De Pl@ntNet à GeoPl@ntNet: nouvelles approches d'IA pour le monitoring de la biodiversité

Alexis Joly, Antoine Affouard, Rémi Palard, Maxime FromeHoltz, Benjamin Deneu,
Hervé Goëau, J.C. Lombardo, Mathias Chouet, Hugo Gresse, Cesar Leblanc,
Maximilien Servajean, François Munoz, Pierre Bonnet

Pl@ntNet

A citizen science platform that uses AI to help people identify plants with their mobile phones

Pl@ntNet app

25 Million users
200+ countries
Up to 2M identifications per day

Personal Usage

Nature, walks

Gardening

Phytotherapy

Professional Usage

Agro-ecology

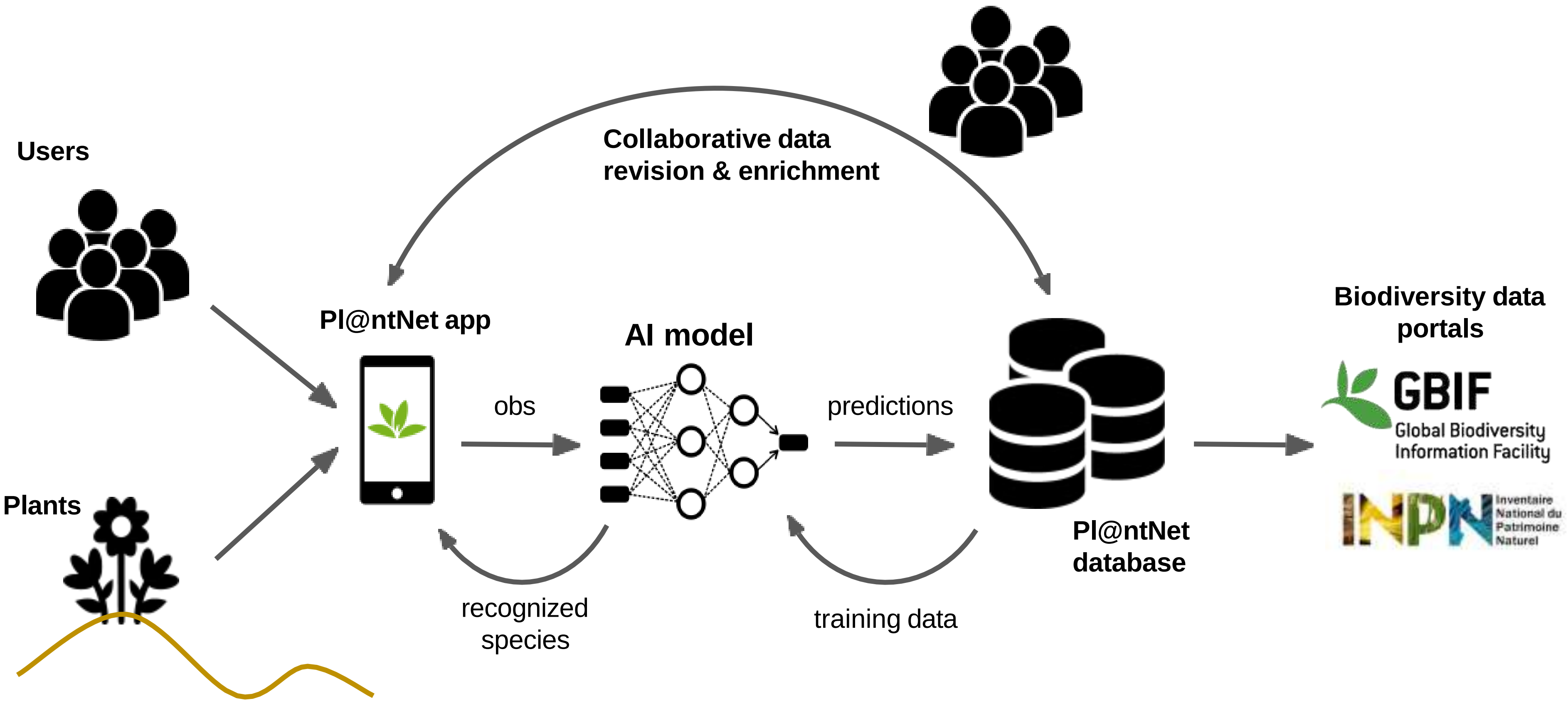
Natural Areas Management

Education, animation

Tourism

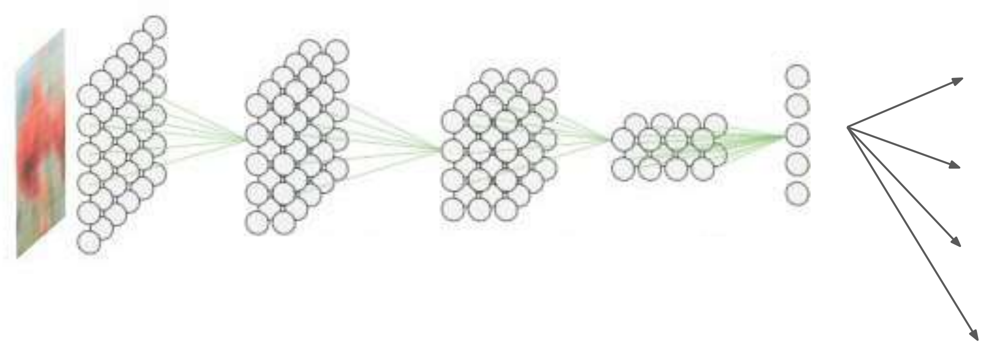
Trade

Key concept of Pl@ntNet: Collaborative AI



Pl@ntNet AI model

Multi-head model trained on **Jean Zay super-computer** on a big dataset of **8M valid observations** (5-6 days of training)



$$\sigma_{species}(f(x))$$ **= 50K classes**

$$\sigma_{organs}(f(x))$$

$$\sigma_{genus}(f(x))$$

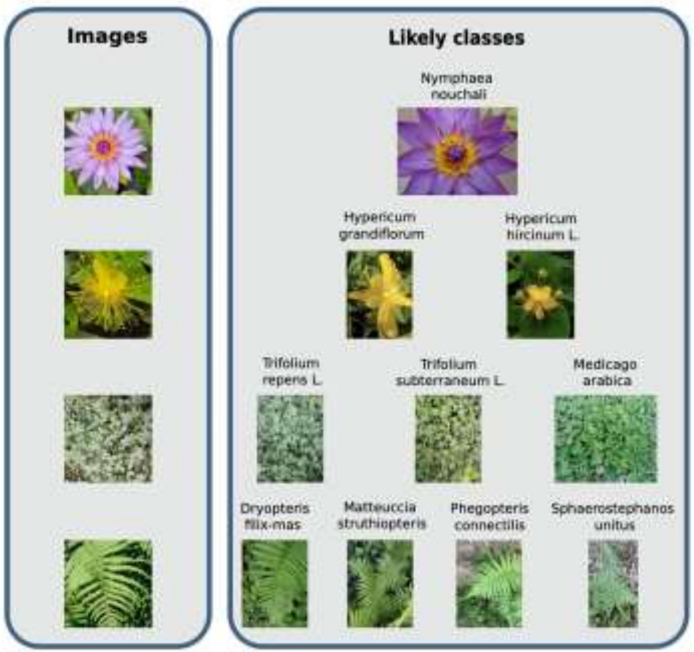
$$\sigma_{disease}(f(x))$$

Model = Vision transformer DinoV2

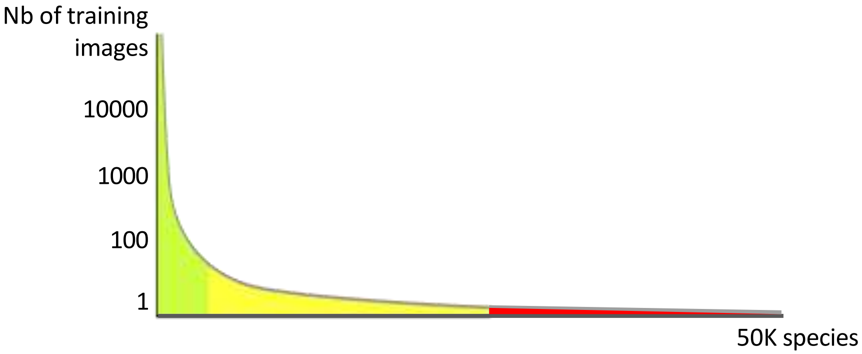
- Backbone pre-trained on 100M images using SSL (by Meta/Inria)
- Final multi-head model fine-tuned on 8M Pl@ntNet images (by Pl@ntNet team)

A difficult problem: uncertainty

Irreducible uncertainty
Species ambiguity



Model uncertainty
Increased by long-tail distribution



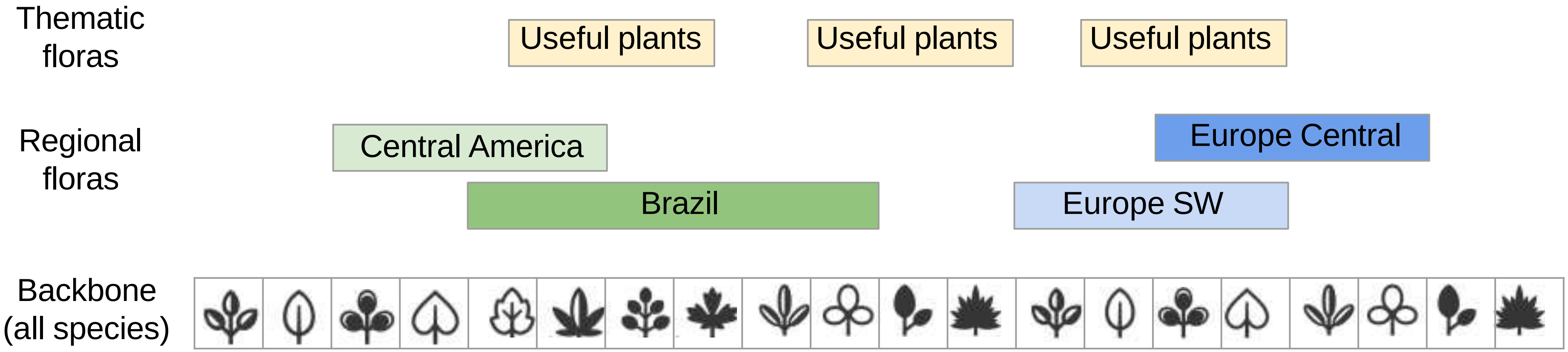
Top1 Identification accuracy:

