From Pl@ntNet to GeoPl@ntNet: new AI-based approaches for monitoring plant biodiversity

Alexis Joly, Pierre Bonnet, Hervé Goëau, Antoine Affouard, J.C. Lombardo, Mathias Chouet, Hugo Gresse, Christophe Botella, Titouan Lorieul, Vincent Espitalier, Benjamin Deneu, Joaquim Estopinan, Cesar Leblanc, Camille Garcin, Diego Marcos, Maximilien Servajean, François Munoz, Joseph Salmon, Christophe Botella

PART I

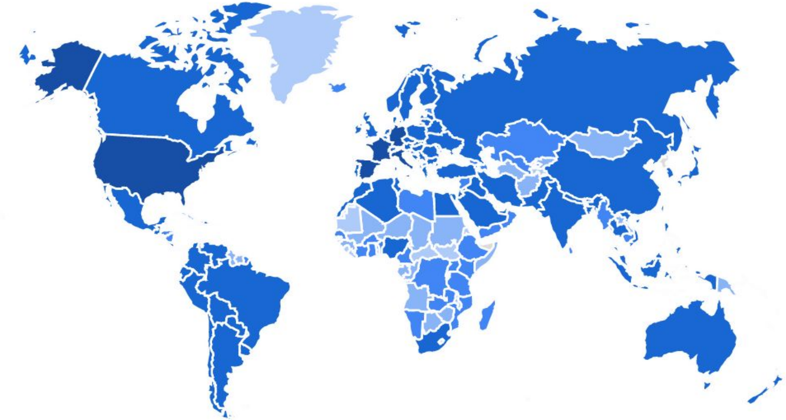
Pl@ntNet under the hood

Pl@ntNet

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

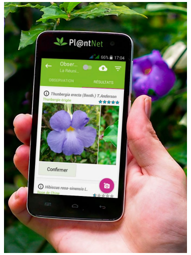
Pl@ntNet app

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



Personal Usage

Nature, walks



Gardening

Phytotherapy

Professional Usage


Agro-ecology


Natural Areas Management

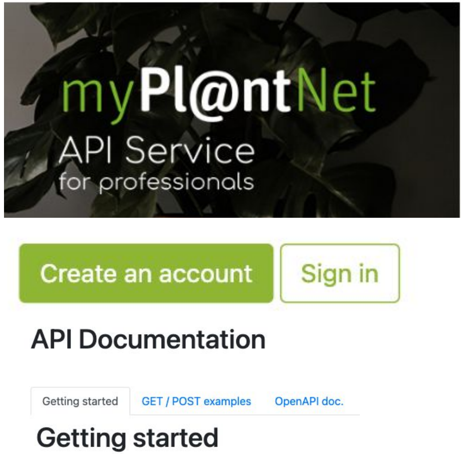

Education, animation

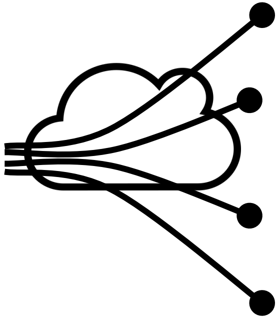
Tourism

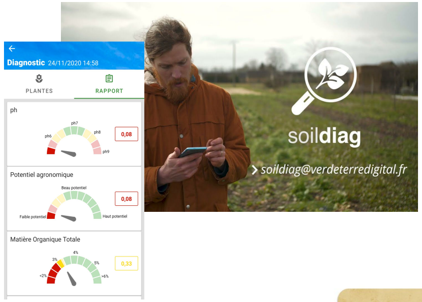
Trade

Pl@ntNet API

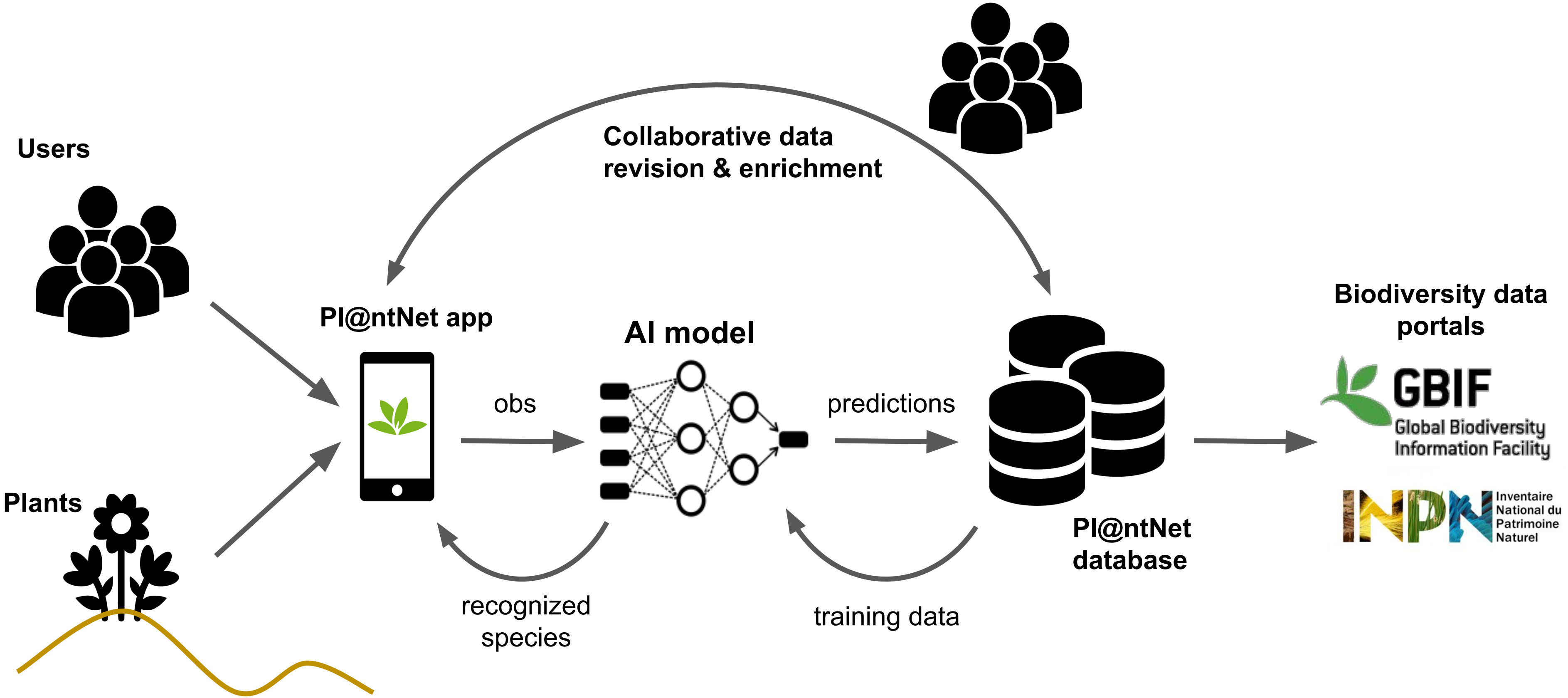
- A secured API providing developers programmatic access to Pl@ntNet engine
- **8K developer accounts** (companies, researchers, citizen observatories)
- Integrated in **European Open Science Cloud** (EOSC)





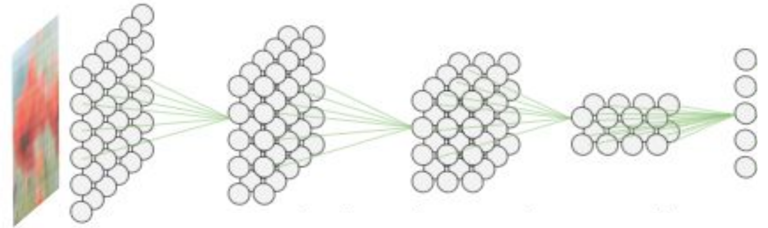


Key concept of Pl@ntNet: Collaborative AI



Pl@ntNet AI model

Model trained on **Jean Zay super-computer** (cross-entropy loss) on a big dataset of valid observations (5-6 days of training)



Softmax output (46K-dimensional)

$$\longrightarrow \quad \sigma(f(x))$$

current version: **Vision transformer (DinoV2)** → **Top1 accuracy = 0.73**

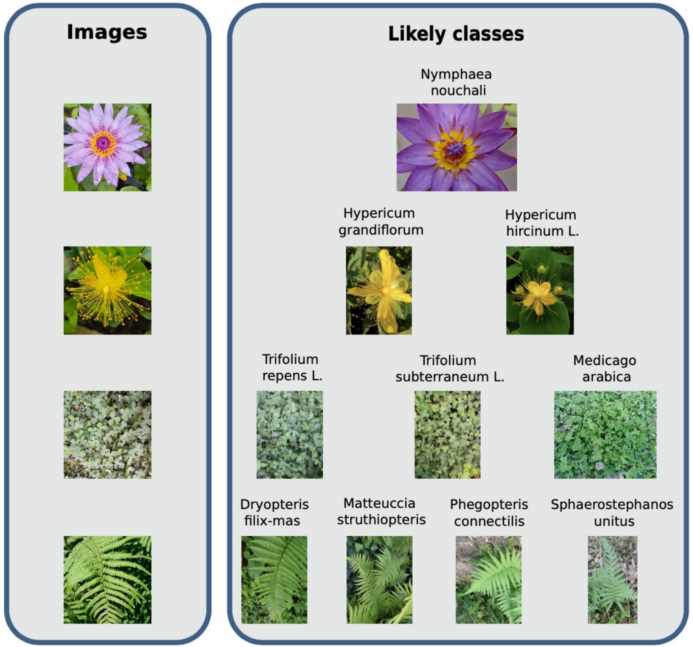
previous version: Convolutional Neural Network (IV3) → Top1 accuracy = 0.70

46K species (+ reject classes)

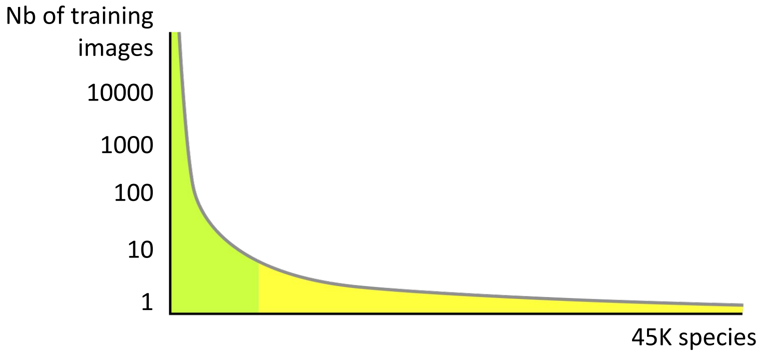
6.5M training images (undersampling for classes > 1000 images)

A difficult problem: uncertainty

Aleatoric uncertainty
Ambiguity (irreducible)



Epistemic uncertainty
Long-tail distribution



Top1 accuracy > Macro-average Top1 accuracy

0.73 > **0.59**

Pl@ntNet Returned results: set-valued

Pointwise error control

Threshold the **accumulated probability**

$$\sum_i \sigma_i(f(x)) > \theta$$

***Papaver rhoeas* L.** **0.63**
+ *Papaver somniferum* L. **0.76**
+ *Papaver californicum* A. **0.87**
+ *Glaucium corniculatum L.* **0.94**
θ=0.95
+ Glaucium flavum L. **0.98**

Average set size control

Threshold the **probability** so as to return less than **K classes on average**

$$\sigma_i(f(x)) > \theta'$$

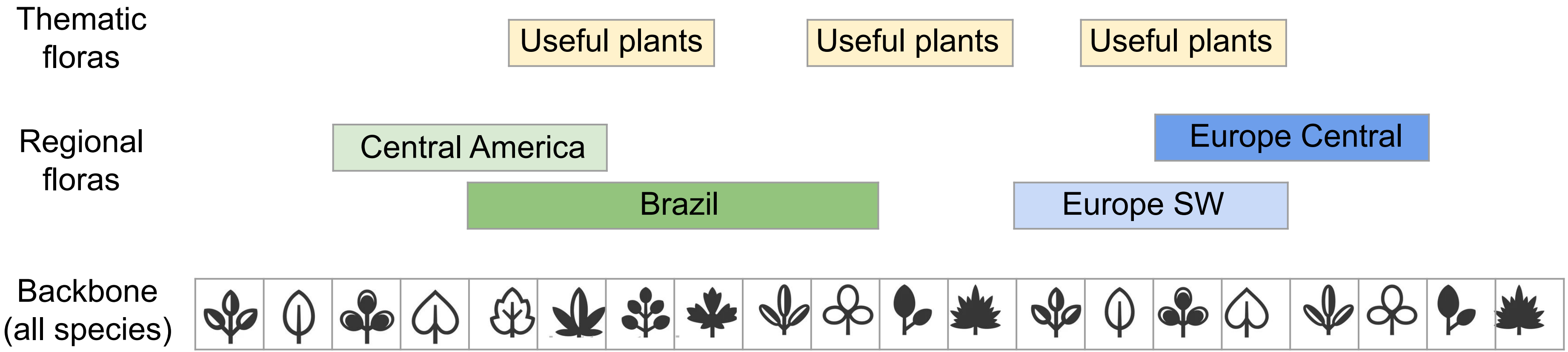
***Papaver rhoeas* L.** **0.63**
***Papaver somniferum* L.** **0.13**
***Papaver californicum* A.** **0.11**
θ'=0.1
Glaucium corniculatum L. **0.07**
Glaucium flavum L. **0.04**

