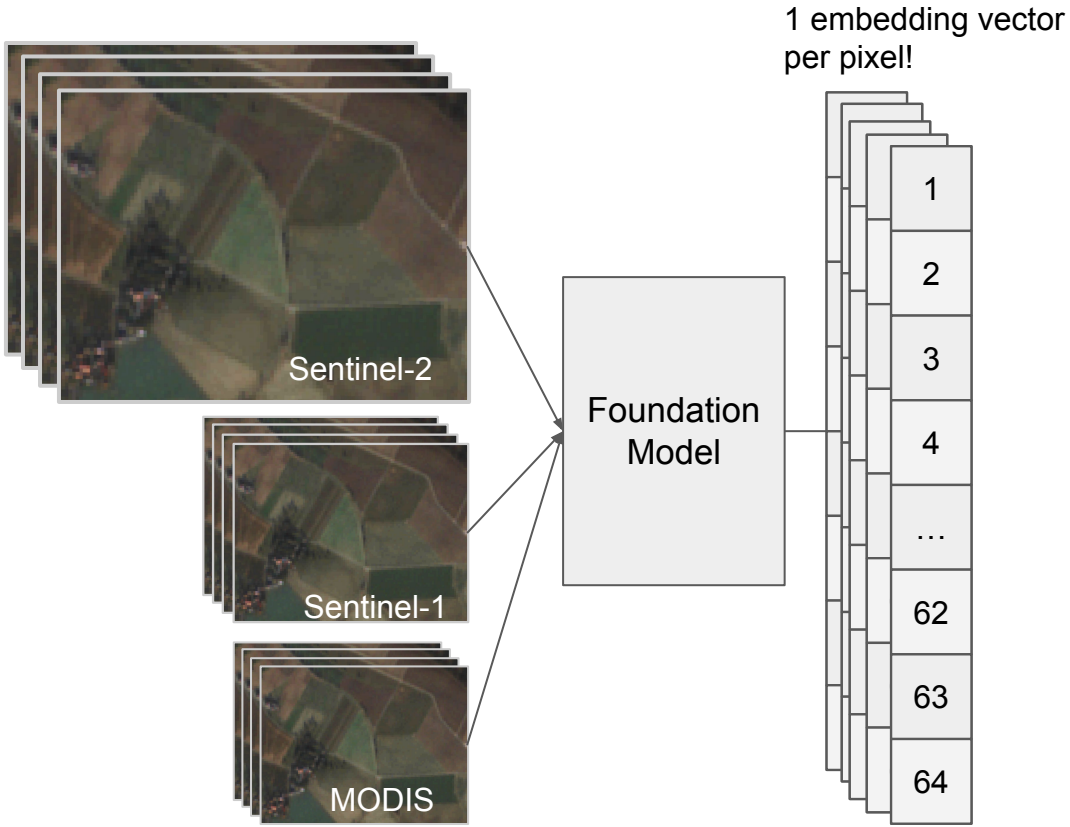
Images from Valerie Pasquarella,
AlphaEarth Foundations and the Satellite Embedding Dataset: A New Foundation for Geospatial Analysis, IEEE GRSS

What's in an Embedding?

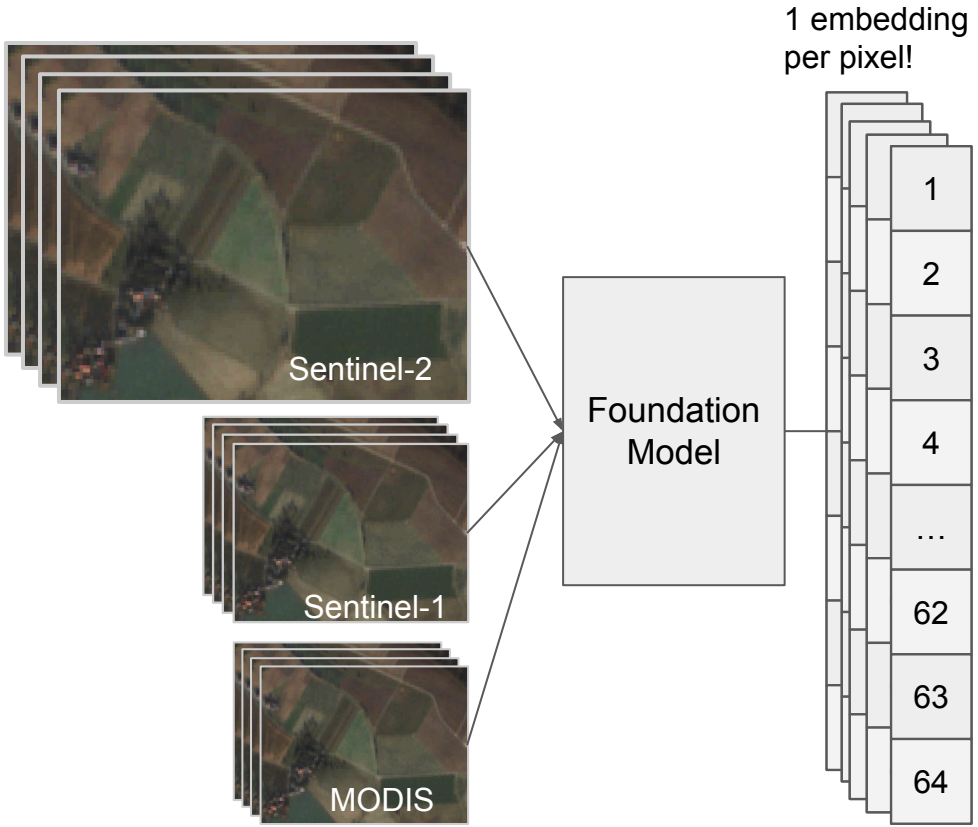


Brown et al, 2025, AlphaEarth Foundations

What's in an Embedding?



