Deep Species Distribution Models

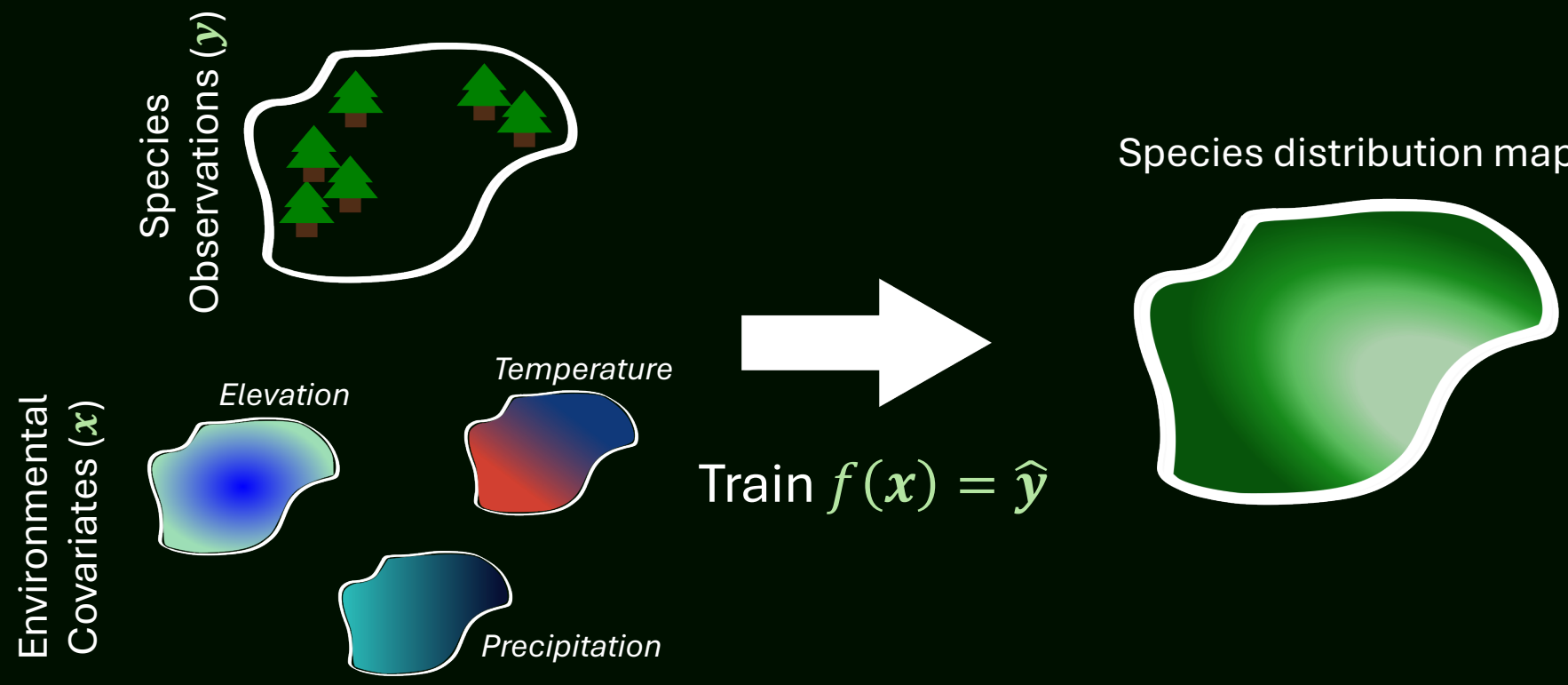
Pablo Ubilla Pavez, Diego Marcos

Inria

UNIVERSITÉ DE MONTPELLIER

Species Distribution Model

Learn a mapping function f from environmental covariates $\boldsymbol{x}$ to a species response $\boldsymbol{y}$.



Species Distribution Model

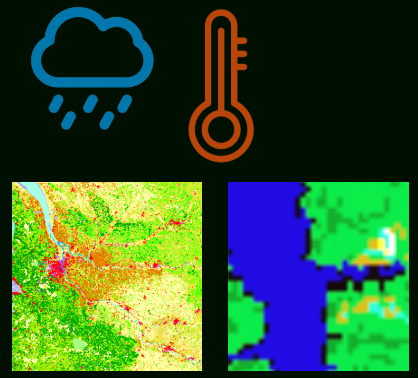
Learn a mapping function f from environmental covariates $\boldsymbol{x}$ to a species response $\boldsymbol{y}$.