Common species	= ~**90%**
Average species	= ~**70%**
Rare species	= ~**40%**

Use of regional or thematic floras

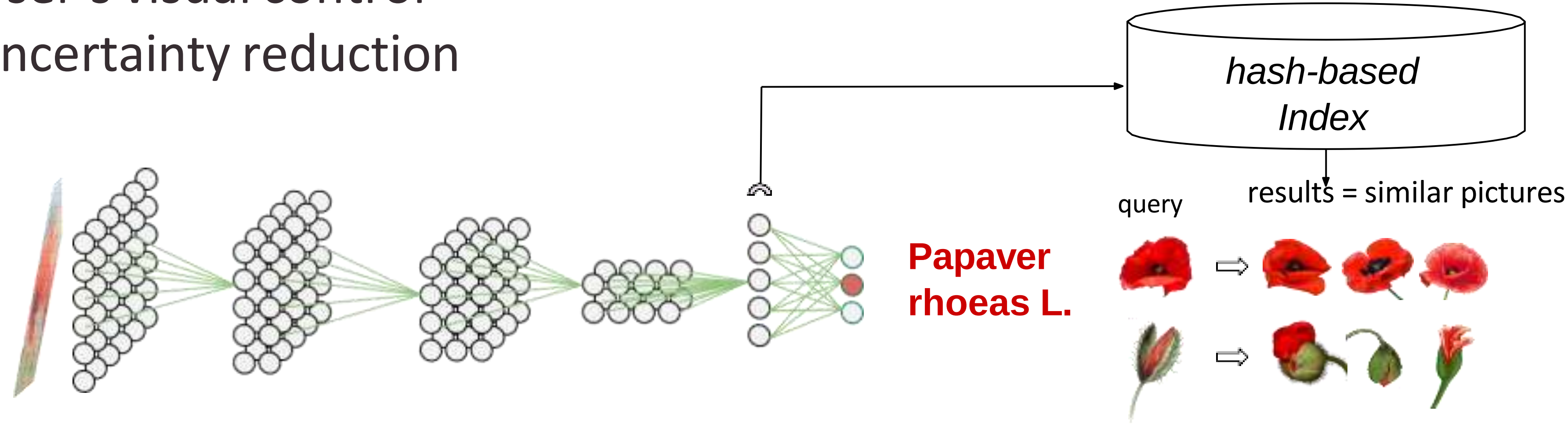
Restricting the hypothesis space to a particular flora allows improving the identification accuracy

$$p(\underset{\text{species}}{y}|\underset{\text{image}}{x}, flora) \geq p(\underset{\text{species}}{y}|\underset{\text{image}}{x})$$



Pl@ntNet Similarity search

User's visual control =
uncertainty reduction



Deep neural network

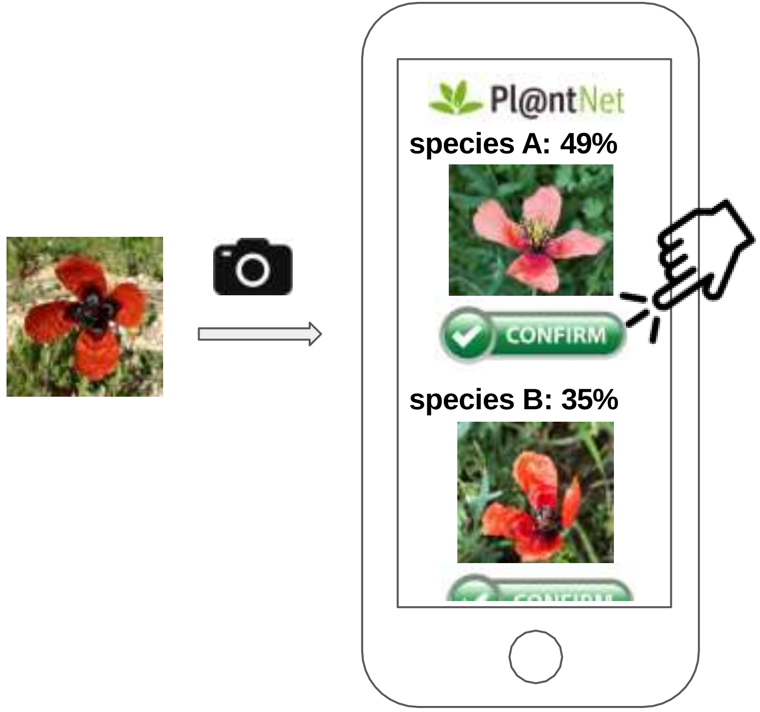
Similarity search engine
9M images

→ Sub-linear algorithm based on locality sensitive hashing

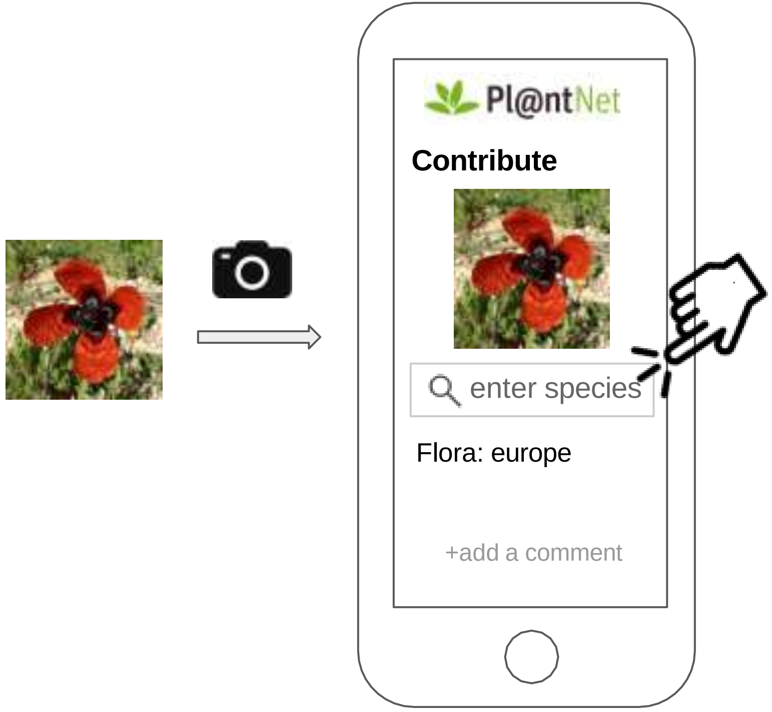
Joly, A., & Buisson, O. (2011, June). Random maximum margin hashing. In CVPR 2011 (pp. 873-880). IEEE.

User’s contributions

Users can contribute their observations



OR



User's revisions

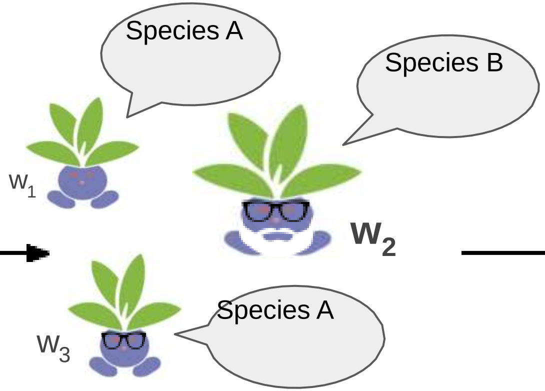
Users can revise observations of other users.

Cooperative Learning algorithm

The most probable label of an observation is determined with a weighted majority voting rule:

An observation

x_i



$$y_i = \arg\max_k \sum_{u \in U_i} w_u \mathbf{1}(y_i^u = k)$$

$U_i =$ Set of users who provided a label y_i^u for the observation x_i

Cooperative Learning algorithm

The weight of a user in Pl@ntNet is a function of the **estimated number of species** he is able to identify

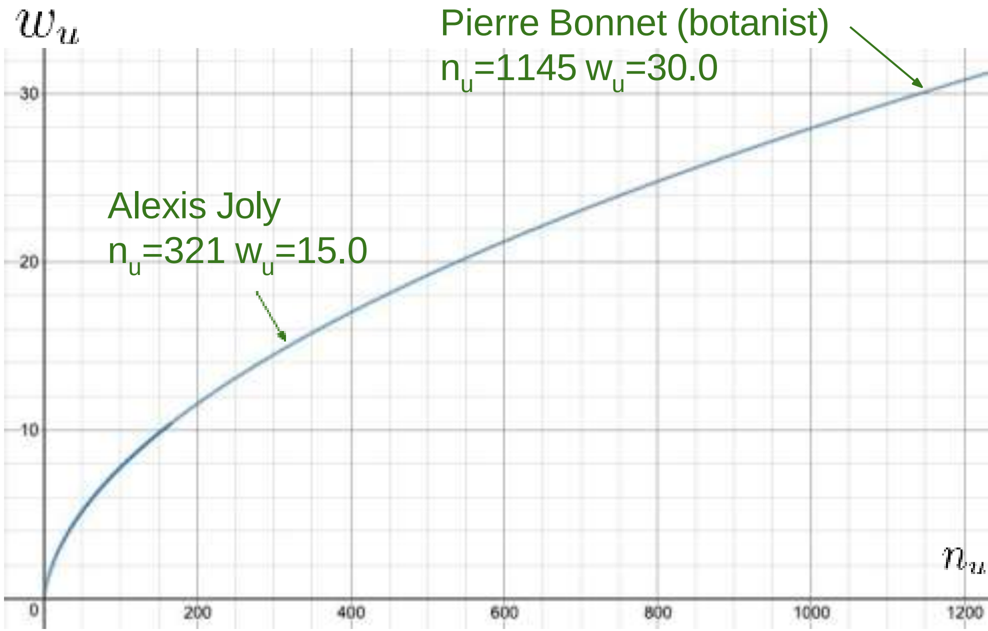
$$w_u = g(n_u) \qquad n_u = \left|\{j : \exists i \; y_i^u = y_i\}\right|$$

w_u

Early career contributor
$n_u=5$ $w_u=1.2$

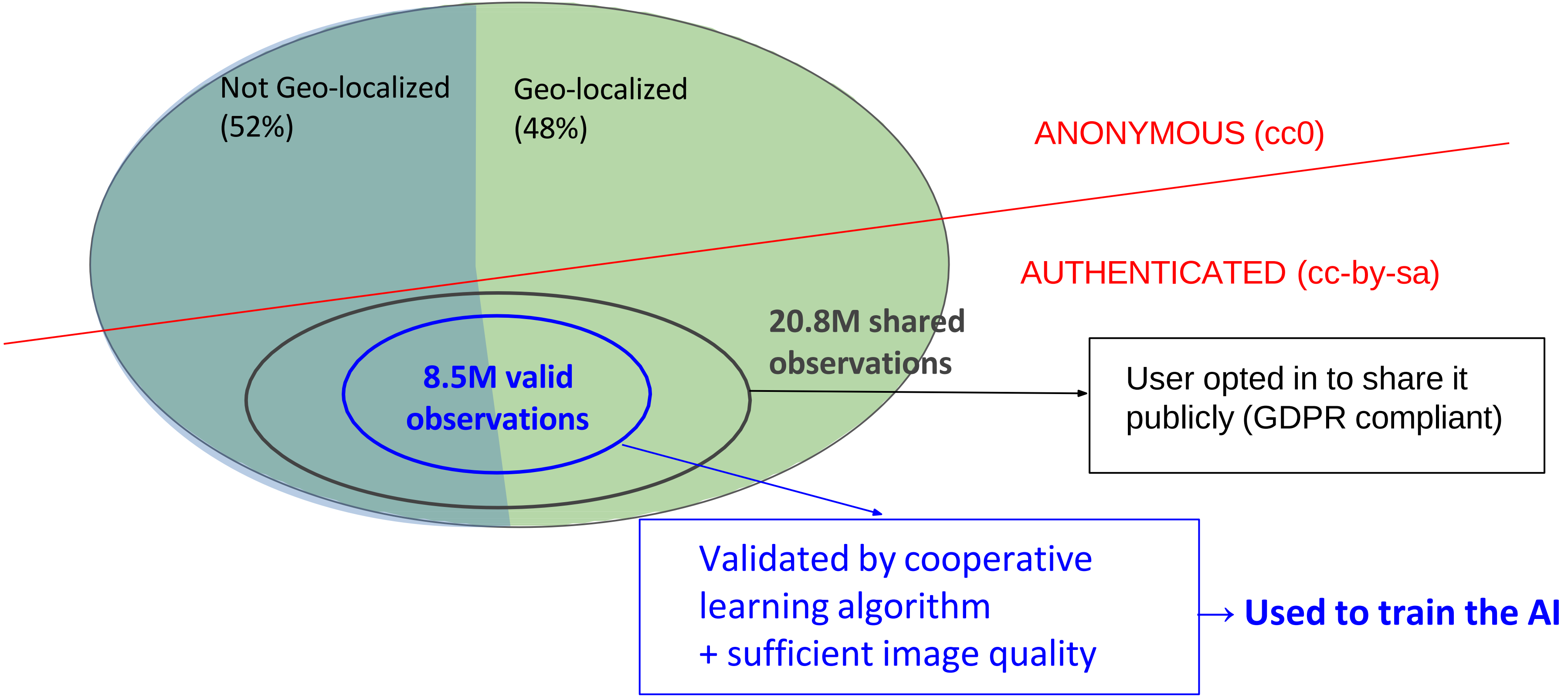
New user
$n_u=0$ $w_0=0.3$

n_u



Pl@ntNet Data

940M raw observations (=queries)