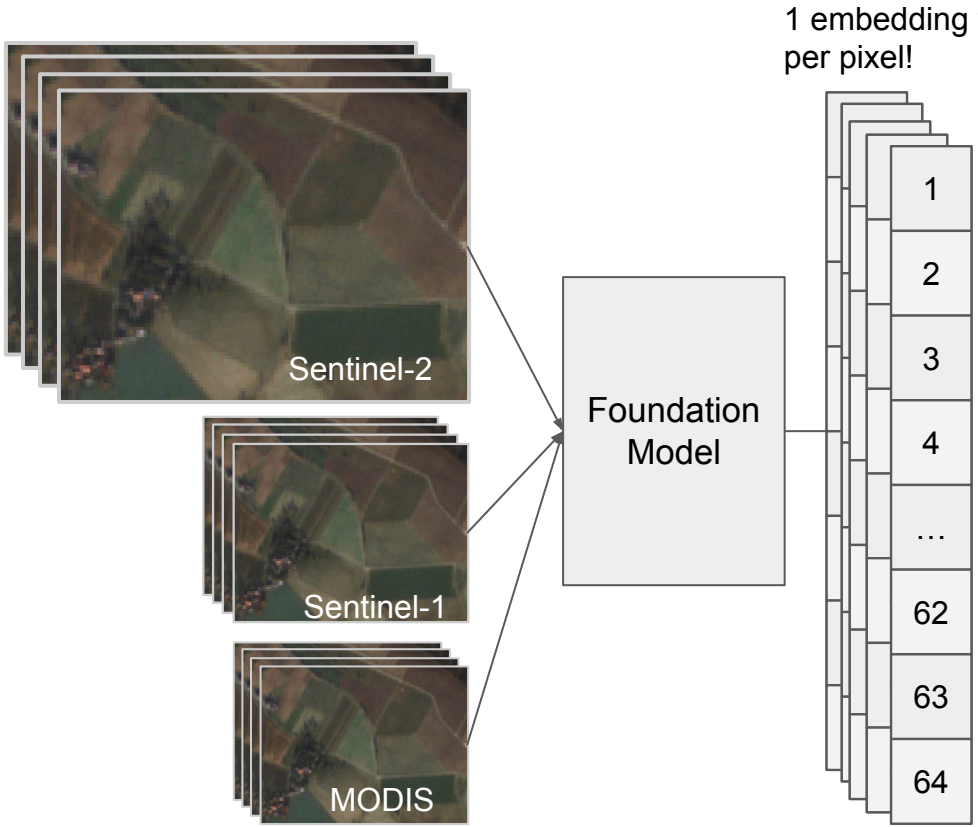
What's in an Embedding?



Latent Space Representation

Learned coordinate system where "nearby" points mean "similar in the way the model cares about."

What's in an Embedding?



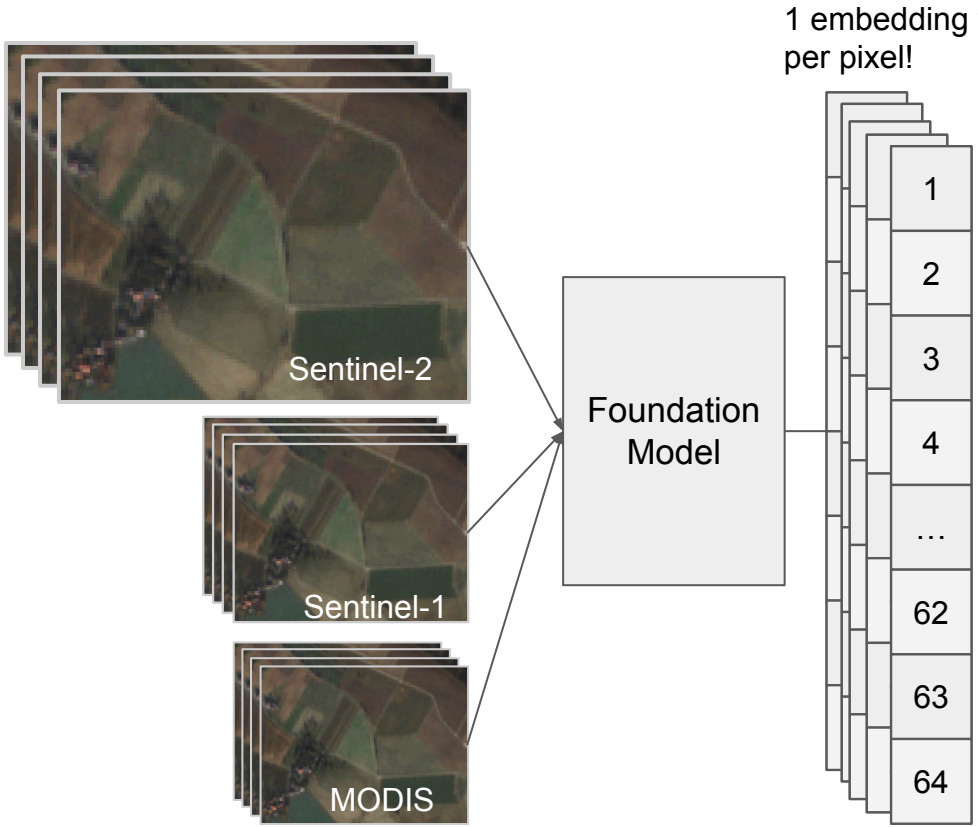
Latent Space Representation

Learned coordinate system where "nearby" points mean "similar in the way the model cares about."