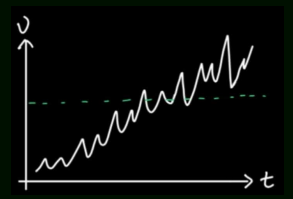
- Mechanistic: Relationship is learned from physiological information
- Correlative: Relationship is learned from observation data
  - Generalized Linear Models
  - Generalized Additive Models
  - MAXENT
  - Random Forest
  - Artificial Neural Networks

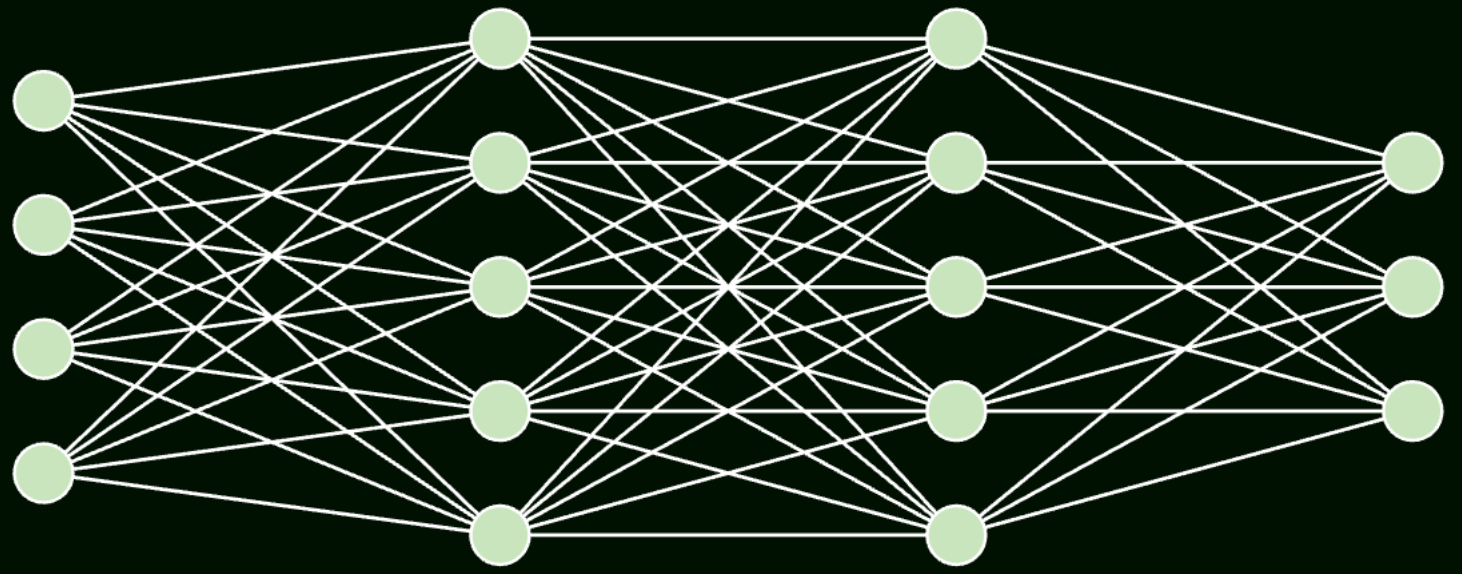
Deep Species Distribution Models

Function f is a Neural Network

- Can learn non-linear relationships
- Can handle multiple modalities as input
- Can learn multiple species at the same time