→ Average-K classification
(proof of consistency)

Use of regional or thematic floras

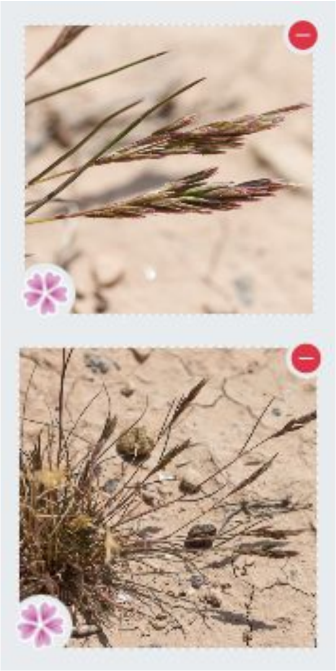
Restricting the hypothesis space to a particular flora allows improving the identification accuracy

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

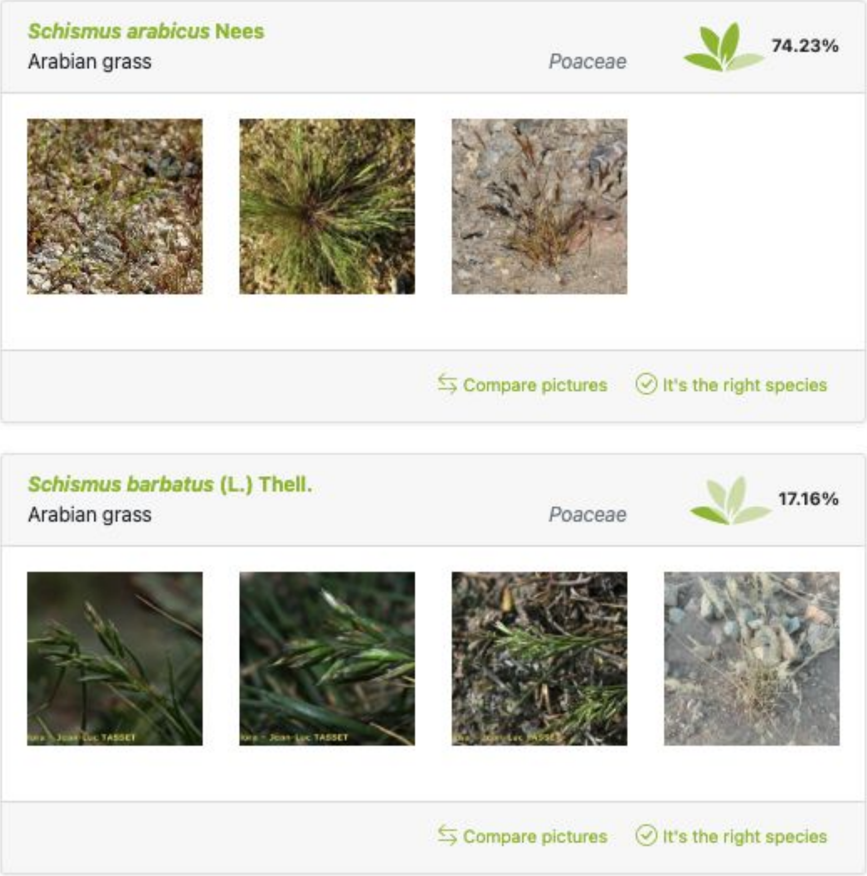


Use of regional or thematic floras

Query

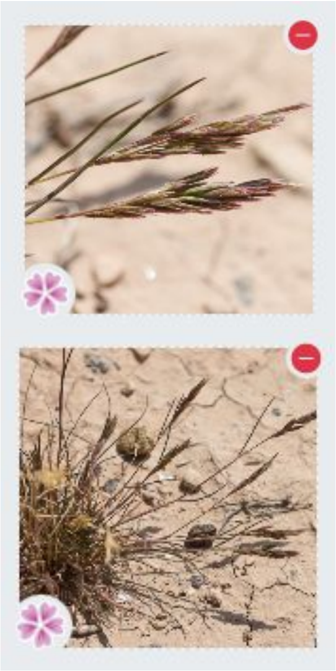


Identify in **World flora**

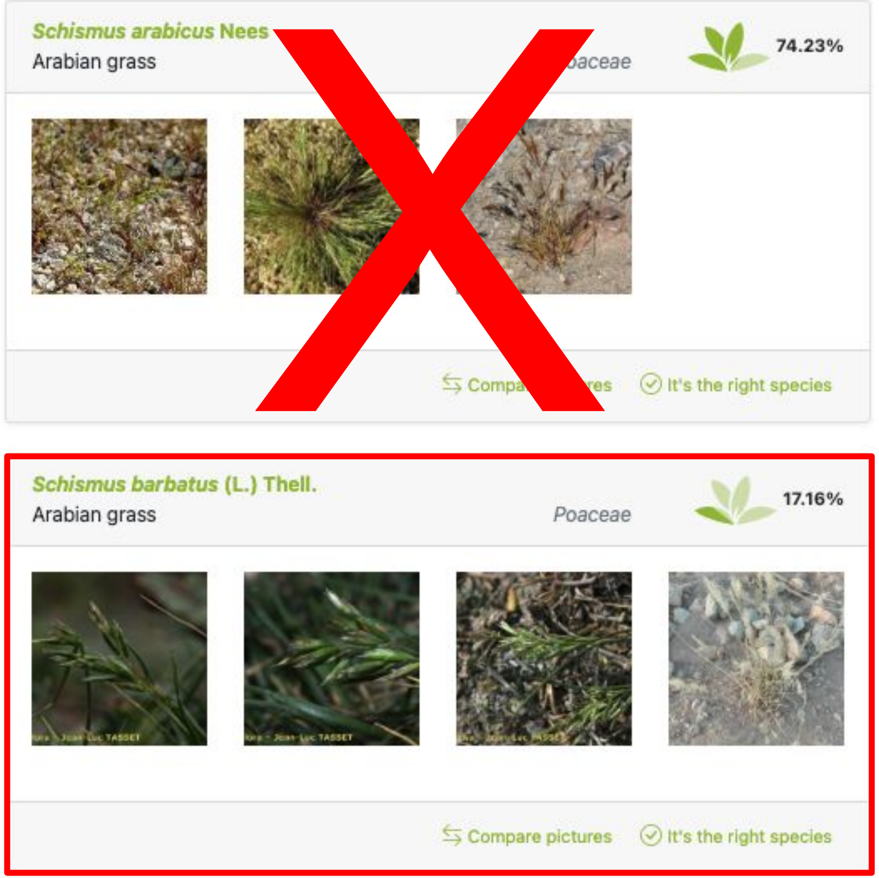


Use of regional or thematic floras

Query

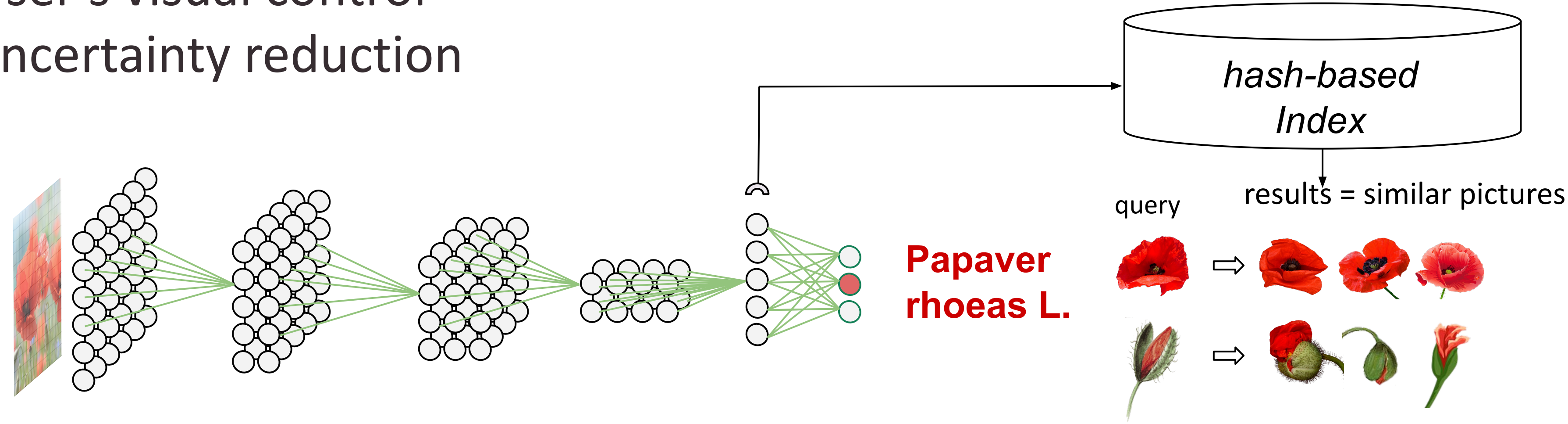


Identify in **West Europe**



Pl@ntNet Similarity search

User's visual control =
uncertainty reduction



Deep neural network

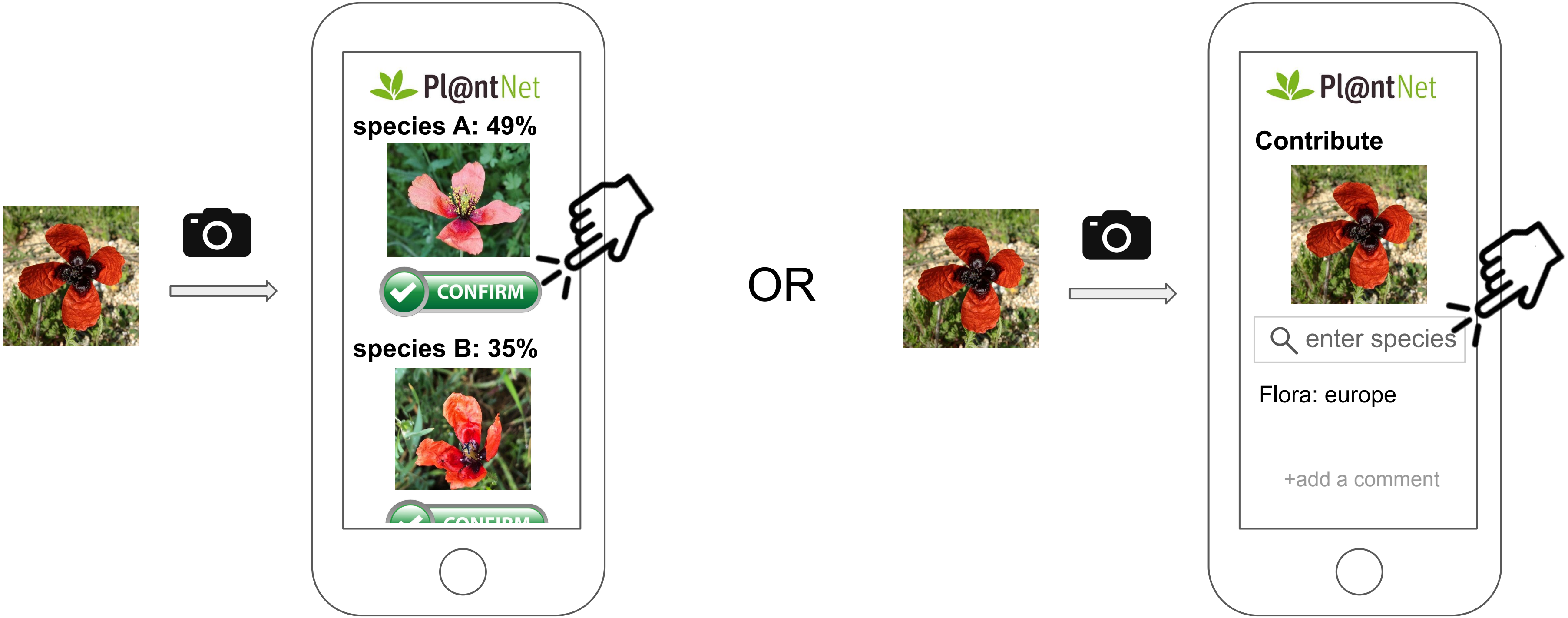
Similarity search engine
9M images

→ 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.

