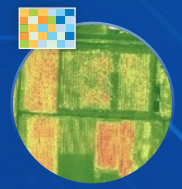
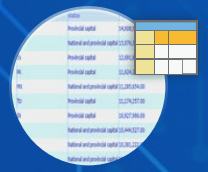
GeoAI in the age of Foundation Models

Rohit Singh
Director, Esri R&D Center New Delhi

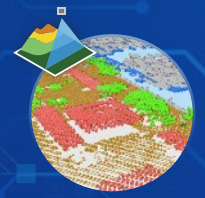
GeoAI



Imagery



Tabular



Lidar / Point Cloud

Vector

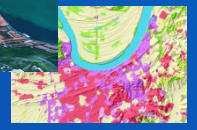
Unstructured

Full-Motion Video

3D

Extract Features

Building and Road Extraction



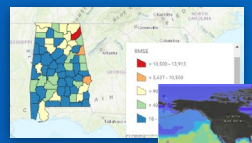
Land Cover Classification

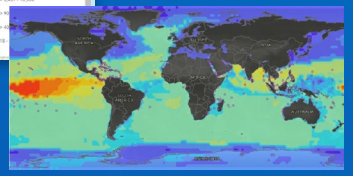
3D Object Detection

Point Cloud Classification

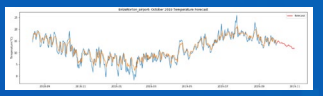
Make Predictions

Regression & Classification





Climate Forecasting



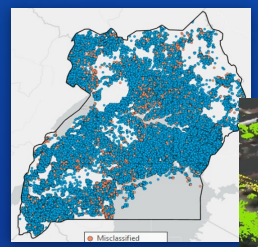
Time Series Forecasting

Unlock Data

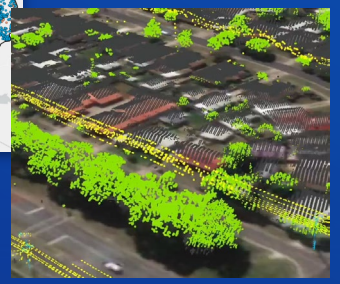
Entity Extraction

Data Extraction from PDFs

Automate Analysis

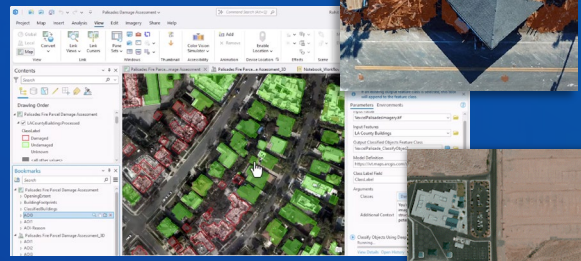


AutoML & AutoDL

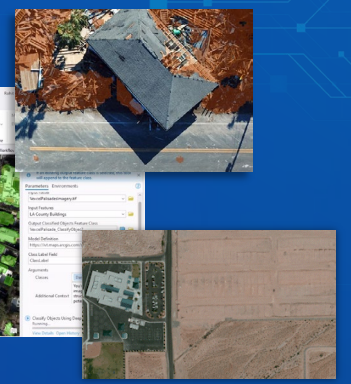


Vegetation Encroachment Analysis

Derive Insights



Damage Assessment



Change Detection

AI

Models pretrained and ready-to-use…

Tools to train and use models…

AI Models

Pretrained & Ready-to-Use

Task-specific Models

- Agricultural Fields
- Buildings
- Cars
- Cloud Masks
- Common Objects
- Cooling Towers
- Crowd Counting
- Face Blurring
- Humans
- Insulator Defects
- Land Cover
- License Plates
- Mangroves
- Map Simplification
- Object Tracking
- Oil Spills
- Oil Tanks
- OCR
- Palm Trees
- Parcels
- Parking Lots
- Pavement Cracks
- Pedestrian Infrastructure
- Plant Leaf Disease
- Swimming Pools
- Power Lines
- Pylons
- Roads
- Scene Text Parsing
- Ships / Shipwrecks
- Solar Panels / Parks
- Trees
- Water Bodies
- Well Pads
- Wildfires
- Wind Turbines

Foundation Models

- CLIP Zero-Shot Classifier
- Depth Anything
- GroundingDINO
- Prithvi
- Prompt-Based Segmentation
- Segment Anything Model (SAM)
- Text SAM
- Time Series Forecasting (Time-MoE)

Integration

Hugging Face

- Image Inpainting
- Object Classification & Detection
- Pixel Classification
- Text Classification & Translation
- Text & Visual QA

Cloud AI Services

- Image Interrogation
- Vision Language Context-Based Classification

Land Cover

Trees

Building Footprints

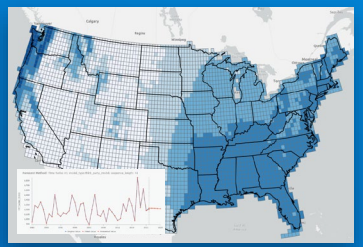
Crop Classification (Prithvi)

SAM3 (Esri NZ)

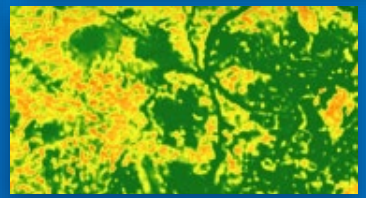
Pedestrian Infrastructure Classification

Cars (Text SAM)

HF Visual Question Answering



Time-MoE



Canopy Height Estimation (Meta)

Global Roads

. . . Available as deep learning packages on the Living Atlas

AI Tools & Models

Advancing the Science of GIS Using AI

GIS DATA

Object Detection

Land Cover Classification

Feature Classification

Regression

Object Tracking

Pixel Classification

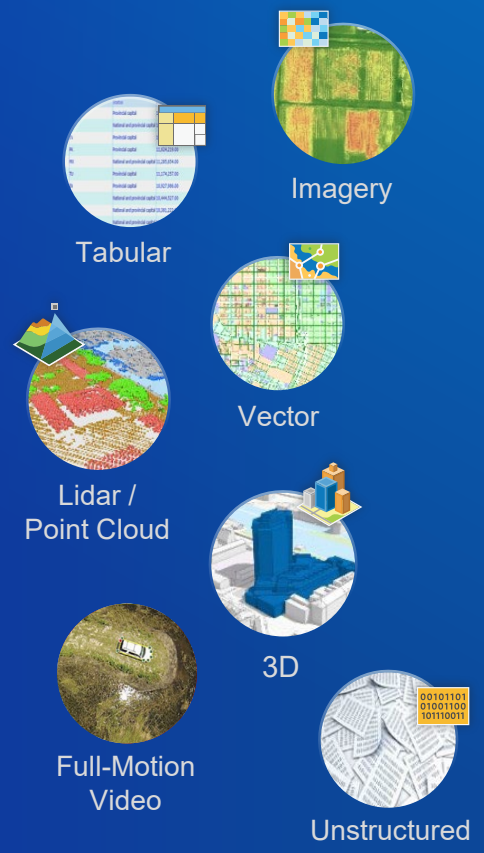
Object Counting

Entity Extraction

Text Classification

GIS TASKS

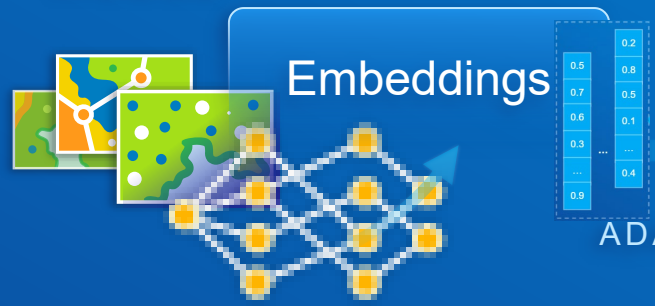
Geospatial Foundation Models



EXPANSIVE
GIS DATA

TRAINING

FOUNDATION
MODEL



Large-scale
Neural Network

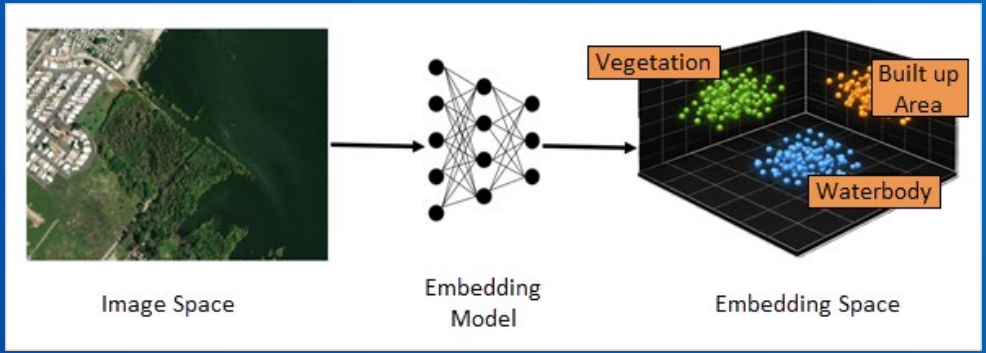
ADAPTATION

Object Detection
Land Cover Classification
Feature Classification
Regression
Object Tracking
Pixel Classification
Object Counting
Entity Extraction
Text Classification
Similarity Search
Change Detection

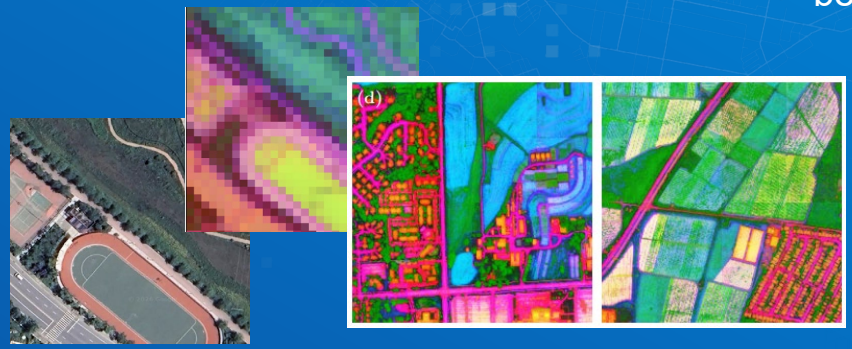
RANGE OF
GIS TASKS

Remote Sensing Foundation Models
Location Encoder Models
Geospatial Vision Language Models

What are Embeddings?