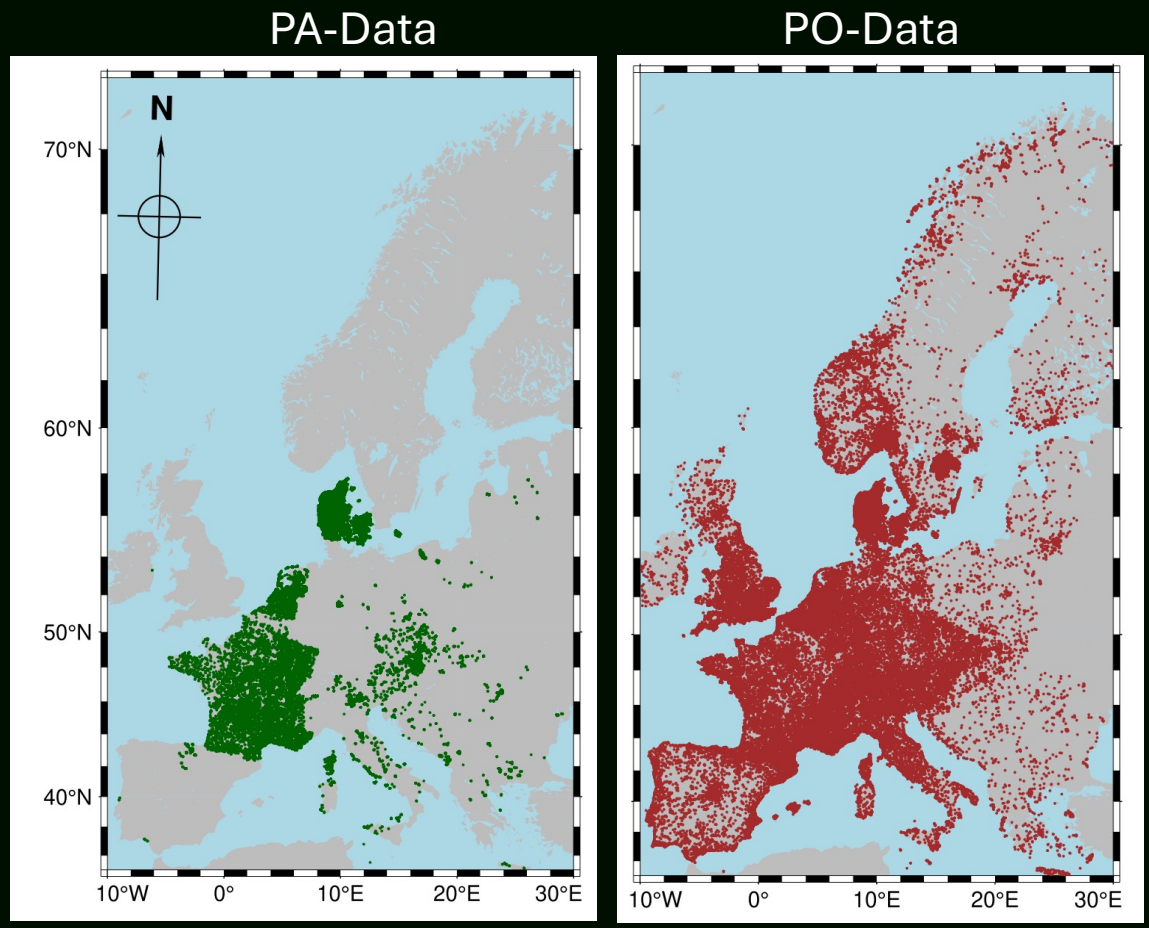






GeoPlant

- Dataset provided by the Pl@ntNet initiative to train SDMs and Deep-SDMs
  - 5M heterogeneous **Presence-Only records**
  - 90k exhaustive **Presence-Absence survey records**
  - Test set is proposed in PA to allow for validation of the models



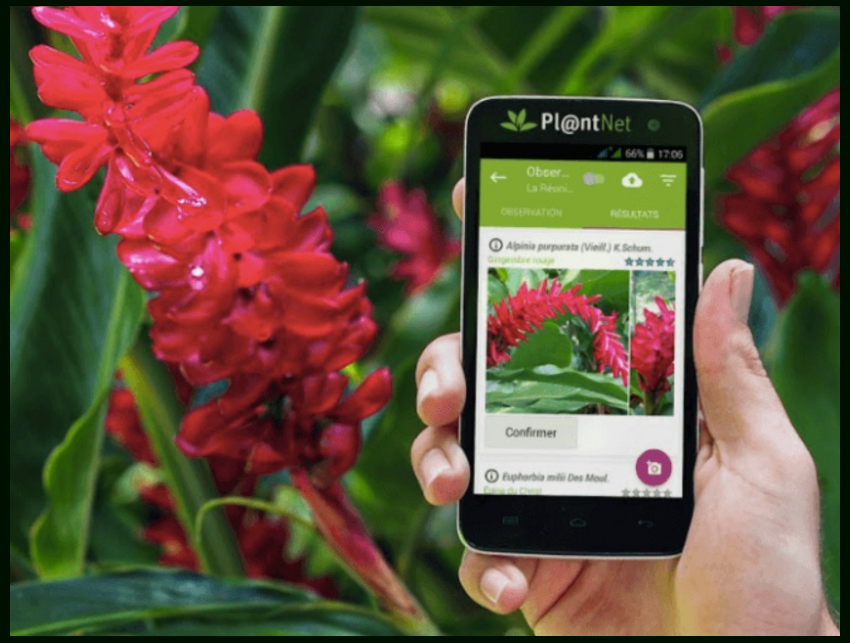
GeoLifeClef 2023 (Botella et al.)

Presence Absence (PA) Data

- Experts report whether species j is present or absent in site i. Then $y_{ij} \in \{0,1\}$
- Standardized, reproducible protocol: area to sample, survey effort, revisit frequency, site selection
- Ideal for correctly testing models

Presence Only (PO) Data

- Opportunistic observations, which only reports presences. Then: $y_{ij} \in \{-, 1\}$
- Tends to be biased geographically
- Largest collections come from **citizen science platforms** like Pl@ntNet
- We will focus on models trained on PO



Model training with PO

- When training a model without absences, we cannot really expect to recover the **presence probability of a species**.

- What can be feasible is to recover the **relative probability of a site given a species**

- This translates into a **Softmax** across sites (observations), instead of species (classes)

