esa Φ-lab

EARTH SYSTEMS LAB

TRILLIUM EUROPE

ESLAB.AI

FOUNDATION MODELS IN EXTREME ENVIRONMENTS

SHRUG-FM: Reliability-Aware Foundation Models for Earth Observation

DATA

DECISIONS

FDL
ESA Φ-LAB
GOOGLE CLOUD / NVIDIA / SCAN AI
FM IN EXTREME ENVIRONMENTS

FDL.AI NETWORK

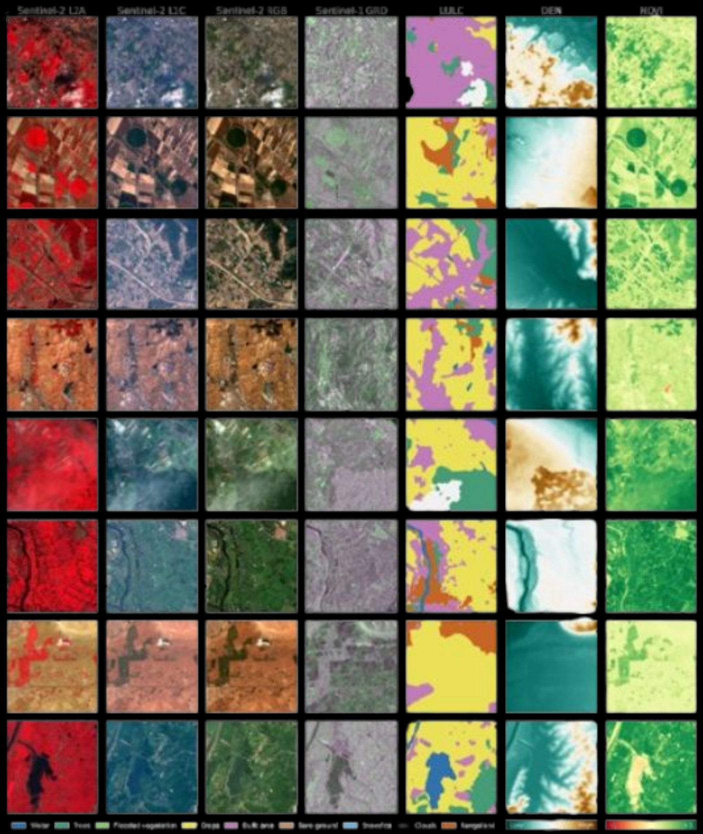
10 YEARS OF APPLIED AI RESEARCH FOR ALL HUMANKIND

Google Cloud NVIDIA SCAN AI

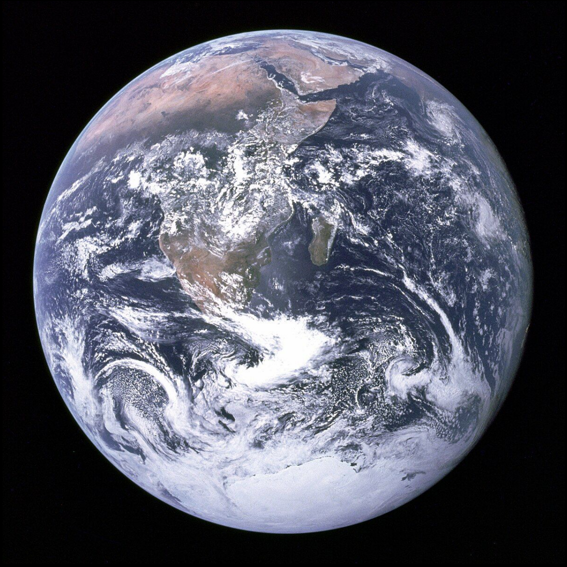
UNIVERSITY OF OXFORD PASTEUR ISI

AD ASTRA PER ALGORITHMOS

TRILLIUM EUROPE



source: TerraMesh, Blumenstiel et al.



Many different foundation models for Remote Sensing….

SatlasNet

GFM-Swin

RemoteCLIP

Prithvi

DOFA

CROMA

… and more

Many different foundation models for Remote Sensing….

SatlasNet

GFM-Swin

RemoteCLIP

Prithvi

DOFA

CROMA

… and more

Many different foundation models for Remote Sensing….