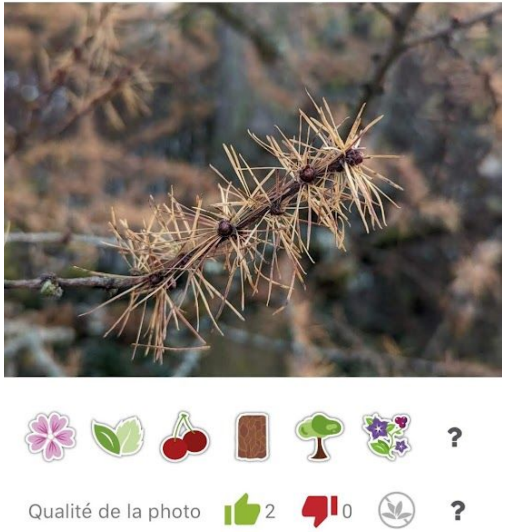
User’s contributions

Users can contribute their observations



User's revisions

Users can revise observations of other users.

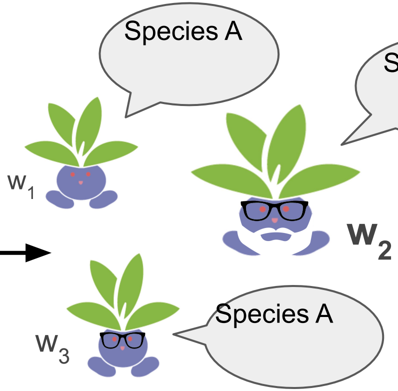


Cooperative Learning algorithm

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

An observation

x_i



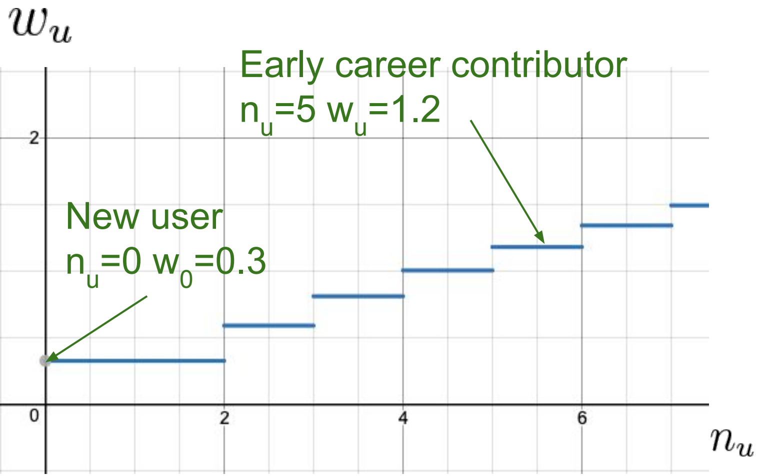
$$y_i = \arg\max_k \sum_{u \in U_i} w_u \mathbf{1}(y_i^u = k)$$

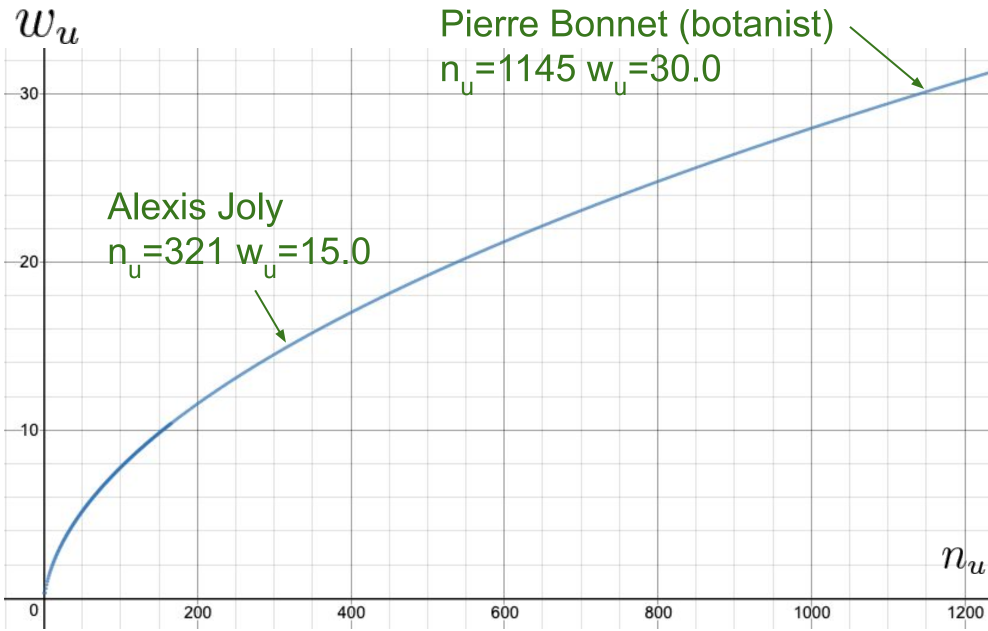
$U_i =$ Set of users who provided a label y_i^u for the observation x_i

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

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





Cooperative Learning algorithm

Practically, n_u is estimated from the set of **valid observations** for which the user has suggested the correct species first

$$n_u = |\{j : \exists i \; y_i^u = \hat{y}_i | \, v(x_i) = 1\}|$$

Where $v(x_i)$ is a function that determines if an observation is valid or not:

$$v(x_i) = \begin{cases} 1 & if \; s_{y_i}(x_i) > \theta, \eta_{y_i}(x_i) > \theta_\eta \\ 0 & otherwise \end{cases}$$

Confidence score (~ quantity of votes)

$$s_{y_i}(x_i) = \sum_{u \in U_i} w_u \mathbf{1}(y_i^u = y_i)$$

Agreement score (~ species proba)

$$\eta_{y_i}(x_i) = \frac{s_{y_i}(x_i)}{\sum_k s_k(x_i)}$$

Cooperative Learning algorithm

Parameters are estimated through an expectation-maximisation algorithm

Initialization:

$$w_u = w_0$$

Same weight for all users

Repeat until convergence:

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

Step 1: Estimate **most likely species** for **all observations**

$$v(x_i) = \begin{cases} 1 & if\ s_{y_i}(x_i) > \theta, \eta_{y_i}(x_i) > \theta_\eta \\ 0 & otherwise \end{cases}$$

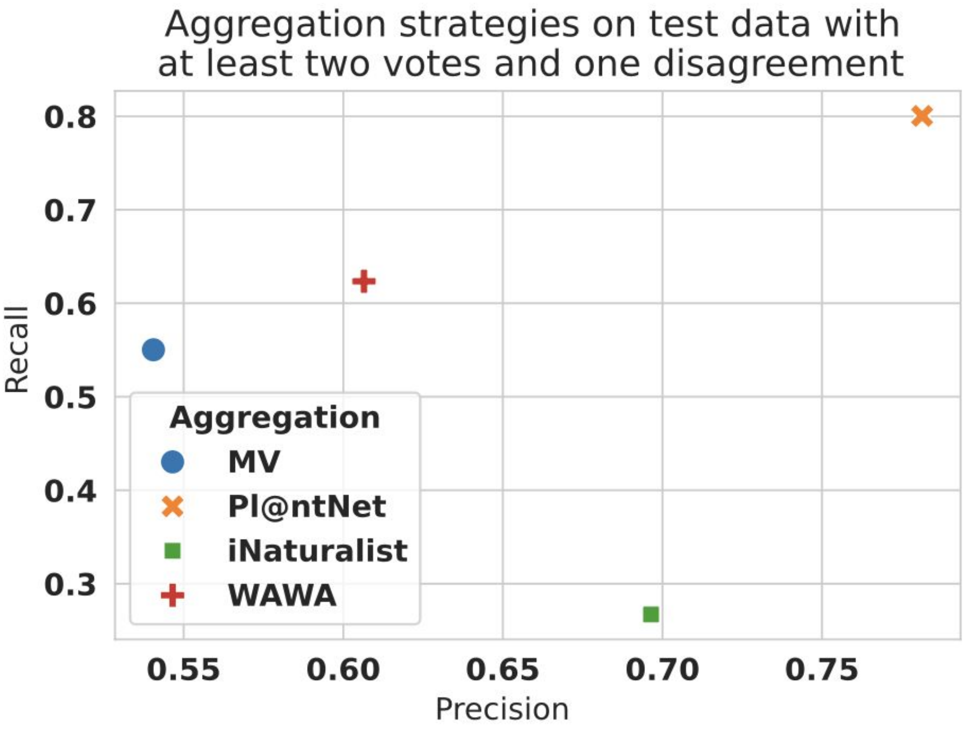
Step 2: Determine **valid observations** based on quantity of votes and probability

$$w_u = g(|\{j : \exists i\ y_i^u = \hat{y}_i | \, v(x_i) = 1\}|)$$

Step 3: update **weights** of **users** based on their **number of valid species**

Cooperative Learning algorithm

Algorithm evaluation (on a subset of observations with ground truth labels)



Majority Vote (MV)

$$\mathrm{MV}(i, \{y_i^u\}_u) = \arg\max_{k\in[K]} \sum_{u\in U_i} \mathbb{1}(y_i^u = k)$$

Worker agreement with aggregate (WAWA)

$$\mathrm{WAWA}(i, \mathcal{D}) = \arg\max_{k\in[K]} \sum_{u\in U_i} w_u \mathbb{1}(y_i^u = k)$$

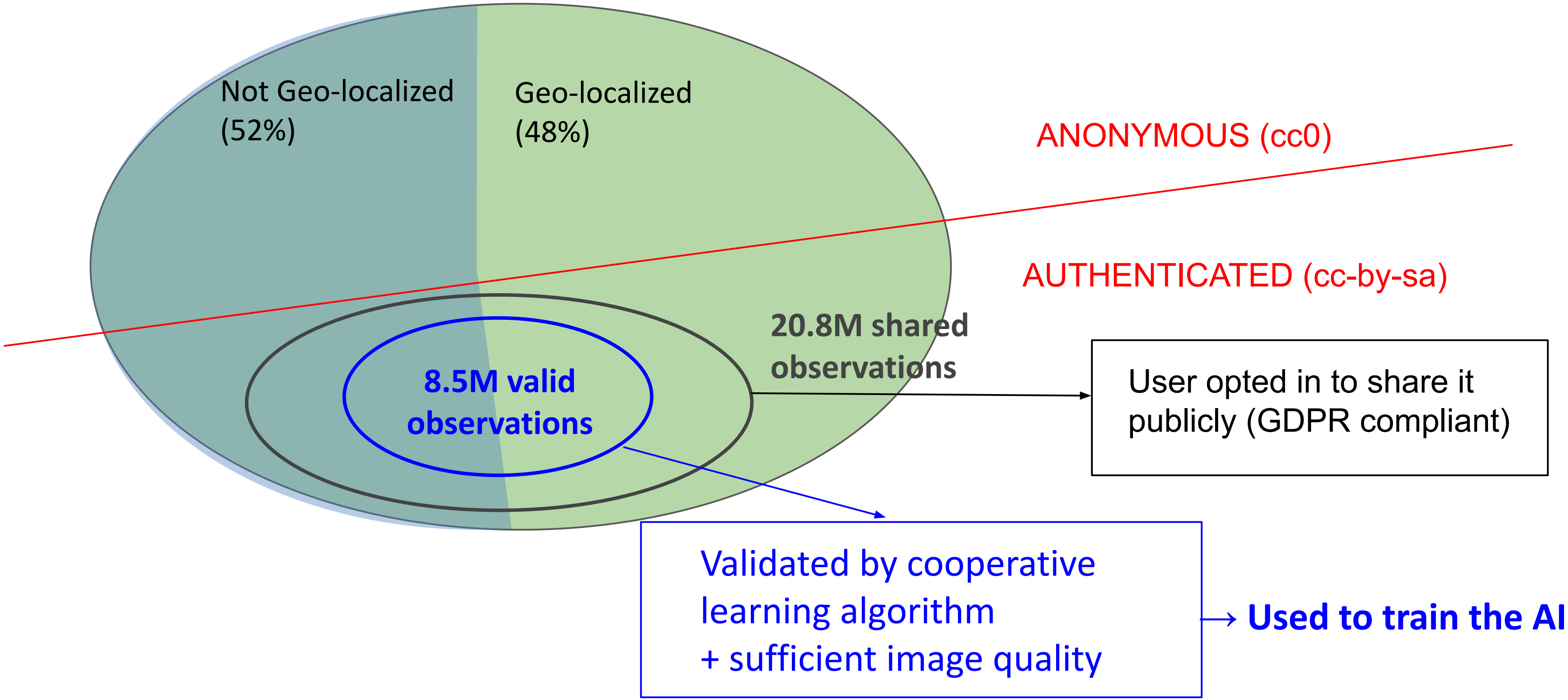
$$\text{with } w_u = \frac{1}{|\{y_{i'}^u\}_{i'}|} \sum_{i'=1}^{|\mathcal{D}|} \mathbb{1}\left(y_{i'}^u = \mathrm{MV}(\{y_{i'}^u\}_u)\right)$$

iNaturalist (Van Horn et al., 2018):

$$\mathrm{iNaturalist}(i, \{y_i^u\}_u) = \begin{cases} \mathrm{MV}(i, \{y_i^u\}_u) & \text{if } \max_{k\in[K]} \sum_{u\in U_i} \mathbb{1}(y_i^u = k) \geq \frac{2}{3} \\ \text{undefined} & \text{otherwise} \end{cases} .$$