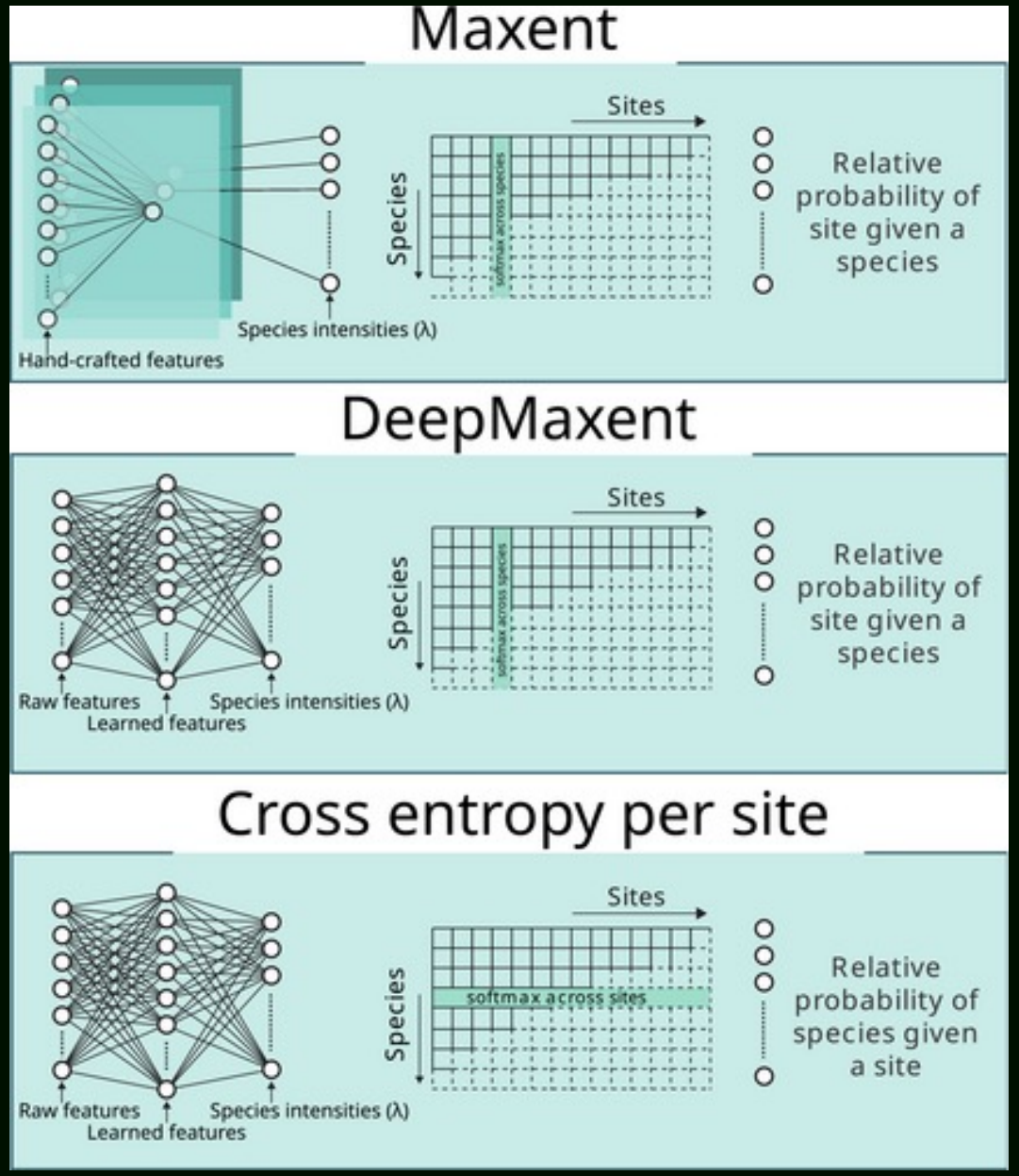
Model training with PO

Consider the intensities proportion of species j in a batch B :

$$\tilde{\lambda}_{ij} = \frac{\exp(-f_\theta(\boldsymbol{x}_i)_j)}{\sum_{b \in \mathcal{B}} \exp(-f_\theta(\boldsymbol{x}_b)_j)}$$

Use DeepMaxent Loss (Ryckewaert et al., 2026.)

$$L(\tilde{\boldsymbol{\lambda}}, \boldsymbol{y}) = \sum_{i \in \mathcal{B}, j \in \mathcal{S}} y_{ij} \cdot \log(\tilde{\lambda}_{ij})$$

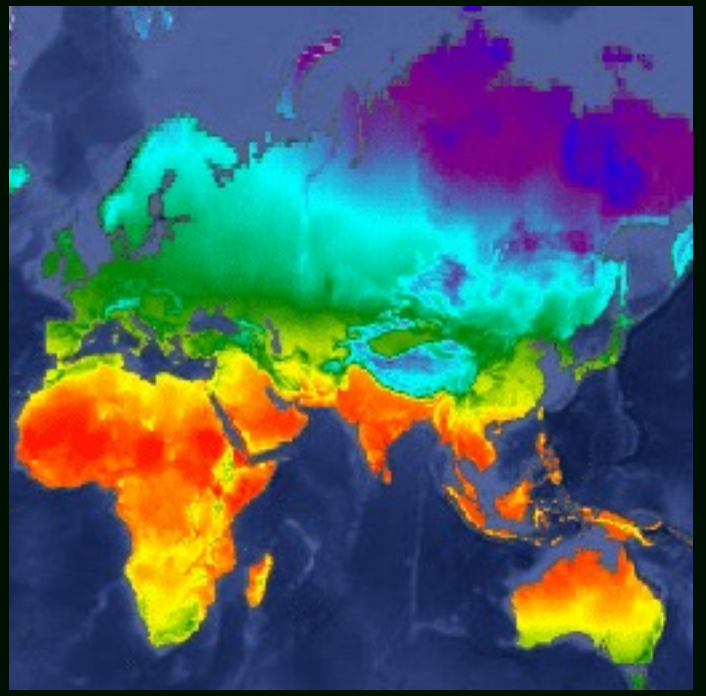


Which data is use for training?