Imagery embeddings

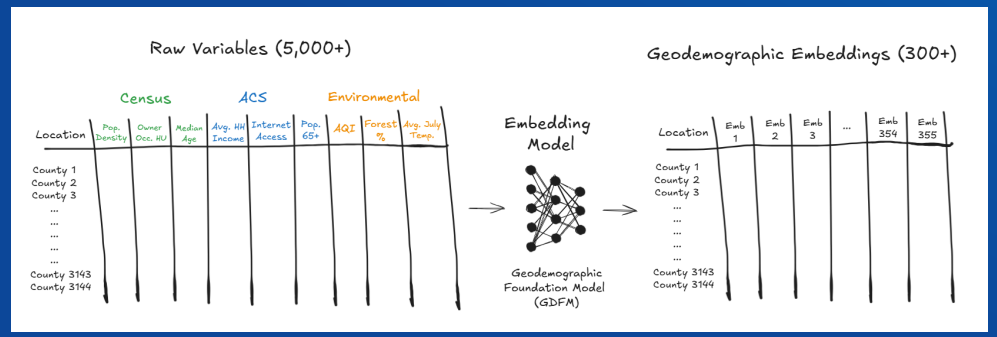


Imagery embeddings capture geographic features

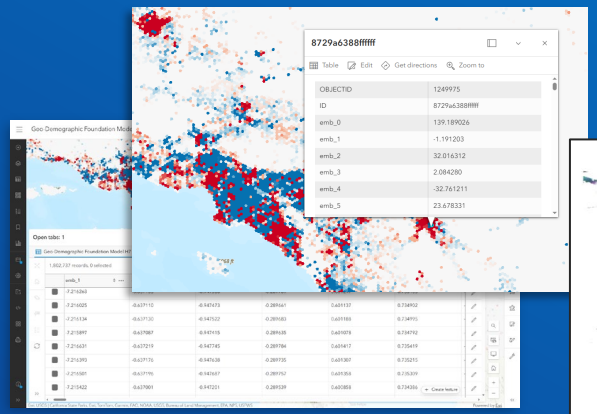
Represent the essence of a place, beyond it's coordinates

Enable better predictive modeling

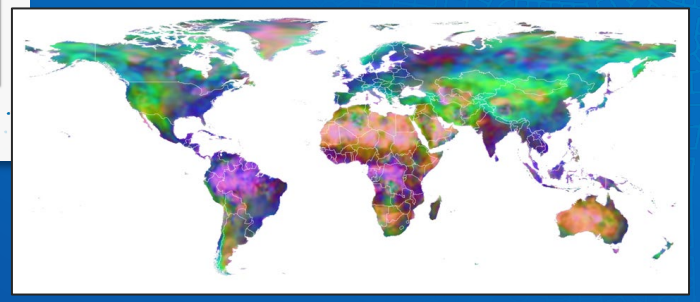
Power advanced geospatial tasks



Location embeddings



Location embeddings encode key characteristics of places

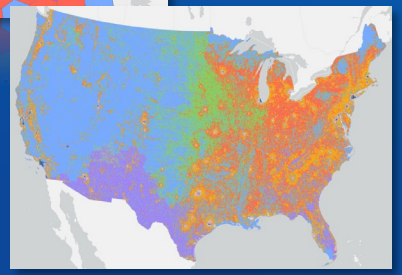


Visualization of Global Location Embeddings

What can you do with Embeddings?

Clustering

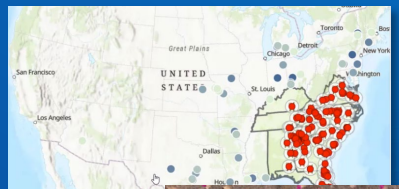
Lead exposure risk clusters



Geodemographic clusters
(K means) for all 1.8 million bins

Similarity Search

New restaurant locations

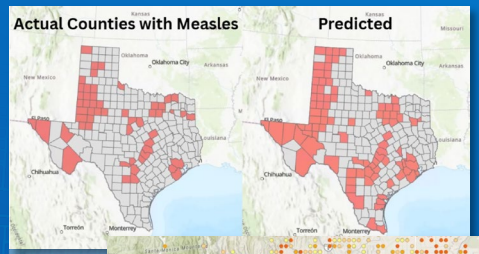


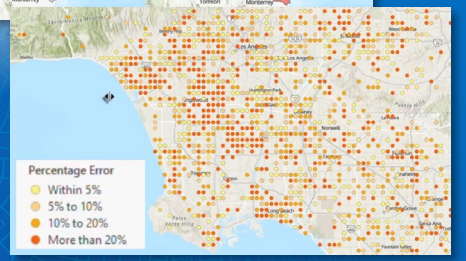
Agricultural pits

"Copper mines" – TerraByte AI

Enhance Predictions

Measles outbreaks



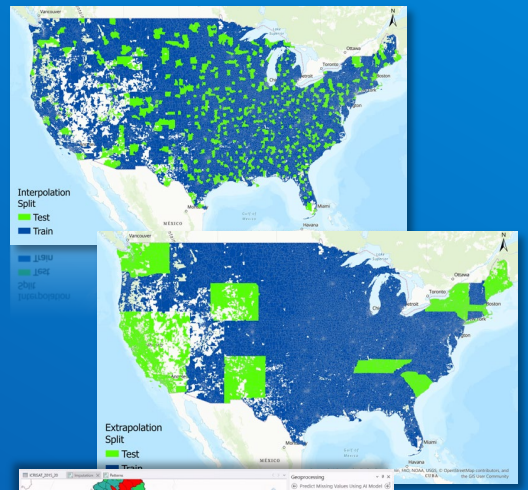


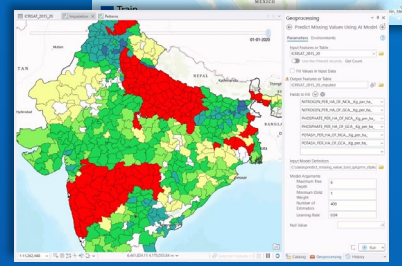
House price prediction

Change Detection - DeltaBit

Fill Missing Values

Interpolation



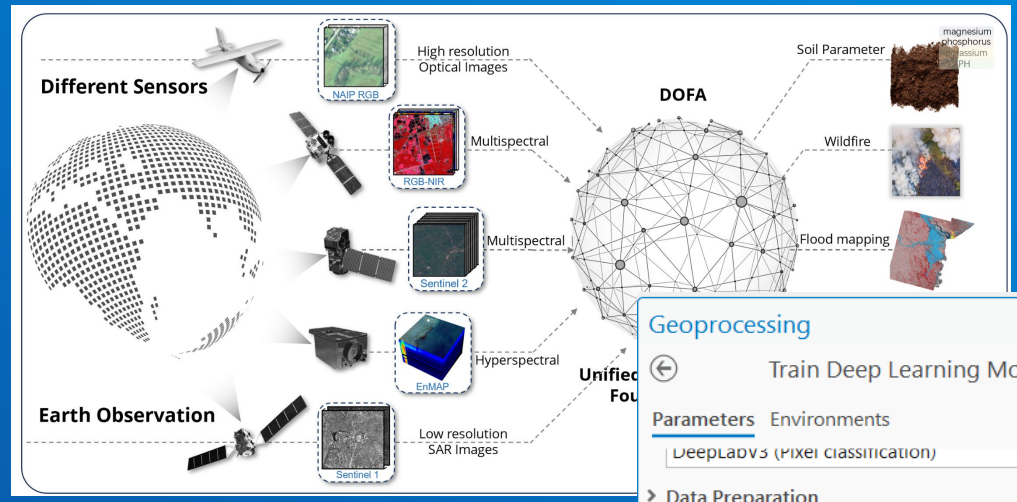


Extrapolation

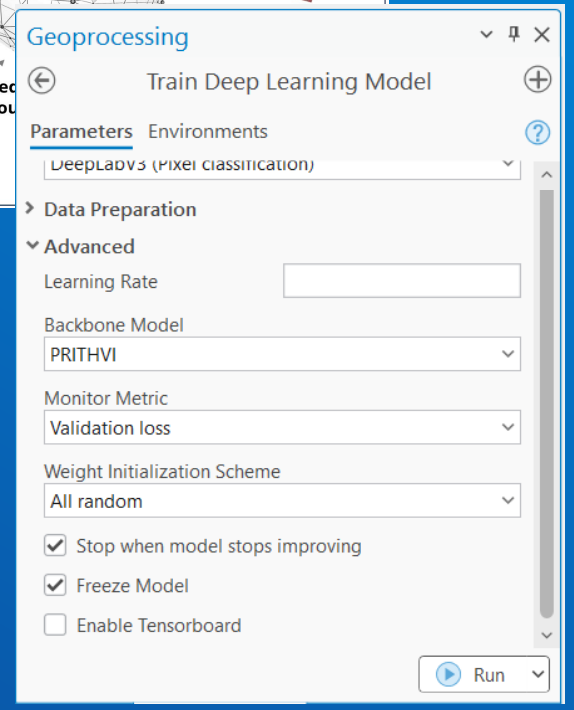
Geospatial Foundation Models

Remote Sensing Foundation Models

- Higher accuracy for computer vision tasks on remote sensing imagery
- Applications:
  - Backbones
  - Embeddings
- Models:
  - Prithvi EO 2.0, Clay FM, DOFA, TeeraMind, Tessera, RAMEN (Multispectral)
  - DinoV3 (RGB)
  - Multispectral + RGB…?