Pl@ntNet Data visualisation tools

Bryonia cretica L.

White bryony, Cretan bryony, مار دارو، فاشرا

Common name(s)

White bryony

Cretan bryony

مار دارو، فاشرا

View all / Edit

Uses

MEDICINE

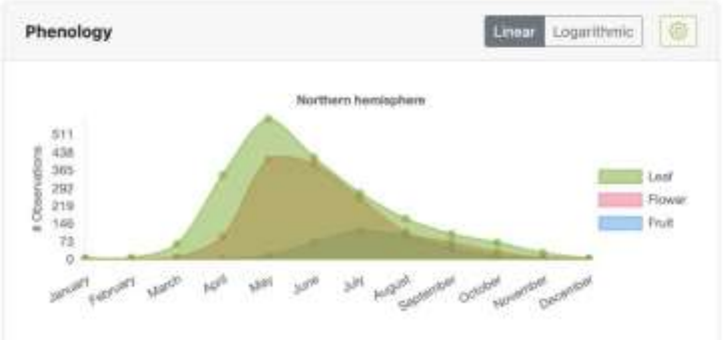
folklore

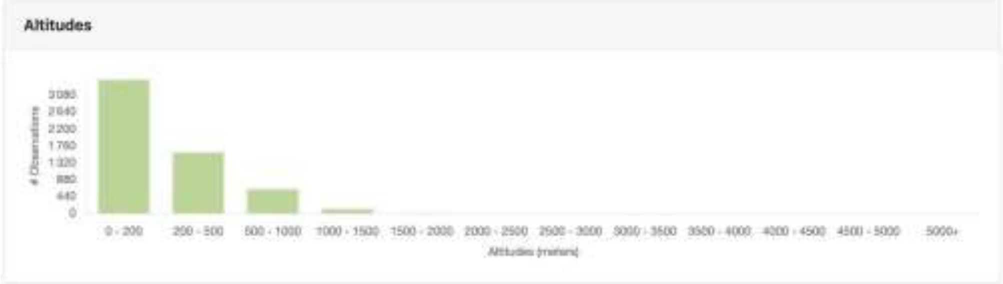
Additional information

Pl@ntUse

GBIF

Royal Botanic Gardens Kew | Plants of the World Online





Pl@ntNet Data shared in GBIF

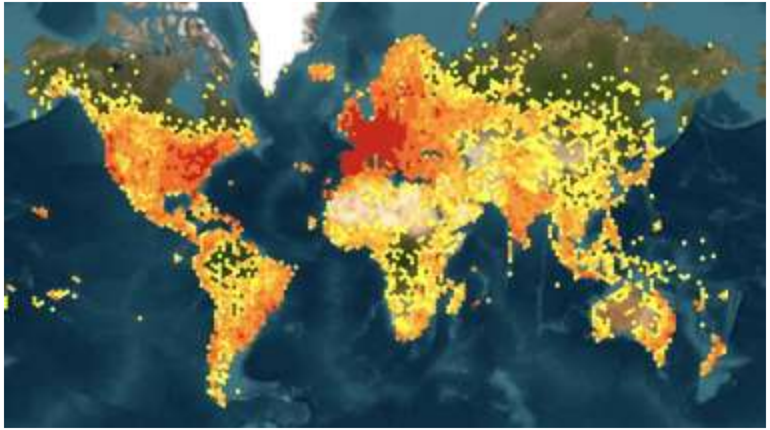
Top-5 data provider to GBIF (world's largest infrastructure for biodiversity data)

- Shared data = revised observations + trusted queries identified by the AI (AI score>0.95)
- Quality filters: potted & cultivated plants removal, region-based filtering (Kew POWO)

GBIF

13 856 500 OCCURRENCES

(87% identified by AI, 13% by humans)



https://doi.org/10.15468/mma2ec

632 citations

RED LIST

nature

PLOS ONE

ANNALS OF BOTANY

WILEY Publishers Since 1807

ELSEVIER

Objective: which species are present in a given location and why ?

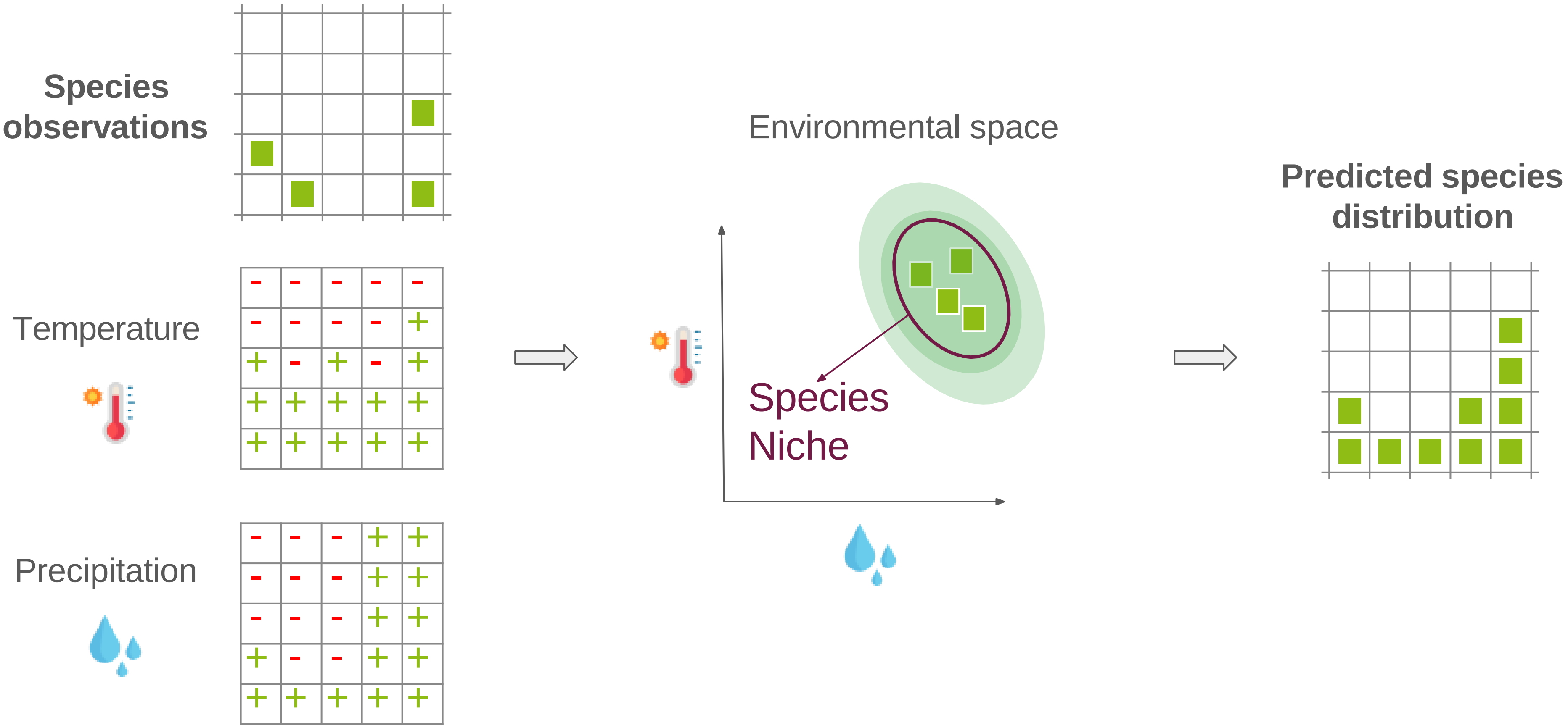
Raw species occurrence data needs to be interpolated in space and time:

Many plant occurrences at world scale

GBIF

But very few locally for most species

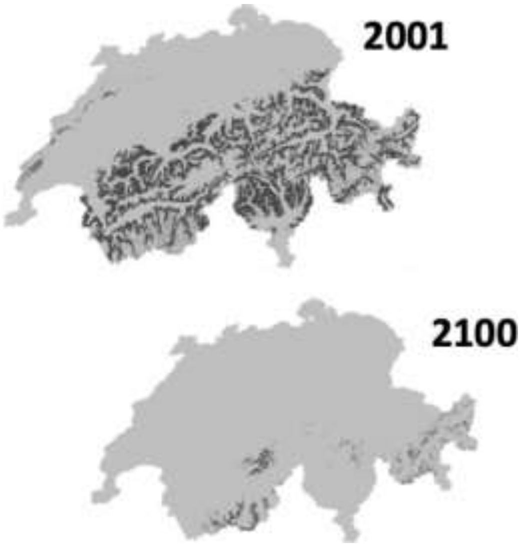
Species Distribution Models (SDM)



Species Distribution Models (SDM)

Motivations

- Help conservation/ plans
- Invasive plant monitoring
- Simulation under climate change
- Learn about species preferences



Credits: “Introduction to species distribution modelling (SDM) in R”, Damaris Zurell

Different types of SDMs

Niche models (e.g. GLM, MAXENT)

- Input: **low-dimensional** (e.g. temperature, precipitation)
- Purpose: **interpretability**, explicability

ML models (e.g. Random Forest, XGBoost)

- Input: **high-dimensional vectors** (e.g. 100 environmental variables)
- Purpose: **performance**, easy to use

Deep SDMs (e.g. CNNs, transformers)

- Input: **complex signals** (e.g. remote sensing images, time series)
- Purpose: **performance** on **large number of species**, **very high resolution**