Environmental predictors

- Traditionally, SDMs are trained with environmental predictors such as:
  - Land Cover: MODIS
  - Vegetation: NDVI
  - Climate: BioClim

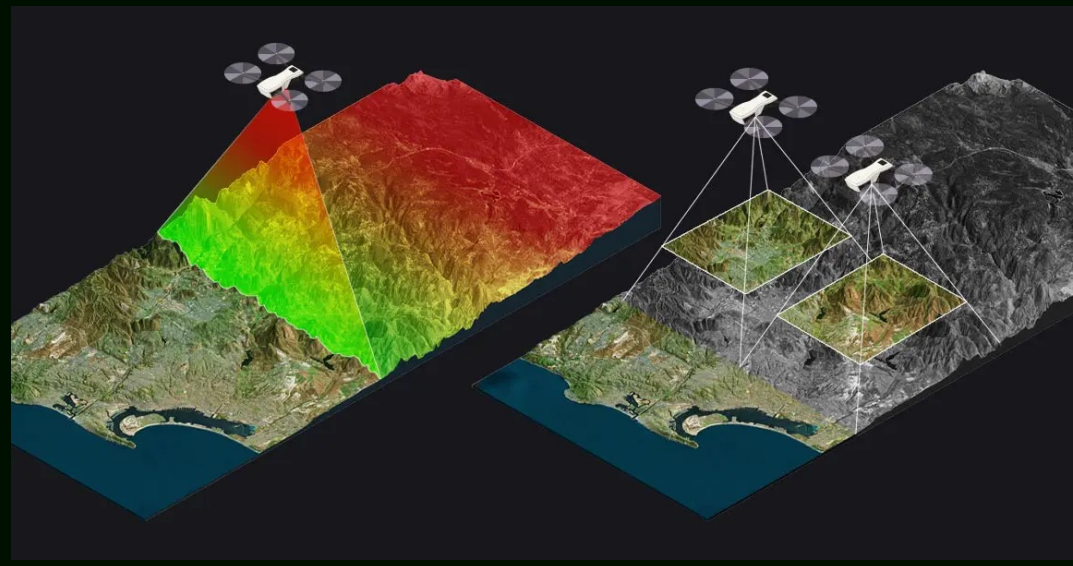
- These can be easily structured as **tabular data**



Other modalities

- Deep SDM could thrive by using more complex modalities that might contain richer information
  - Satellite Images: Landsat, Sentinel
  - Audio
  - Time-series
  - LiDAR