Characteristics of Embeddings

- Values are not directly observed, but are (hidden) features that the model learns
- "Compression" of different data modalities & dates per pixel
- Individual dimensions usually don't have clean human labels
- Directions can encode attributes

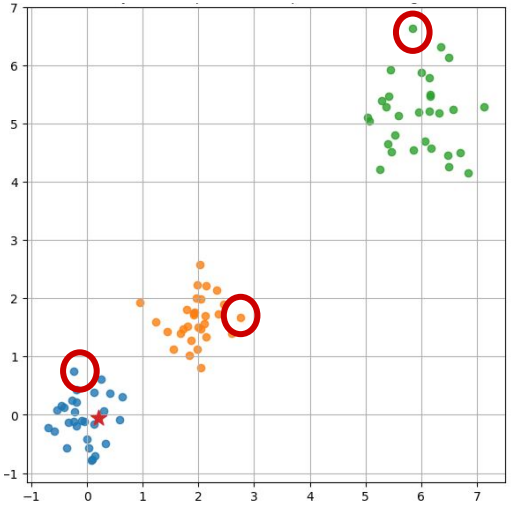
What's in an Embedding?



Latent Space Representation

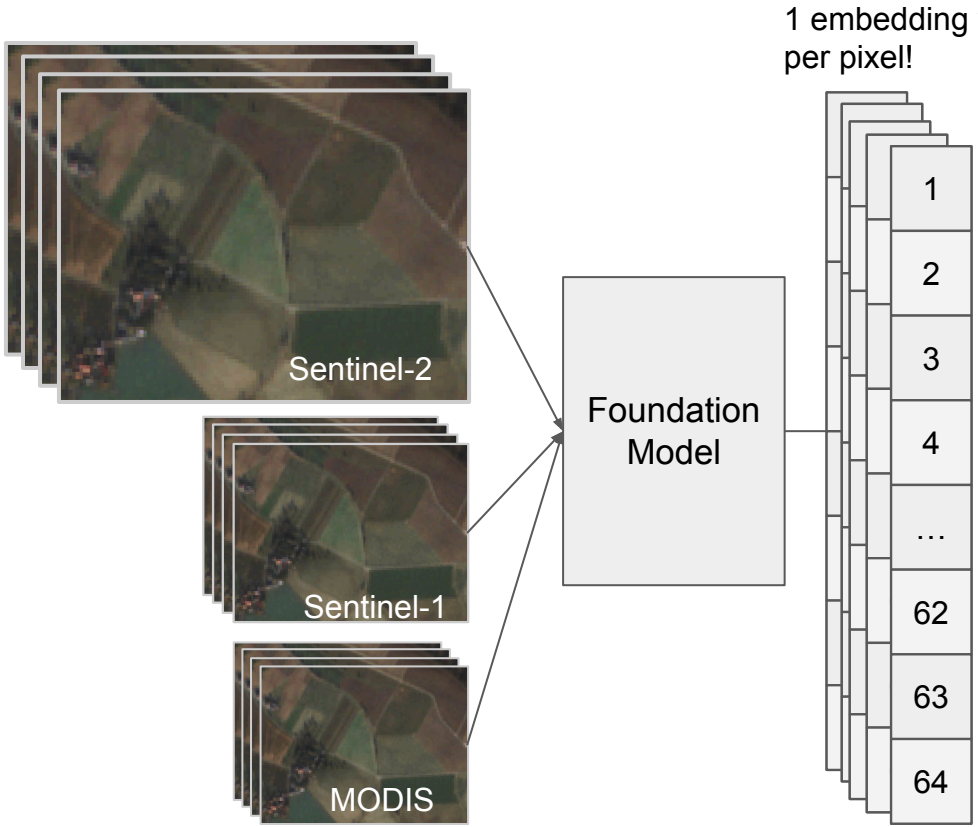
Learned coordinate system where "nearby" points mean "similar in the way the model cares about."

Similarity of Representations



UMAP dimensionality reduction of embeddings. Star = Wheat field

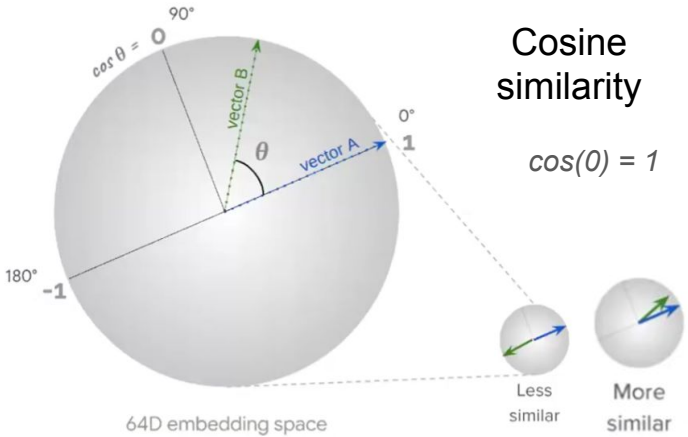
What's in an Embedding?