Remote sensing based SDM

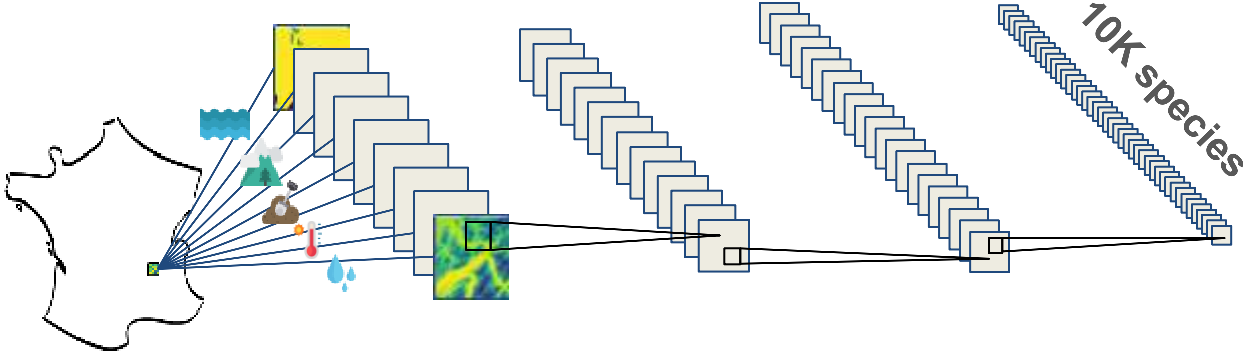
PLOS COMPUTATIONAL BIOLOGY

Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment

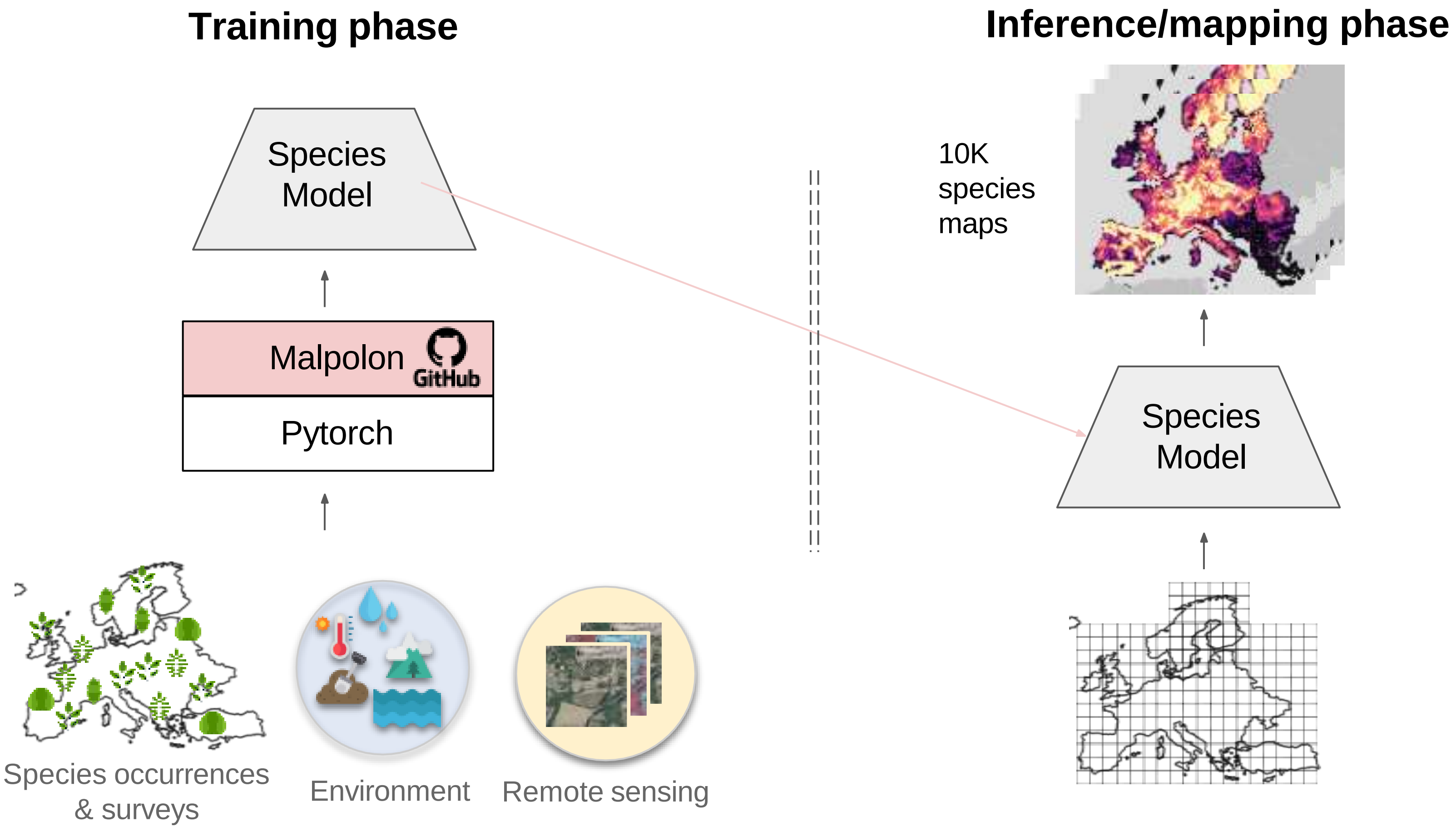
Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, Christophe Botella, François Munoz, Alexis Joly

Model Input =
data cubes

Model Output =
Suitability score of each species



From models to species mapping



Different tasks vs. available data

Input data: x target: y

- **Abundance data** (very hard to produce)

  Task: predict $\hat{y} = f_\theta(x) \in \mathbb{R}^d$

0	12	0	4	0	0	32	0

- **Presence / absence data** (hard to produce)

  Task: predict $\hat{y} = f_\theta(x) \in [0, 1]^d$

0	1	0	1	0	0	1	0

- **Presence only data** (more data available)

  Task: predict $\hat{y} = f_\theta(x) \in \{1, \ldots, d\}$

1

Limitations of models trained on presence-only data

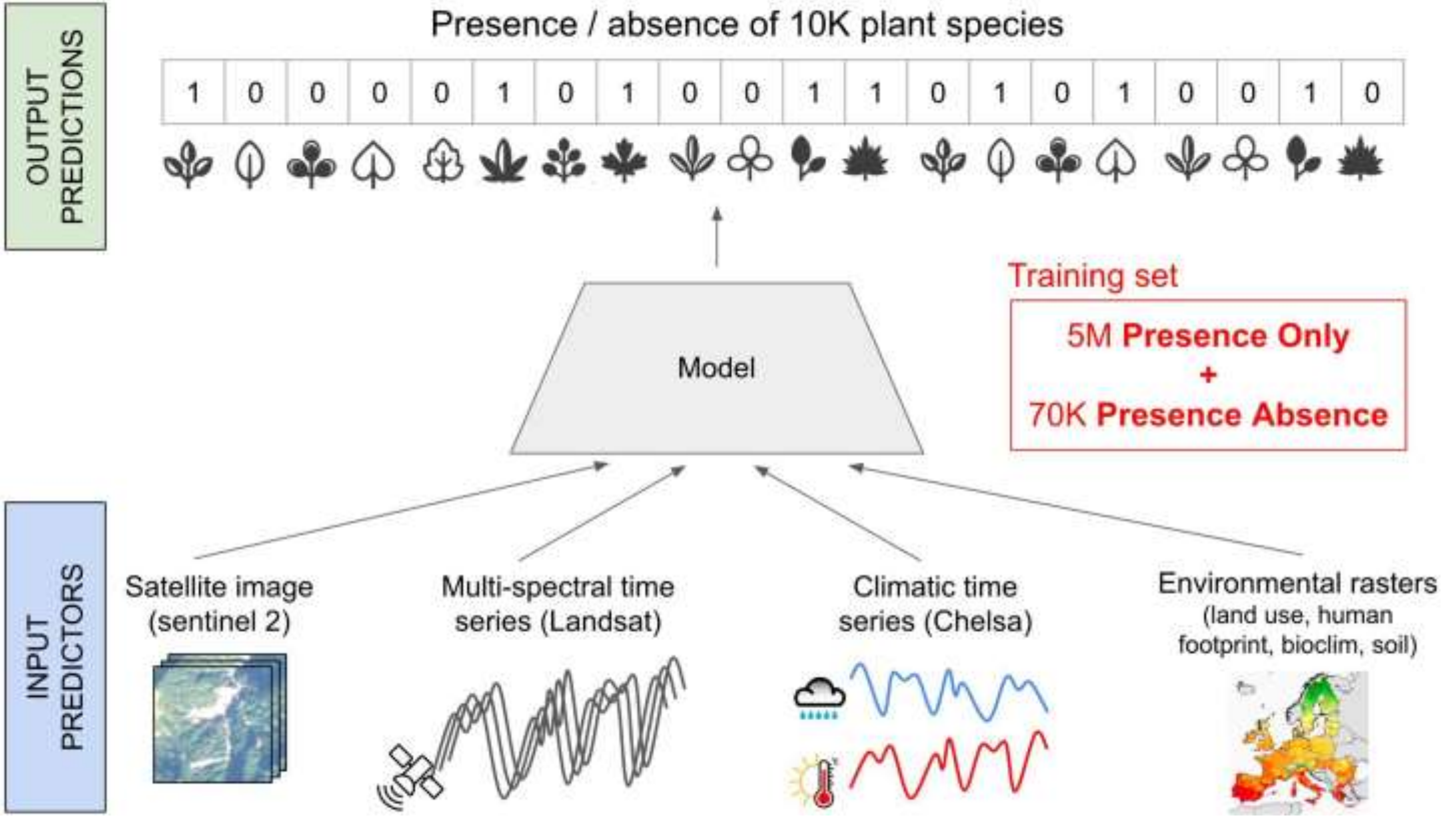
Sensitive to **taxonomic reporting bias**