SatlasNet

GFM-Swin

RemoteCLIP

Prithvi

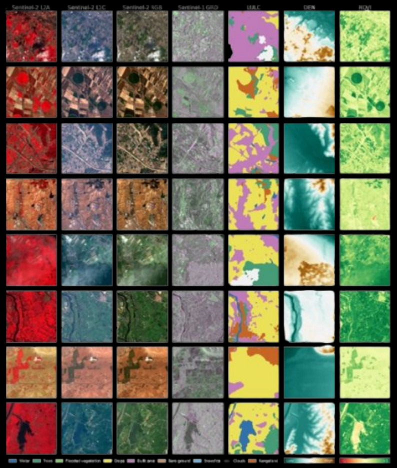
DOFA

CROMA

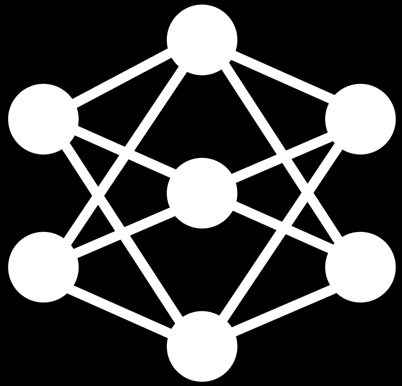
… and more

→ IDK

The data doesn't know

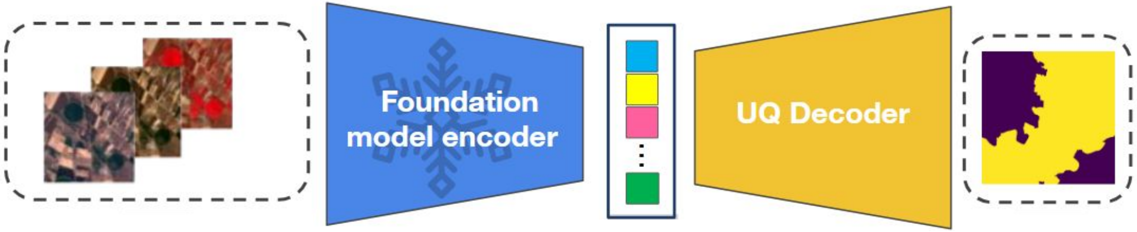


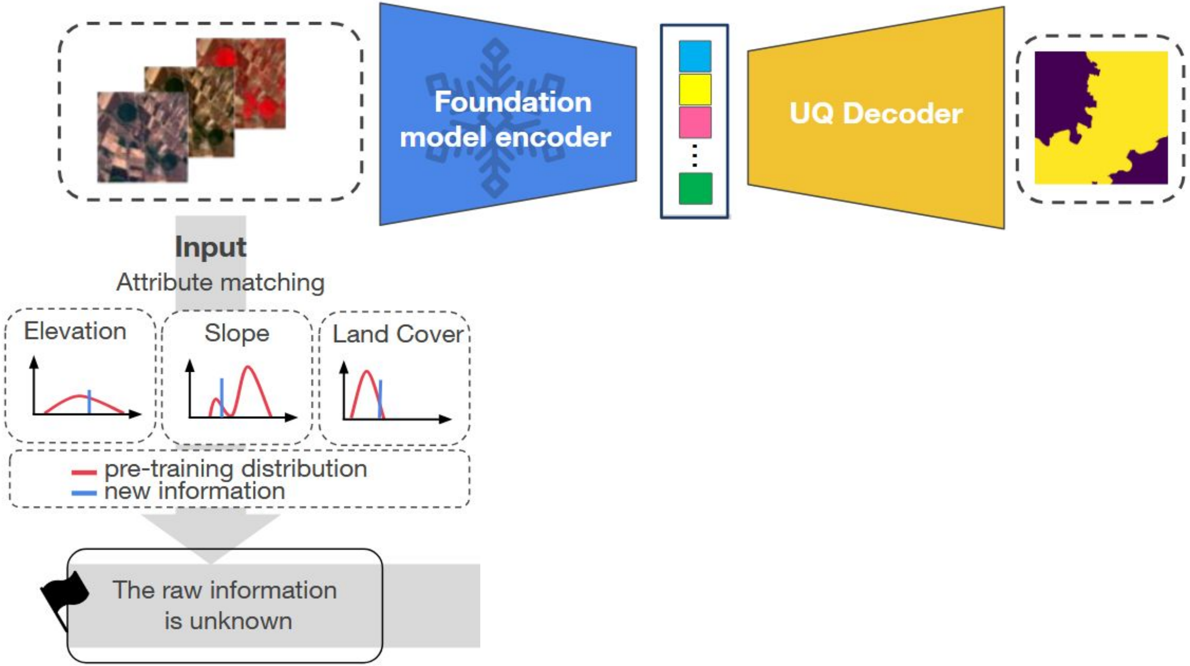
The model doesn't know

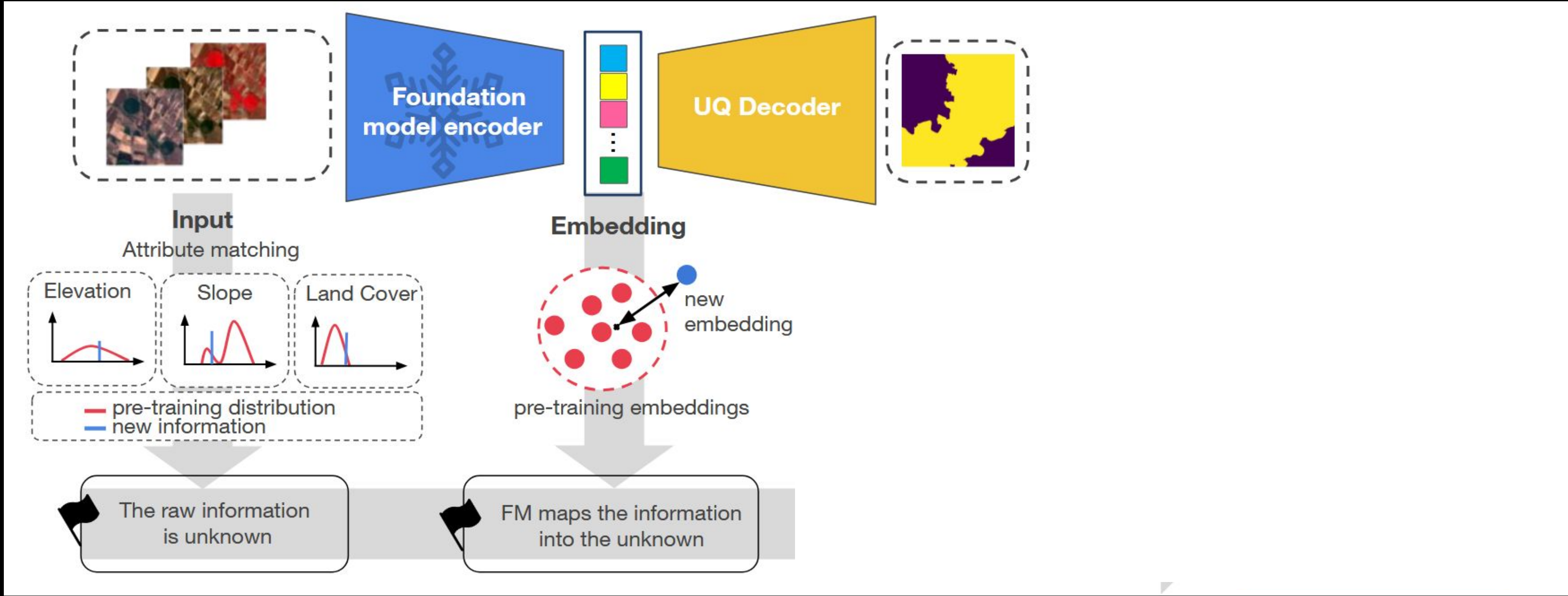


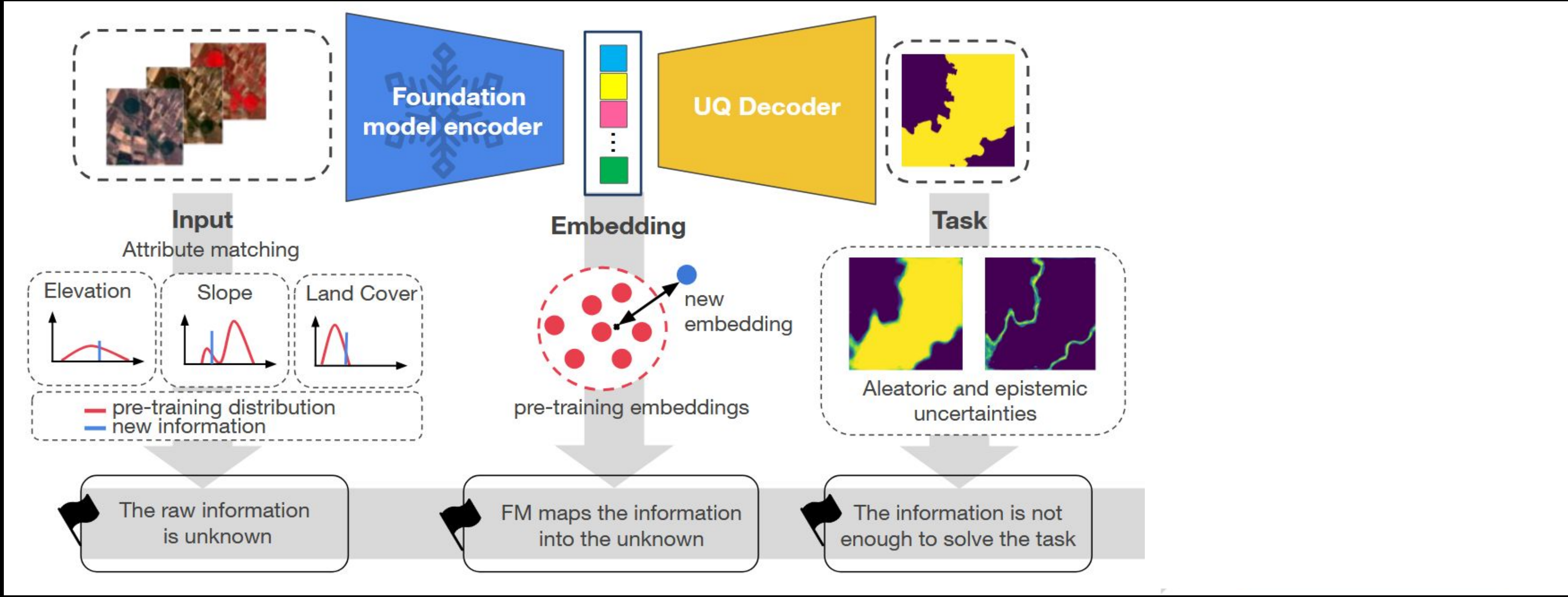
IDK

flagging and selective predictions

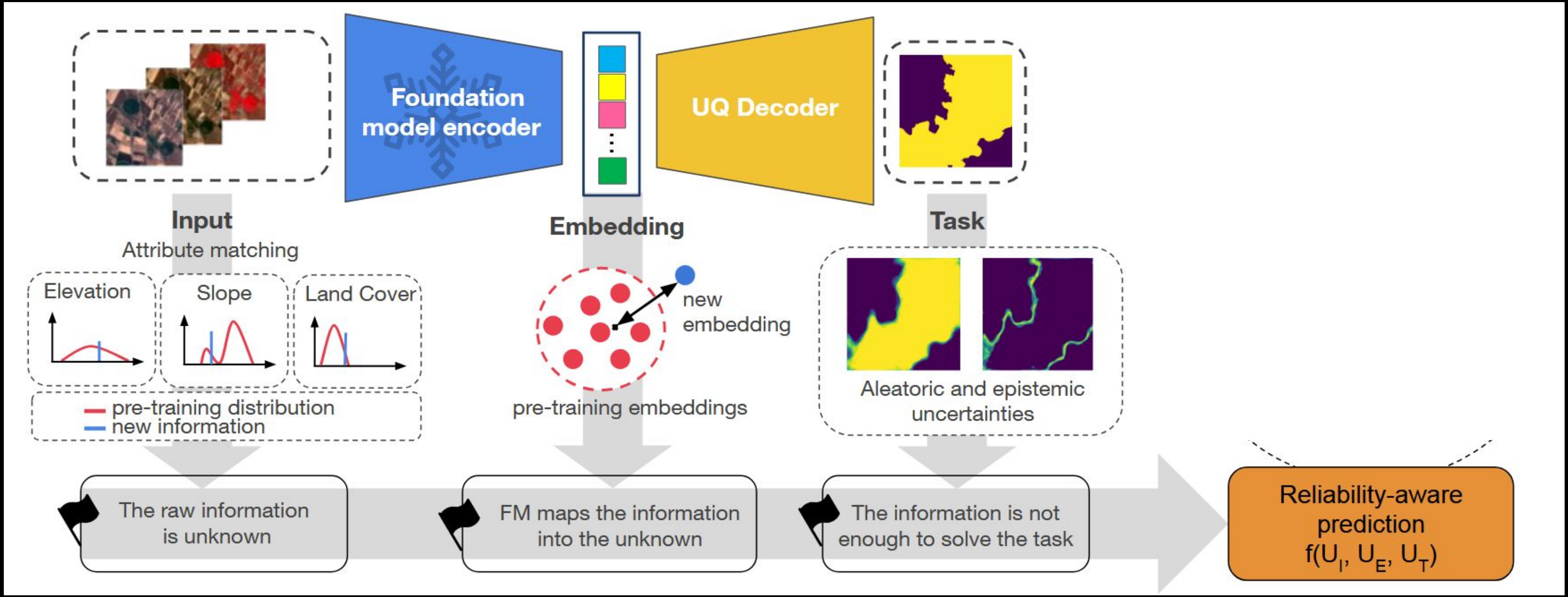

Foundation
model encoder
UQ Decoder

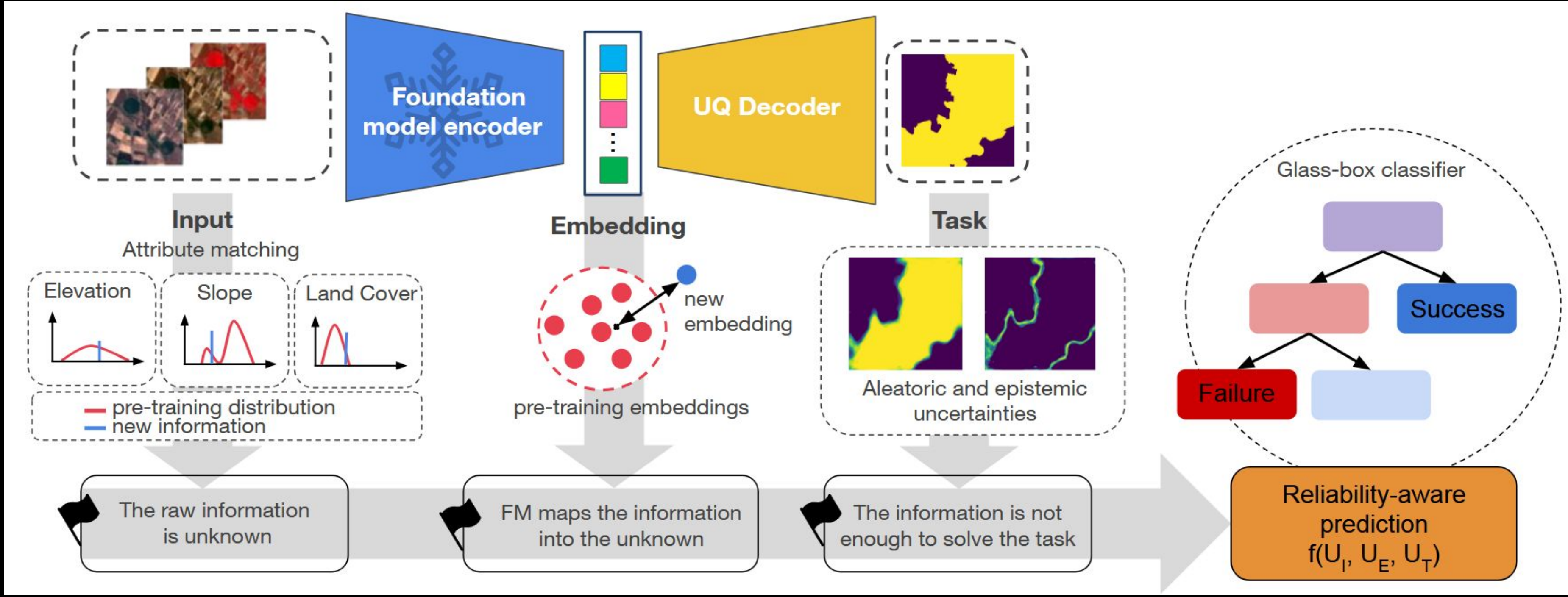

Foundation
model encoder
UQ Decoder
Input
Attribute matching
Elevation
Slope
Land Cover
pre-training distribution
new information
The raw information
is unknown