Pl@ntNet Data

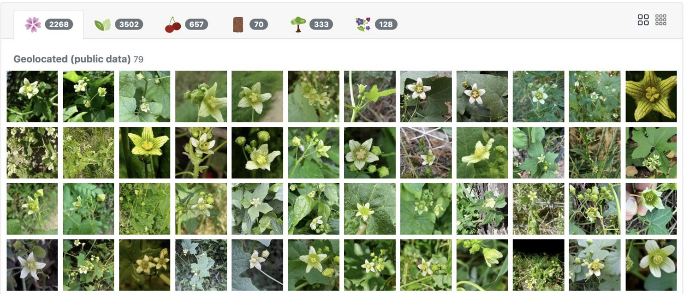
940M raw observations (=queries)

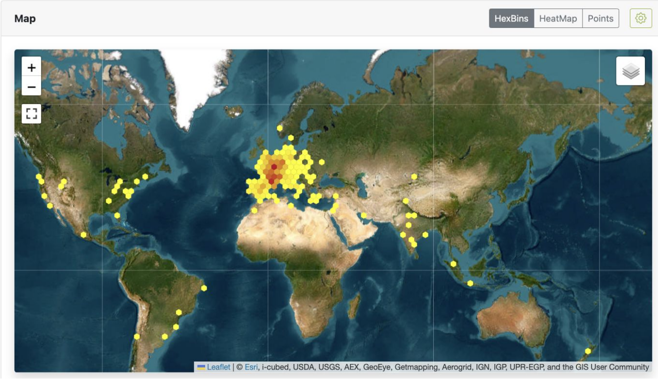


Pl@ntNet Data visualisation tools

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

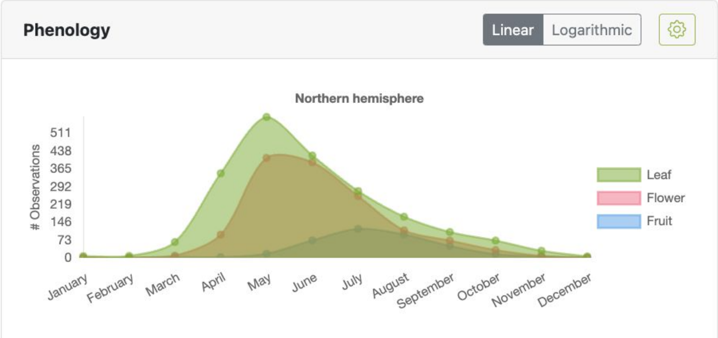
folklore

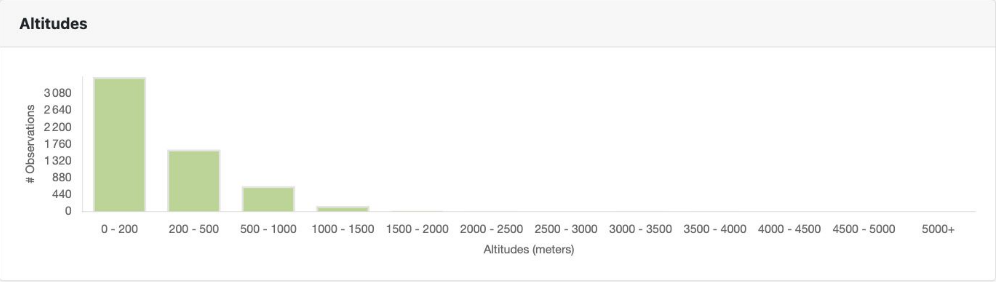
Additional information

Pl@ntUse

GBIF

Royal Botanic Gardens Kew | Plants of the World Online





Pl@ntNet Data shared in GBIF

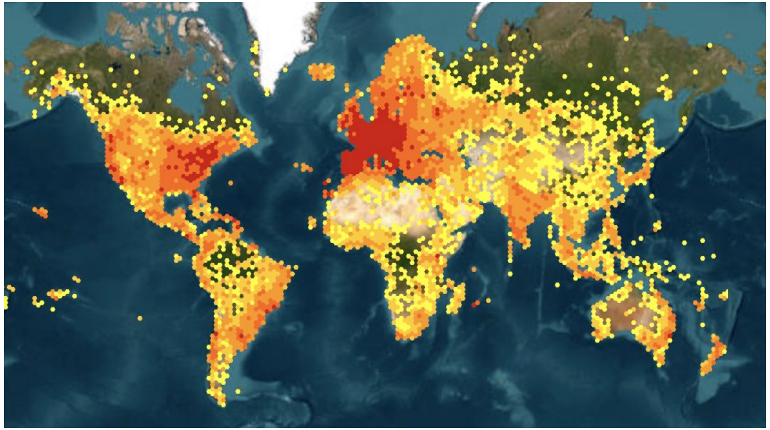
Top-5 data provider to GBIF (world's largest infrastructure for biodiversity data)

- 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)

632 citations



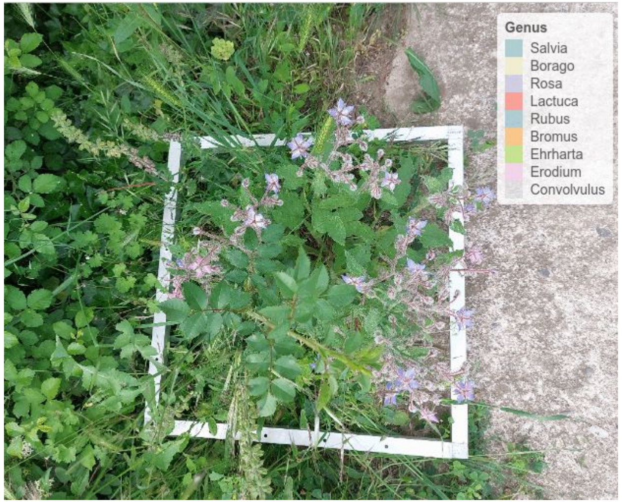
https://doi.org/10.15468/mma2ec

RED LIST nature PLOS ONE ANNALS OF BOTANY WILEY ELSEVIER

PART II
From individual plants to plant communities monitoring

Multi-specimen images for community-level monitoring

- Quadrat images for the monitoring of vulnerable habitats or fields biodiversity (e.g. VigieFlore)
- Vegetation cover images (e.g. terrestrial robots, drones, smartphones)
- Landscape views (e.g. car views for the monitoring of invasive species)



Weakly-supervised multi-label classification

Training data

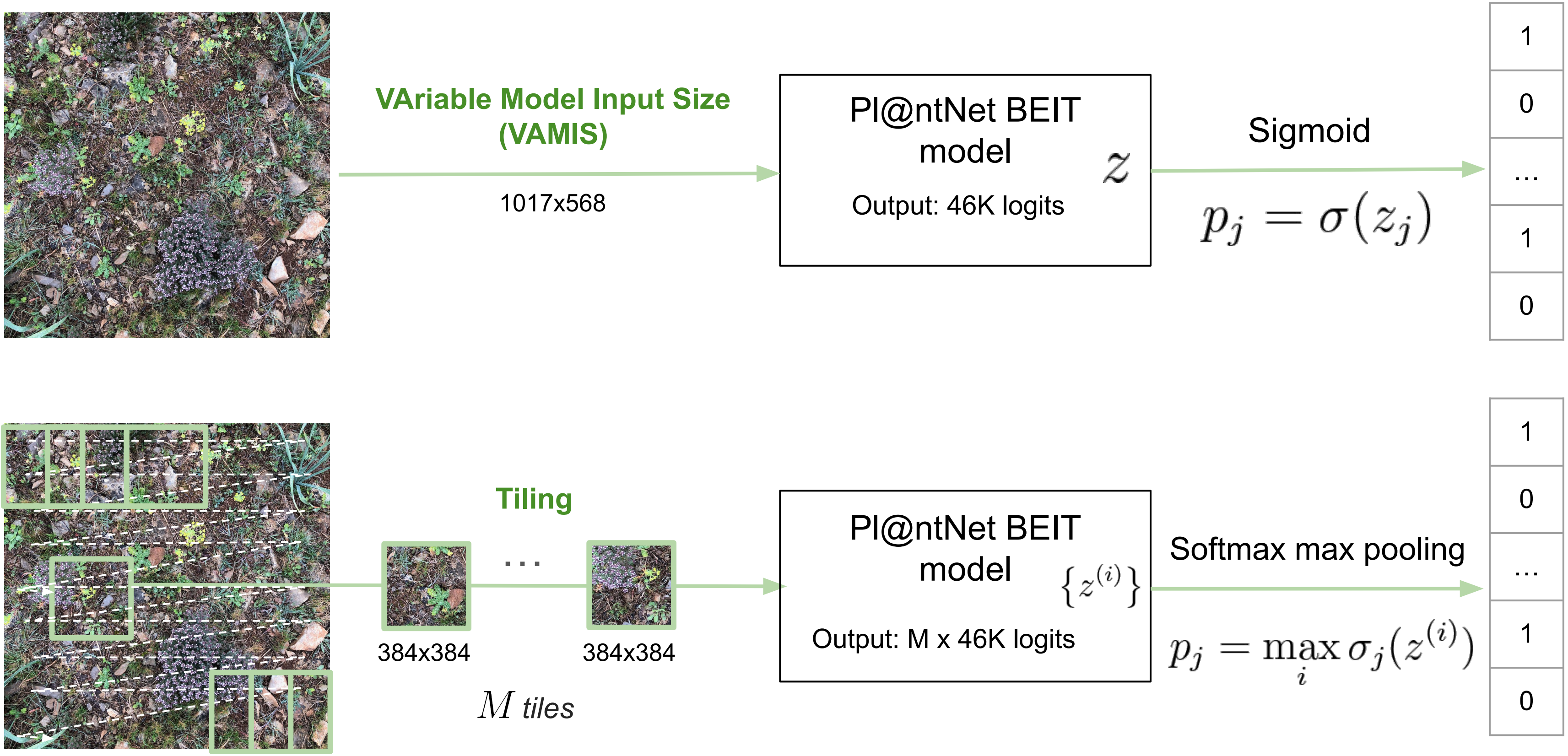
1

Pl@ntNet database

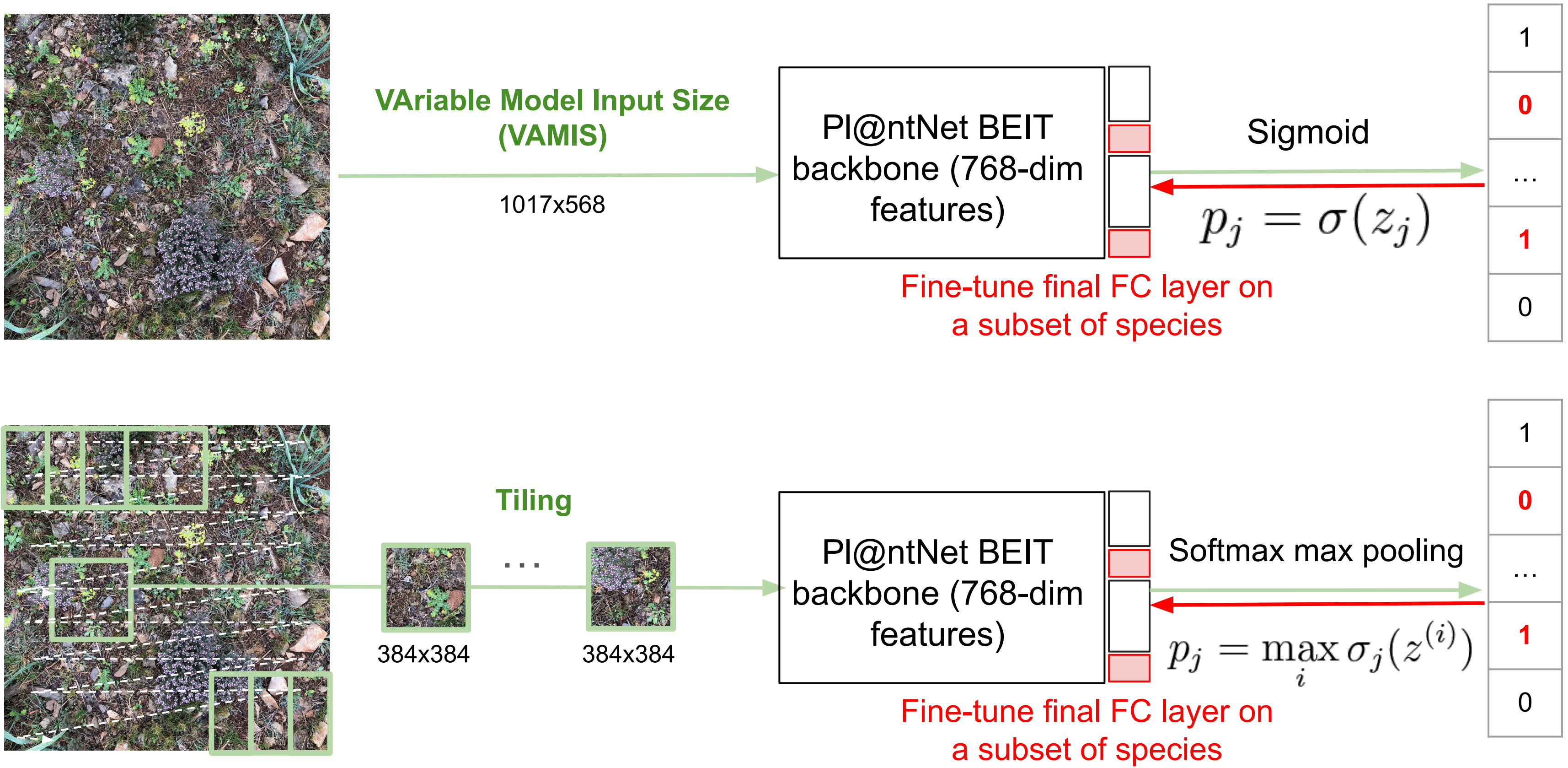
Test data

0	1	0	1	0	1	1	0

Zero-shot multi-label classification (no fine-tuning)



Few-shot multi-label classification (with fine-tuning)



Weakly-supervised multi-label classification

Evaluation on Danish road dataset

- Seven invasive species annotated
- 8.4K images with 1 to 3 invasive species

Dyrmann, M., Mortensen, A. K., Linneberg, L., Høye, T. T., & Bjerge, K. (2021). Camera assisted roadside monitoring for invasive alien plant species using deep learning. *Sensors*, *21*(18), 6126.