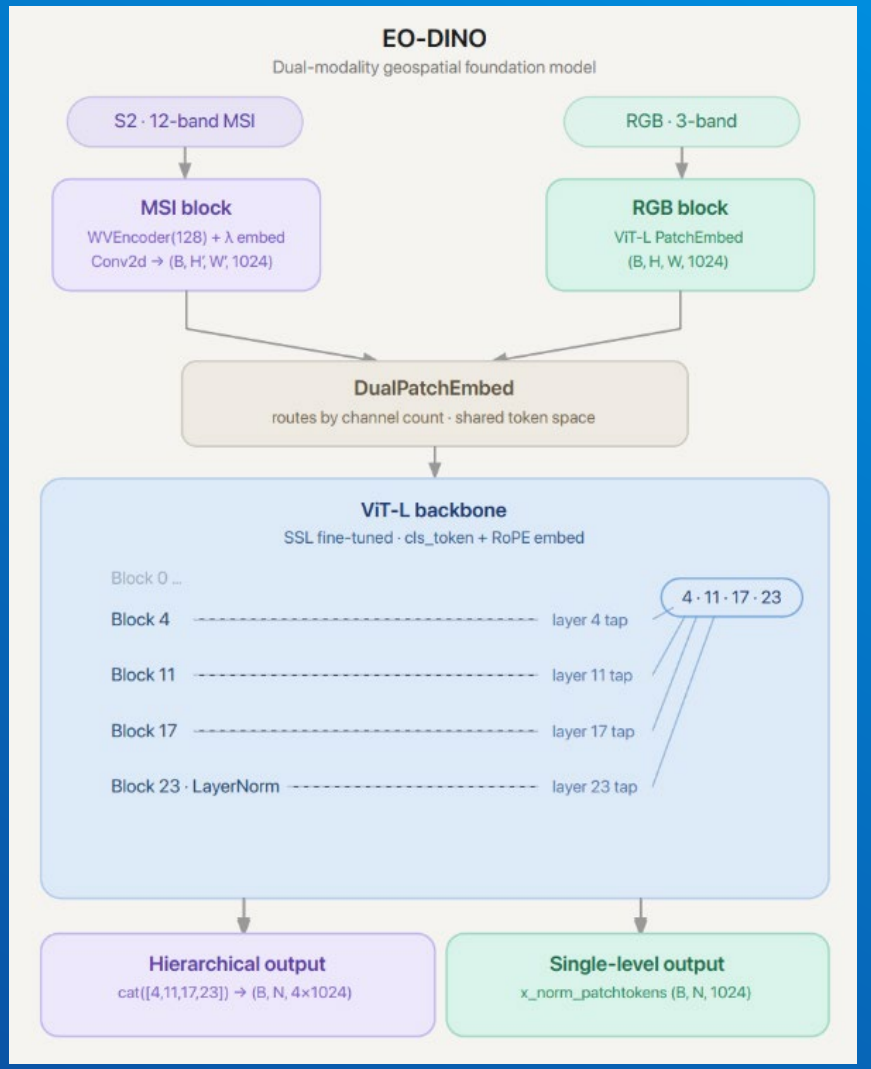
Dynamic-one-for-all (DOFA) model



Integrated as backbones

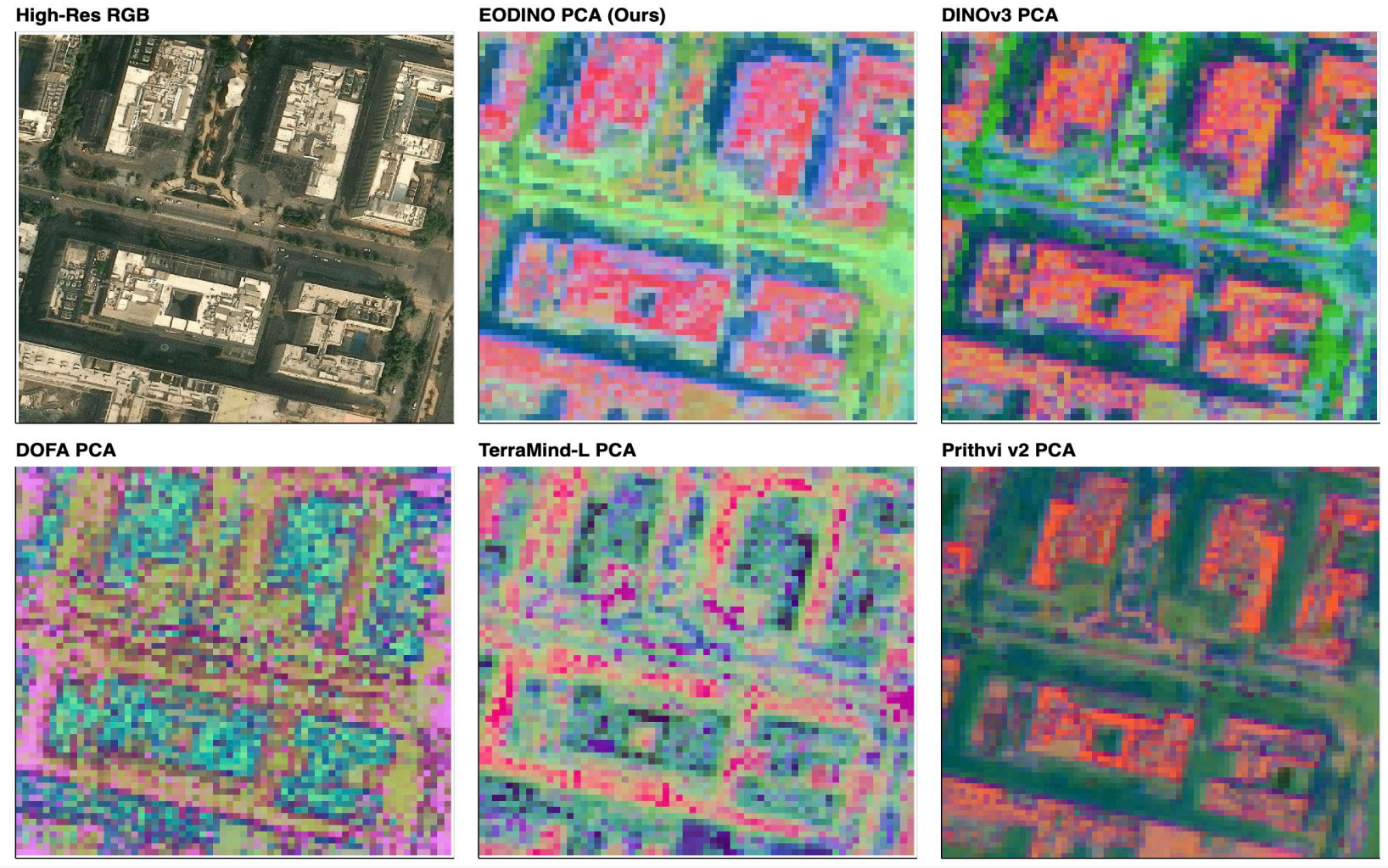
Esri's EO-DINO* = DINO + DOFA

- Architecture & losses
  - MSI Block (DOFA style)
  - RGB Block (patchify Layer)
  - Traditional ViT-L
  - DINO SSL Training & Losses
- Training data
  - Sentinel-2 (10TB) + Hi-res RGB (1TB)