Foundation model encoder
UQ Decoder
Input
Attribute matching
Elevation
Slope
Land Cover
pre-training distribution
new information
Embedding
new embedding
pre-training embeddings
The raw information is unknown
FM maps the information into the unknown



Foundation model encoder
UQ Decoder
Input
Attribute matching
Elevation
Slope
Land Cover
pre-training distribution
new information
Embedding
new embedding
pre-training embeddings
Task
Aleatoric and epistemic uncertainties
The raw information is unknown
FM maps the information into the unknown
The information is not enough to solve the task


Foundation model encoder
UQ Decoder
Input
Attribute matching
Elevation
Slope
Land Cover
pre-training distribution
new information
Embedding
new embedding
pre-training embeddings
Task
Aleatoric and epistemic uncertainties
The raw information is unknown
FM maps the information into the unknown
The information is not enough to solve the task
Reliability-aware prediction
f(U_I, U_E, U_T)



Foundation model encoder
UQ Decoder
Input
Attribute matching
Elevation
Slope
Land Cover
pre-training distribution
new information
Embedding
new embedding
pre-training embeddings
Task
Aleatoric and epistemic uncertainties
Glass-box classifier
Success
Failure
The raw information is unknown
FM maps the information into the unknown
The information is not enough to solve the task
Reliability-aware prediction
$f(U_I, U_E, U_T)$

TRILLIUM EUROPE

FM IN EXTREME ENVIRONMENTS

FOUNDATION MODELS IN EXTREME ENVIRONMENTS 2025

#makeFMsayIDK

RESEARCHERS | FACULTY | ADVISORS | esa Φ-lab


Kai-Hendrik Cohrs


Maria Gonzalez


Joppe Massant


Vasileios Sitokonstantinou


Steffen Knoblauch


Patrick Ebel

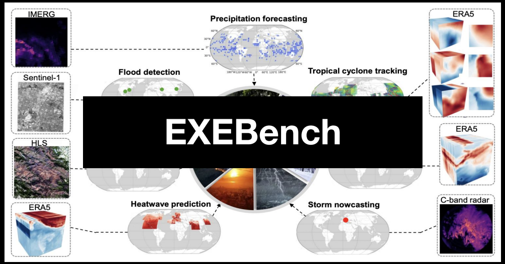

Zuzanna Osika

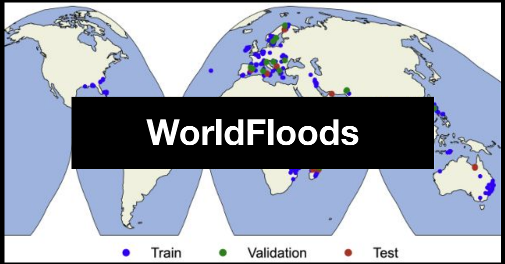

Vishal Nedungadi

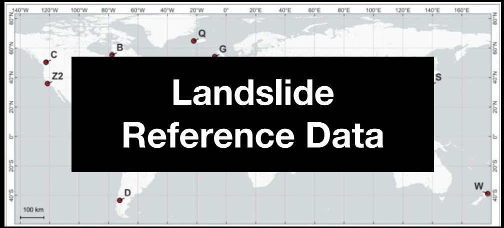

Shruti Nath

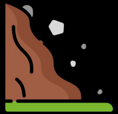

Ruben Cartuyvels

Datasets

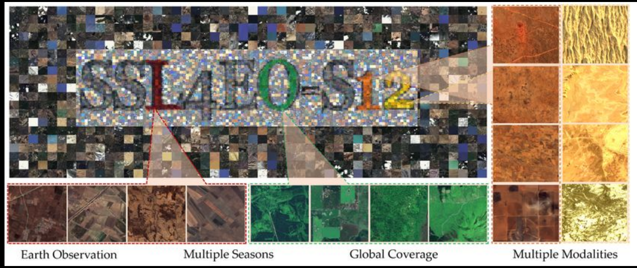








Foundation Model

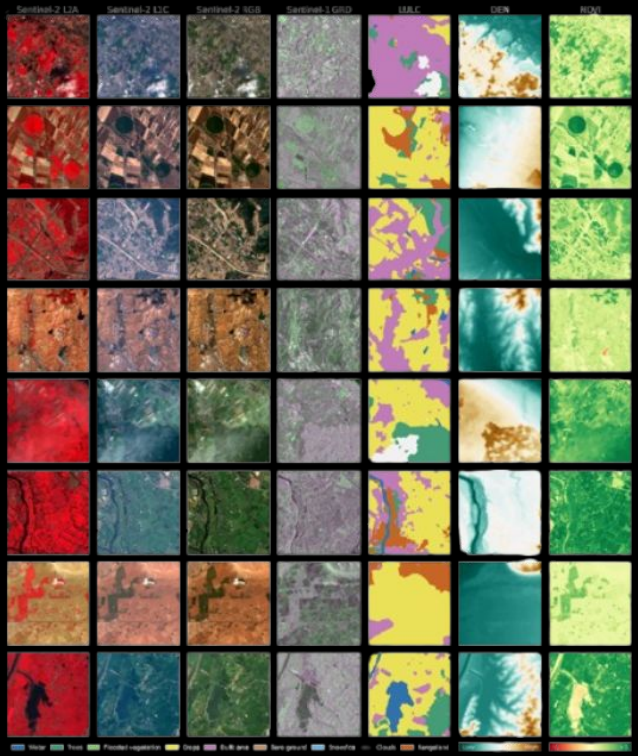


Pre-trained models

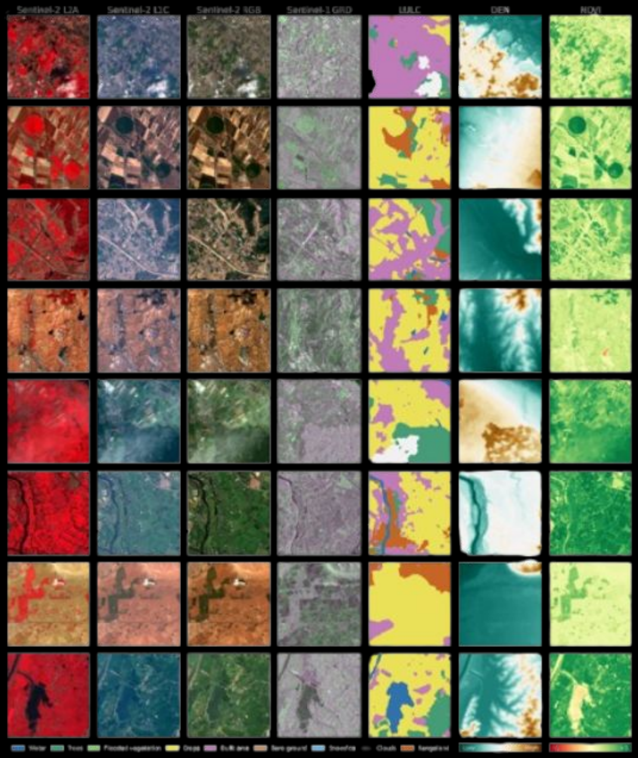
The pre-trained models with different SSL methods are provided as follows (13 bands of S2-L1C, 100 epochs, input clip to [0,1] by dividing 10000).

SSL method	Arch	BigEarthNet*	EuroSAT	So2Sat-LCZ42	Download			Usage
MoCo	ResNet50	91.8%	99.1%	60.9%	full ckpt	backbone	logs	define model, load weights
MoCo	ViT-S/16	89.9%	98.6%	61.6%	full ckpt	backbone	logs	define model, load
DINO	ResNet50	90.7%	99.1%	63.6%	full ckpt	backbone	logs	define model, load weights
DINO	ViT-S/16	90.5%	99.0%	62.2%	full ckpt	backbone	logs	model, load weights
MAE	ViT-S/16	88.9%	98.7%	63.9%	full ckpt	backbone	logs	define model, load weights
Data2vec	ViT-S/16	90.3%	99.1%	64.8%	full ckpt	backbone	logs	define model, load weights