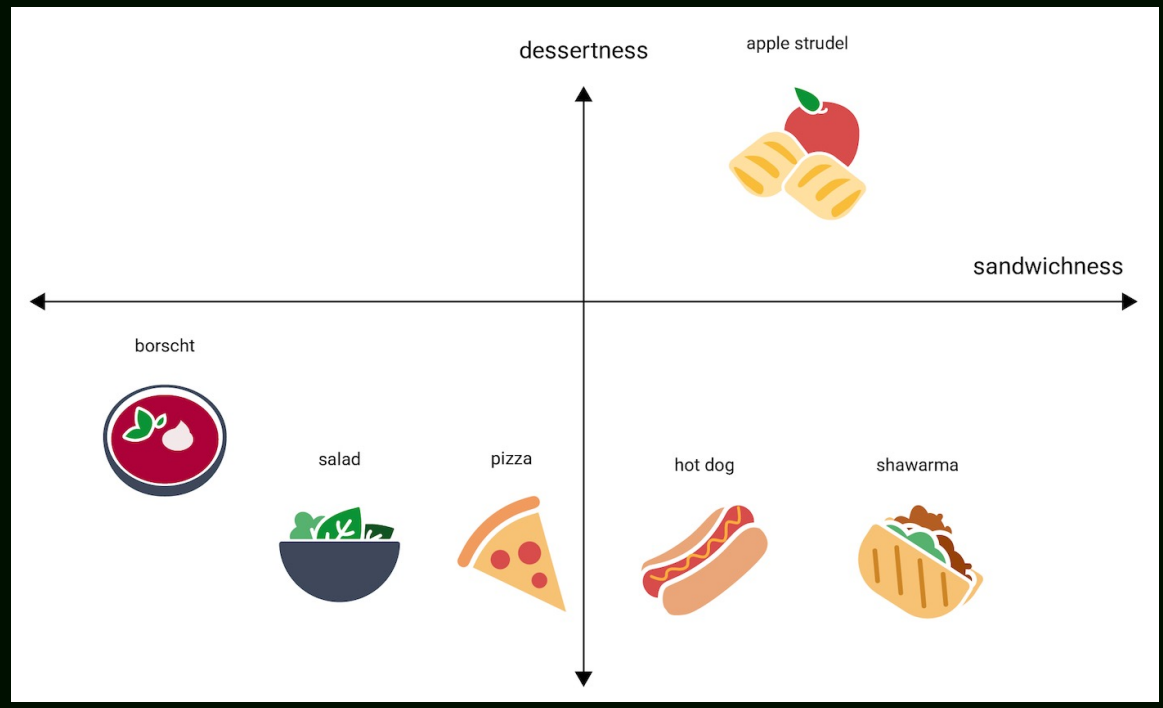
- One issue: high-computational cost



Today we will look at an alternative: Earth Embeddings

Embedding

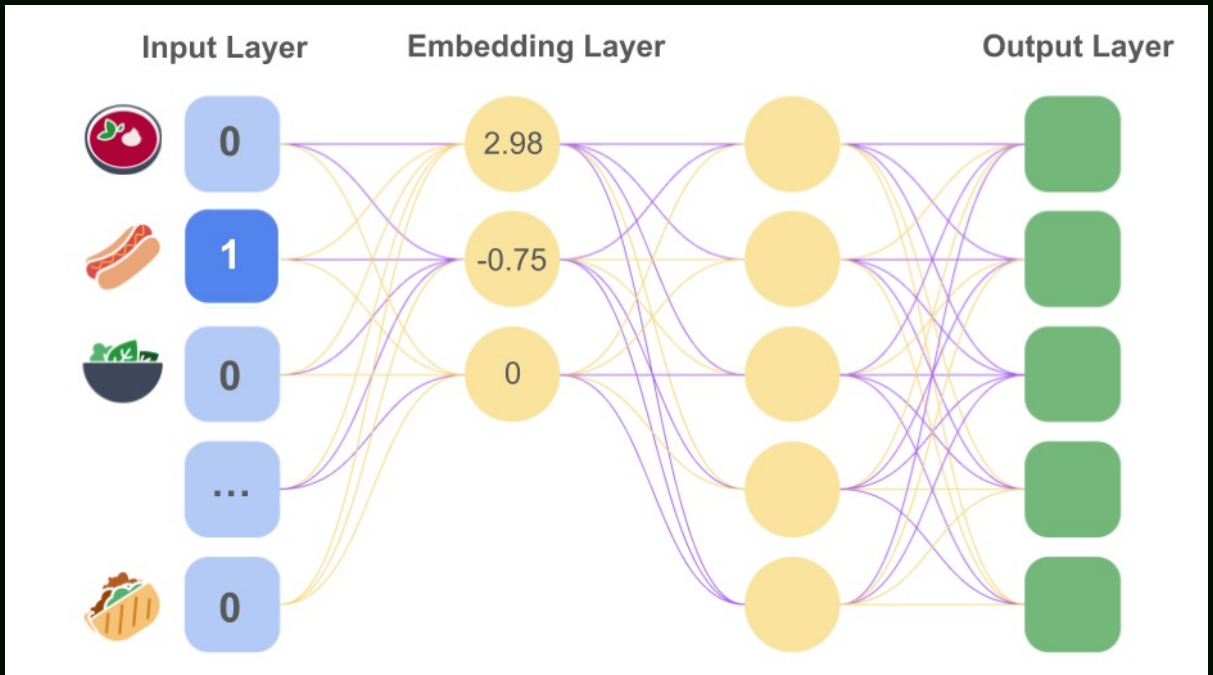
- Representation learning technique
- Maps complex high-dimensional information into a (lower-dimensional) vector space.