(a) (b)

Results

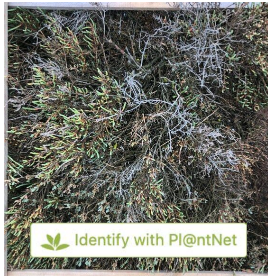
	Zero-shot (no fine-tuning)		With fine-tuning	
	VAMIS	Tiling	VAMIS	Tiling
AUC	75.52	**91.58**	96.49	**<u>96.50</u>**
F1	36.45	**63.39**	74.28	**<u>76.46</u>**

Application

Tiling approach integrated in Pl@ntNet (without fine-tuning so far)

- Beta version of a front-end dedicated to plot images in Pl@ntNet web app

Create a plot



Species

	Species
1	*Halimione portulacoides* (L.) Aellen
2	*Hornungia procumbens* (L.) Hayek
3	*Parapholis filiformis* (Roth) C.E.Hubb.
4	*Plantago coronopus* L.
5	*Sarcocornia fruticosa* (L.) A.J.Scott
6	*Sphenopus divaricatus* (Gouan) Rchb.

- API (my.plantnet.org): used for our participation to Xprize (Brazilian team, finalist)

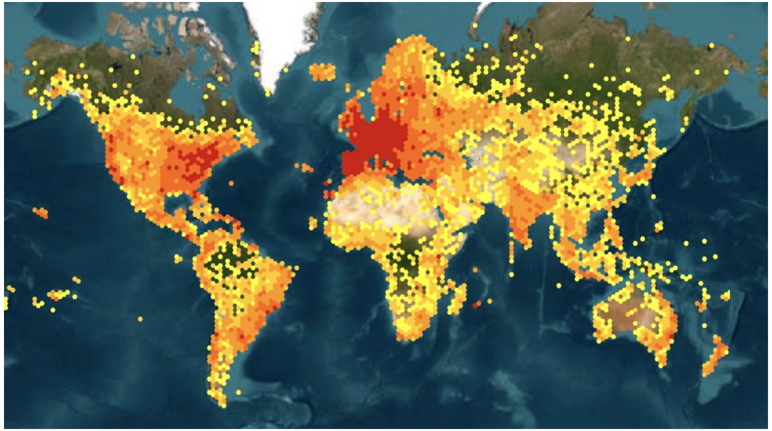
PART III

GeoPl@ntNet: from field observations to mapping tools and decision support applications

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

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

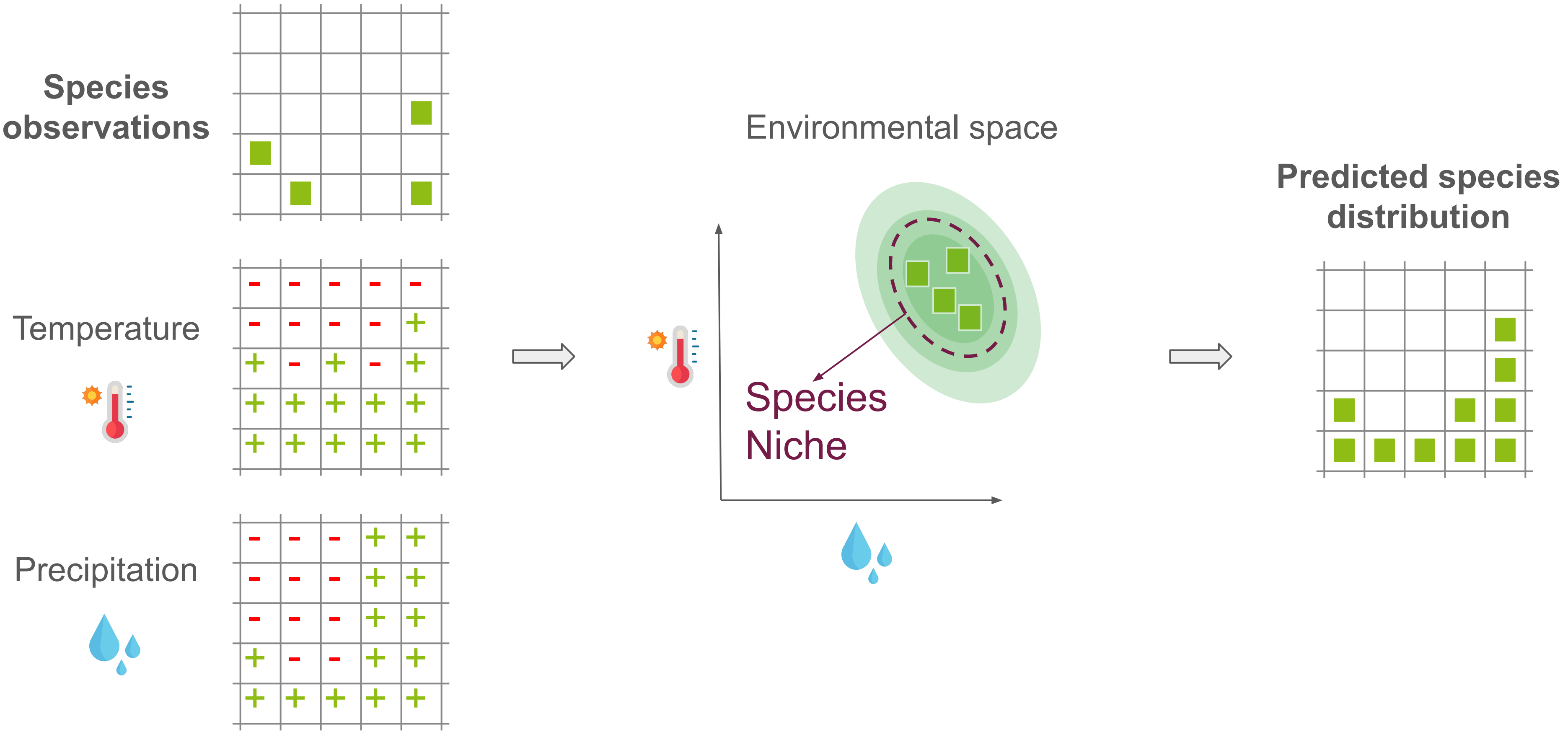
Many plant occurrences at world scale



GBIF

But very few locally for most species

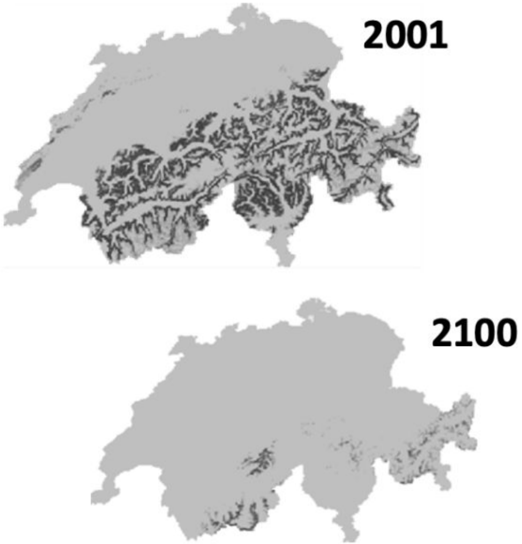
Species Distribution Models (SDM)



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**

How to train SDMs ?

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 $\hat{y} = f_\theta(x) \in \{1, ..., d\}$

1

Predicting species assemblages from presence only data

Given presence-only occurrences

$$(x_1, y_1), \ldots, (x_{n_t}, y_{n_t}) \text{ sampled from } \mathbb{P}_{X,Y}$$

The **assemblage of species** likely to be present conditionally to x can be defined as:

$$S^*_\lambda(x) := \{k \in \mathcal{Y} : \mathbb{P}_{X,Y}(Y = k | X = x) \geq \lambda\}$$

Can be estimated by **thresholding the softmax** output of a DNN (with **CE loss**):

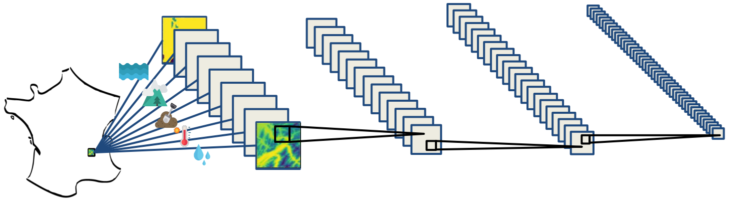
$$S_\lambda(x) := \{k \in \mathcal{Y} : \hat{\eta}_k(x) > \lambda\} \quad \text{with} \quad \hat{\eta}_k(x) = \frac{exp(f_\theta^k(x))}{\sum_j exp(f_\theta^j(x))}$$

We did that in several works using CNNs

PLOS COMPUTATIONAL BIOLOGY

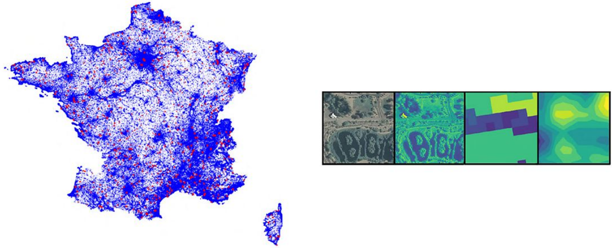
Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment

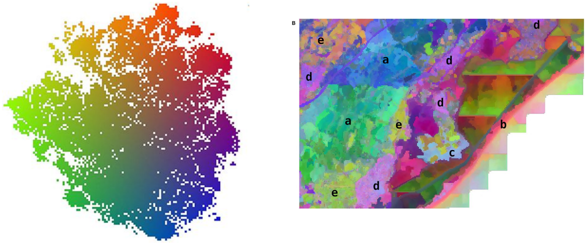
Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, Christophe Botella, François Munoz, Alexis Joly



frontiers in plant science

Very High Resolution Species Distribution Modeling Based on Remote Sensing Imagery

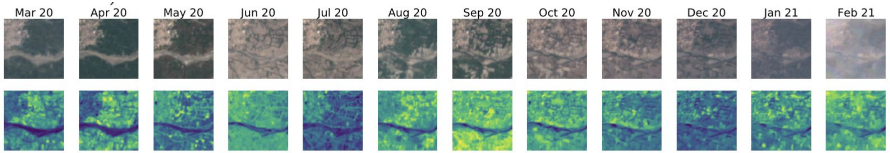




frontiers in plant science

Deep Species Distribution Modeling From Sentinel-2 Image Time-Series: A Global Scale Analysis on the Orchid Family

Joaquim Estopinan [1,2], Maximilien Servajean [2,3], Pierre Bonnet [4,5], François Munoz [6], Alexis Joly [1,2]



Limitations

Very sensitive to **taxonomic reporting bias**

8,548 observations

Centaurea jacea

vs.