1. The raw information is unknown

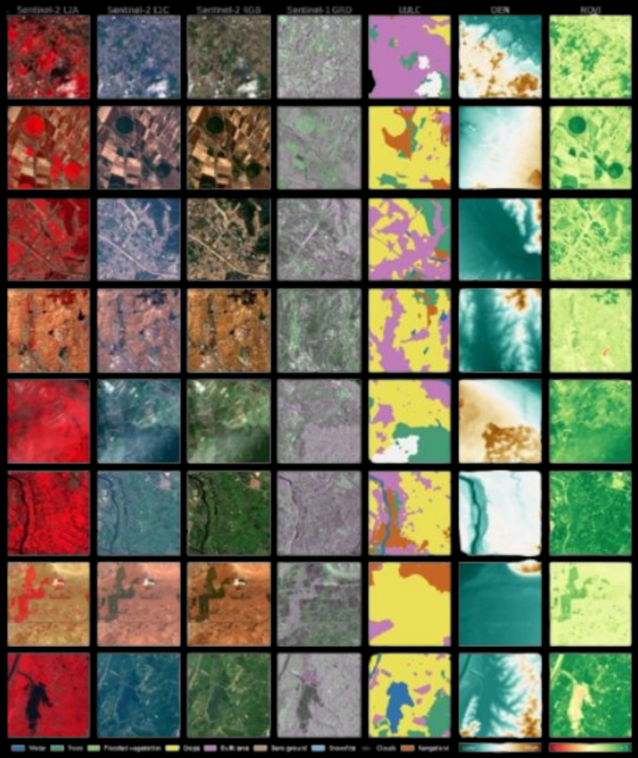


1. The raw information is unknown



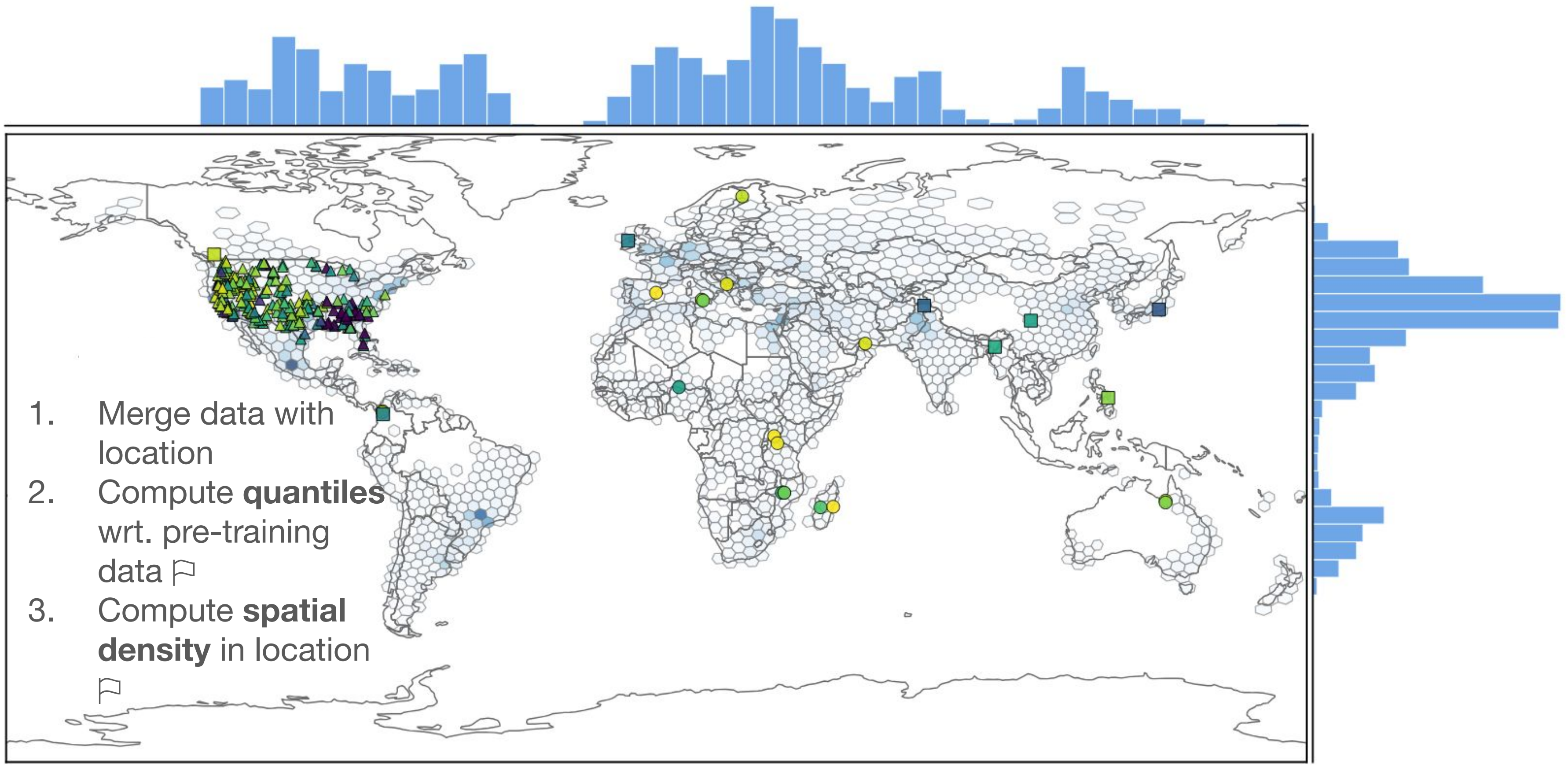
I've never seen anything like this before!

1. The raw information is unknown



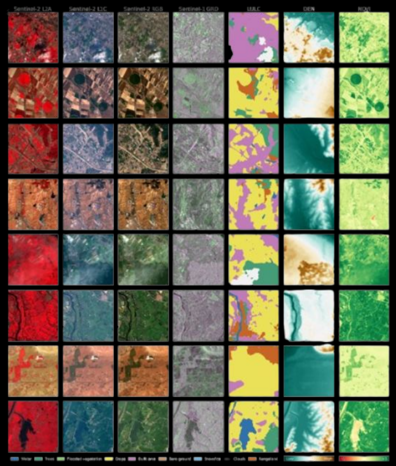
Dataset	Resolution	Included Features
HydroATLAS [23]	500m	Hydro-environmental & land cover attributes
Copernicus DEM [9]	30m	Elevation

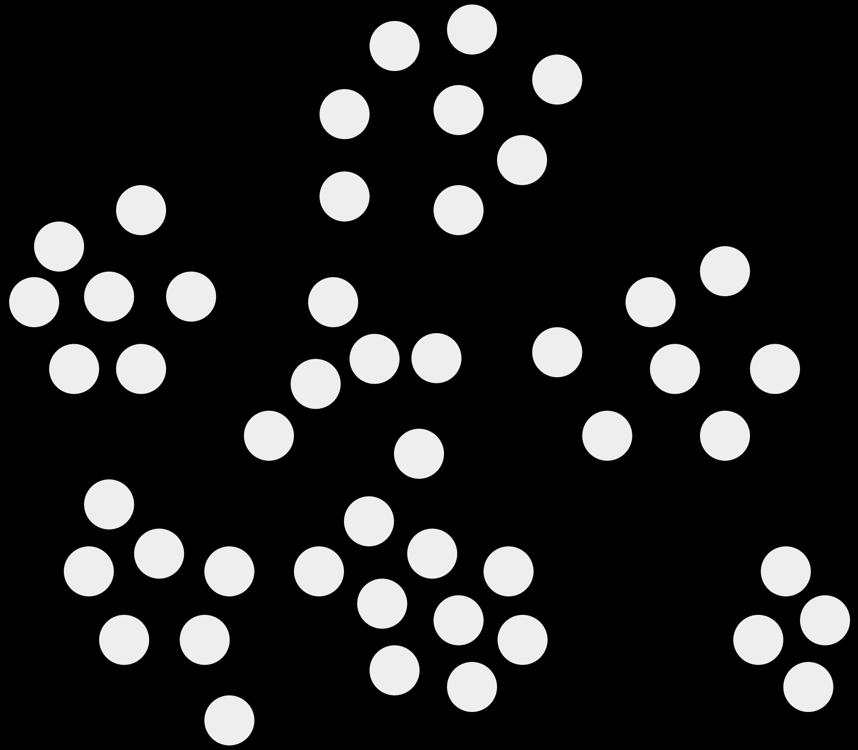
Category	Feature	Description
HydroATLAS	`inu_pc_smn`	Min. annual inundation extent (% cover).
	`inu_pc_smx`	Max. annual inundation extent (% cover).
	`ria_ha_ssu`	Total river area (hectares).
	`slp_dg_sav`	Average terrain slope (degrees × 10).
	`snw_pc_syr`	Avg. annual snow cover extent (% cover).
	`snw_pc_smx`	Max. annual snow cover extent (% cover).
	`glc_cl_smj`	Land cover classes (spatial majority).
	`wet_pc_sg1`	Wetland extent (grouping 1, % cover).
	`wet_pc_sg2`	Wetland extent (grouping 2, % cover).
	`for_pc_sse`	Average forest cover extent (% cover).
	`crp_pc_sse`	Average cropland extent (% cover).
	`pst_pc_sse`	Average pasture extent (% cover).
	`ire_pc_sse`	Average irrigated area extent (% cover).
	`urb_pc_sse`	Average urban extent (% cover).
	`hft_ix_s09`	Human Footprint index (year 2009).
DEM	`elev_mean`	Average elevation (meters a.s.l.).
	`elev_min`	Minimum elevation (meters a.s.l.).
	`elev_max`	Maximum elevation (meters a.s.l.).
Spatial	`density`	Spatial density of pre-training data.



1. Merge data with location
2. Compute **quantiles** wrt. pre-training data 🏳
3. Compute **spatial density** in location 🏳

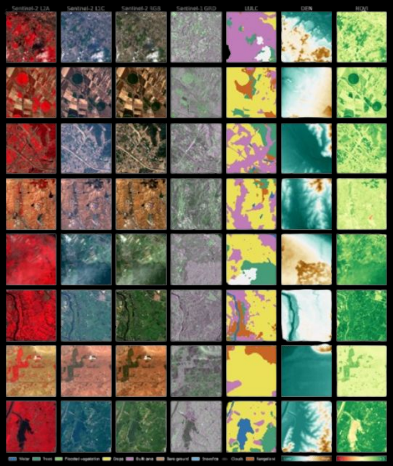
2. FM maps the information into the unknown