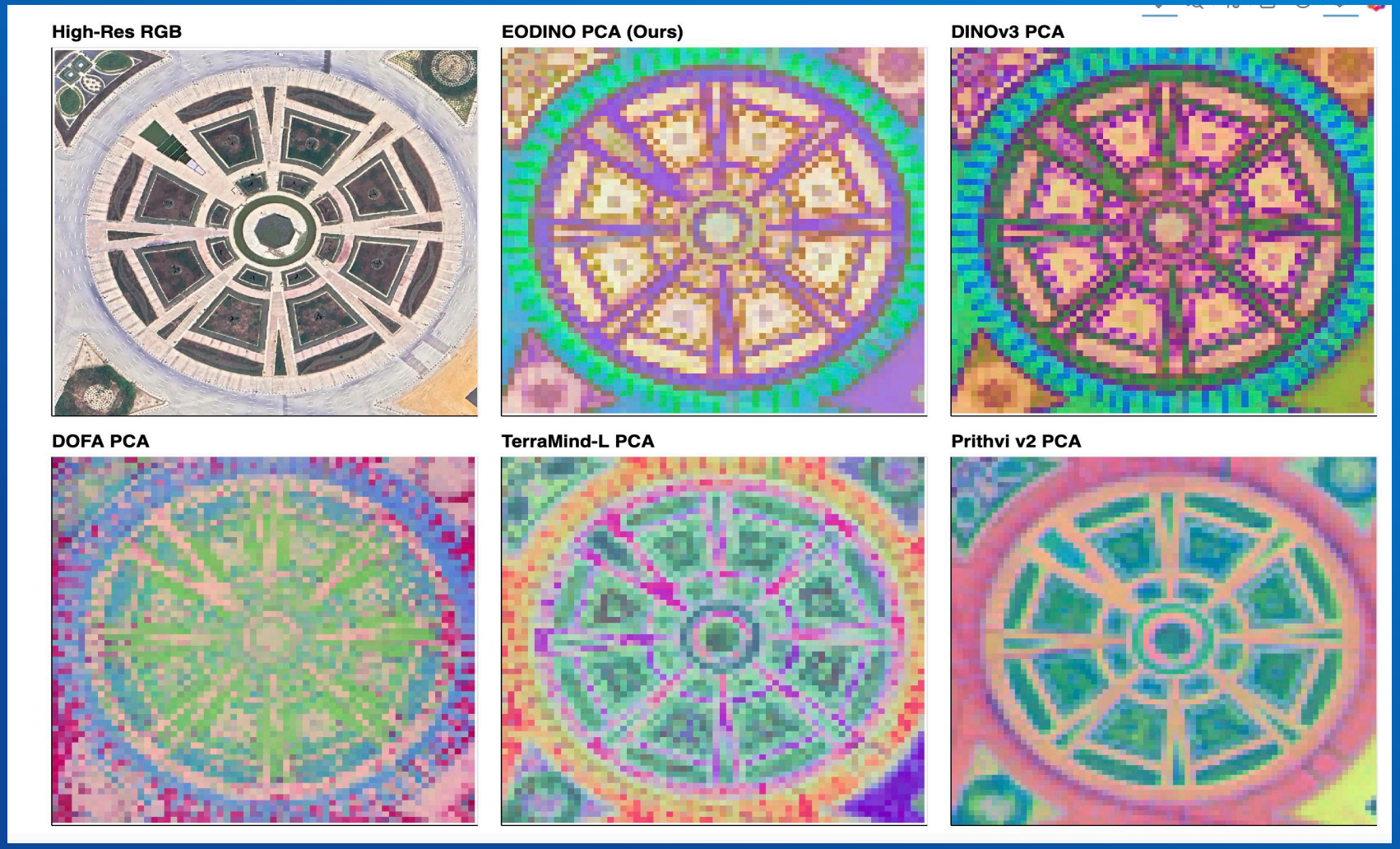


93K	~1M	12	10m	0.44m	10.8 TB
S2 patches	temporal frames	spectral bands	S2 resolution	RGB resolution	total disk
~1M temporal frames	across all patches	Sentinel-2 MSI	spatial GSD	WorldImagery	9.6 + 1.2 TB

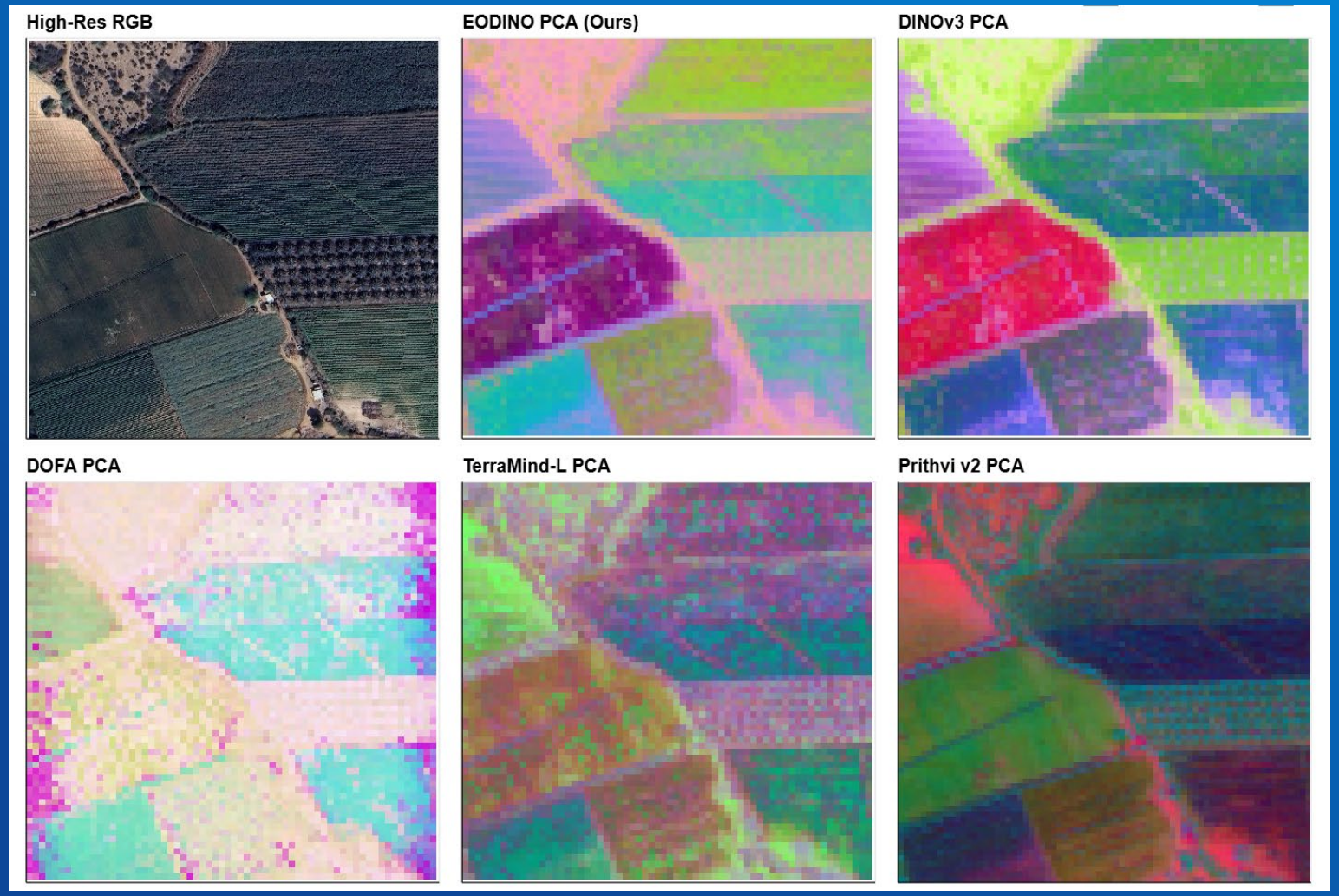
PCA Visualization



PCA Visualization

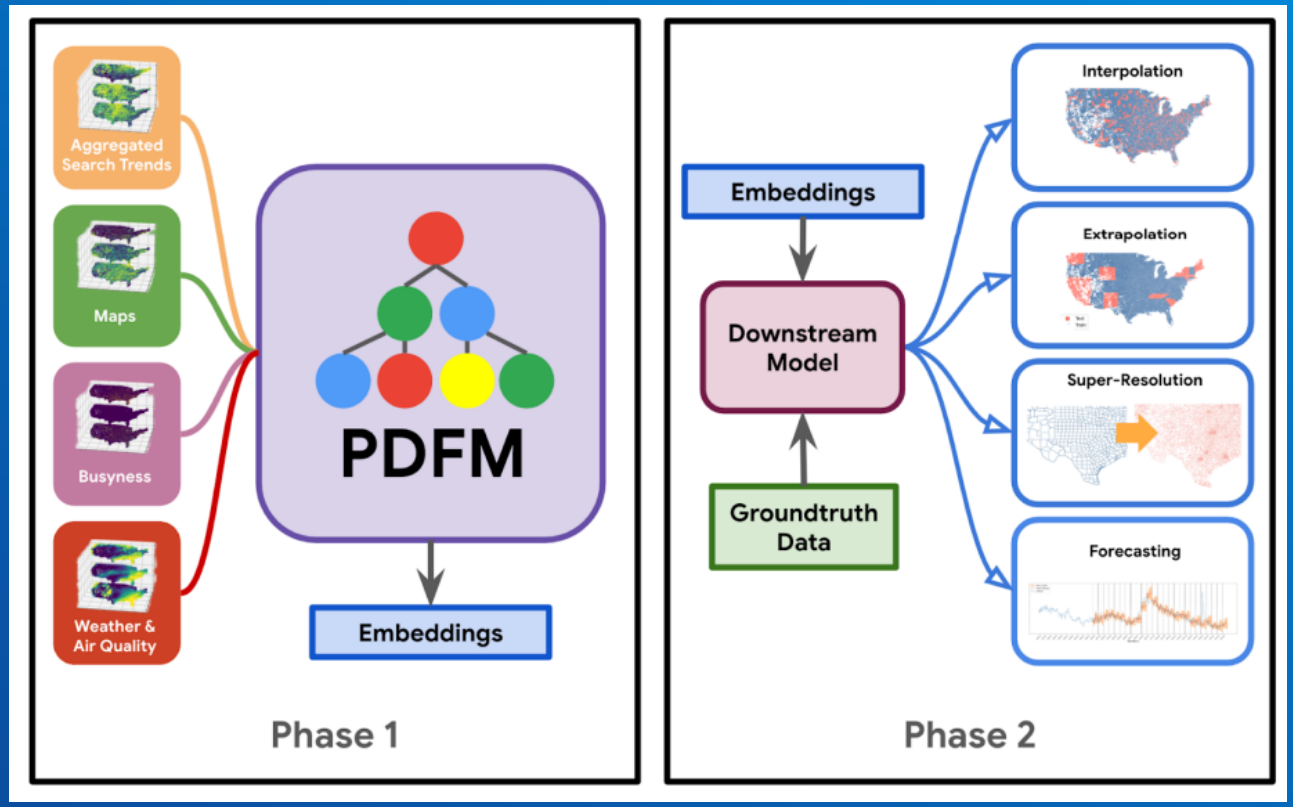


PCA Visualization

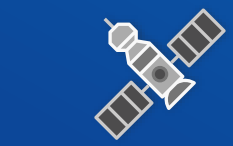


Location Encoder Models

- Encode geographic coordinates into rich feature representations
  - Capture spatial context, human - environment interactions, and geographic patterns
  - Improve socioeconomic predictions, health and environmental domains
- Examples:
  - SatCLIP
  - Population Dynamics FM (PDFM)
- Technical details:
  - Contrastive learning or GNNs

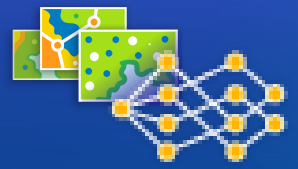


Esri’s Global Location Encoder (Sentinel-2)