Google, Advanced ML Models

Embedding

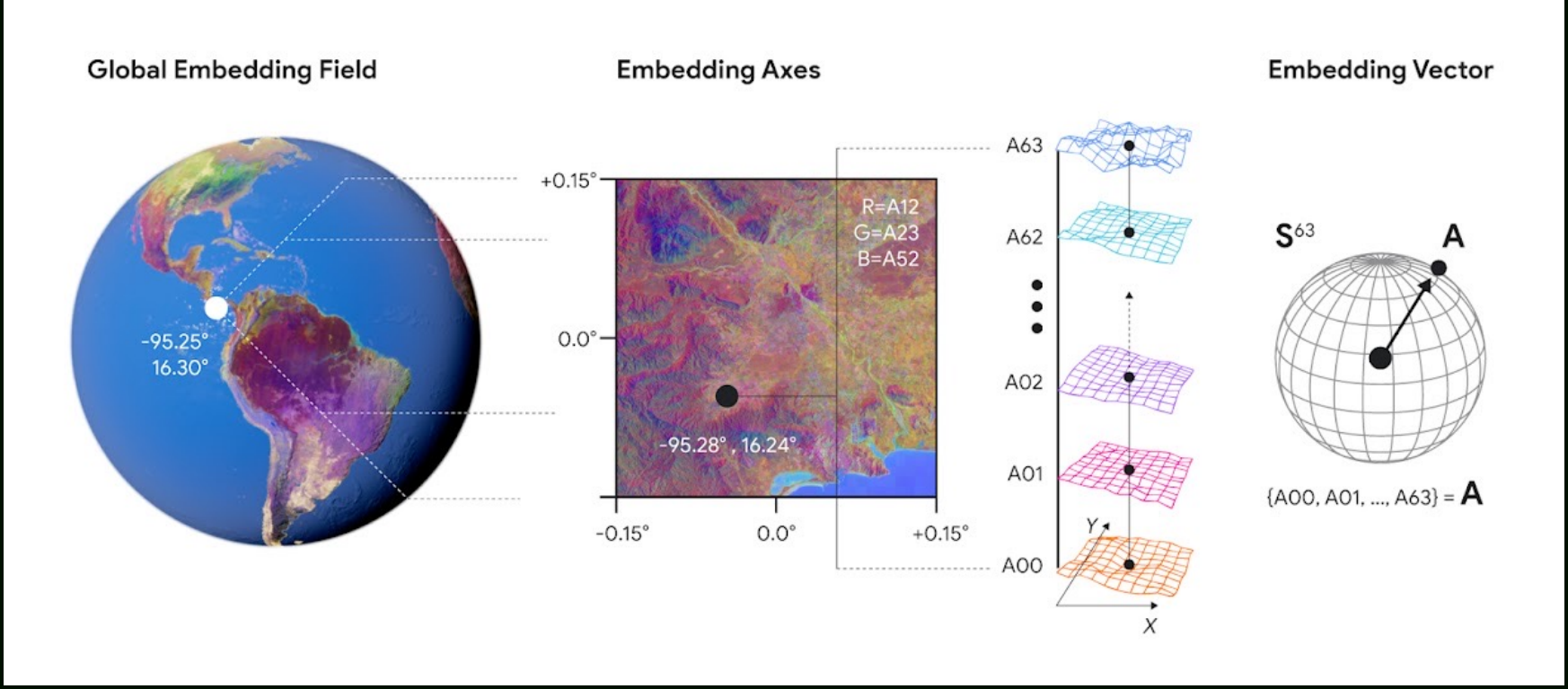
- An embedding can be obtained by **training a neural network** for a target task
- One of the middle layers (lower-dimension) can be used to summarize the information of the object



Google, Advanced ML Models

Earth Embedding

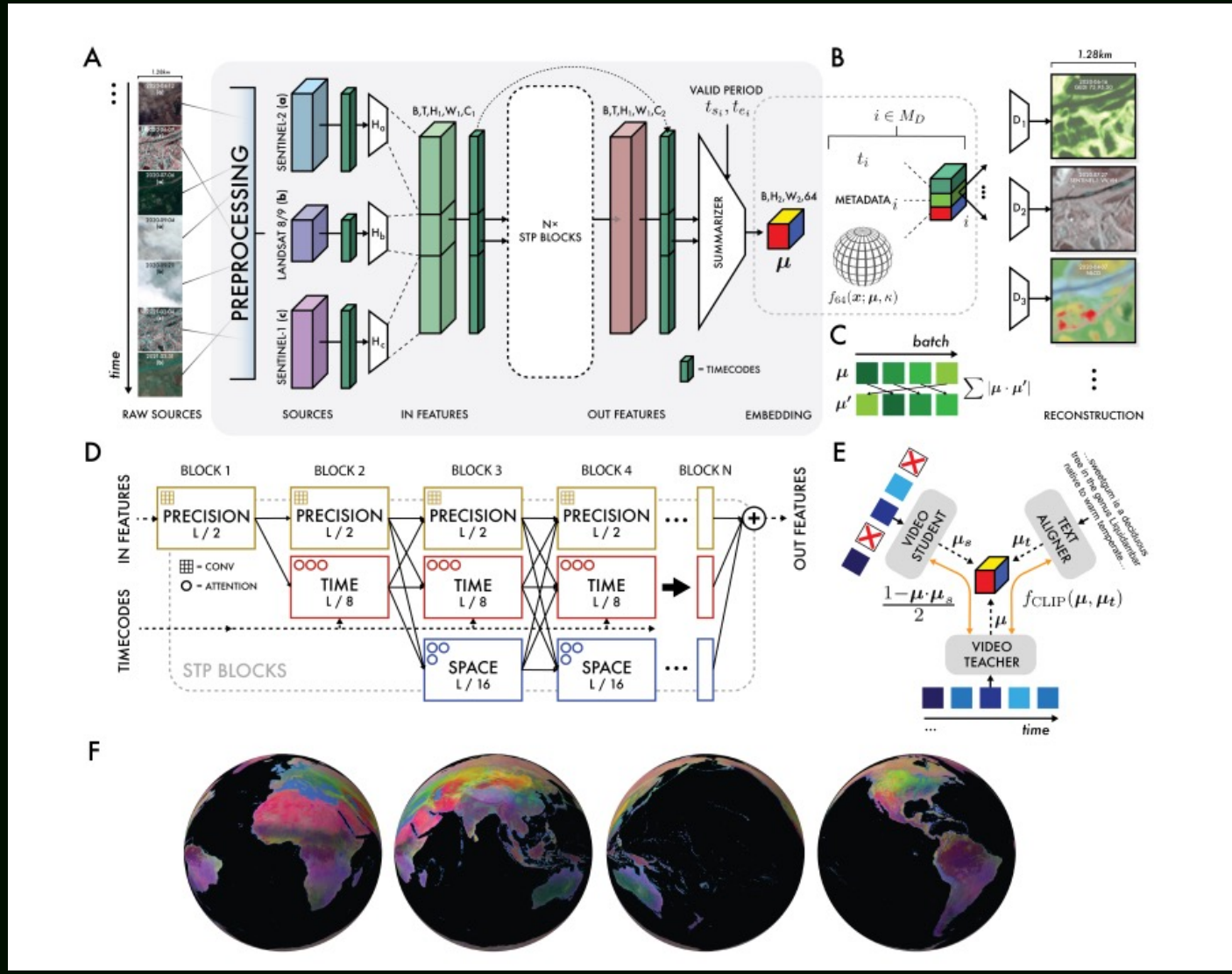
- Attempts to summarize the information (e.g. land cover, vegetation, vegetation) of a location on Earth into a vector.