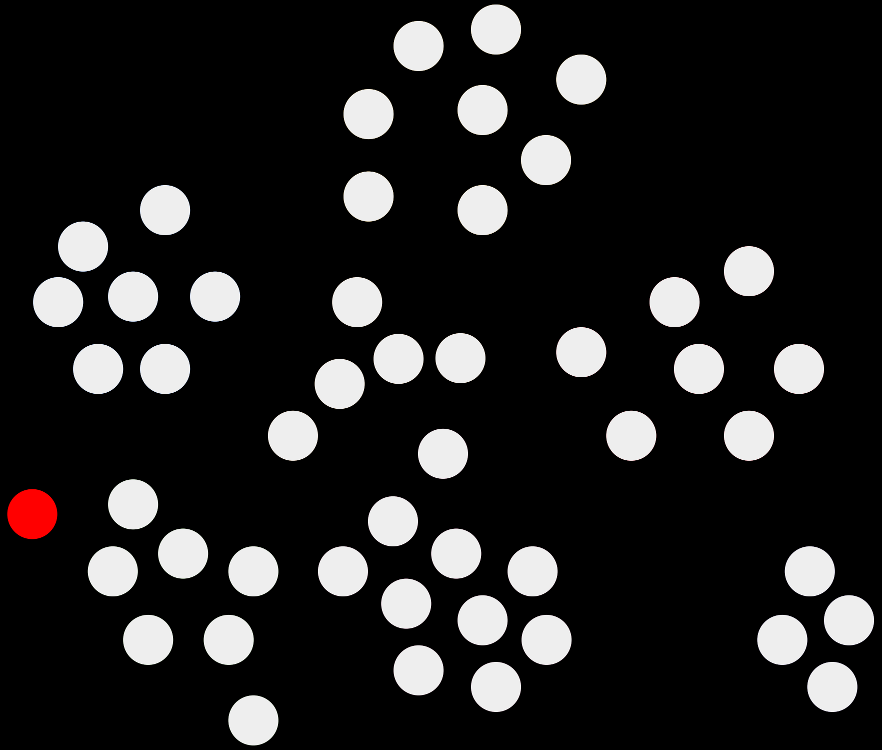




2. FM maps the information into the unknown

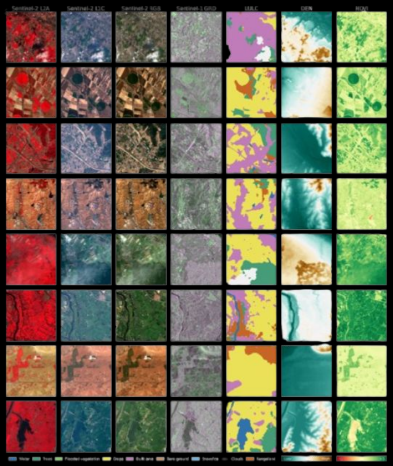
OOD score ≤ threshold

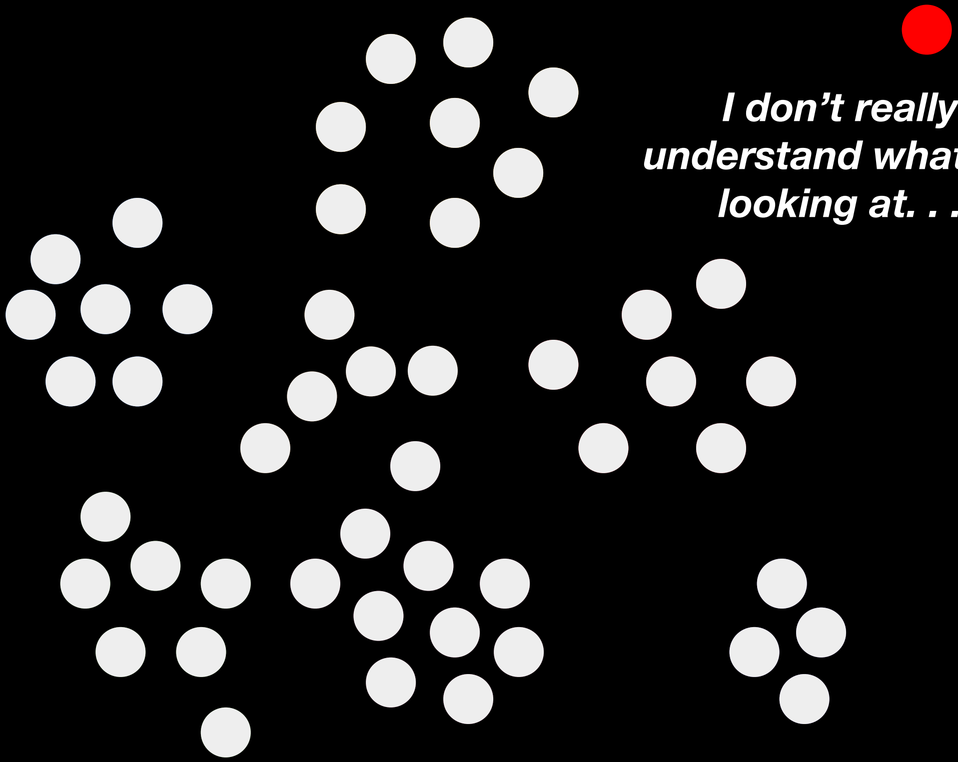




2. FM maps the information into the unknown

OOD score > threshold





I don’t really understand what I’m looking at. . .

Problem Set-up

- **Input space: Raw Images**
  - SSL4EOS12 : Sentinel-2
  - 250k samples, ~550GB
  - 13 bands: 652.288 dims.

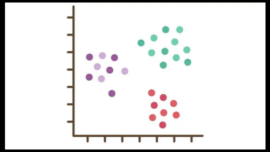
Foundation Model: *ViT small architecture*

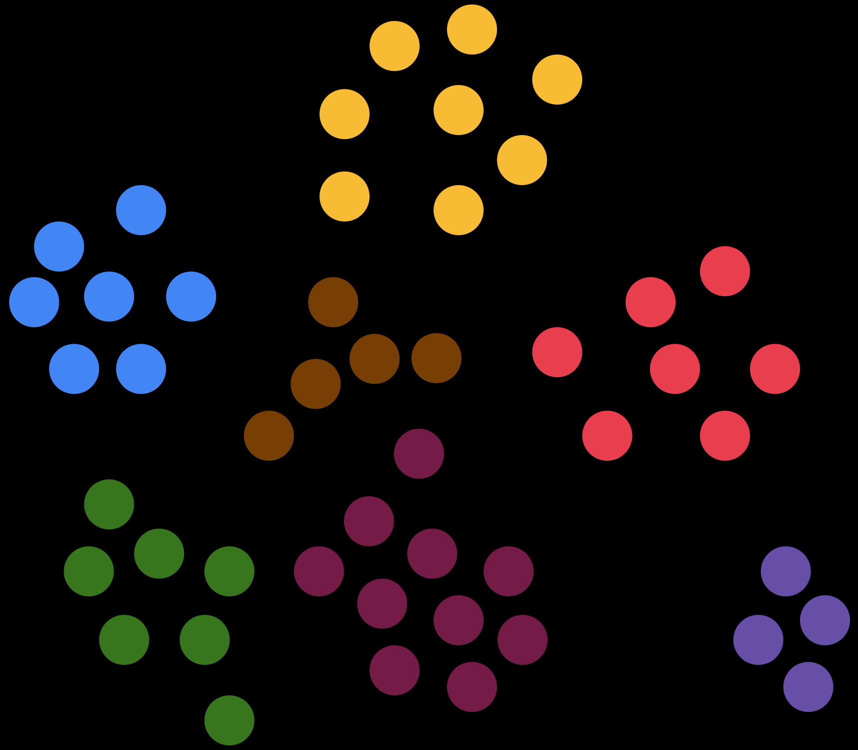
- **Embedding space**
  - ViT small pre-trained on SSL4EOS12
  - Numpy arrays, ~400MB
  - 384 dimensions

`[0.23 2.45 5.66 …… 6.21]`

I don't know where to put this

- K Means
  - Faiss k-means

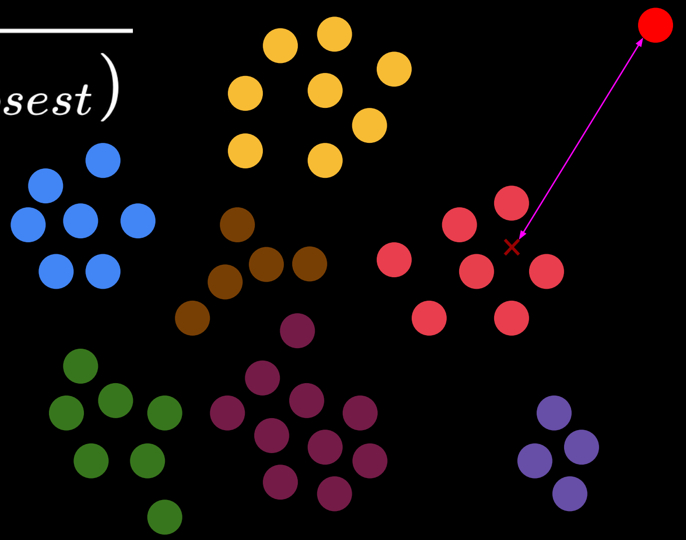




Out-of distribution detection

- Normalized distance:

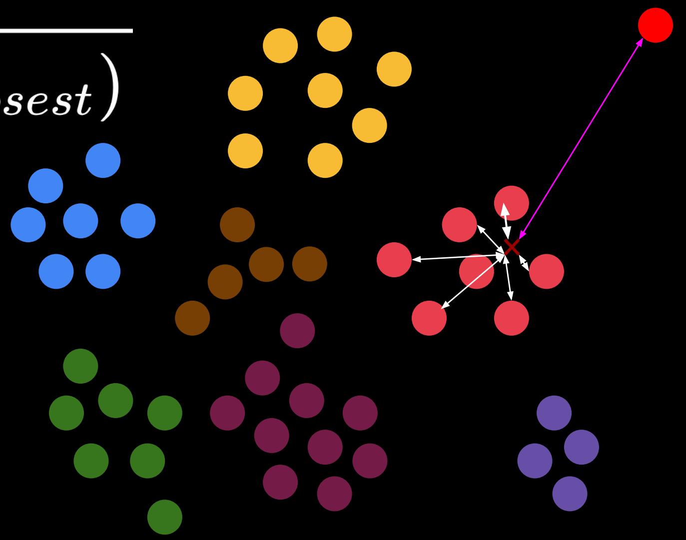
$$\frac{d\left(input, c_{closest}\right)}{\frac{1}{n_{closest}}\sum_{i=0}^{n_{closest}} d\left(x_i, c_{closest}\right)}$$



Out-of distribution detection

- Normalized distance:

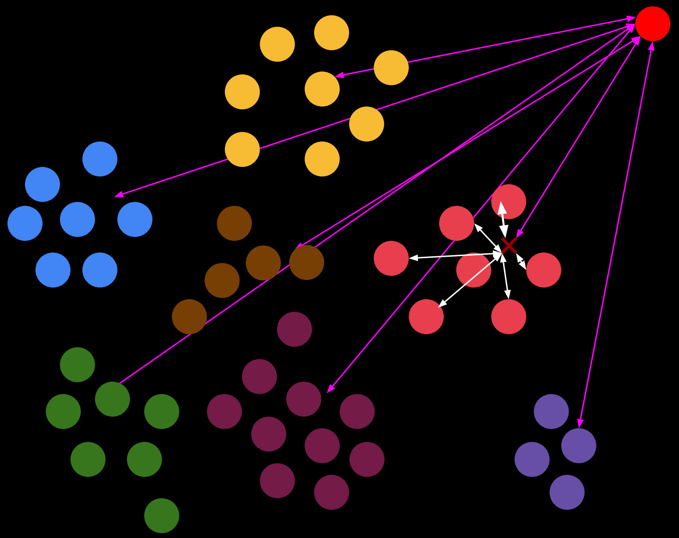
$$\frac{d\left(input, c_{closest}\right)}{\frac{1}{n_{closest}}\sum_{i=0}^{n_{closest}} d\left(x_i, c_{closest}\right)}$$