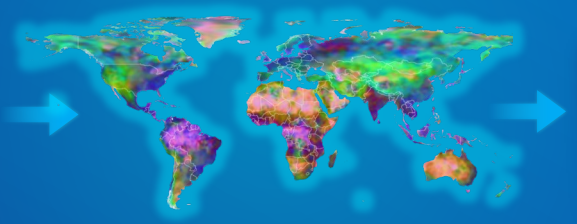


Sentinel-2
Satellite imagery

Global Location Encoder
(Sentinel-2)



SatCLIP architecture



Global Location Embeddings

User Data

Tools /
Models

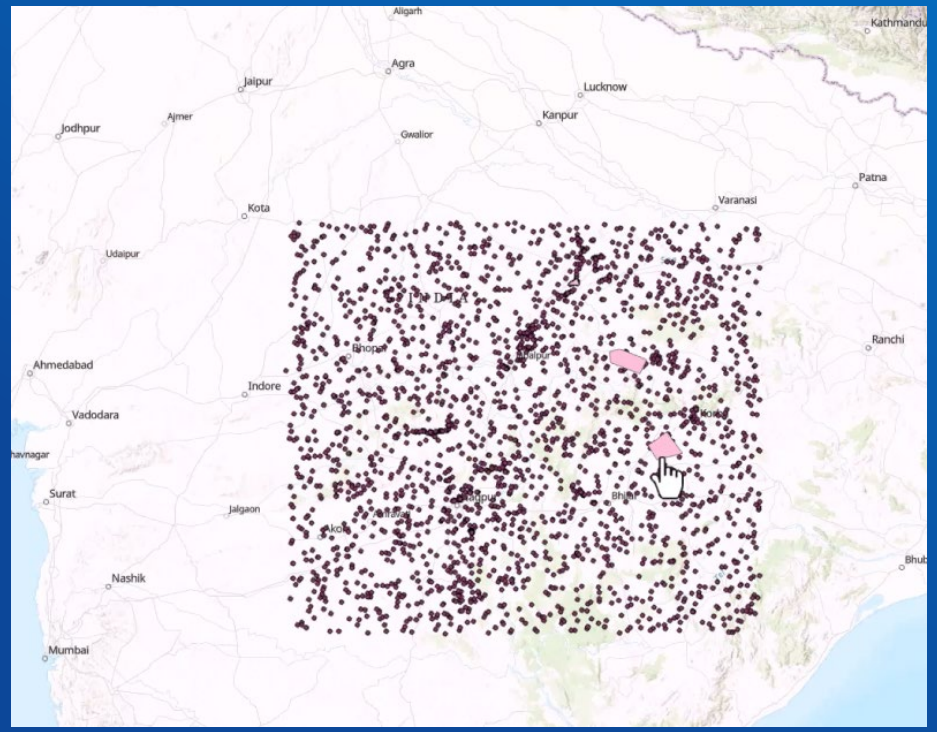
Similarity Search

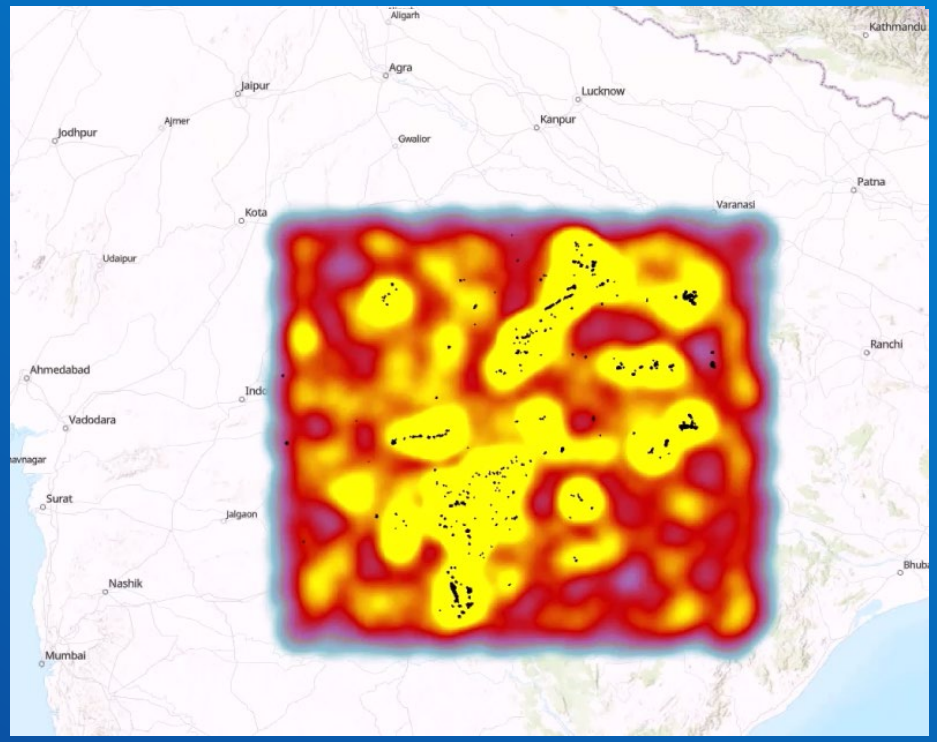
Change Detection

Classification

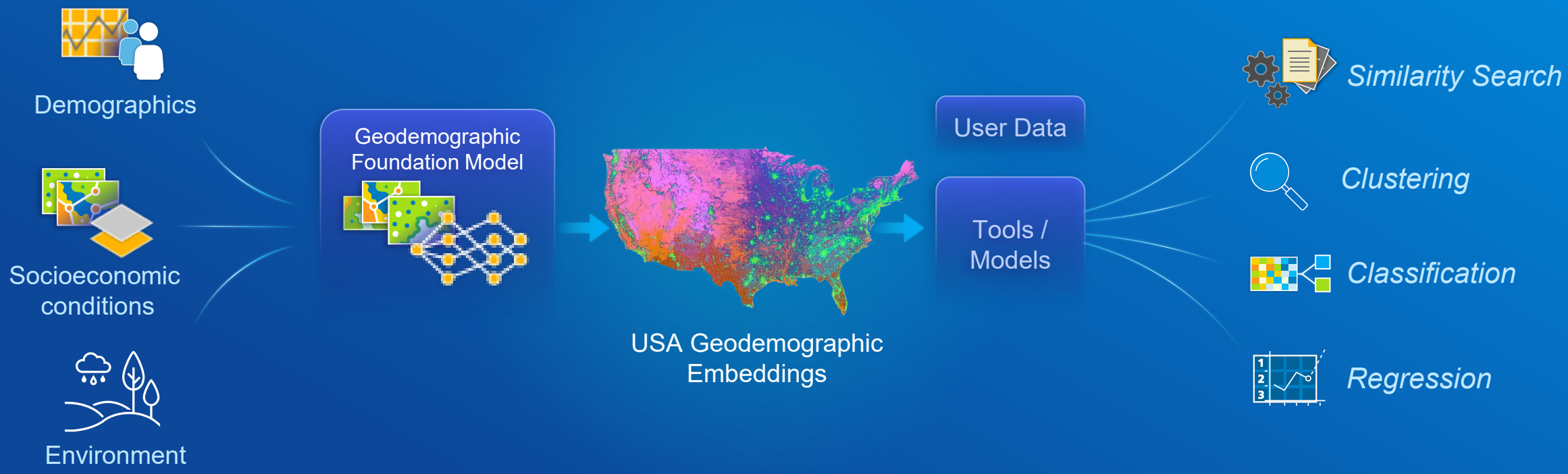
Regression

Finding mining areas using location embeddings



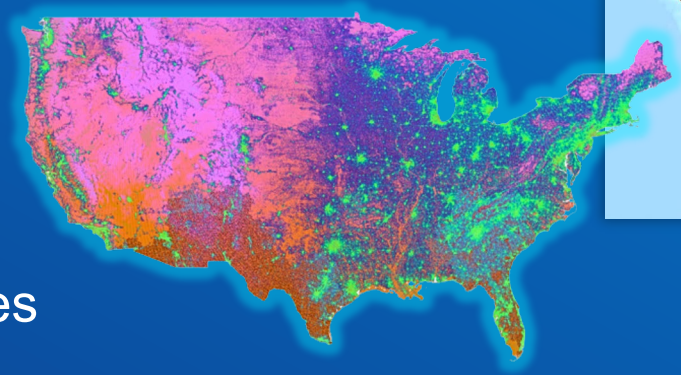


Esri’s Geodemographic Foundation Model (USA)

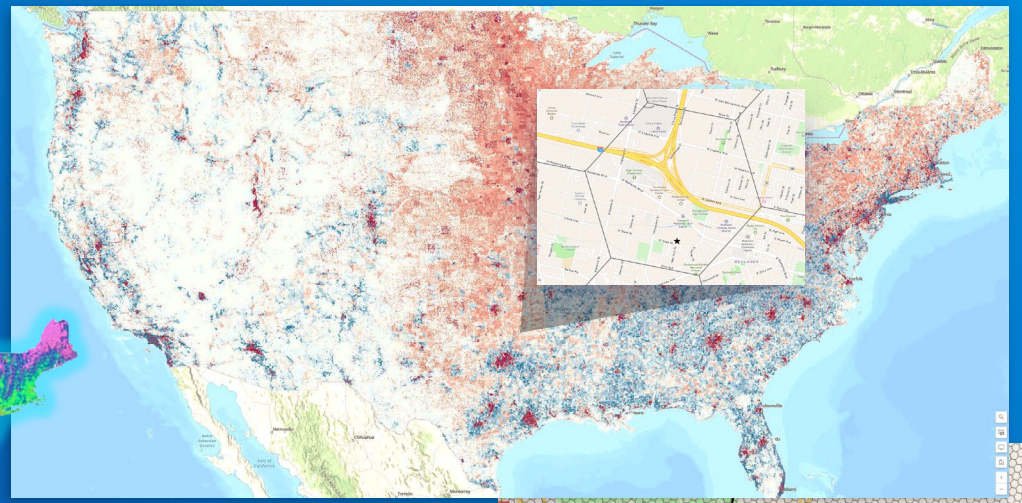


USA Geodemographic Embeddings

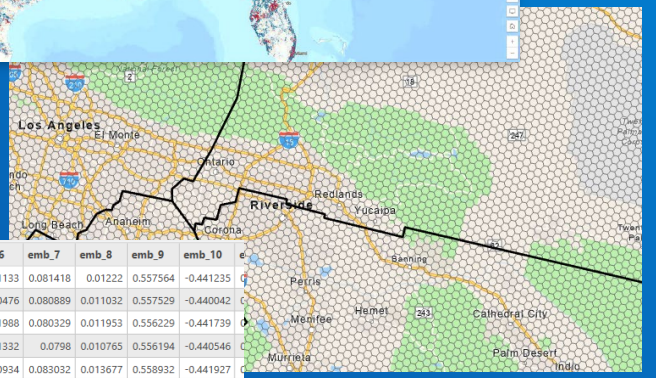
- Available as **beta** feature layer at no cost to ArcGIS Online subscribers (Esri UC)
- 5000+ Input variables
  - US Census
  - ACS
  - Environment
- 300+ embedding attributes