Latent Space Representation

Learned coordinate system where "nearby" points mean "similar in the way the model cares about."

Similarity of Representations

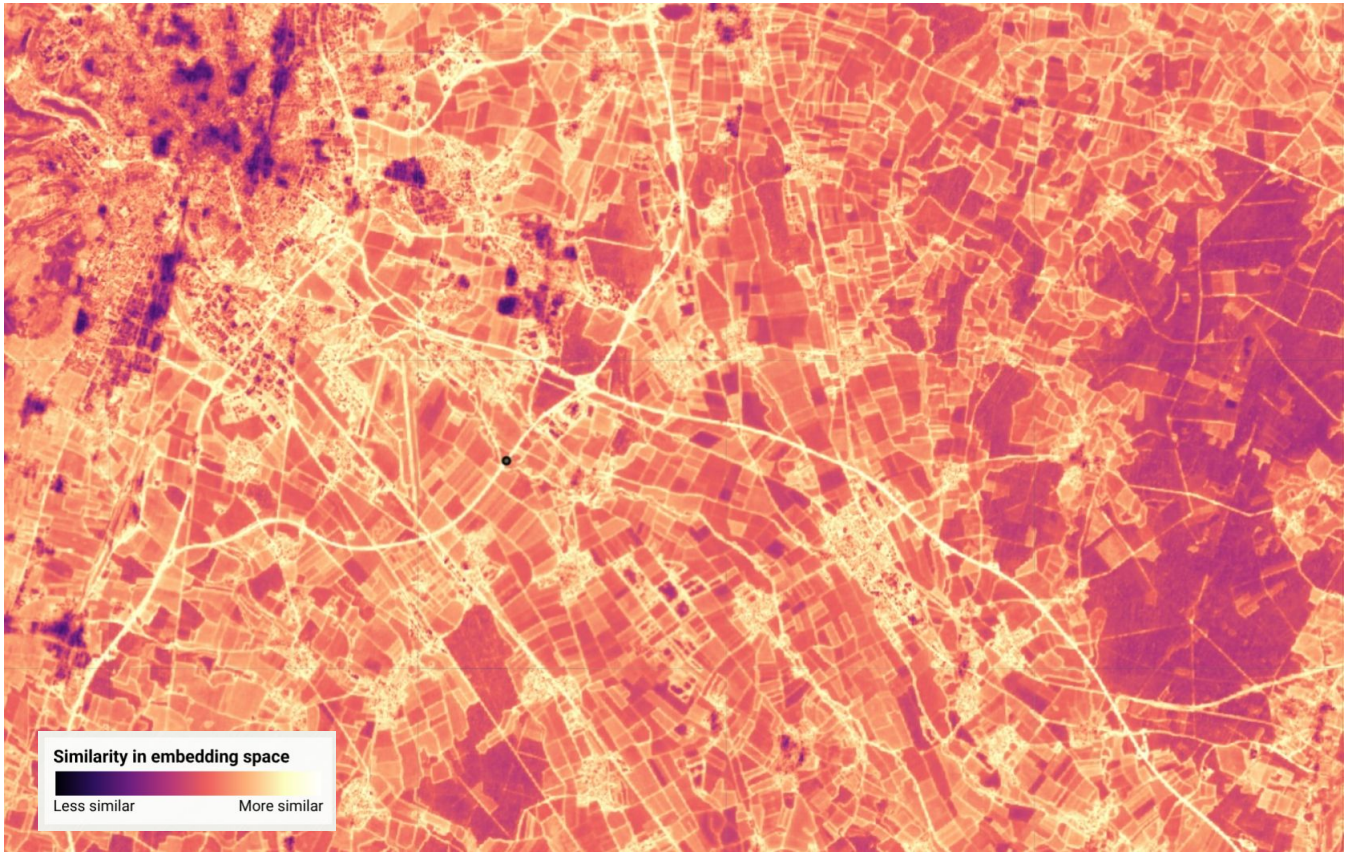


Cosine similarity

cos(0) = 1

Valerie Pasquarella,
AlphaEarth Foundations and the Satellite Embedding Dataset: A New Foundation for Geospatial Analysis, IEEE GRSS

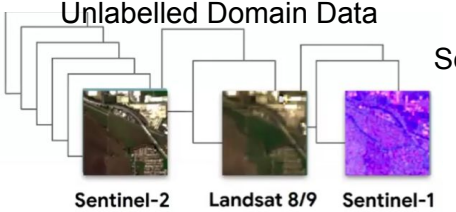
What’s in an Embedding?