Google, AlphaEarth

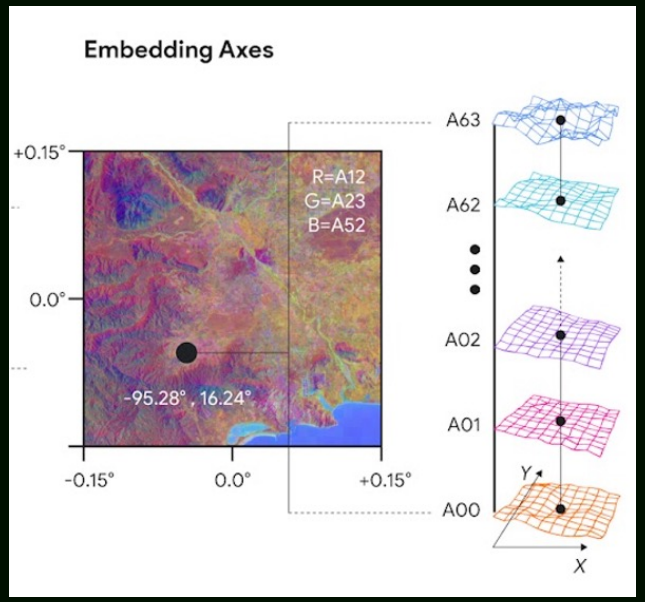
Earth Embedding

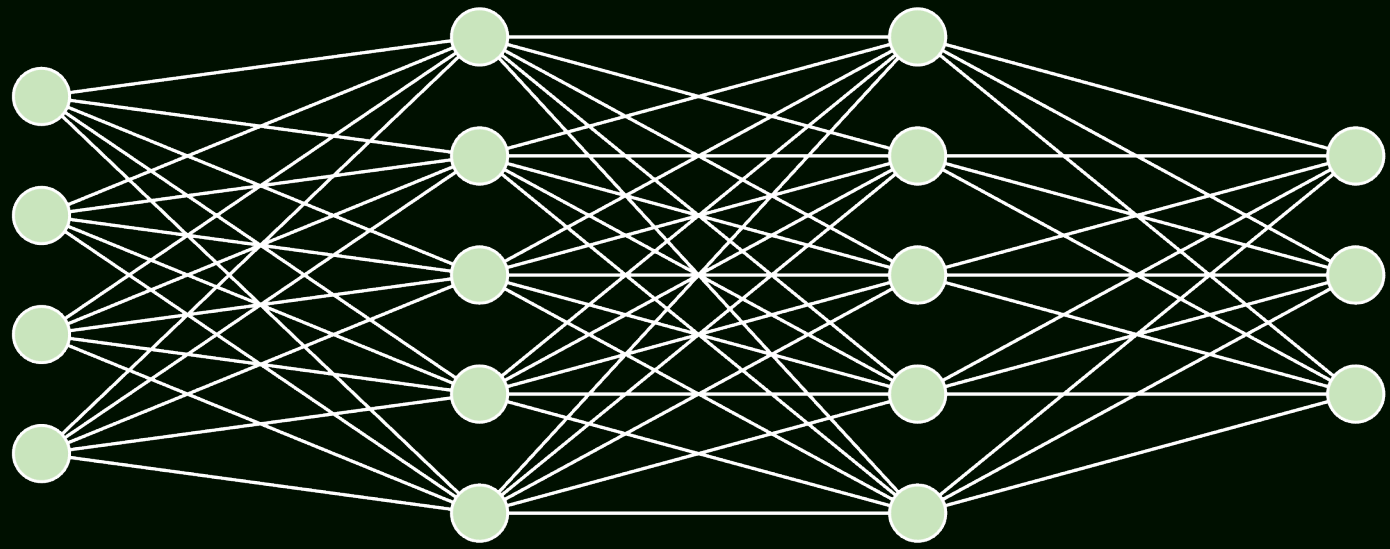
- The training can become highly sophisticated
- In the end the goal is the same: obtain a vector that summarizes information



Google, AlphaEarth

We can train a Deep SDM directly using the Earth Embeddings, which allows to use simpler architectures and reduce computation-cost





Hands on!!!!