6 observations

Cenchrus agrimonioides

$$\hat{\eta}_k(x) = \frac{exp(f_\theta^k(x))}{\sum_j exp(f_\theta^j(x))}$$

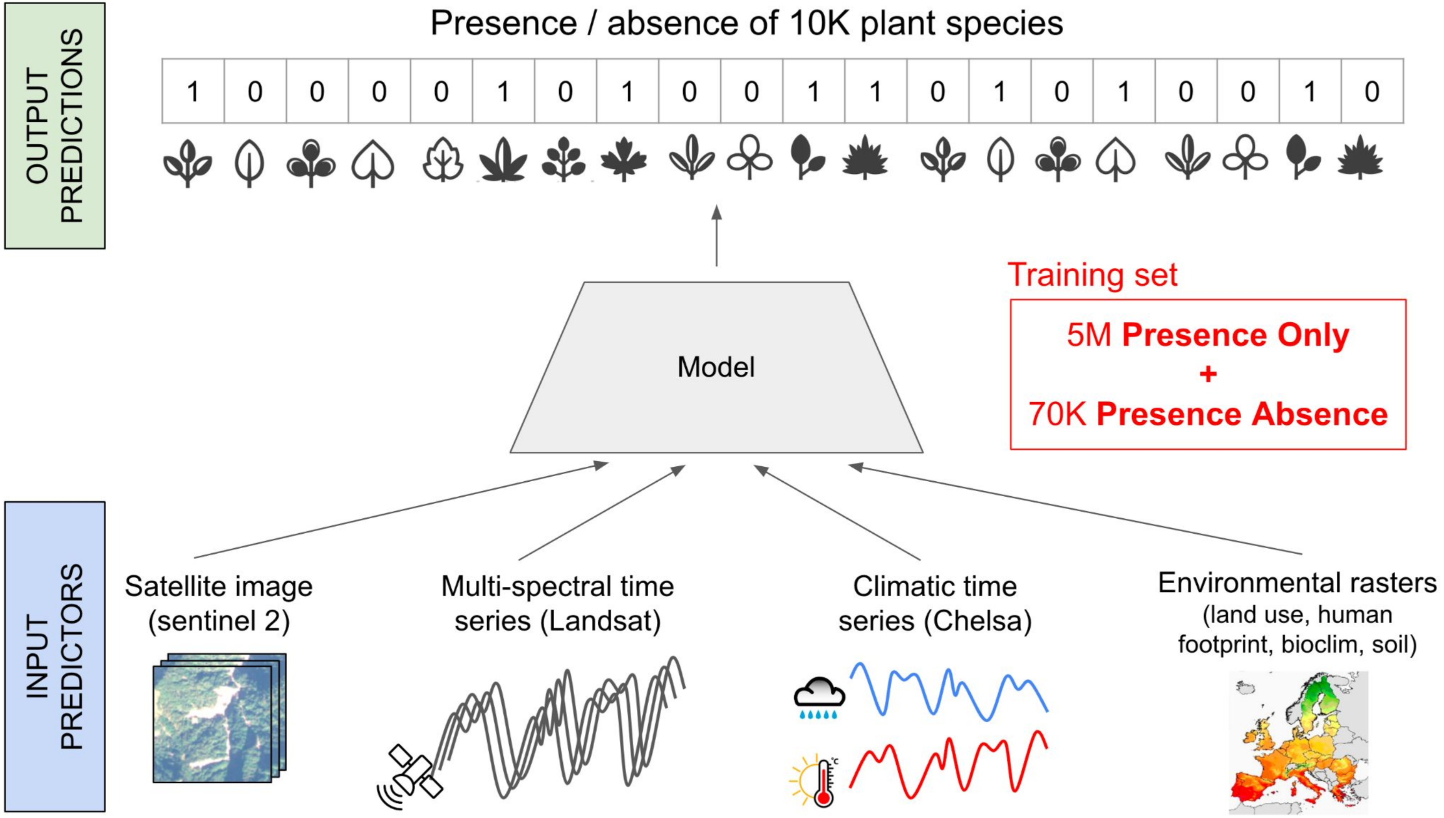
Observation probability ≠ Presence probability

The threshold λ is **arbitrary** (we don't know how many species there are)

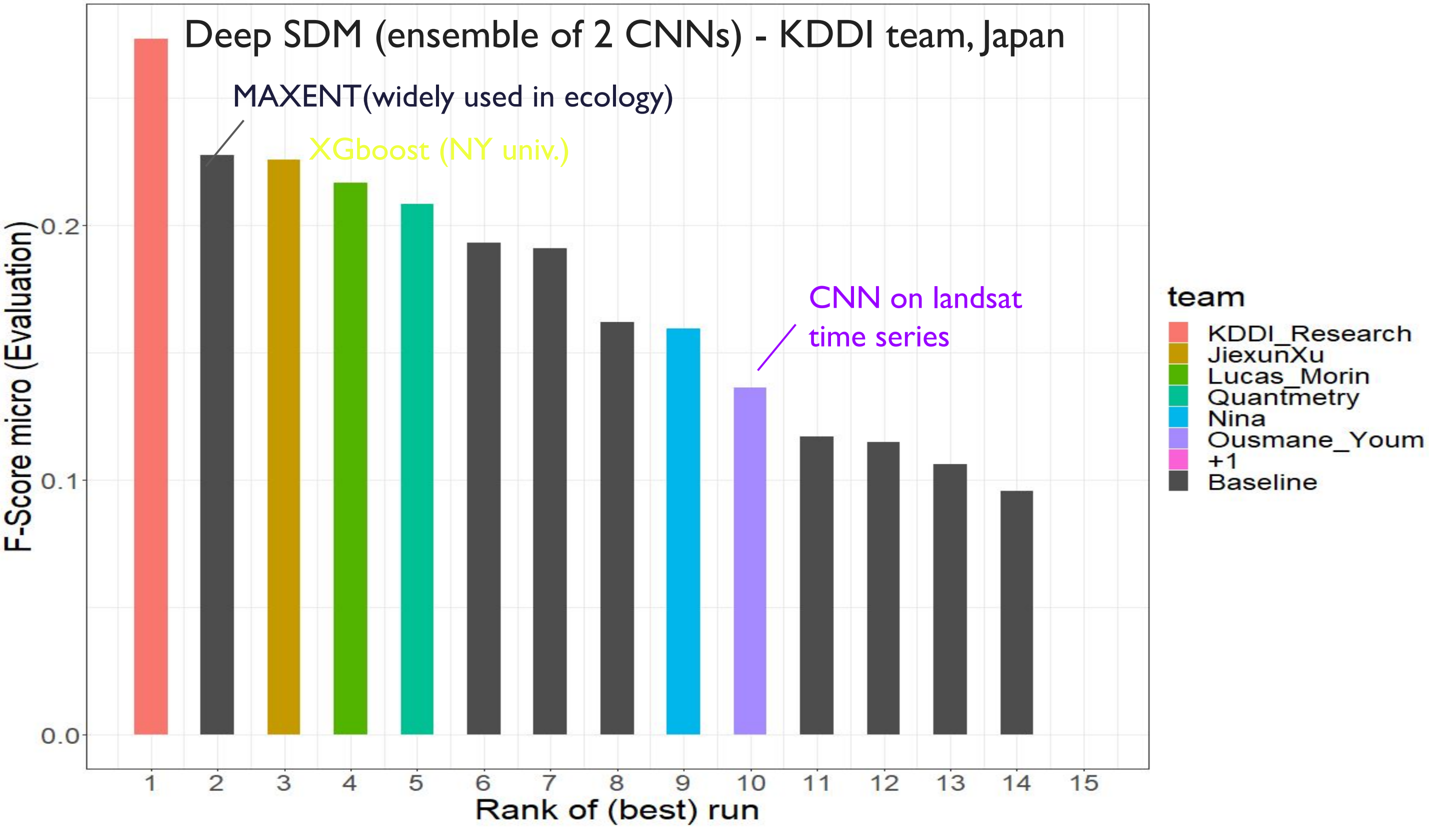
The probability of each species is **relative** to the others and depends on the **number of species** present somewhere

→ this is not appropriate for mapping each species individually

GeoLifeCLEF challenge 2023 & 2024



GeoLifeCLEF challenge 2023 - results



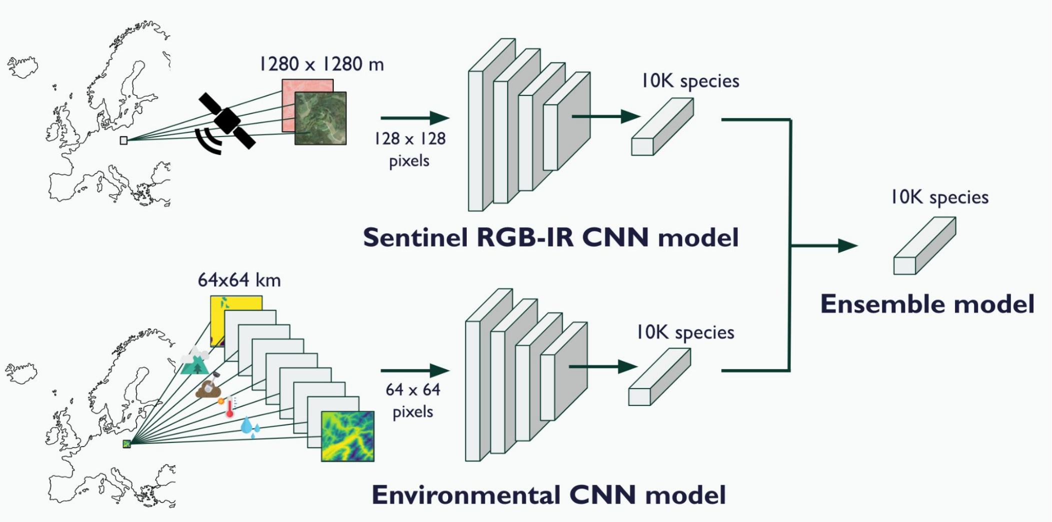
Ranked scores of the best model

GeoLifeCLEF challenge 2023 - best approach

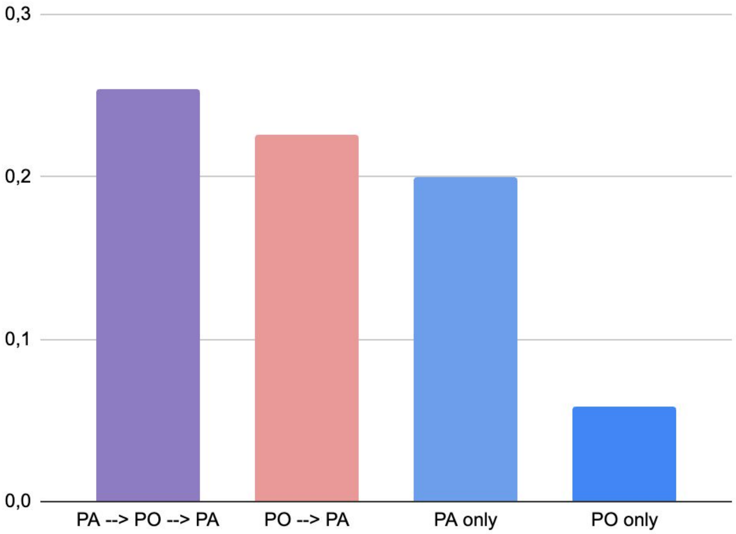
Leverage Samples with Single Positive Labels to Train CNN-based Models For Multi-label Plant Species Prediction

Huy Quang Ung, Ryoichi Kojima, Shinya Wada

Architecture



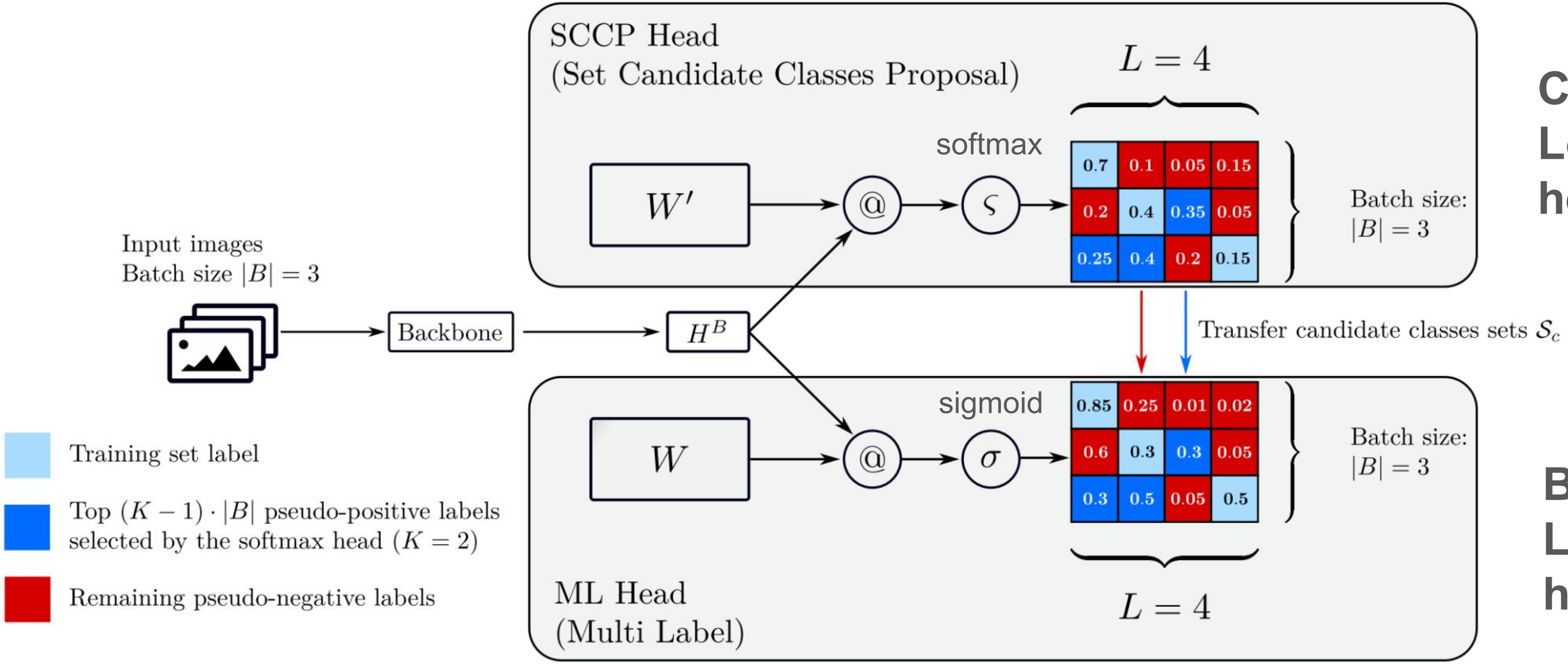
Training strategy



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

Ongoing work: a two-head loss function to improve transfer learning from PO to PA