Try it out yourself:
https://earthengine-ai.projects.earthengine.app/view/embedding-similarity-search#year=2024;zoom=18;lon=5.122757904832156;lat=47.2597373
4720962;clicked=true;

Foundation Models in EO

Global Dataset
Unlabelled Domain Data



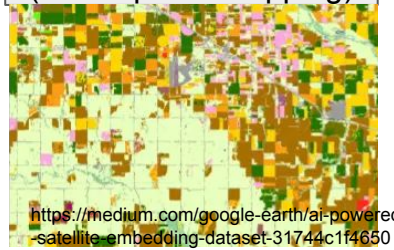
Optical, Hyperspectral, SAR, Precipitation, Climate, Elevation, Location, Text

Self Supervised Learning

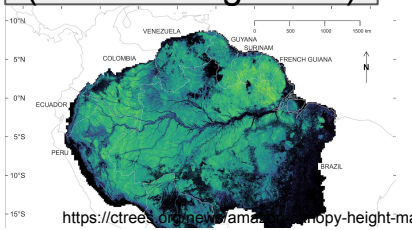
Pretrained Foundation Model

Embedding
(Latent Representation)

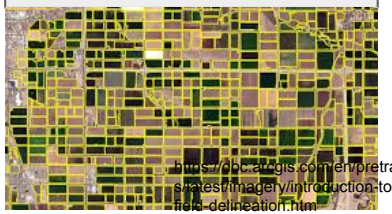
Classifier
(ie. Cropland mapping)



Regressor
(i.e Tree height est.)

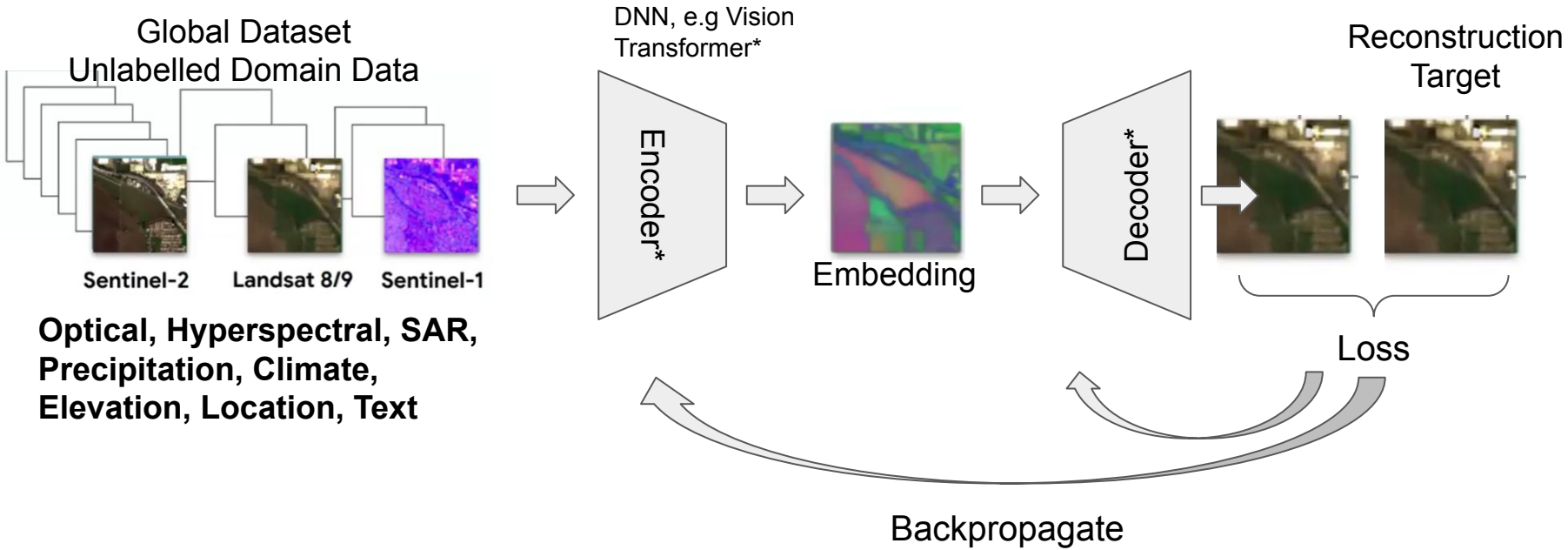


Segmentation



Images from Valerie Pasquarella,
AlphaEarth Foundations and the Satellite Embedding Dataset: A New Foundation for Geospatial Analysis, IEEE GRSS

Foundation Models in EO: Training



*Simplified schematics. Actual model architecture varies significantly by model.

Images from Valerie Pasquarella,
AlphaEarth Foundations and the Satellite Embedding Dataset: A New Foundation for Geospatial Analysis, IEEE GRSS

Self Supervised Learning

"*Self-supervised learning is a way to train a model on **unlabeled data** by creating its **own training signal** from the data itself.*"