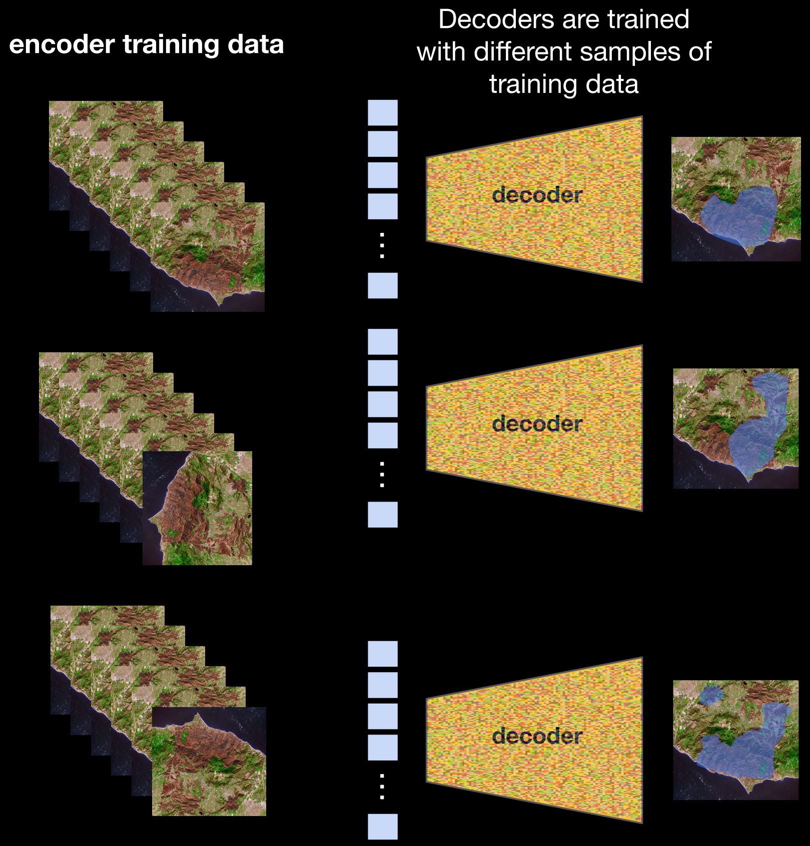
Out-of distribution detection

- Normalized distance
- Nearest Centroid Distance Deficit:

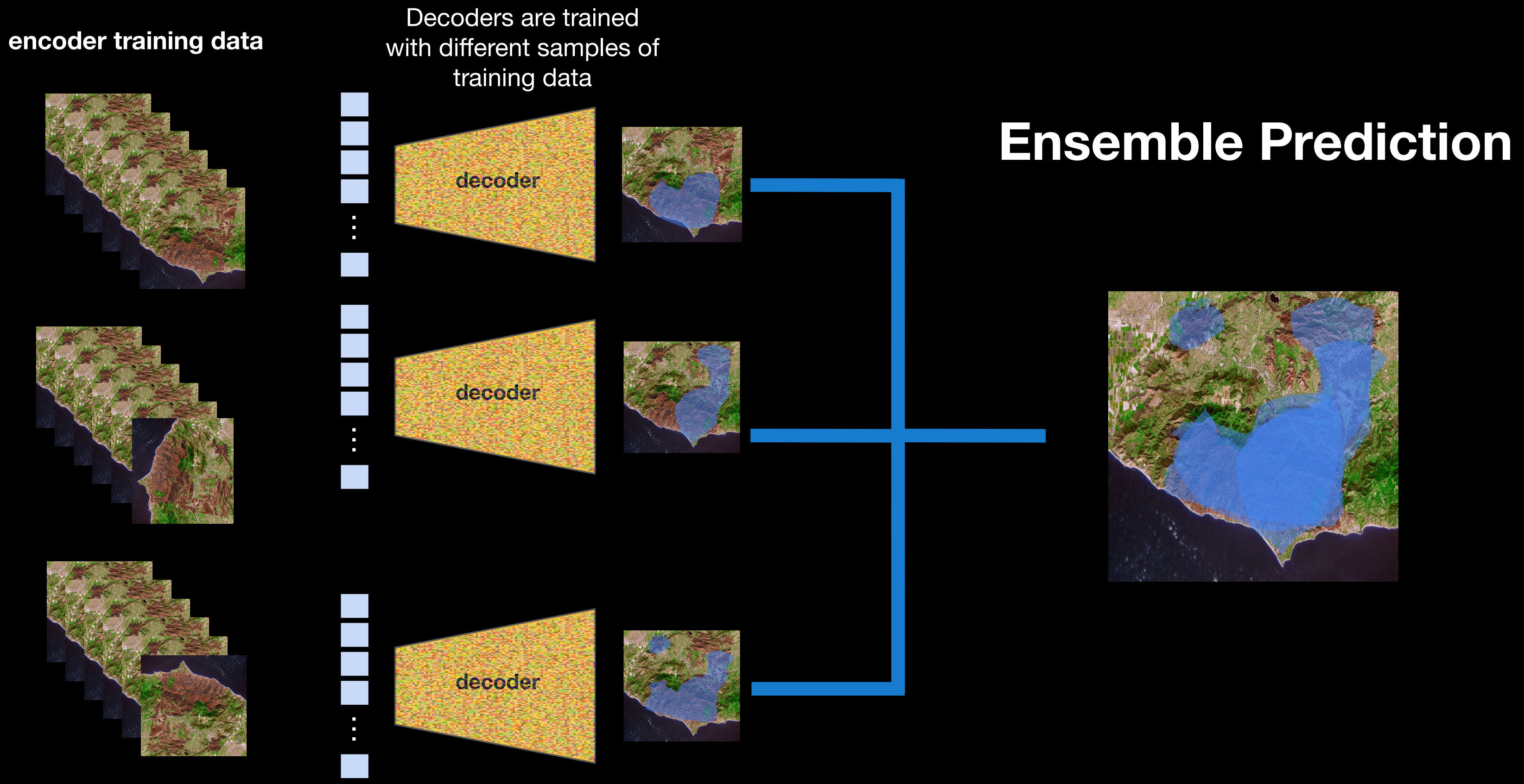
$$\alpha D_{others} - \beta \cdot D_{nearest}$$



3. The information is not enough to solve the task

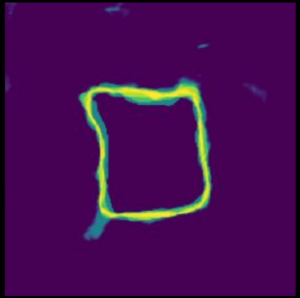


3. The information is not enough to solve the task



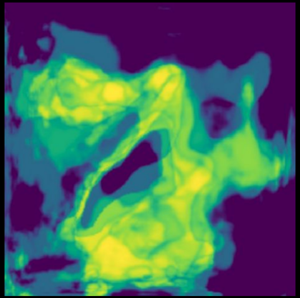
predicted

entropy



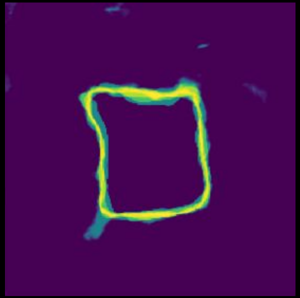
predicted

entropy



predicted

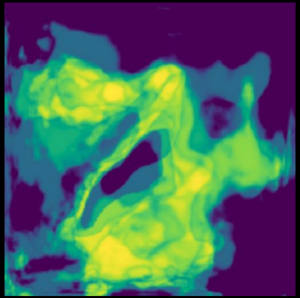
entropy

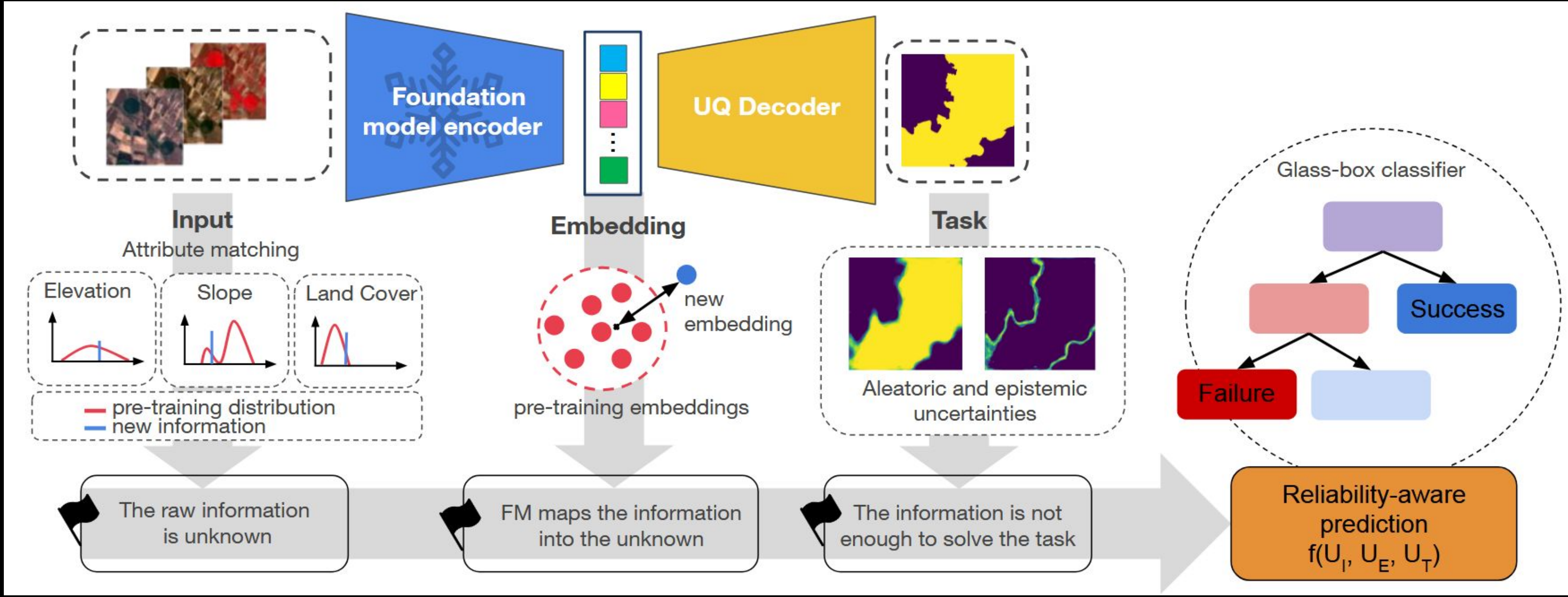


$$\frac{\text{Integral of entropy}}{\text{area of the interest}}$$

predicted

entropy




Foundation model encoder
UQ Decoder
Input
Attribute matching
Embedding
Task
Glass-box classifier
Elevation
Slope
Land Cover
new embedding
pre-training distribution
new information
pre-training embeddings
Aleatoric and epistemic uncertainties
Success
Failure
The raw information is unknown
FM maps the information into the unknown
The information is not enough to solve the task
Reliability-aware prediction
f(U_I, U_E, U_T)

Evaluating Selective Prediction

What constitutes a failure?