Observation probability ≠ Presence probability

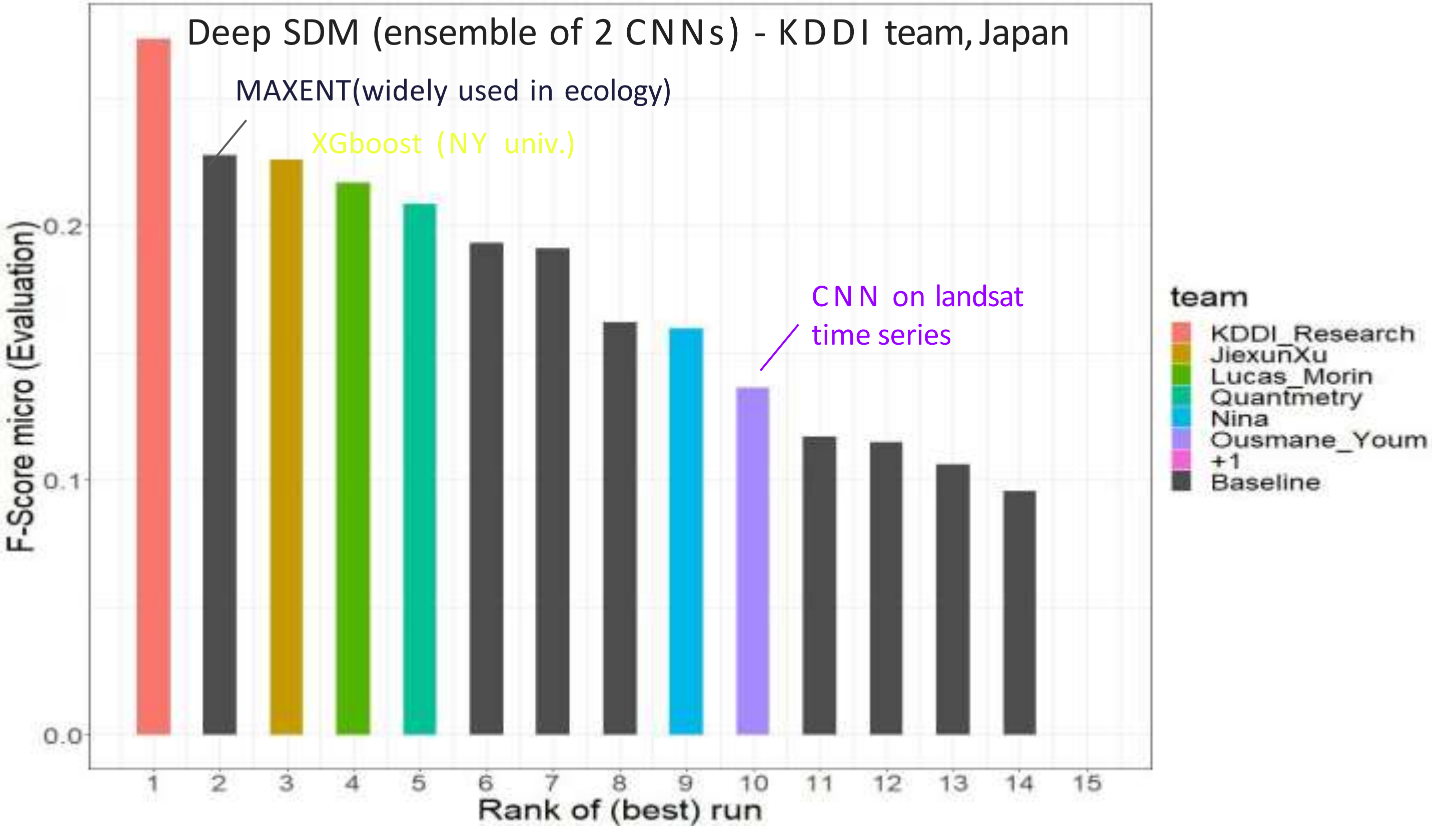
The threshold λ over the estimated probabilities is **hard to set** (we don't know how many species there are)

The probability of each species is **relative** to the others and depends on the **number of species** present somewhere
→ this is not appropriate for mapping each species individually

GeoLifeCLEF challenge 2023 & 2024



GeoLifeCLEF challenge 2023 - results



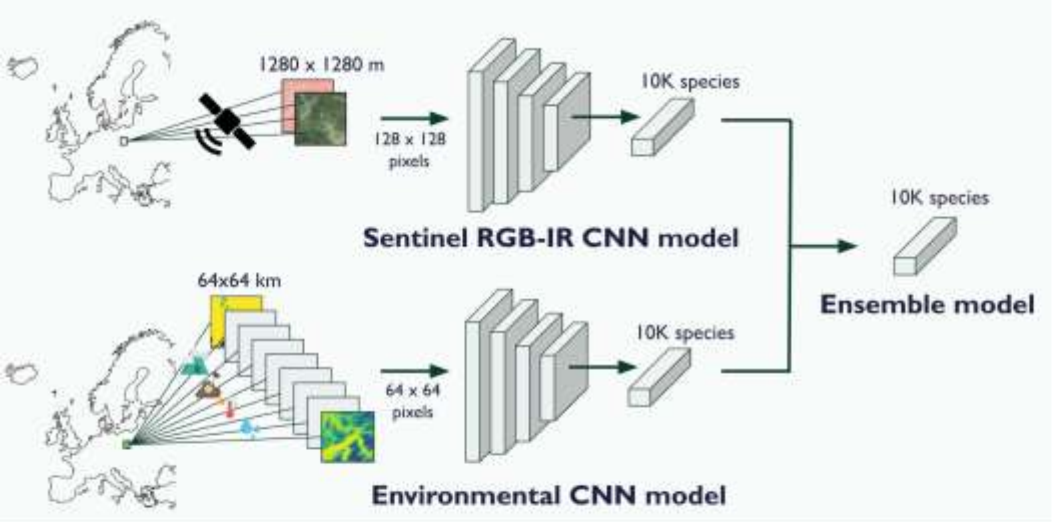
Ranked scores of the best model

GeoLifeCLEF challenge 2023 - best approach

[Leverage Samples with Single Positive Labels to Train CNN-based Models For Multi-label Plant Species Prediction]

Huy Quang Ung, *Ryoichi Kojima*, *Shinya Wada*

Architecture

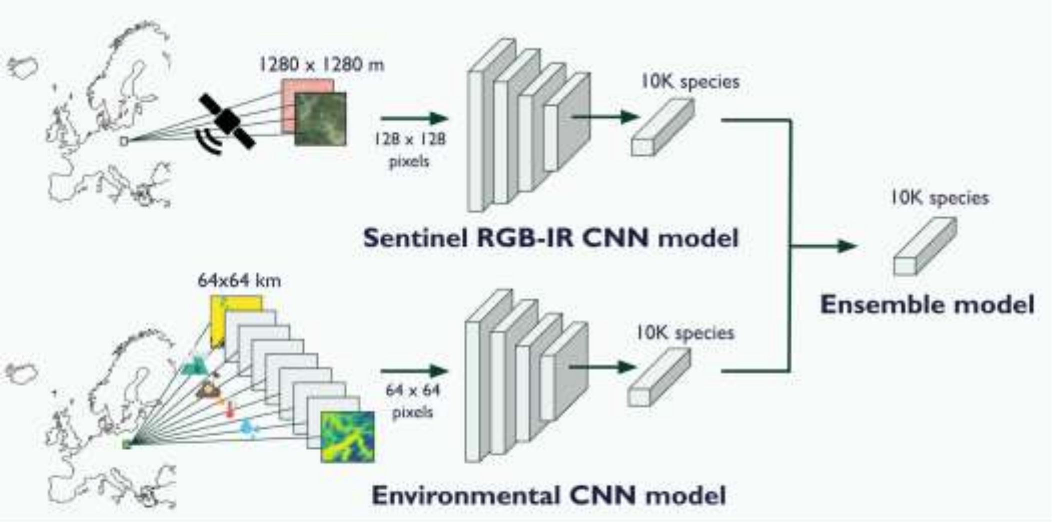


GeoLifeCLEF challenge 2023 - best approach

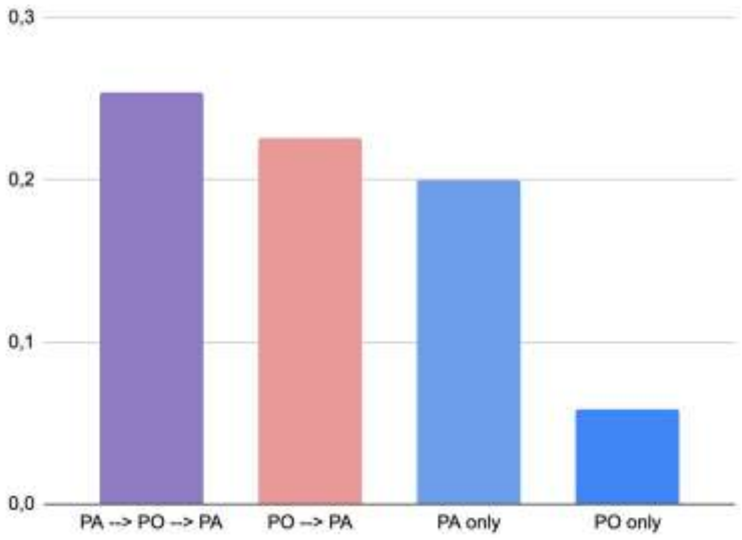
Leverage Samples with Single Positive Labels to Train CNN-based Models For Multi-label Plant Species Prediction

Huy Quang Ung, *Ryoichi Kojima*, *Shinya Wada*

Architecture



Training strategy



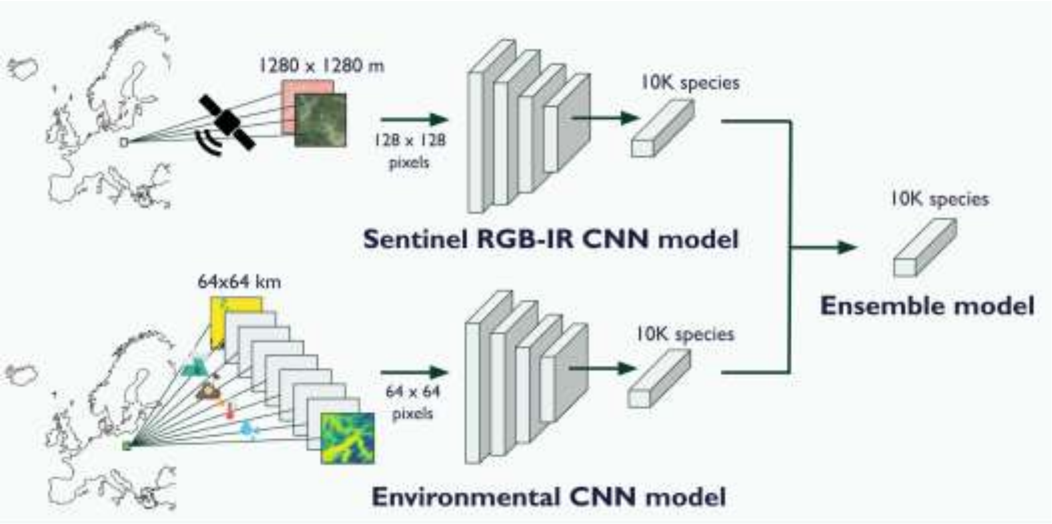
PA = Presence/Absence data (with Binary Cross Entropy loss)
PO = Presence only data (with Cross Entropy loss)

GeoLifeCLEF challenge 2023 - best approach

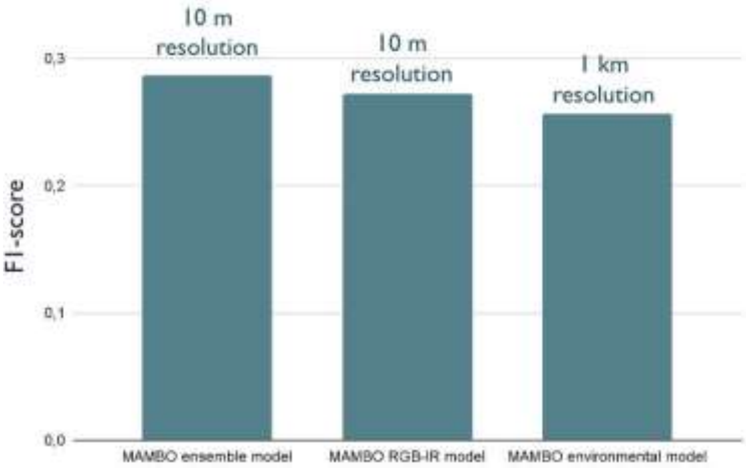
Leverage Samples with Single Positive Labels to Train CNN-based Models For Multi-label Plant Species Prediction