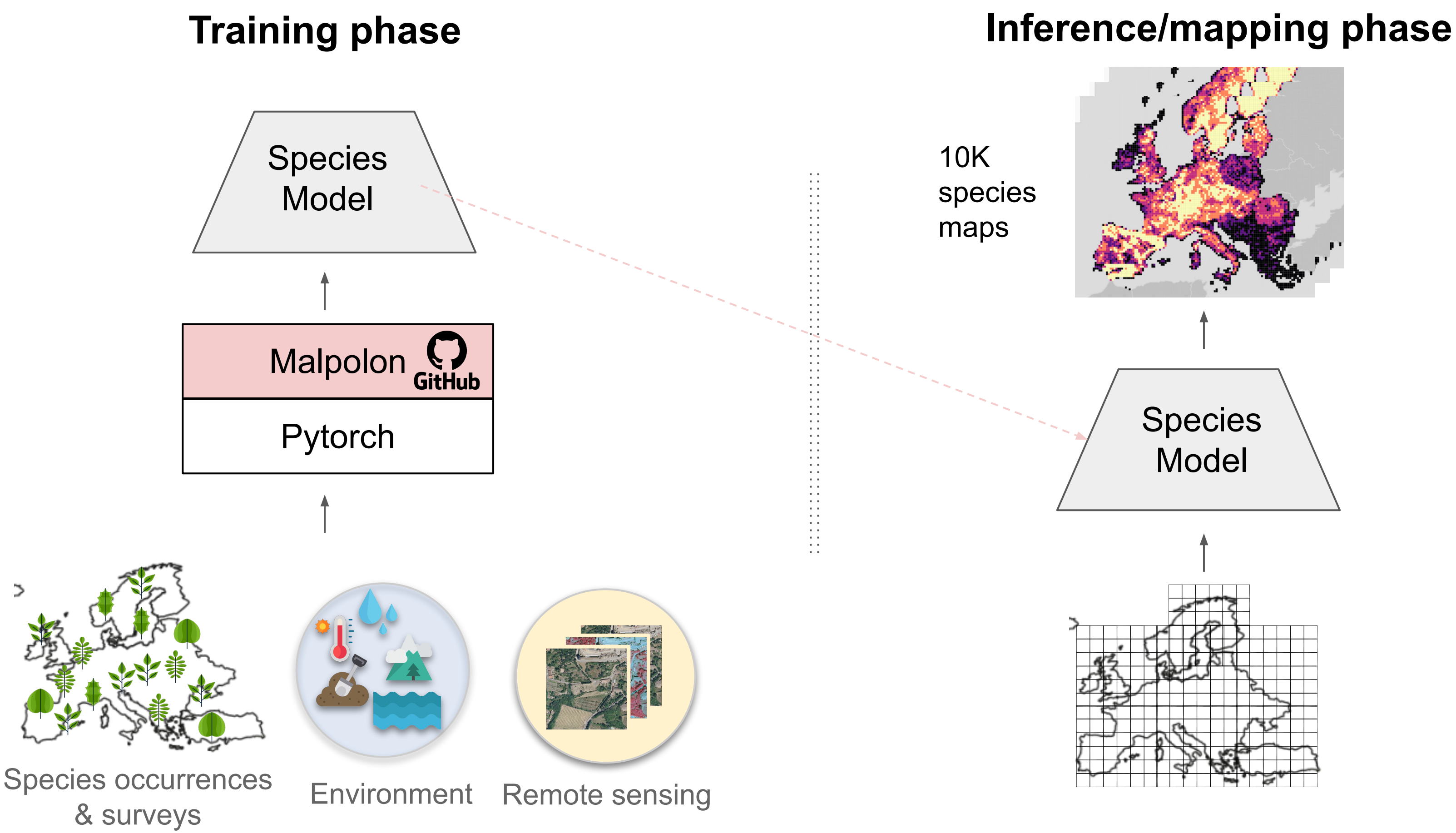
A two-**head** loss function for deep Average-K classification
C **Garcin**, M Servajean, A **Joly**, J Salmon - arXiv preprint arXiv:2303.18118, 2023 - arxiv.org



CE Loss head

BCE Loss head

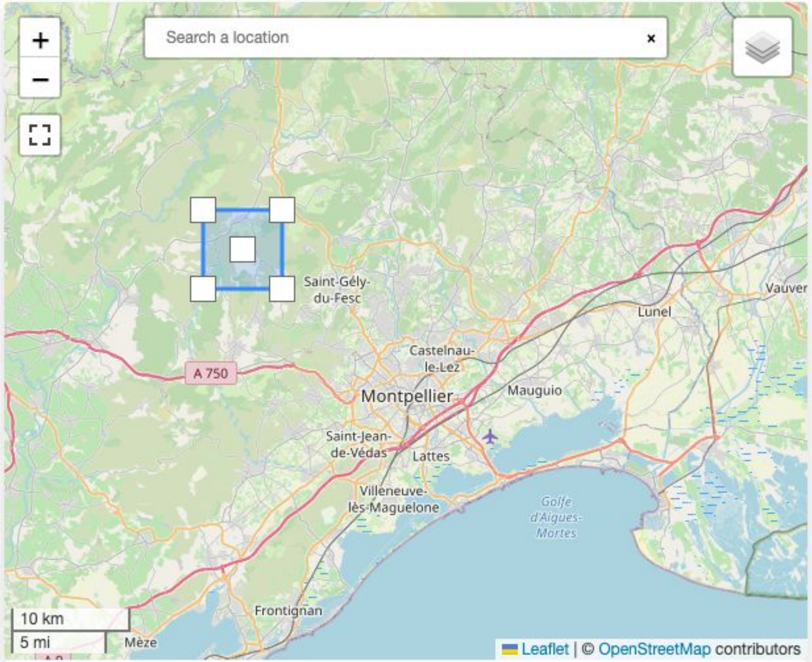
From models to species mapping



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

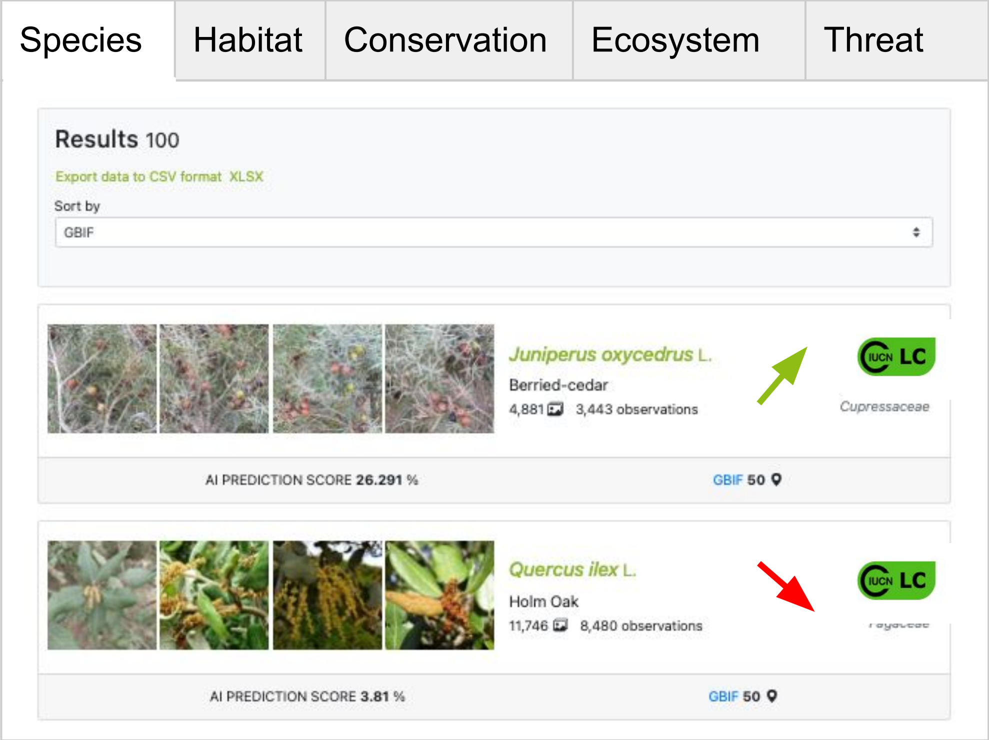
GeoPl@ntNet

Discover plant biodiversity around you



Right click on the map to move the marker (or drag / drop)

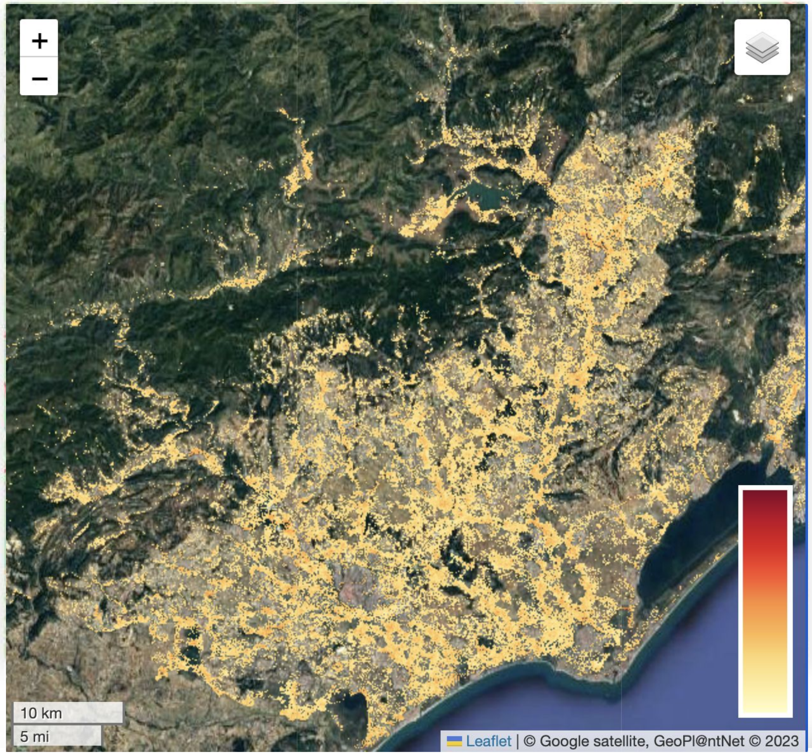
Search



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

GeoPl@ntNet

Discover plant biodiversity around you



Species | Habitat | Conservation | Ecosystem | Threat

Results 100

Export data to CSV format XLSX

Sort by

GBIF

Fraxinus angustifolia Vahl

Narrow-leaved Ash

8,647 6,124 observations

IUCN LC

Oleaceae

AI PREDICTION SCORE **0** %

GBIF **2**

Lysimachia vulgaris L.