1. What is the application?

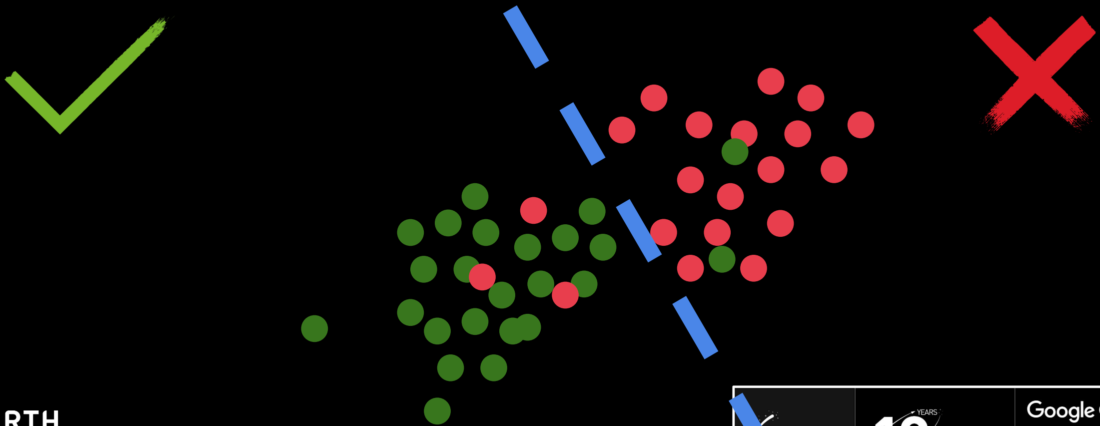
   Burn Scars Segmentation, Flood Segmentation,...

2. What is the specific usage?

   Damage Assessment, Media Information, Reforestation, Rapid Response,...

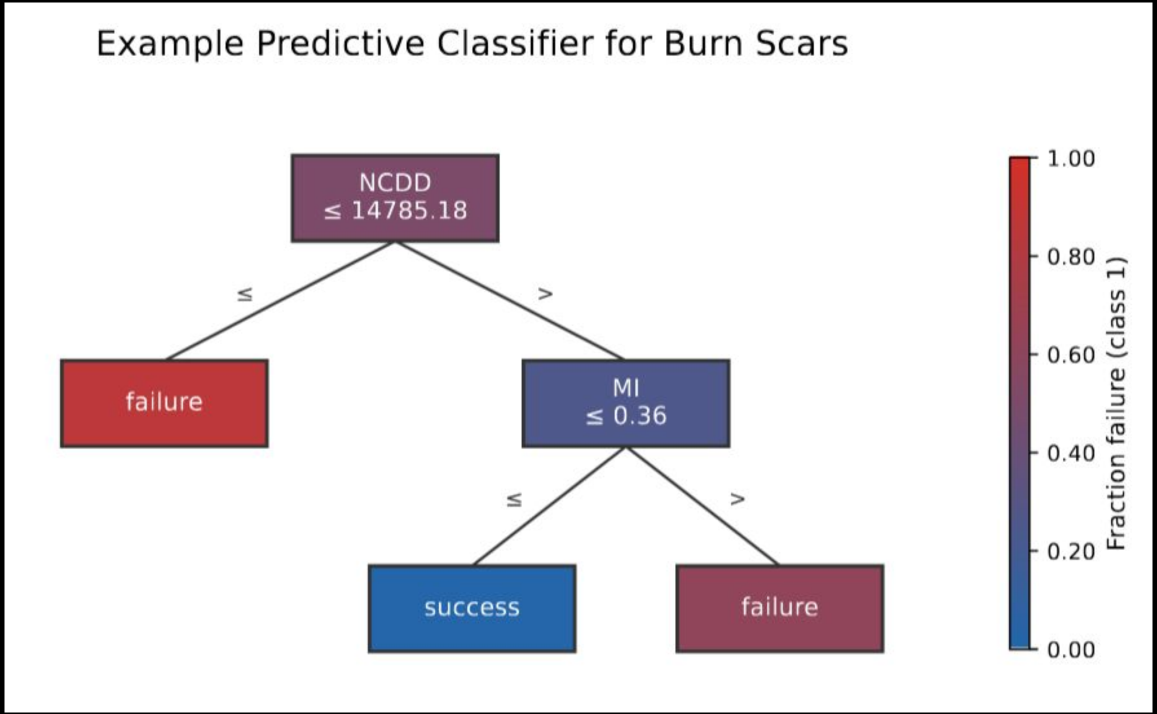
Selective Prediction as a Classification Problem

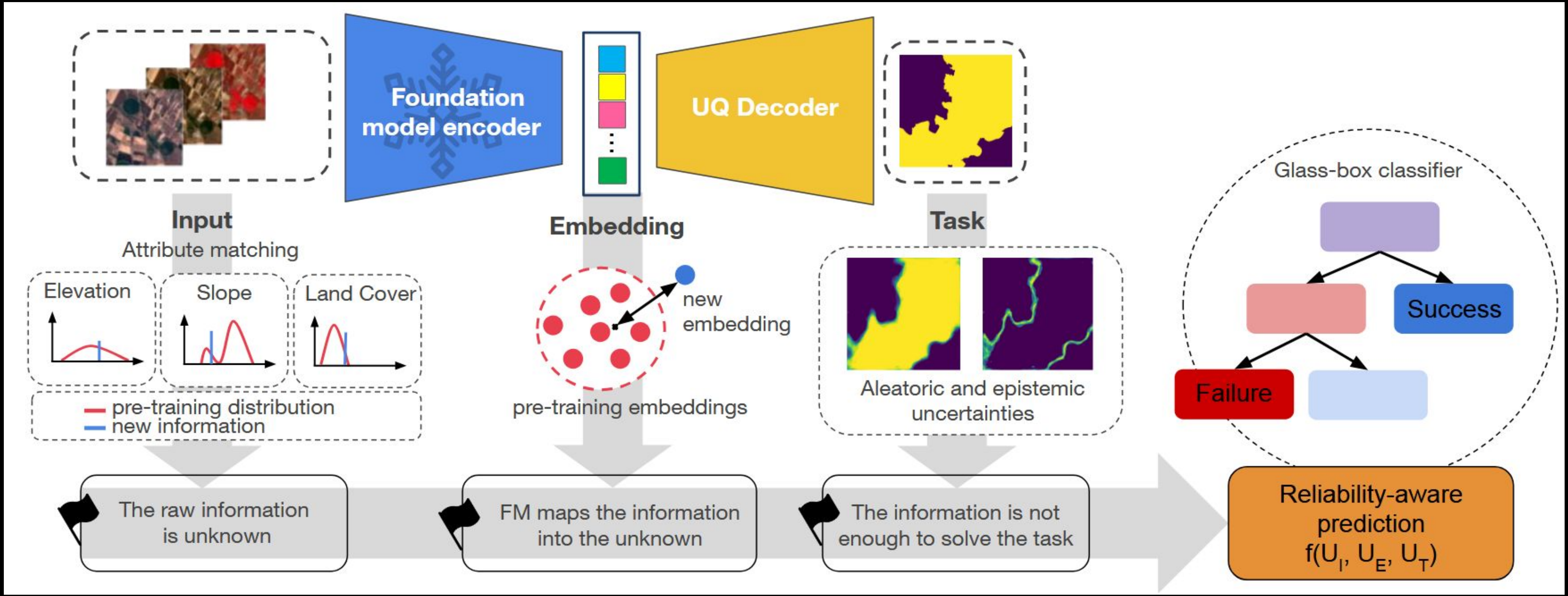
predicted | ground truth | predicted | ground truth



For each task:

1. Automatic feature selection
2. Hyperparameter tuning
3. Training of a final shallow tree for failure classification




Foundation model encoder
UQ Decoder
Glass-box classifier
Input
Attribute matching
Embedding
Task
Elevation
Slope
Land Cover
new embedding
Success
Failure
pre-training distribution
new information
pre-training embeddings
Aleatoric and epistemic uncertainties
The raw information is unknown
FM maps the information into the unknown
The information is not enough to solve the task
Reliability-aware prediction
f(U_I, U_E, U_T)

SHRUG FM, an adaptable framework for FM uncertainty

TRILLIUM EUROPE

https://2025-esl-extreme-environments-demo-rj7dyok9w.vercel.app

esa Φ-lab EARTH SYSTEMS LAB

FDL.AI NETWORK 10 YEARS OF APPLIED AI RESEARCH FOR ALL HUMANKIND Google Cloud NVIDIA SCAN AI UNIVERSITY OF OXFORD PASTEUR AD ASTRA PER ALGORITHMOS