Huy Quang Ung, *Ryoichi Kojima*, *Shinya Wada*

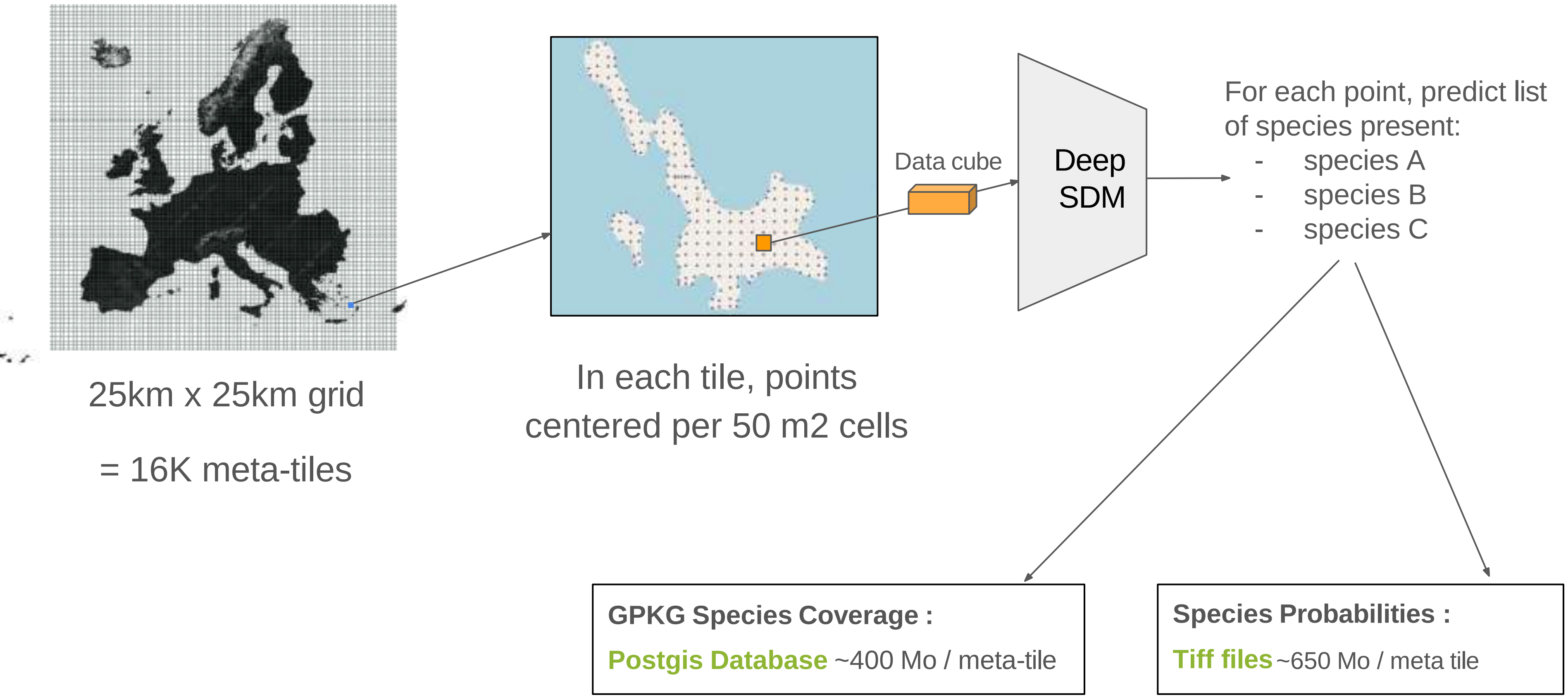
Architecture



Contribution of modalities



Integration in Pl@ntNet for EU-scale species mapping



GeoPl@ntNet Barcelona Cyprus CBN Med Greece About

GUARDEN European Commission

Species predictions

Model: GPN_RGBI_2

Grid: greece

Resolution: 50 m

Species threshold: 50

Fraxinus ornus L. (#6353)

WMS opacity:

Fraxinus ornus

Family: Oleaceae

Genus: Fraxinus

Common Name: Manna

Model prediction score: 13.32

GeoPl@ntNet Barcelona Cyprus CBN Med Greece About

GUARDEN European Commission

Species predictions

Model: GPN_RGBI_2

Grid: cbnmed

Resolution: 50 m

Species threshold: 50

Thymus vulgaris L. (#5404)

WMS opacity:

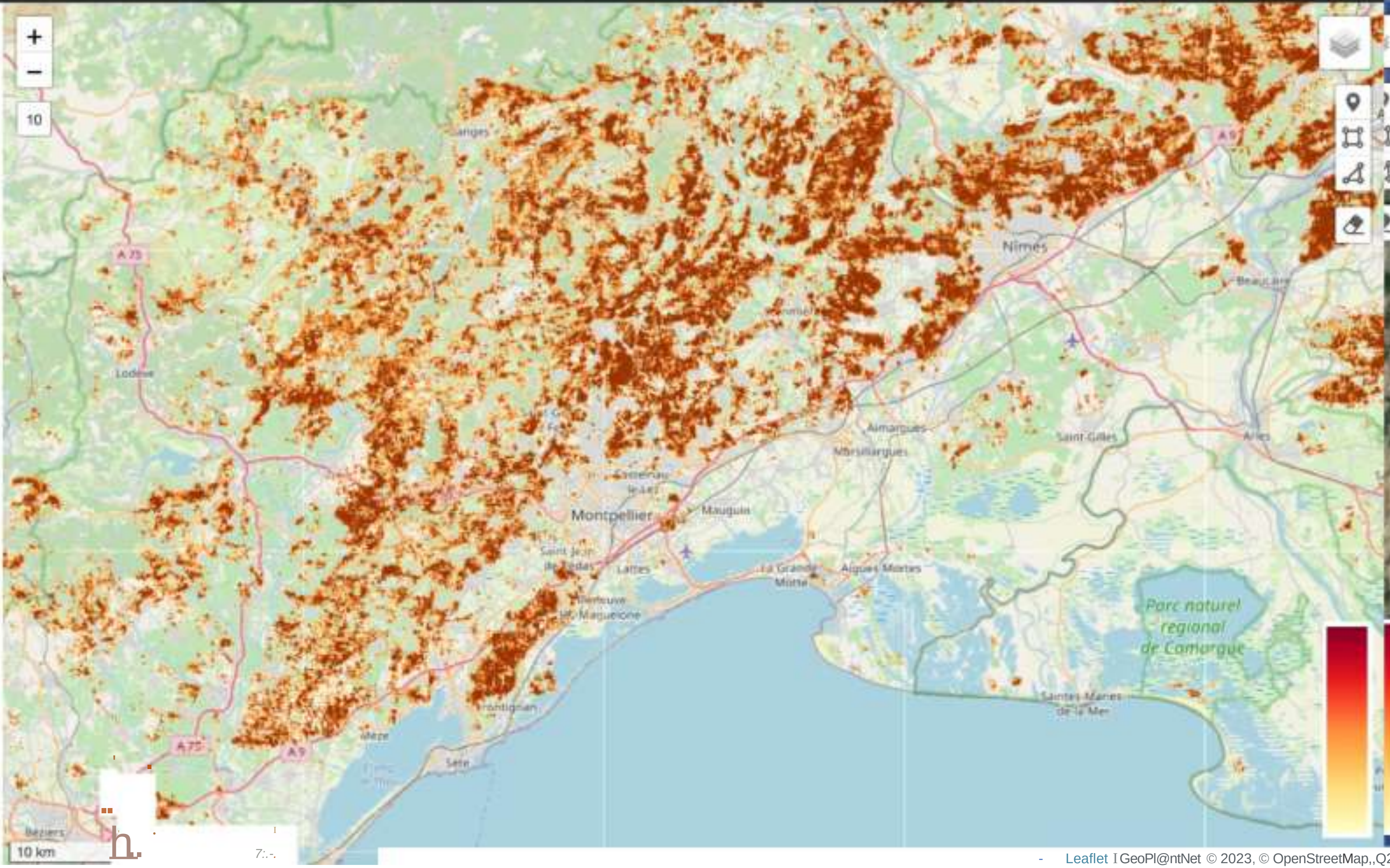
Thymus vulgaris

Family: Lamiaceae

Genus: Thymus

Common Name: Garden thyme

Model prediction score: 13.67



Pl@ntNet Identify Explore Contributions Groups GeoPl@ntNet BETA English alexis joly

GeoPl@ntNet

Anthemis maritima

Family: Asteraceae
Genus: Anthemis
Common Name:
Seaside Chamomile

GeoPl@ntNet

CBN Med

Biodiversity indicators

Shannon Index

WMS opacity:

High Biodiversity

Low Biodiversity

shannon

Band 1 (Gray)
3.901799
2.661361

isoom

GeoPl@ntNet Barcelona

Species List (180)

Get coverage

Id	Name	IUCN	Tree	Invasive	% Coverage
905	Limbarda crithmoides (L,) Dumort				90.57
4216	Salicornia fruticosa (!:J..1.,.				83.82
9752	Phragmites australis (Cav.) Trin. ex Steud.	LC			82.70
3878	Trii:iolium i;iannonicum (Jacg.) Dobrocz.				82.56
7137	Juncus maritimus Lam.				81.70
2212	Limonium vulgare Mill.				77.42
8238	Juncus acutus L.	LC			72.80
4012	Anthemis maritima L.				70.72
7590	Suaeda maritima (bl Dumort.				67.99
9411	Artemisia caerulescens L.				67.39
4846	Suaeda vera Forssk. ex J.F.Gmel.				67.19
7950	Schoenus nigricans L.	LC			63.67
8456	Elaeagnus angustifolia L.	LC	●		63.62

Biodiversity indicators

Shannon Index

WMS opacity:

High Biodiversity

Low Biodiversity

shannon

Band 1 (Gray)