Embeddings capture geodemographic and environmental patterns (PCA)



1.8 million H3 level 7 hex bins with 300+ attributes



ID	emb_0	emb_1	emb_2	emb_3	emb_4	emb_5	emb_6	emb_7	emb_8	emb_9	emb_10
870c00000fffffff	-0.132687	0.266221	0.810085	0.147747	0.291089	0.048743	-0.601133	0.081418	0.01222	0.557564	-0.441235
870c00001fffffff	-0.132573	0.264013	0.809107	0.147624	0.291736	0.049013	-0.600476	0.080889	0.011032	0.557529	-0.440042
870c00002fffffff	-0.13472	0.267533	0.811214	0.147918	0.290712	0.04925	-0.601988	0.080329	0.011953	0.556229	-0.441739
870c00003fffffff	-0.134607	0.265327	0.810234	0.147795	0.291358	0.04952	-0.601332	0.0798	0.010765	0.556194	-0.440546
870c00004fffffff	-0.130766	0.26712	0.809935	0.147697	0.290819	0.047967	-0.600934	0.083032	0.013677	0.558932	-0.441927
870c00005fffffff	-0.130654	0.264909	0.808957	0.147574	0.291467	0.048239	-0.600276	0.082506	0.012486	0.558899	-0.440731
870c00006fffffff	-0.132798	0.268432	0.811065	0.147869	0.290441	0.04847	-0.601791	0.081945	0.013411	0.557597	-0.44243
870c00008fffffff	-0.134378	0.260921	0.808275	0.147545	0.292652	0.050052	-0.600024	0.078736	0.008393	0.556116	-0.438162
870c00009fffffff	-0.134261	0.258723	0.807297	0.147416	0.2933	0.050314	-0.599371	0.078202	0.007209	0.556074	-0.436972
870c0000afffffff	-0.136414	0.262236	0.809399	0.147717	0.292272	0.050563	-0.60088	0.077646	0.008127	0.554781	-0.438668

0 of 1,802,737 selected Filters:

Applications of USA Geodemographic Embeddings

- Improve predictions
- Similarity Search
- Clustering
- Fill Missing Values
  - Interpolation
  - Extrapolation

Health

Predict prevalence of
- Diabetes
- Stroke
- Cancer
- Heart disease
- Chronic kidney disease
- Asthma
- High blood pressure
- Cholesterol
- Arthritis
- Obesity
- Mental health
- Sleep disorders

Socioeconomic

Predict
- Median household income
- Median home value
- Voting pattern
- Population
- Night lights

Environment

Predict
- Ice storm risk
- Drought risk
- Hurricane risk
- Heatwave risk
- Evapotranspiration
- Elevation
- Forest cover

Downstream tasks used for evaluation

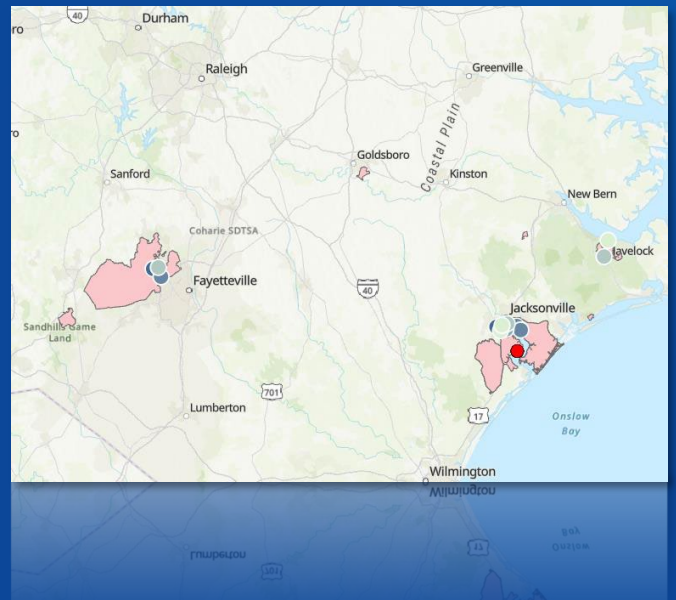
…embeddings boost models where demographics are a genuine driver

Application of Geodemographic Embeddings – Similarity Search

Highest success with unique demographic areas– college towns, military bases

45x retrieval uplift

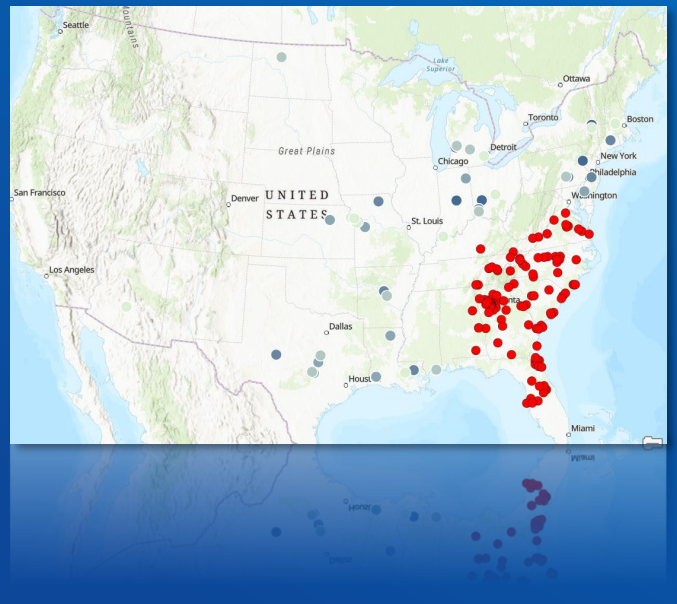
Military bases



Useful when demographic context is important

28x retrieval uplift

New restaurant locations



Not suitable for non-demographic cases and are sensitive to scale

Non-significant uplift

Finding endangered habitat

Cemetries locations

Civil War Battle sites

Application of Geodemographic Embeddings – Prediction

Heart Disease Rates

Great Fit

With Embeddings	0.62
With Curated Variables	0.61
Combined	0.68
Uplift	+0.07

Strong demographic fit — embeddings capture similar population health patterns across regions.

Voter Turnout

Good Fit

With Embeddings	0.39
With Curated Variables	0.76
Combined	0.84
Uplift	+0.08

Partial fit — key non-demographic drivers (voter turnout from previous years) boost the combined model.

Library Square Footage

Poor Fit

With Embeddings	0.03
With Curated Variables	0.59
Combined	0.49
Uplift	-0.10

Poor fit — demographics are not the driver. Funding and policy decisions matter more.

…embeddings boost models where demographics are a genuine driver

Limitations

Lost Explainability

Results are opaque — no interpretability.

No Time Awareness

Embeddings have no sense of time, and past results don't map 1:1 to new data.

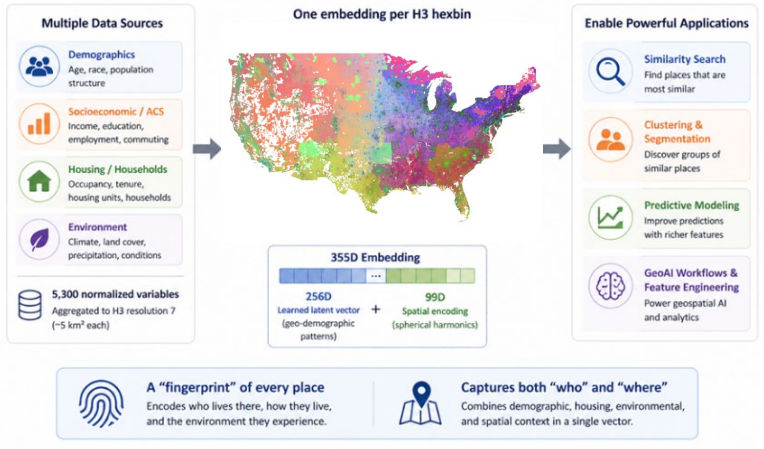
Towards Building A Geo-Demographic Foundation Model

Rutuja Gurav, Pamir Roy, Susanket Sarkar, Rahul Deo, Karthik Dutt, Rohit Singh
Esri

Introduction

Foundation models (FMs) have enabled general-purpose representation learning across domains like computer vision (CV) and natural language processing (NLP), producing state-of-the-art performance on a wide variety of tasks, effectively simplifying the model training, finetuning and inference workflows with a ready-to-use data product in the form of foundational embeddings. The success of FMs models in these domains has garnered interest from geospatial artificial intelligence (GeoAI) community to create such FMs for geospatial data to produce embeddings that can generalize across a variety of tasks which previously would need expert-curated task-specific datasets to model a single target of interest.