1. Masked Reconstruction
2. Multi-Modal Alignment
3. Contrastive Learning
4. ….

Self Supervised Learning

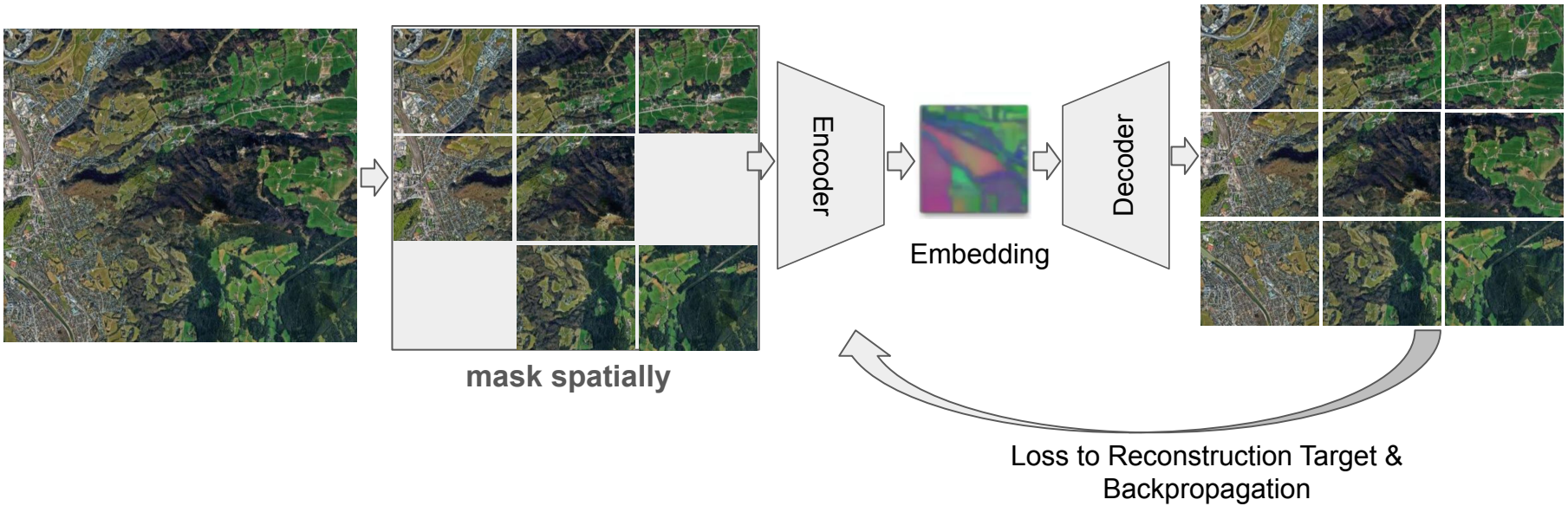
Masked Reconstruction

Self Supervised Learning

Masked Reconstruction

Self Supervised Learning

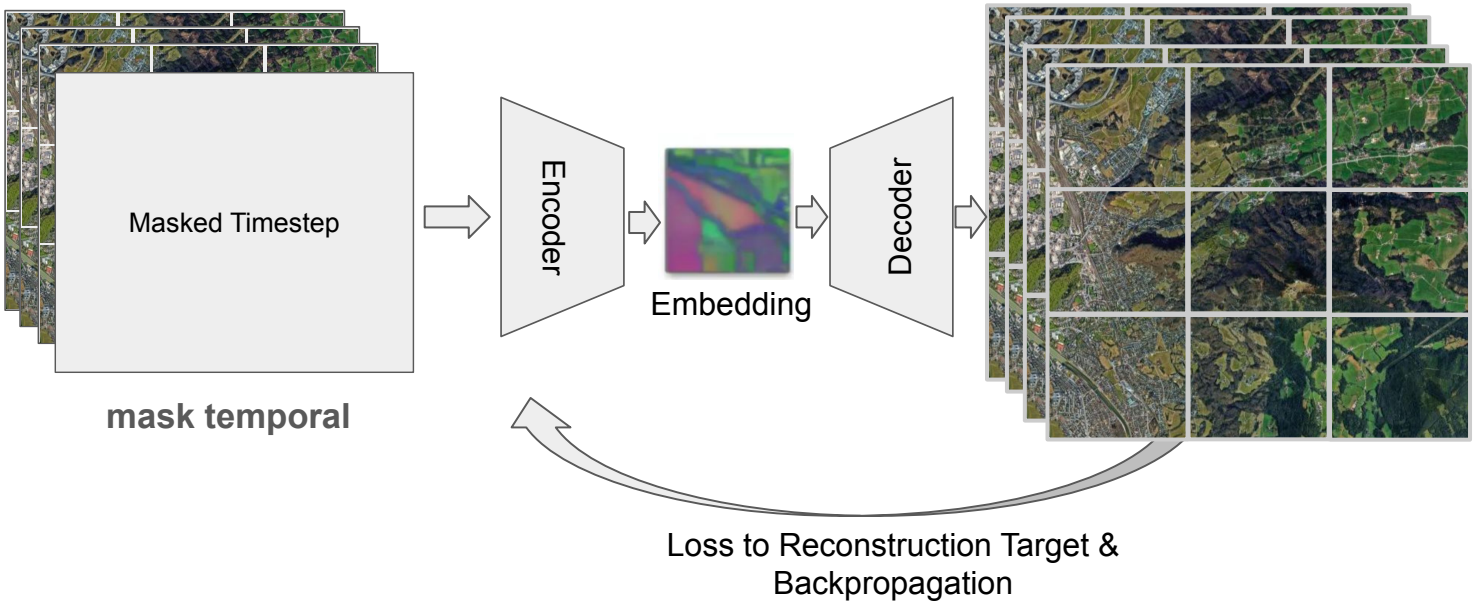
Masked Reconstruction



=> Model learns to reconstruct masked data on the spatial axis.

Self Supervised Learning

Masked Reconstruction



=> Model learns to reconstruct missing data on the temporal axis.

=> Similar process for the spectral bands.

Self Supervised Learning

Multi-Modal Alignment

Data from different sensors available: Multispectral, Hyperspectral, SAR, different resolutions, …

Goal: Make Representations of the same locations with temporal alignment have similar embeddings.
Important: Need to ensure temporal alignment of different sensor modalities!

Different Methods

- Also use Masked Autoencoding for different modalities!
- Separate encoders (+decoders) per modality + (learnt) fusion into latent space embedding
- Contrastive-Style Losses
- ….