Garden Loosestrife

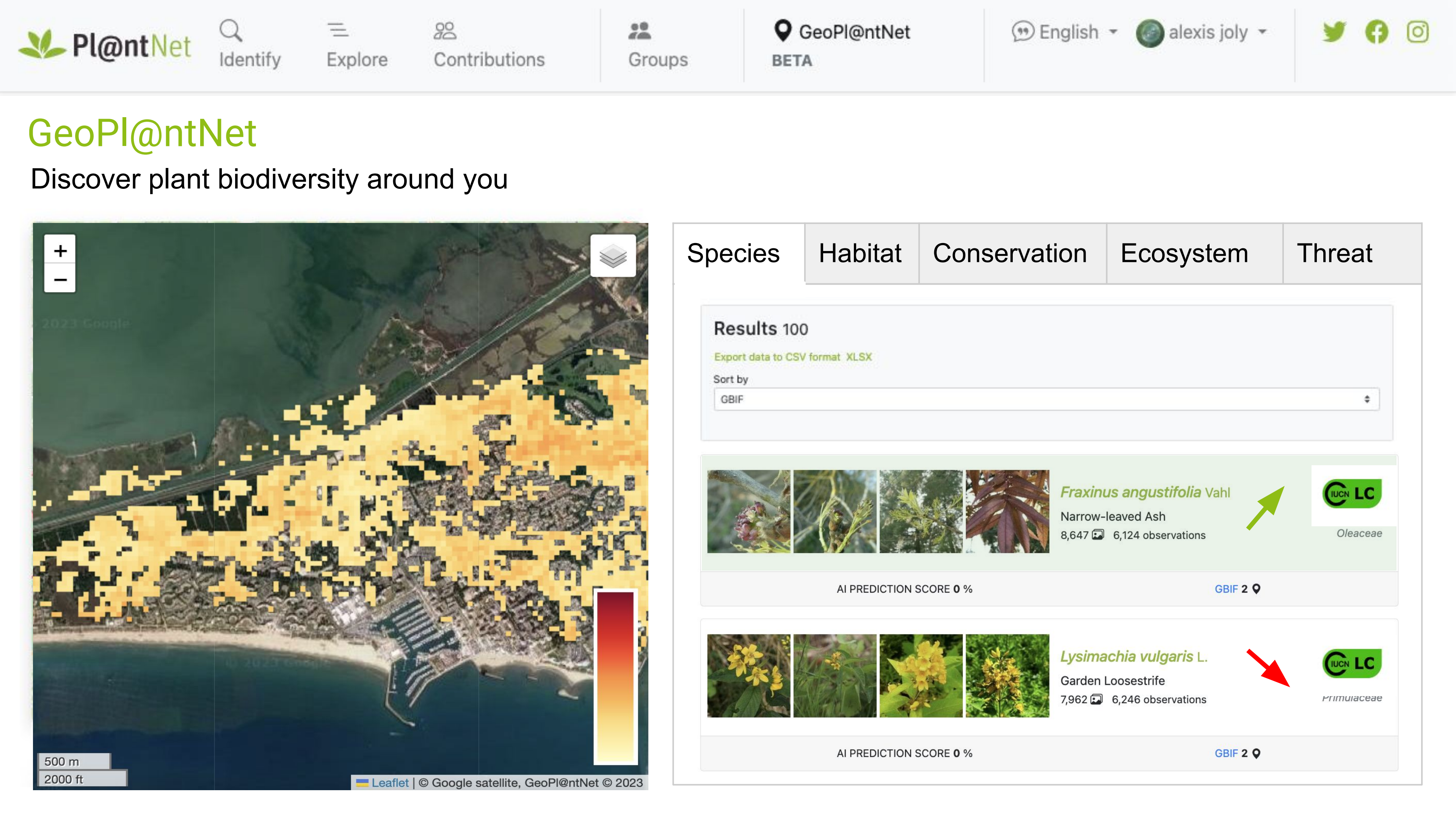
7,962 6,246 observations

IUCN LC

Primulaceae

AI PREDICTION SCORE **0** %

GBIF **2**



Pl@ntNet
Identify
Explore
Contributions
Groups
GeoPl@ntNet
BETA
English
alexis joly
GeoPl@ntNet
Discover plant biodiversity around you
500 m
2000 ft
Leaflet | © Google satellite, GeoPl@ntNet © 2023
Species
Habitat
Conservation
Ecosystem
Threat
Results 100
Export data to CSV format XLSX
Sort by
GBIF
Fraxinus angustifolia Vahl
Narrow-leaved Ash
8,647
6,124 observations
IUCN LC
Oleaceae
AI PREDICTION SCORE 0 %
GBIF 2
Lysimachia vulgaris L.
Garden Loosestrife
7,962
6,246 observations
IUCN LC
Primulaceae
AI PREDICTION SCORE 0 %
GBIF 2

Mapping biodiversity conservation indicators

From the species assemblage

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

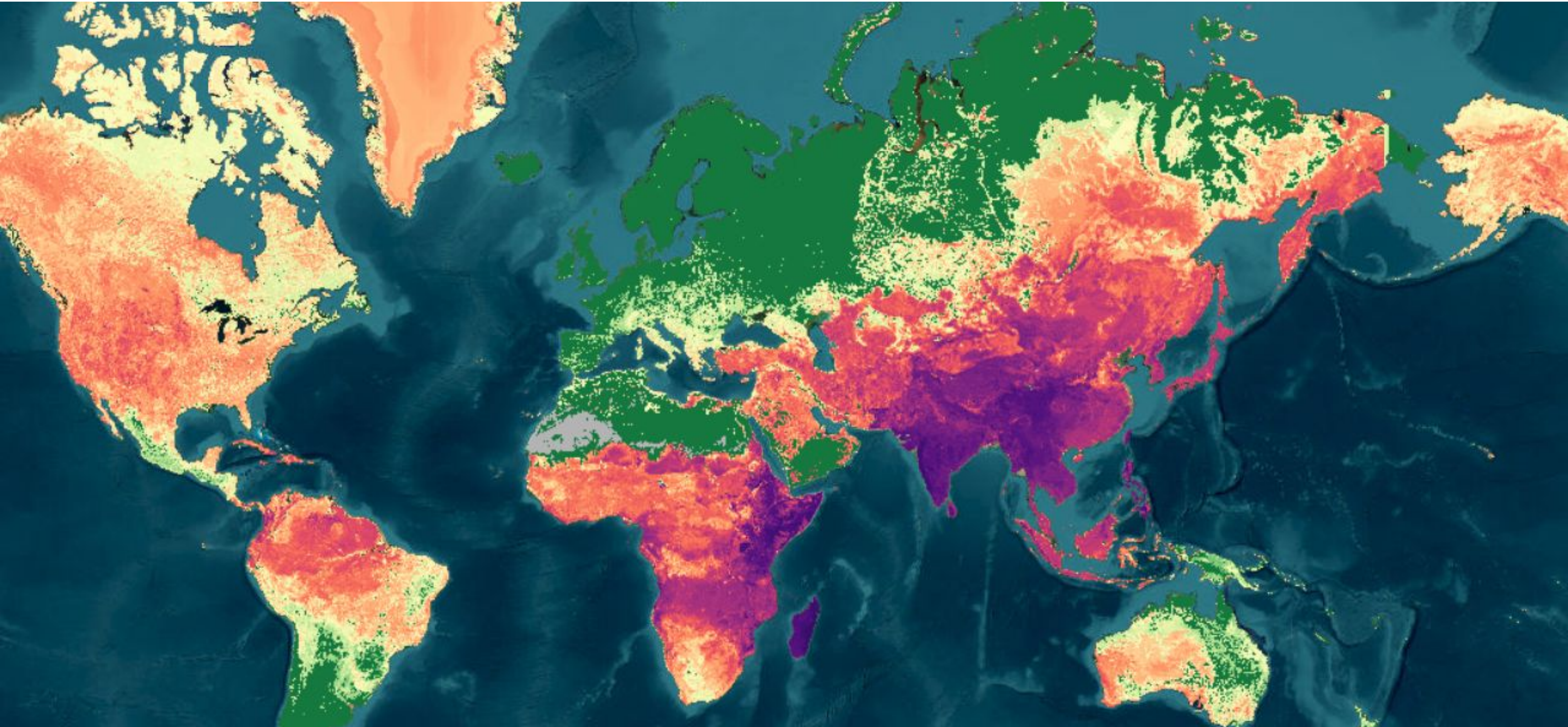
We can compute indicators such as:

- 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

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

Proportion of endangered species (Orchid Family, 14K species)

1x1 km resolution ([view online]()) PhD of Joaquim Estopinan

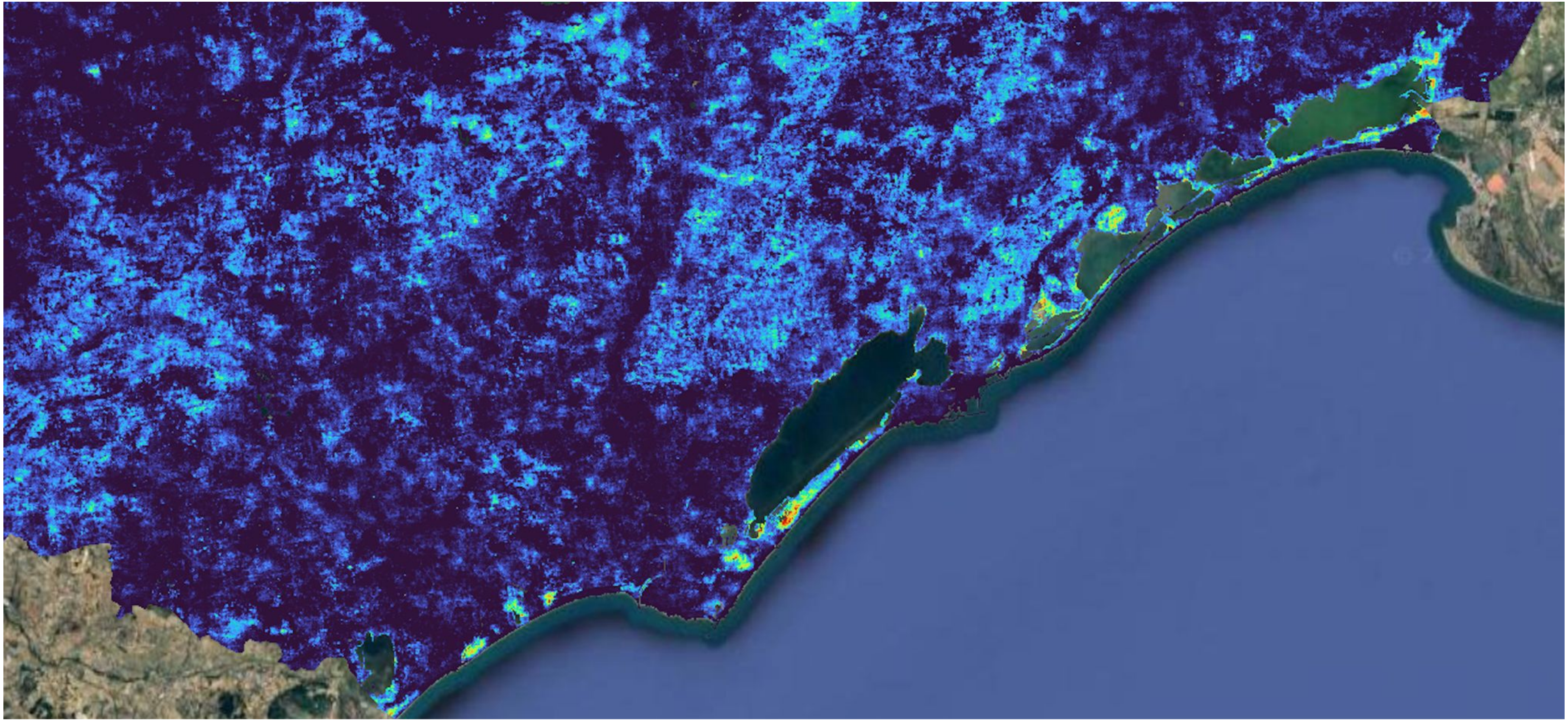


Invasive species number

50x50 m resolution

Rare species number

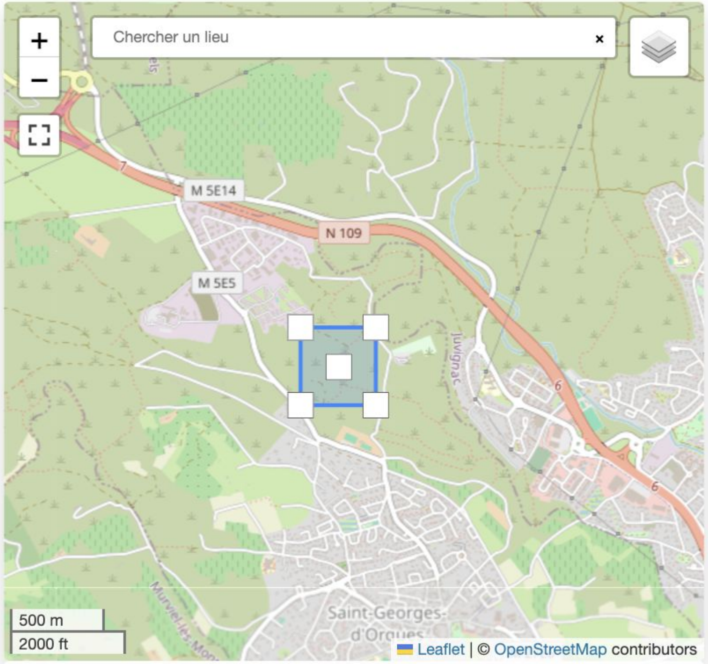
50x50 m resolution



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

GeoPl@ntNet

Discover plant biodiversity close to home and help protect it better

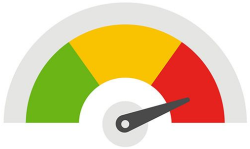


Clic droit sur la carte pour déplacer le marqueur (ou glisser / déposer)

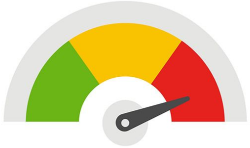
Search

Species	Habitat	Conservation	Ecosystem	Threat

Presence of rare species



Presence of species on the European directive



Humm, maybe we should not construct a new facility here

Thank you