In this work, we present a geo-demographic foundation model (GDFM) which produces embeddings for H3 hexagons spanning the United States at resolution 7 level (approx. 5 sq. km.) using a curated collection of tabular geospatial datasets, including census-derived demographic indicators and physical environmental measurements. We show that these geo-demographic embeddings are useful across multiple downstream tasks, demonstrating their potential as a unified representation for geospatial analysis and reducing sole reliance on bespoke, task-specific datasets and models.

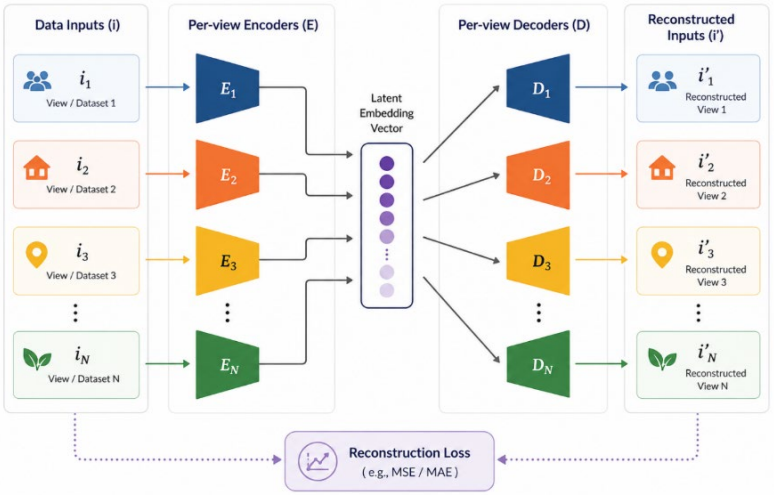


Related Works

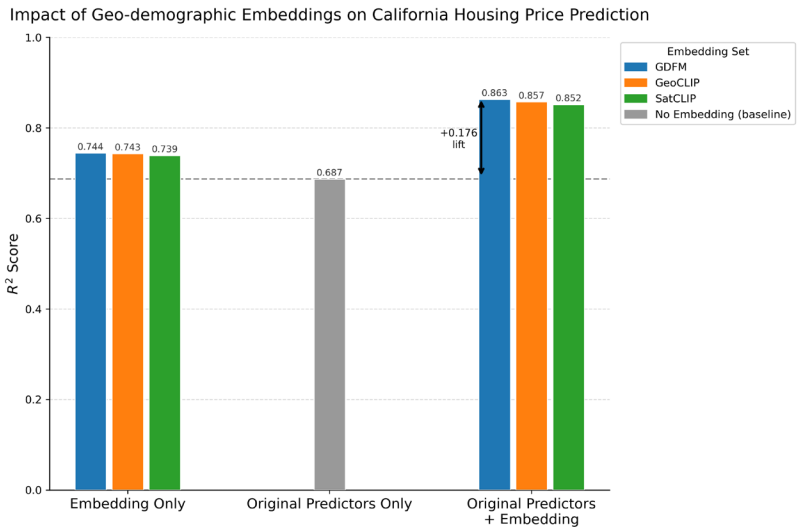
PDFM[1] presents a zipcode-level population dynamics foundation model which provides compact embeddings using tabular datasets of aggregated search trends, POIs and their busyness and weather measurements for training.

Satclip[2] and Geoclip[3] are imagery-based location embeddings models; the former uses satellite imagery and contrastively aligns them with their geolocations on the globe while the latter uses ground-level imagery.

Model Overview

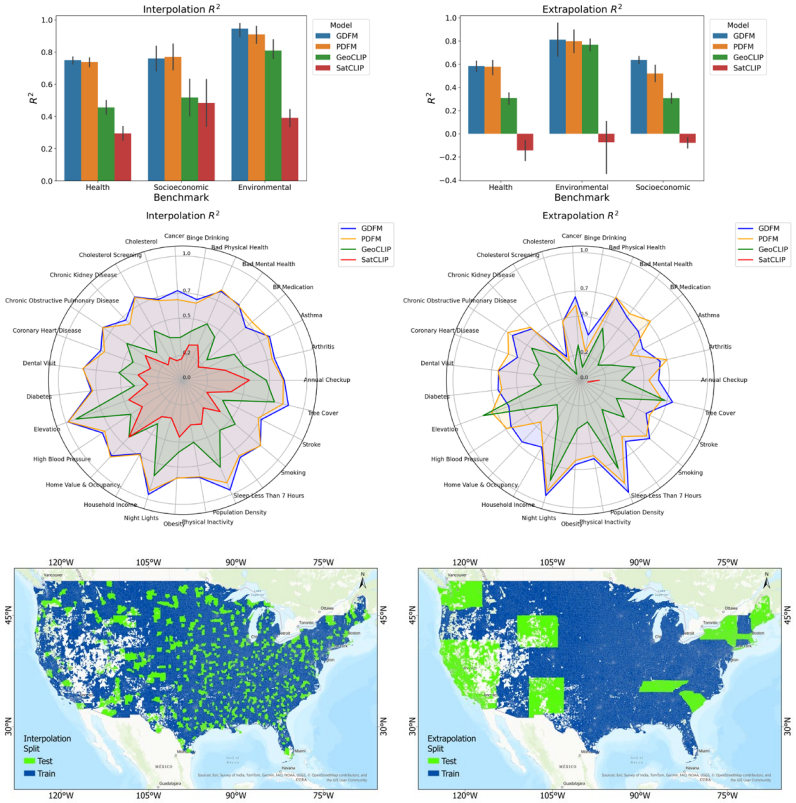


Results: Performance Lift



Results: Benchmarking

*PDFM metrics are as reported in [1]



Conclusions and Future Work

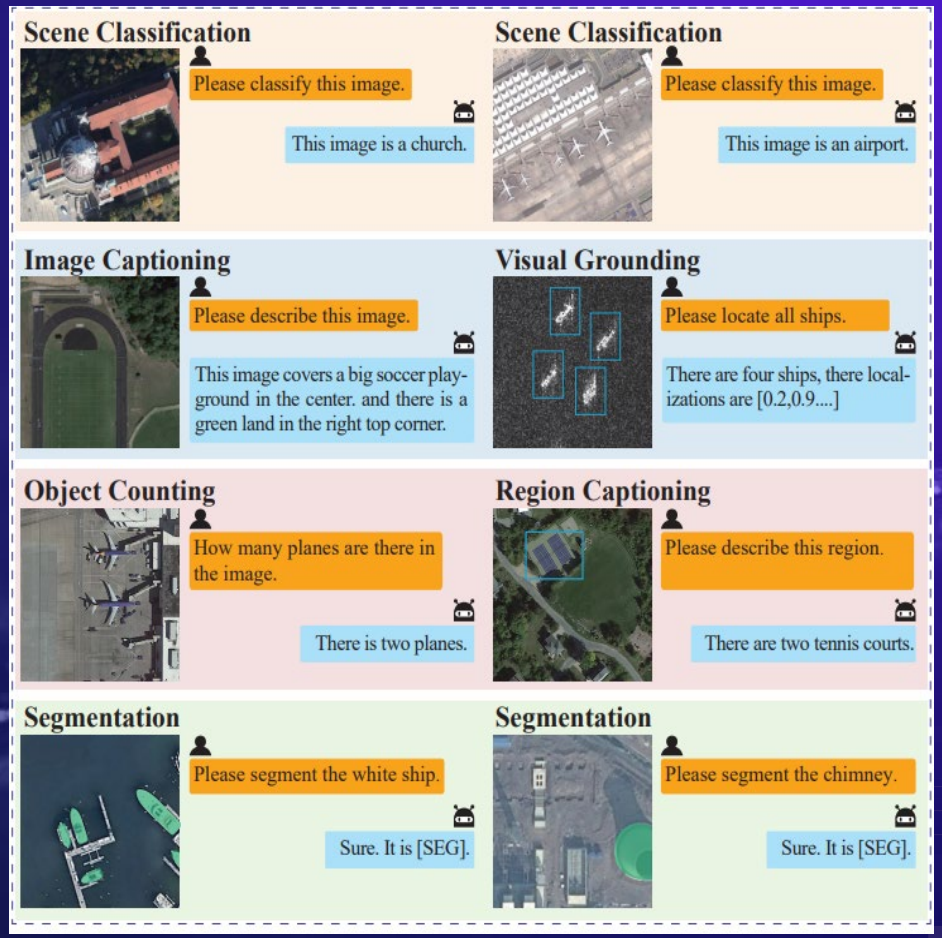
High spatial resolution geo-demographic embeddings like GDFM are a valuable addition to the GeoAI toolkit alongside remote sensing imagery embeddings, providing predictive performance lift and enabling country-scale geospatial analysis. For future work, we plan to extend these embeddings to include explicit spatial modelling using neighborhood context via graphs and incorporate more modalities, beyond tabular data, like text and imagery.

References

[1] Agarwal, Mohit, et al. "General geospatial inference with a population dynamics foundation model." arXiv preprint arXiv:2411.07207 (2024).
[2] Klemmer, Konstantin, et al. "Satclip: Global, general-purpose location embeddings with satellite imagery." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. No. 4. 2025.
[3] Vivanco Cepeda, Vicente, Gaurav Kumar Nayak, and Mubarak Shah. "Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization." Advances in Neural Information Processing Systems 36 (2023): 8690-8701.