Self Supervised Learning

Pre-Training for EO Foundation Models differs from model to model!

Some great further reads with details:

- Szwarcman, Daniela, et al. "Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications." IEEE Transactions on Geoscience and Remote Sensing (2025).
- Brown, Christopher F., et al. "Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data." arXiv preprint arXiv:2507.22291 (2025).
- Tseng, Gabriel, et al. "Galileo: Learning Global & Local Features of Many Remote Sensing Modalities." Forty-second International Conference on Machine Learning.

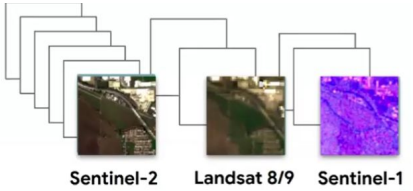
Foundation Models in Practice

Goal: Learn how to apply a pre-trained foundation model for an earth observation task for sustainable development.

Applying EO Foundation Models

Unsupervised - Zero Shot (No Labels required)

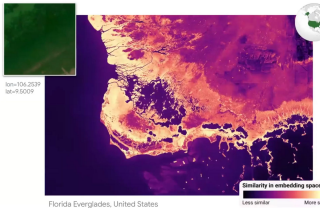
Imagery of Region of Interest



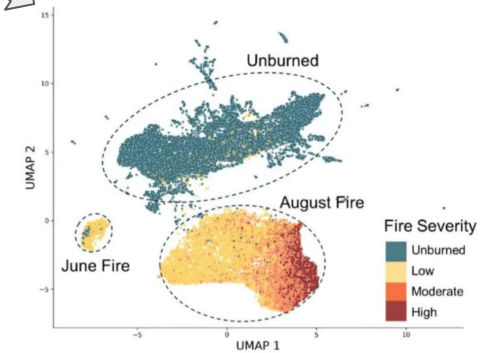
Pretrained Encoder*

Embeddings for Region of Interest

Similarity Search



Dimensionality Reduction & Clustering



*Decoder dropped! Most Geo-FMs do this. However, somel model might have an architecture that uses both are a different architecture altogether..

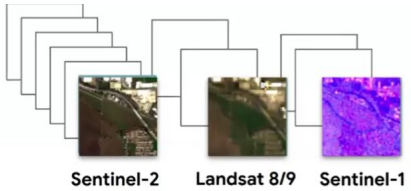
Madeline Lisaius, Tessera: Building a Global Open-Source Geospatial Foundation Model

Applying EO Foundation Models: Supervised

Supervised - Few Shot

Embeddings are used as features to train standard ML models.

Few samples of locations that are relevant for the task.



+ Labels

Embedding
(Latent Representation)

Task-specific heads

- Classifier
- Regressor
- Segmentation

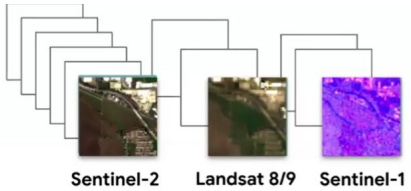
Option A: End-to-End finetuning (Train the task-head + finetune the FM)
Option B: Only train the task-head and freeze the foundation model (backbone).

Applying EO Foundation Models: Supervised

Supervised - Few Shot

Embeddings are used as features to train standard ML models.

Few samples of locations that are relevant for the task.



+ Labels

Embedding
(Latent Representation)

Task-specific heads

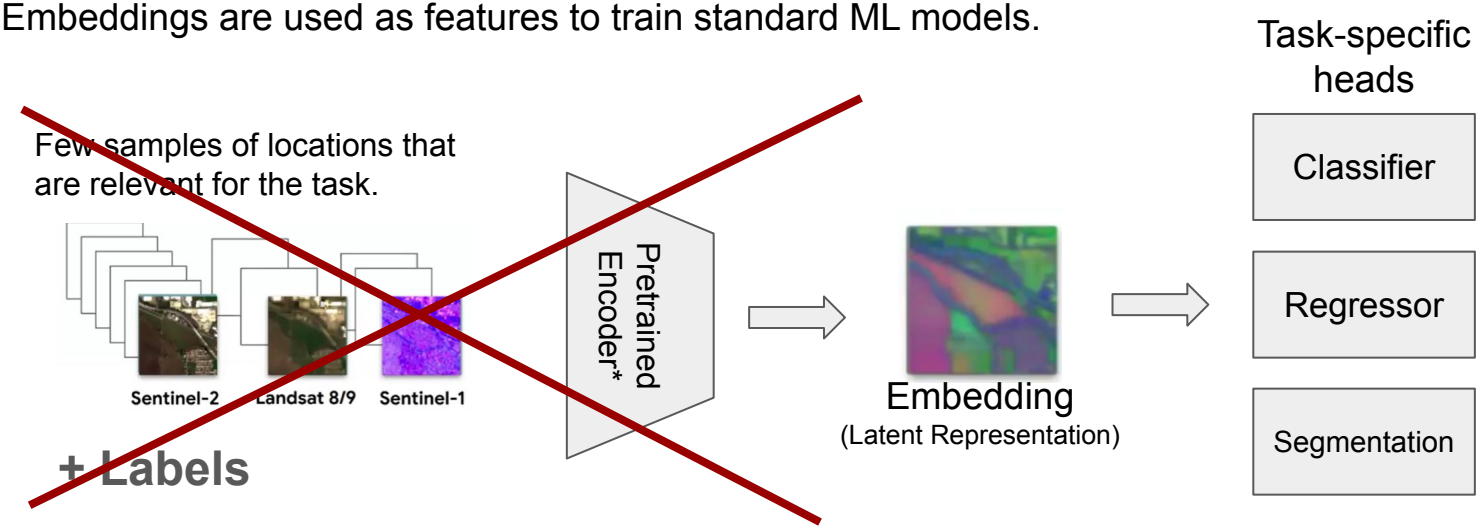
- Classifier
- Regressor
- Segmentation

Problem: Still requires access to all the satellite imagery + need all the different modalities to infer the embeddings. Imagery can be multiple TBs of data + Compute on GPUs is expensive!