ESLAB.AI

TRILLIUM EUROPE

ESLAB.AI

#makeFMsayIDK

esa Φ-lab

EARTH SYSTEMS LAB

FDL.AI NETWORK

10 YEARS OF APPLIED AI RESEARCH FOR ALL HUMANKIND

Google Cloud NVIDIA SCAN AI

UNIVERSITY OF OXFORD PASTEUR iSi

AD ASTRA PER ALGORITHMOS

FDL EARTH SYSTEMS LAB

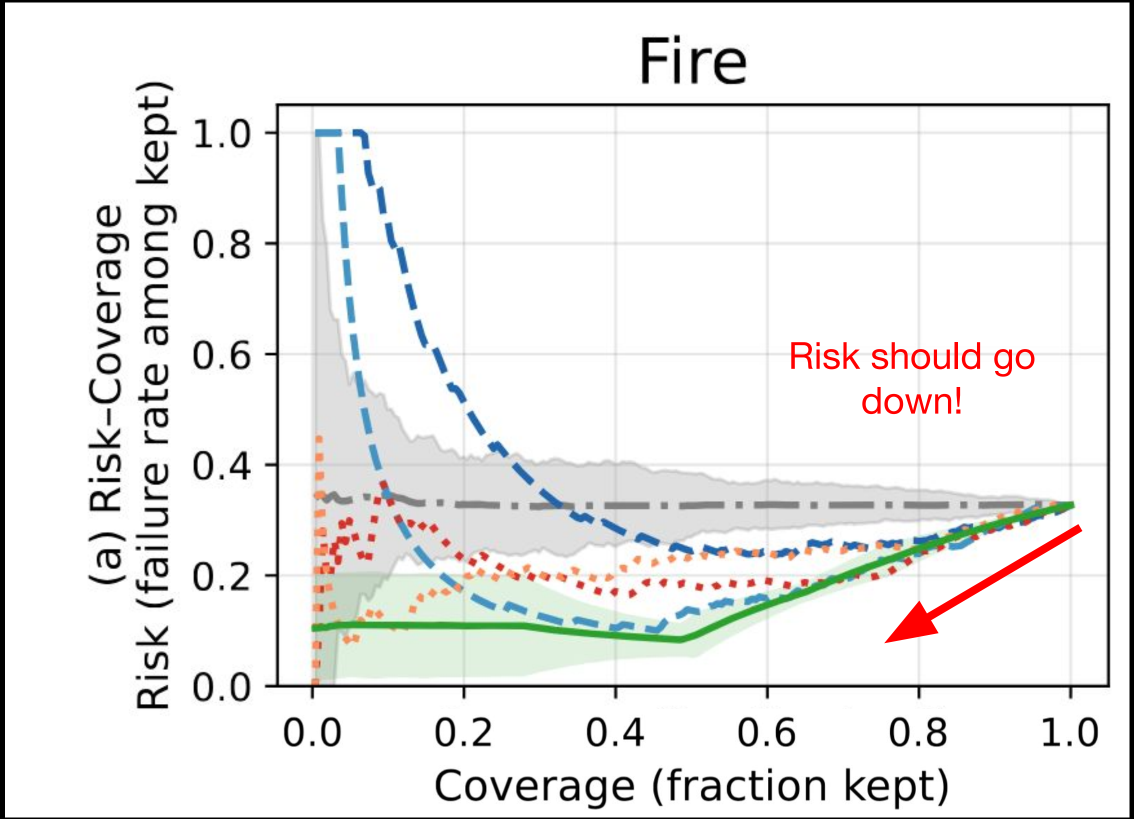
#makeFMsayIDK

SHRUG-FM

SYSTEMATIC HANDLING OF REAL-WORLD UNCERTAINTY FOR GEOSPATIAL FMs

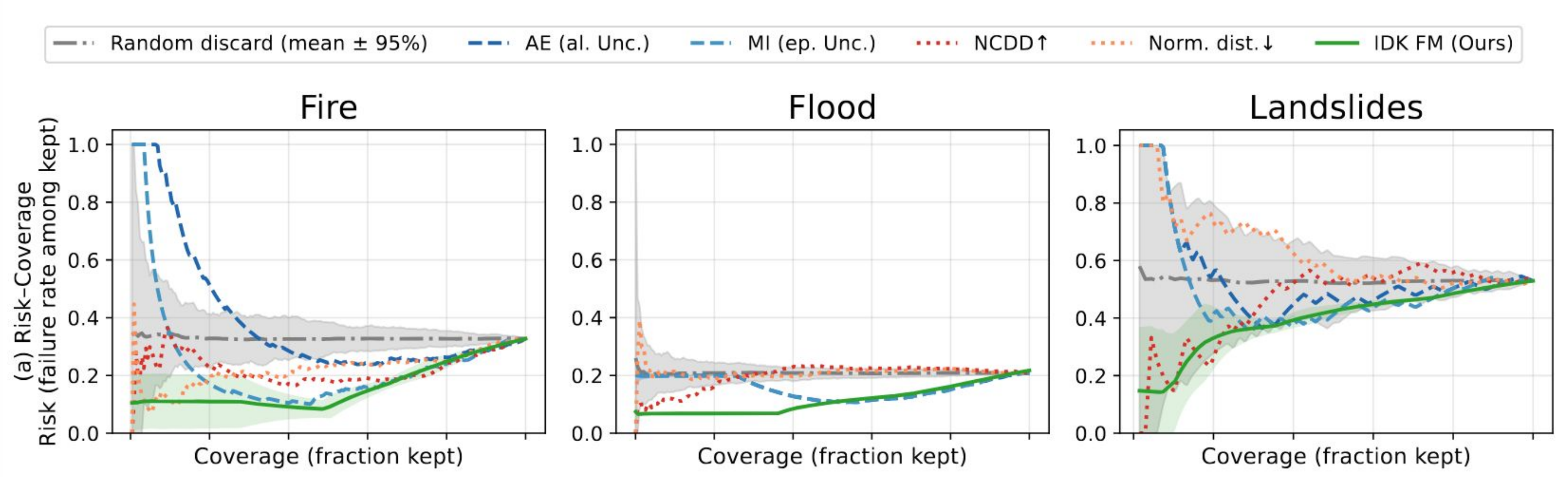
Risk-Coverage Curve

Should reduce the percentage of failures

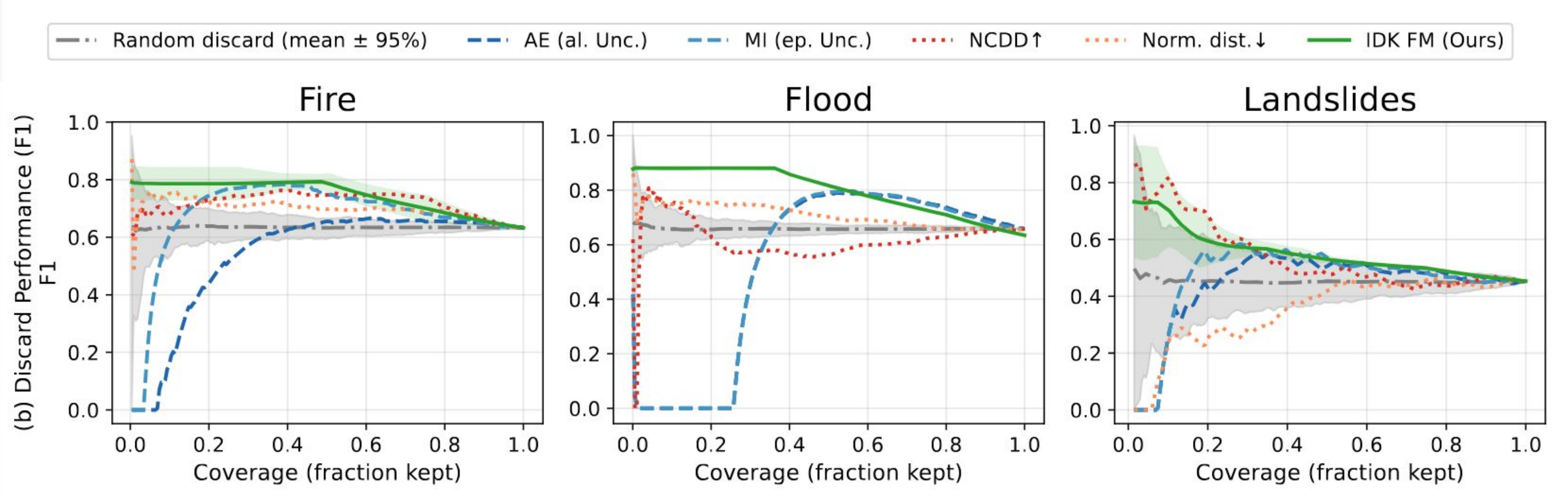


Dropping highest failure score samples

Results: Risk-Coverage Curve



Results: Discard Performance



Results: Discard Performance

	Fire			Flood			Landslides		
		Risk@			Risk@			Risk@	
Score	AUC↓	0.5↓	0.75↓	AUC↓	0.5↓	0.75↓	AUC↓	0.5↓	0.75↓
AE (al. U.)	0.391 ± 0.000	0.244 ± 0.000	0.259 ± 0.000	0.160 ± 0.001	0.110 ± 0.000	0.142 ± 0.000	0.526 ± 0.000	0.471 ± 0.000	0.508 ± 0.000
MI (ep. U.)	0.240 ± 0.000	0.137 ± 0.000	0.228 ± 0.000	0.159 ± 0.001	0.108 ± 0.000	**0.139 ± 0.000**	0.491 ± 0.000	**0.412 ± 0.000**	**0.471 ± 0.000**
NCDD↑	0.225 ± 0.000	0.183 ± 0.000	**0.203 ± 0.000**	0.198 ± 0.000	0.228 ± 0.000	0.223 ± 0.000	0.451 ± 0.000	0.529 ± 0.000	0.569 ± 0.000
Norm. dist.↓	0.224 ± 0.000	0.236 ± 0.000	0.253 ± 0.000	0.210 ± 0.000	0.211 ± 0.000	0.219 ± 0.000	0.614 ± 0.000	0.559 ± 0.000	0.510 ± 0.000
IDK-FM (Ours)	**0.160 ± 0.045**	**0.090 ± 0.038**	0.223 ± 0.029	**0.115 ± 0.001**	**0.107 ± 0.004**	0.149 ± 0.000	**0.392 ± 0.035**	0.425 ± 0.021	0.471 ± 0.000

Features chosen by Task (Over 30 runs with different seeds)

feature	Fire	Flood	Landslide
Average_Entropy (0.05)	—	100.0%	16.7%
Average_Entropy (0.3)	90.0%	—	—
Mutual Information (0.05)	—	100.0%	100.0%
Mutual Information (0.1)	100.0%	—	—
density	53.3%	100.0%	—
ncdd	30.0%	—	70.0%
normalized_distance	—	—	100.0%
ria_ha_ssu	—	—	3.3%

	Classifier ROC AUC↑			Risk–Coverage AUC↓		
	Fire	Flood	Landslides	Fire	Flood	Landslides
All	0.791 ± 0.0635	0.725 ± 0.00378	0.654 ± 0.0263	0.16 ± 0.0435	0.114 ± 0.00124	0.392 ± 0.0329
No Input Feature Signals	$\mathbf{0.84 \pm 0.00944}$	$\mathbf{0.739 \pm 0.00106}$	$\mathbf{0.661 \pm 0.0382}$	$\mathbf{0.13 \pm 0.0105}$	0.104 ± 0.000206	$\mathbf{0.381 \pm 0.0385}$
No Embedding OOD Signals	0.77 ± 0.0224	0.724 ± 0.00365	0.622 ± 0.0373	0.173 ± 0.0258	0.114 ± 0.0012	0.423 ± 0.0823
No UQ (MI, AE) Signals	0.706 ± 0.0242	0.524 ± 0.024	0.548 ± 0.0248	0.227 ± 0.0282	0.207 ± 0.00605	0.499 ± 0.00608
UQ only	0.8 ± 0.00788	0.739 ± 0.000477	0.633 ± 0.0104	0.138 ± 0.0102	$\mathbf{0.104 \pm 0.000229}$	0.412 ± 0.039
Dist only	0.739 ± 0.0465	0.495 ± 0.00921	0.562 ± 0.0181	0.198 ± 0.0382	0.21 ± 0.00418	0.494 ± 0.00732
Input only	0.473 ± 0.0556	0.533 ± 0.0056	0.417 ± 0.0529	0.351 ± 0.0404	0.205 ± 0.00193	0.604 ± 0.0495