3.901799

2.661361

Mapping biodiversity conservation indicators

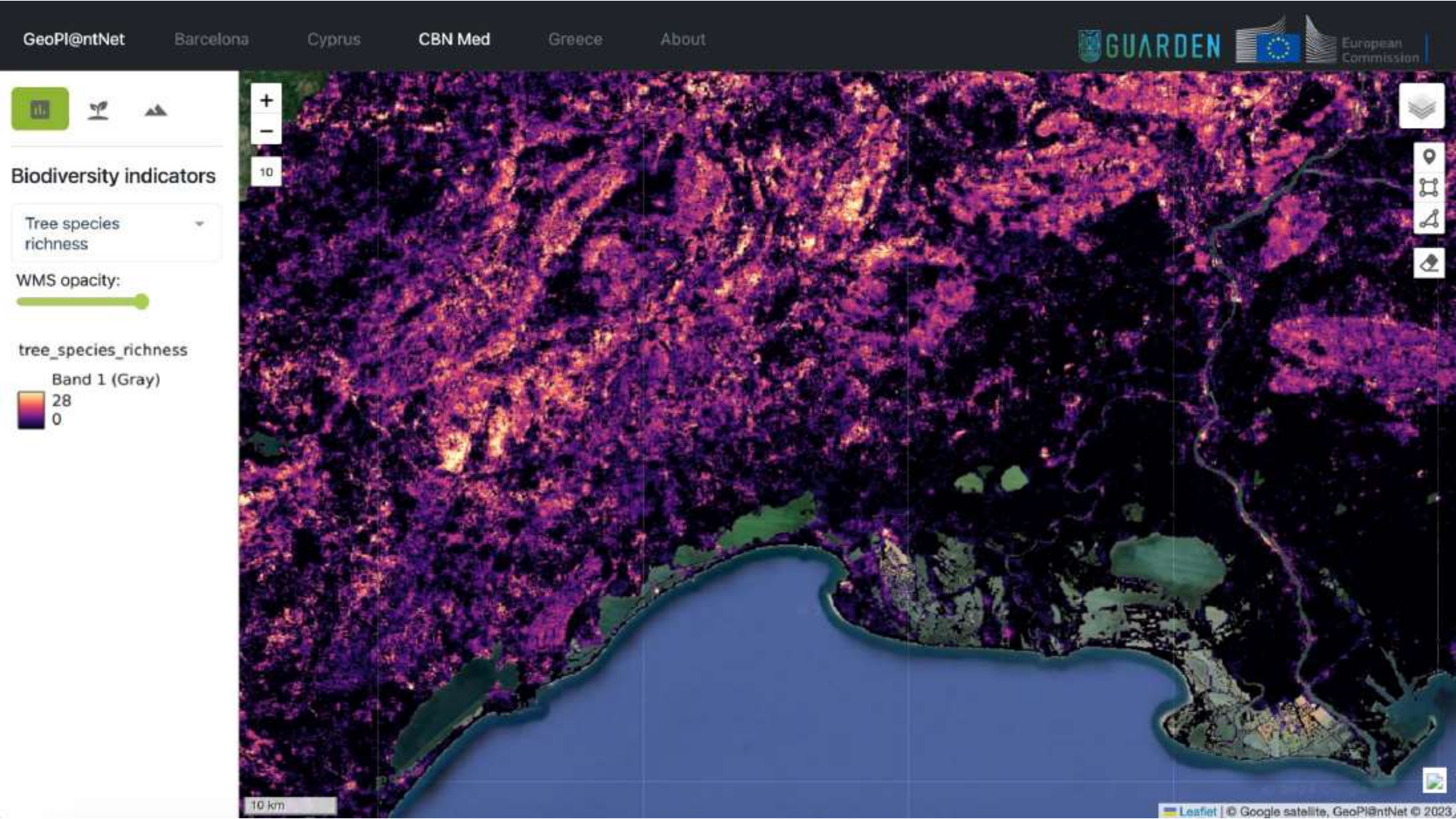
From the species assemblage predicted at each point

$$S_\lambda(x) := \{k \in \mathcal{Y} : \hat{\eta}_k(x) > \lambda\}$$

We can compute indicators such as:

- The number of endangered species (e.g. on IUCN red list)
- The proportion of woody species (carbon capture)
- The diversity of species (e.g. Shanon index)
- The number or rare species
- The EUNIS habitat (using a species-to-habitat model)

We can construct maps of such indicators at very high resolution by computing $S_\lambda(x)$ for all x_i on a dense spatial grid


GeoPl@ntNet
Barcelona
Cyprus
CBN Med
Greece
About
GUARDEN
European Commission
Biodiversity indicators
Tree species richness
WMS opacity:
tree_species_richness
Band 1 (Gray)
28
0
10
10 km
Leaflet | © Google satellite, GeoPl@ntNet © 2023

GeoPl@ntNet Barcelona Cyprus CBN Med Greece About

GUARDEN European Commission

Biodiversity indicators

Invasives species

WMS opacity:

invasive

Band 1 (Gray)
63
0

GeoPl@ntNet Barcelona Cyprus CBN Med Greece About

GUARDEN European Commission

Biodiversity indicators

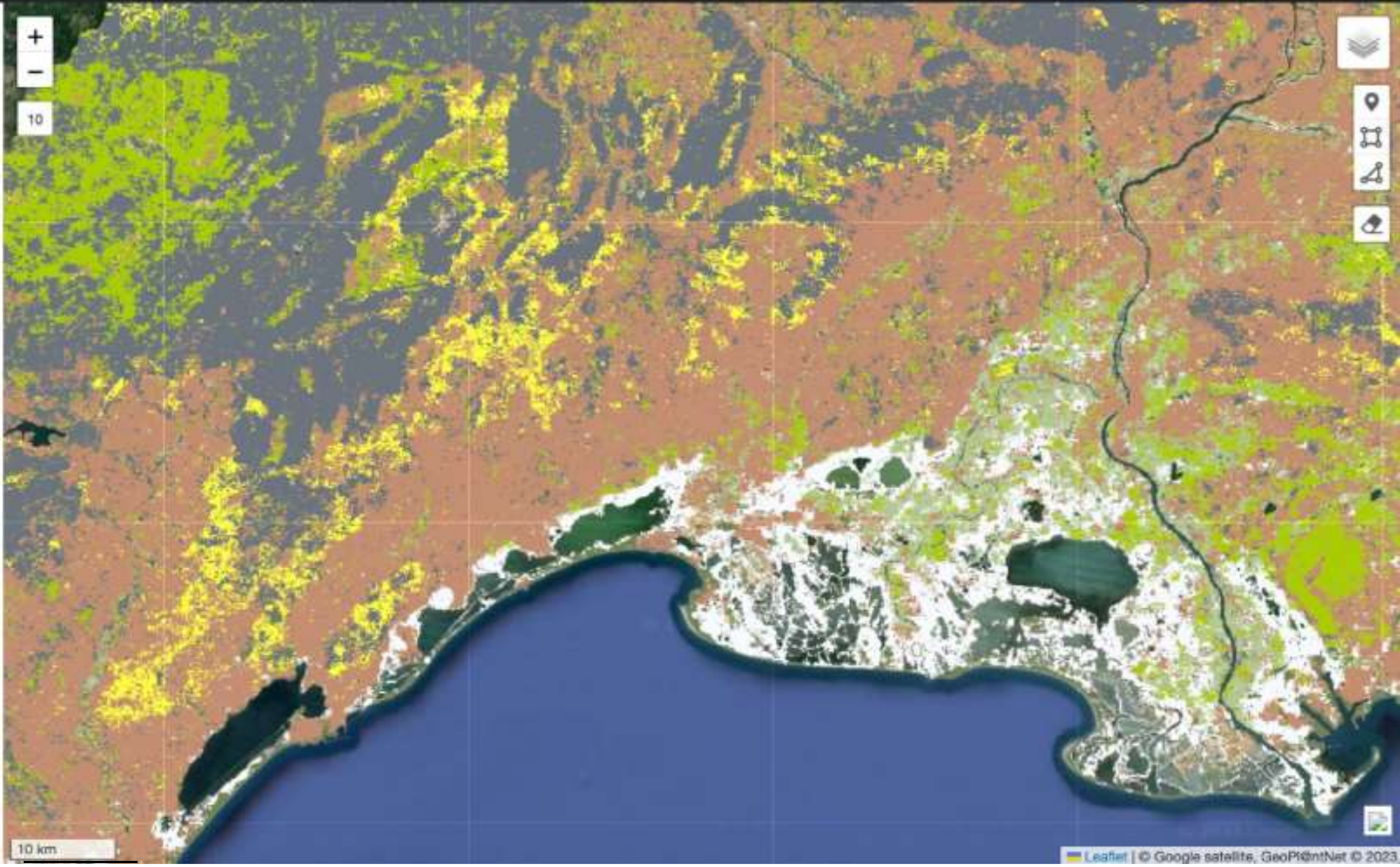
Habitat

WMS opacity:

habitat

Band 1 (Gray)

0 M A 2
0 M A 2
0 M A 2
0 M A 2
D M A 2
D M A 2
0 M A 2
0 M A 2
0 M A 2
0 M A 2
0 M A 2
- N
- N
- N
- N
- N
- N
- N
- N
- N



GeoPl@ntNet Barcelona Cyprus CBN Med Greece About

Biodiversity indicators

Specialization

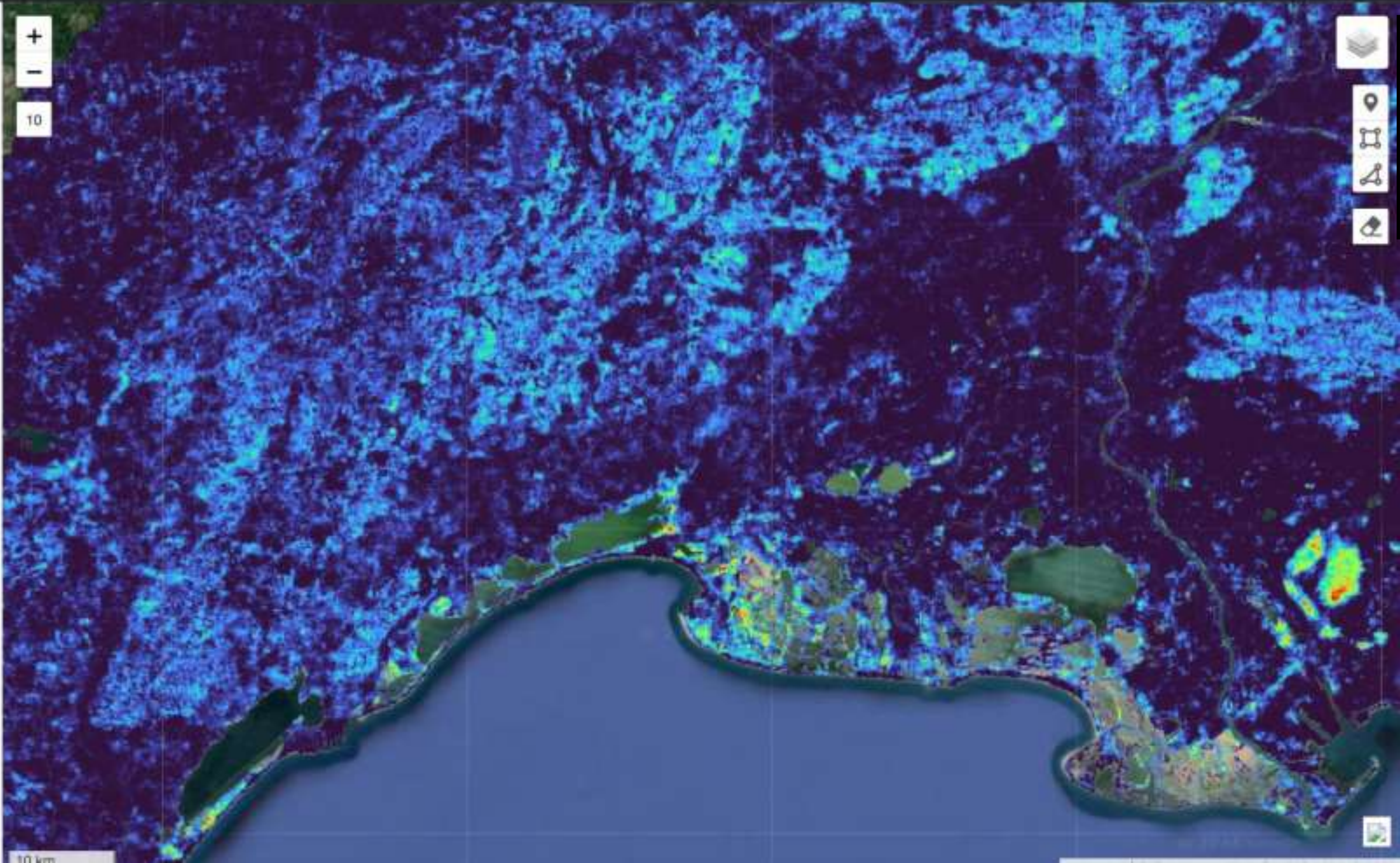
WMS opacity:

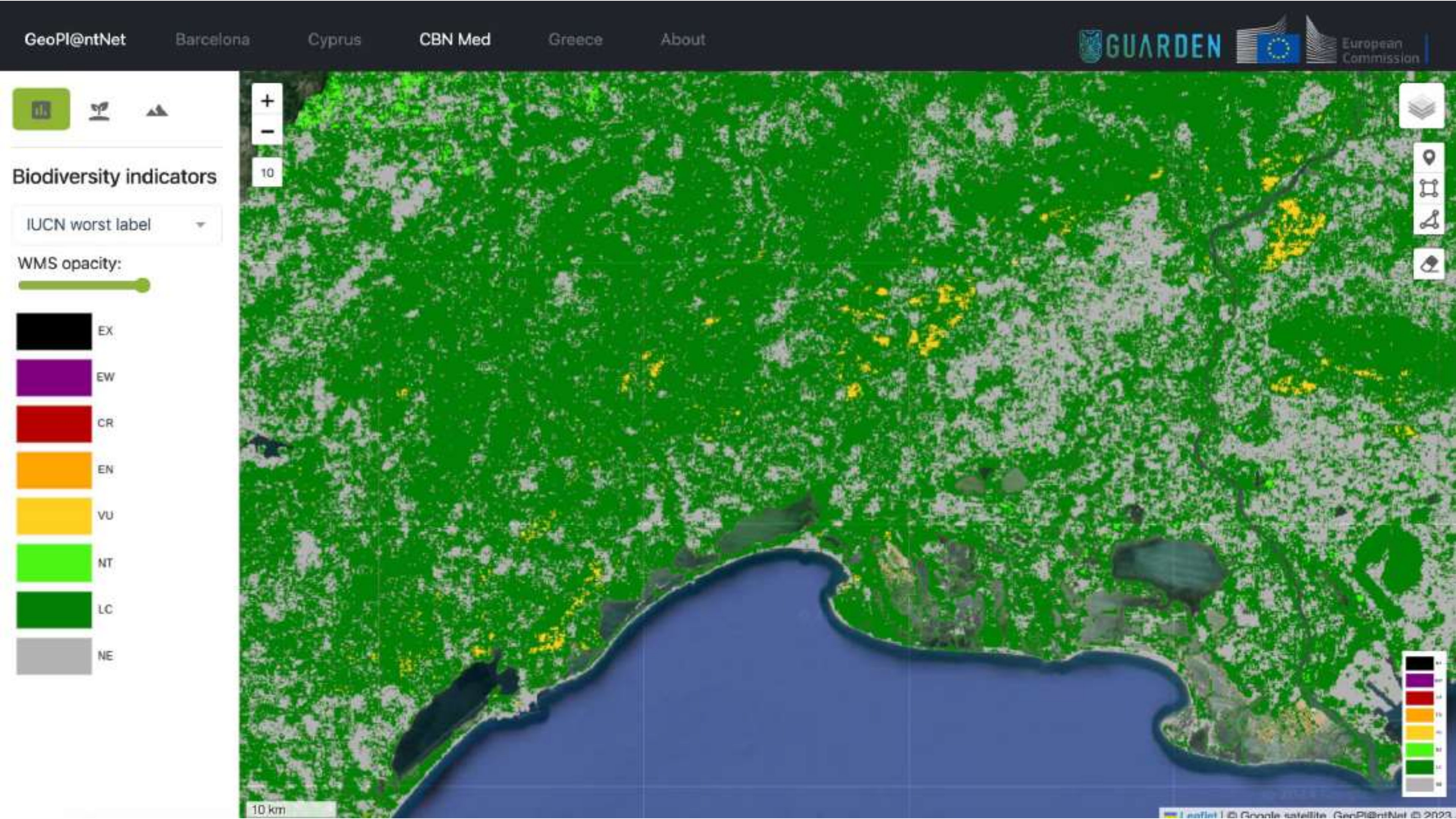
endemism

Band 1 (Gray)

203

0




GeoPl@ntNet
Barcelona
Cyprus
CBN Med
Greece
About
GUARDEN
European Commission
Biodiversity indicators
IUCN worst label
WMS opacity:
EX
EW
CR
EN
VU
NT
LC
NE
10
10 km

GeoPl@ntNet Barcelona Cyprus CBN Med Greece About

GUARDEN European Commission

Biodiversity indicators

EU directive

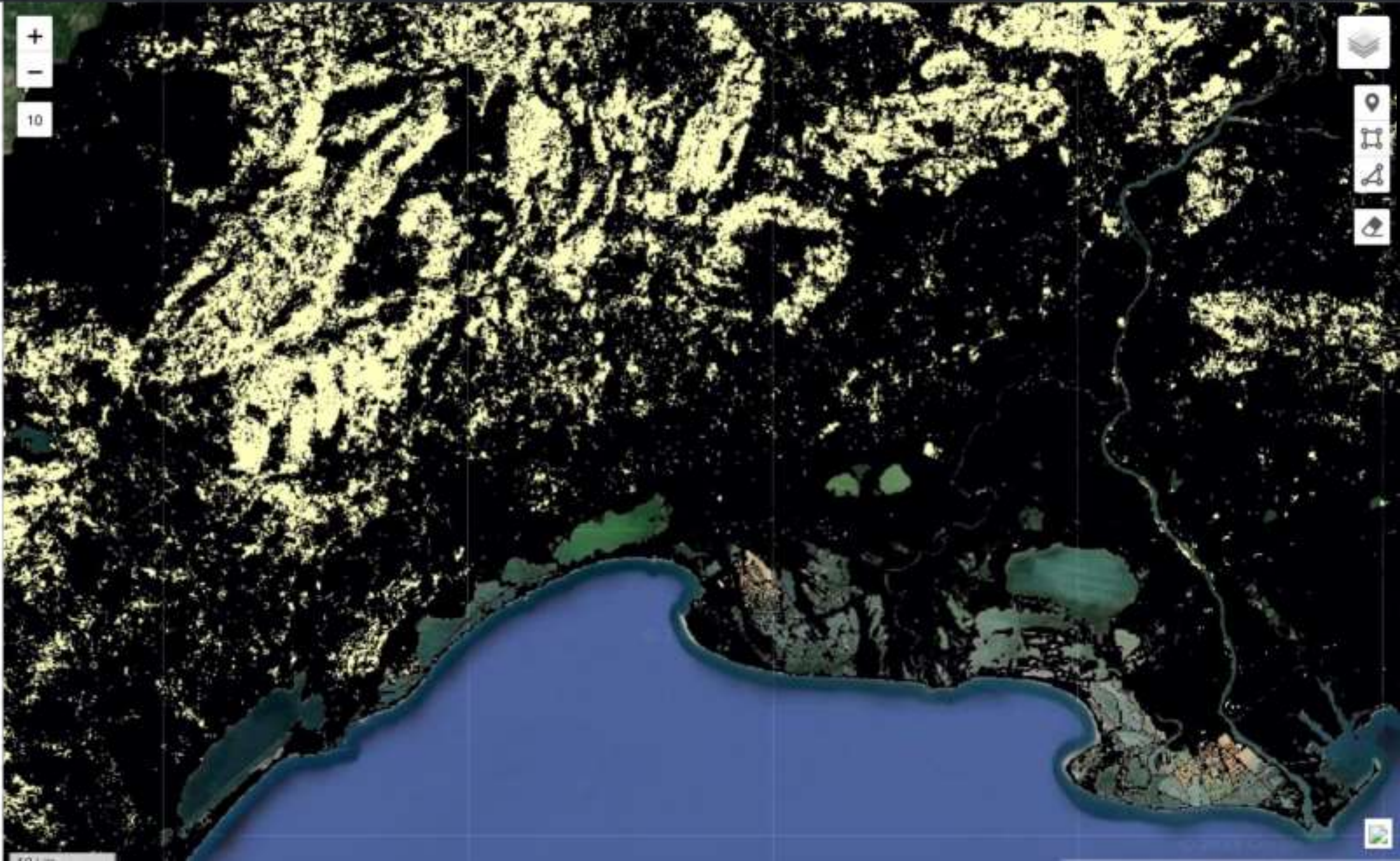
WMS opacity:

eu_directive

Band 1 (Gray)

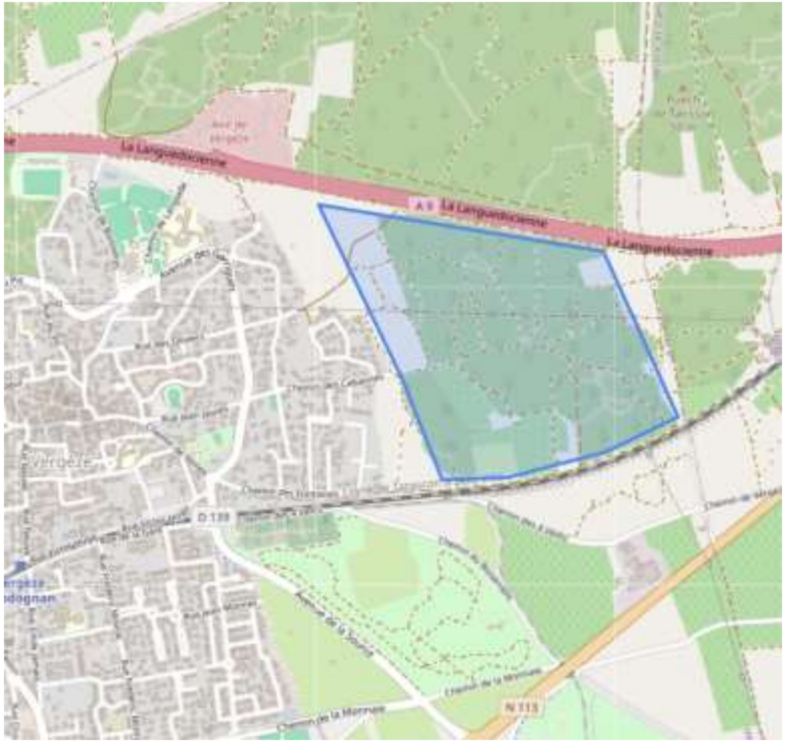
1

0



GeoPl@ntNet

Discover plant biodiversity close to home and help protect it better



Species	Habitat	Conservation	Ecosystem	Threat

Presence of rare species

Presence of species on the European directive

Humm, maybe we should not construct our new facility here

Thank you