Applying EO Foundation Models: Supervised

Supervised - Few Shot

Embeddings are used as features to train standard ML models.



Problem: Still requires access to all the satellite imagery + need all the different modalities to infer the embeddings. Imagery can be multiple TBs of data + Compute on GPUs is expensive!

=> Available Stores (Datasets) with free access to **precomputed** embeddings. Democratizes access to EO data. E.g Google Earth Engine, Source Cooperative

Hands-On Session Instructions

Goal: Learn how to use pre-computed embeddings for your work!

Train your own Model for

- 1: Crop Type Mapping: **https://shorturl.at/wEY4V**
- 2: Advanced / Exercise: Deforestation Detection https://shorturl.at/auXW3

Step 1: Open the link and sign into Colab
Step 2: Run the code cell by cell and try to understand what is happening
Step 3: Once you have the model working, try to think of ways how to improve it and play around with the parameters. What changes when you change the number of training samples?

Pitfalls & Limitations of GeoFMs

Fast Changing Space with many Unanswered Questions:

- “Black-Box Models”
- High price- and energy costs for pre-training
- Timely availability of pre-computed embeddings
- Seasonality and temporal resolution in inputs
- Hard to quantify uncertainty
- Standard RS+ML Methods often achieve competitive performance
    - Limited understanding of when they work well

Crop	Scheme	AEF				RS			
		RF		XGB		RF		XGB	
		R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Corn	County-Year	0.43	2.28	0.44	2.27	0.45	2.24	**0.46**	**2.24**
	Yearly	0.31	2.54	0.30	2.55	**0.33**	2.50	**0.33**	**2.49**
	Scale-Transfer	0.24	2.67	0.25	2.64	0.37	2.43	**0.38**	**2.40**
Soybean	County-Year	0.36	0.87	**0.37**	**0.87**	0.35	0.90	0.35	0.90
	Yearly	0.29	0.92	**0.31**	**0.91**	0.28	0.93	0.25	0.95
	Scale-Transfer	0.13	1.02	0.11	1.03	0.22	0.97	**0.24**	**0.95**
Winter Wheat	County-Year	**0.43**	1.32	**0.43**	**1.31**	0.42	1.30	0.41	1.31
	Yearly	0.38	1.34	**0.40**	**1.32**	0.36	1.35	0.36	1.36
	Scale-Transfer	-0.18	1.85	-0.04	1.73	0.20	1.52	**0.22**	**1.50**

Task	Scale	Crop	Scheme	AEF-based RF		RS-based RF	
				R^2	RMSE	R^2	RMSE
Yield Prediction	County	Corn	East → West	0.02	2.71	**0.66**	**1.61**
			West → East	0.51	1.23	**0.60**	**1.12**
		Soybean	East → West	-0.28	0.94	**0.53**	**0.56**
			West → East	0.39	0.44	**0.59**	**0.37**
	Field	Corn	East → West	0.25	2.65	**0.38**	**2.41**
			West → East	0.33	2.48	**0.34**	**2.45**
		Soybean	East → West	0.08	0.99	**0.19**	**0.93**
			West → East	0.23	1.00	**0.28**	**0.97**

Ma et al, 2025, Harvesting AlphaEarth: Benchmarking the Geospatial Foundation Model for Agricultural Downstream Tasks

Future of Foundation Models in EO

- Dealing with Seasonality
- Pipelines for easier inference & finetuning
- LoRa & Adapter based finetuning
- Physics constrained FMs
- Explainability of Predictions & Quantifying Uncertainty
- More pre-computed embeddings

Vision-Language Models for EO

- ChatGPT-style interaction with EO data

…

Time to work on your Paper Presentation.
(~20min, groups)

References & Further Material

Literature Recommendations

- Brown et Al, 2025, Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data
- Tseng et Al, 2023, Lightweight, pre-trained transformers for remote sensing timeseries
- Xiao et Al, 2025, Foundation models for remote sensing and earth observation: A survey
- Vaswani et Al, 2017, Attention is all you need
- Herzog et Al, 2025, OlmoEarth: Stable Latent Image Modeling for Multimodal Earth Observation
- Lu et al, 2025, Vision Foundation Models in Remote Sensing: A survey
- Ma et al, 2025, Harvesting AlphaEarth: Benchmarking the Geospatial Foundation Model for Agricultural Downstream Tasks

References and Sources

- https://www.youtube.com/watch?v=yg5l9ED05-s&t=2387s
- https://medium.com/google-earth/ai-powered-pixels-introducing-googles-satellite-embedding-dataset-31744c1f4650
- https://deepmind.google/blog/alphaearth-foundations-helps-map-our-planet-in-unprecedented-detail/

Interested in doing an Internship with NASA Harvest @ Unistra?