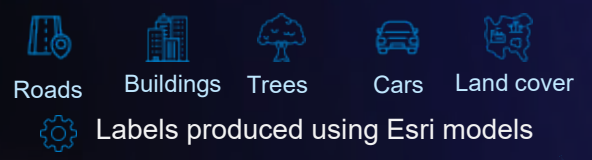
Geospatial Vision Language Models

- Connect satellite imagery with natural language for intuitive interaction
  - Enable querying, captioning, and reasoning over geospatial imagery using prompts
  - One model for all Remote Sensing tasks
- Examples:
  - EarthGPT, EarthDialRemoteSAM, GeoChat, GeoPixel, GeoVLM-R1
- Technical details:
  - Trained on paired imagery-text datasets
  - Combine LLMs with geospatial RS FMs
  - Preserve linguistic, reasoning and visual understanding of base models



Geospatial VLMs handle a wide spectrum of earth observation tasks through prompts

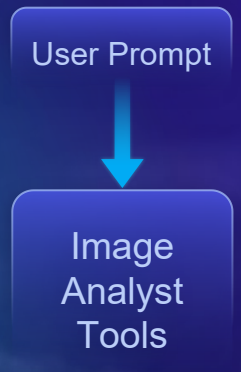
Esri’s Geospatial Vision Language Model (GeoVLM)



Labels produced using Esri models

USA (NAIP) + Europe aerial imagery

Open datasets



Pixel Classification

Object Detection

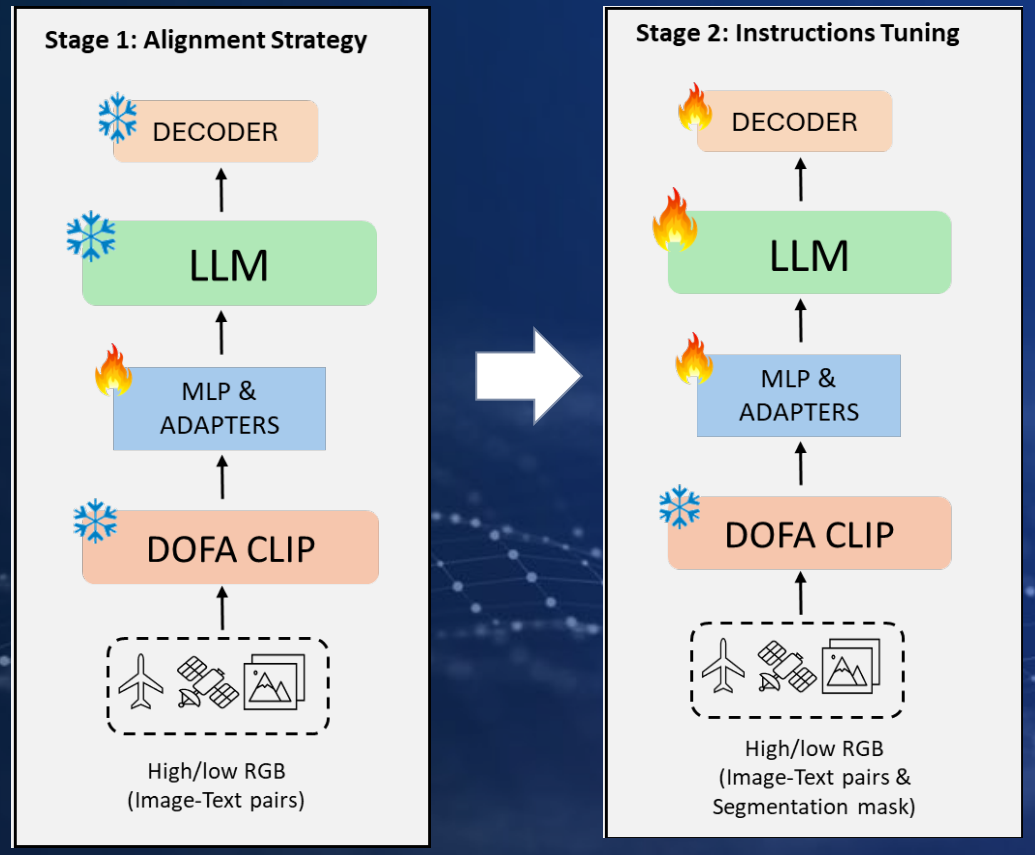
Object Classification

Object Counting

Question Answering

Captioning

Geospatial-VLM Alignment & Finetuning Strategy



Frozen weights Trainable weights Shared weights

Implementation Details

Vision Encoder: DOFA-CLIP

LLM: Qwen2.5-1.5B

Training Setup:
- 8 x A100 40GB GPUs
- 4 Days (1 Epoch)

Loss Functions:
- Autoregressive Cross-Entropy Loss
- BCE Loss
- Dice Loss
- KL Divergence Loss

Additional Enhancements:
- DOFA CLIP Adapters
- Spatial Attention Prompting
- Text-Guided Attention

Benchmarking: how GeoVLM stands among peers

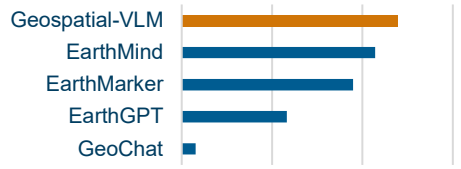
Image-level context

	RESISC	AID	UCMerced	RSVQA-HRBEN
EarthGPT	93.84	88.76	92.42	72.06
MSGNet	**94.57**	-	-	-
GeoVLM-R1	-	88.46	**97.81**	**75.27**
EarthMind		97.2	95.0	74.0
Geospatial-VLM	91.03	**98.23**	95.85	**75.21**

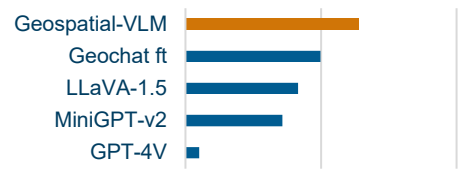
Best on AID; competitive on UCMerced.

Region-level grounding and captioning

RSVG region caption (CIDEr)



VRSBench visual grounding (Precision@0.5)



Referring image segmentation (mIOU)

	RRSIS-D	RefSegRS	RISBench
EarthMind	82.2	62.6	-
GeoPixel	67.3	-	-
RemoteSAM	71.75	-	71.46
Geospatial-VLM	**82.31**	**64.85**	**73.89**

State-of-the-art grounded pixel results.

Semantic segmentation (mIOU/mF1) – Zero-shot

	Vaihingen	Potsdam	LoveDA R	LoveDA U	WHU Aerial
RemoteSAM	16.95/26.73	**38.99/49.20**	7.80/14.11	17.75/27.54	47.81/64.69
EarthMind	**26.30/34.90**	24/31.70	29.16/43.16	32.07/47.44	62.67/77.05
Geospatial-VLM	25.19/34.37	28.94/35.75	**30.26/45.12**	**41.03/56.64**	**71.08/83.09**

Best on LoveDA Rural, Urban and WHU Aerial; competitive on Vaihingen.

Best in references on region captioning and visual grounding.

Best in referring segmentation across RRSIS-D, RefSegRS, and RISBench.

Competitive on image-level and semantic segmentation benchmarks.

Geospatial-VLM Results

Supports a unified set of EO tasks, from global scene understanding to localized grounding and pixel-accurate masks.

IMAGE-LEVEL TASKS (Global Understanding)

1. Scene Classification

Classify the given image.

Urban.

2. Image Captioning

Describe the image in detail.

The image, sourced, shows a high-res. view of a recreational area featuring multiple tennis ....

3. VQA

Is this an urban or rural area?

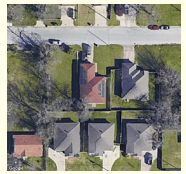
Rural.

4. Object Counting

Count the number of airplanes in this image.

5

5. Multi-label Classification



Classify the image. Use few words or multiple classes.

Buildings, Cars, Trees, Roads

6. Similarity Search

Container cranes near ships and docks

REGION-LEVEL TASKS (Local Understanding with Regions)

1. Region Classification

Classify the region inside the box

Bridge

2. Region Captioning

Please describe the region in the box.

A red car in the middle

3. Grounding

Locate the windmill in the image.

Bounding box <x1=402,y1=108, x2=840,y2=720>

4. Spatial Reasoning

What do you see next to basketball court?

A red tennis court.

PIXEL-LEVEL TASKS (Fine-Grained Understanding with Masks)

1. Semantic Segmentation

Segment the solar panels.

Sure, it's a [SEG]

2. Referring Image Segmentation

Segment the airplane at the top.

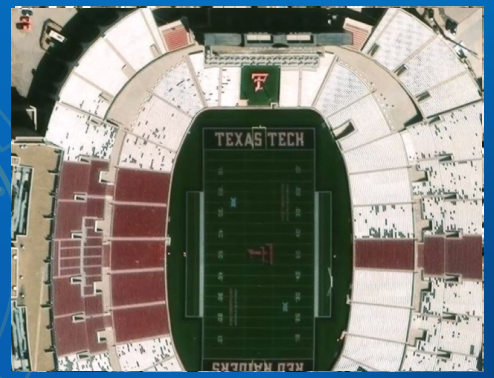
Sure, it's a [SEG]

Geospatial Vision Language Model (GeoVLM)

GeoVLM Answers Human Questions About Places

“Segment roads / buildings”

- Extract GIS features



What do you see?

The high-resolution image from Aerial planes shows a large stadium with a distinctive red and green color scheme. The stadium features a large green field with markings for a football game, and the word " TEXAS TECH" is prominently displayed on the stadium's facade. Surrounding the stadium is a vast parking area filled with numerous vehicles, indicating a significant event or gathering.

“Baseball field at bottom right”

- Identify and visually ground features

“Describe the region”

- Identify and describe areas within an image

GeoAI

Helping Address Our Most Complex Challenges

Agriculture
Infrastructure
Economic Development
Pollution
Nature Conservation
Health
Energy
Business
Natural Resources
Climate Change
National Security
Science & Technology
Environment
Water Resources
Social Conflicts
Urbanization & Development
Transportation
Equity & Access
Natural Disasters
Networks & 5G

Understanding Our World

Enhancing Productivity

Innovating for a Sustainable Future