Cooperative learning for biodiversity monitoring: what's new and what's next in Pl@ntNet ?

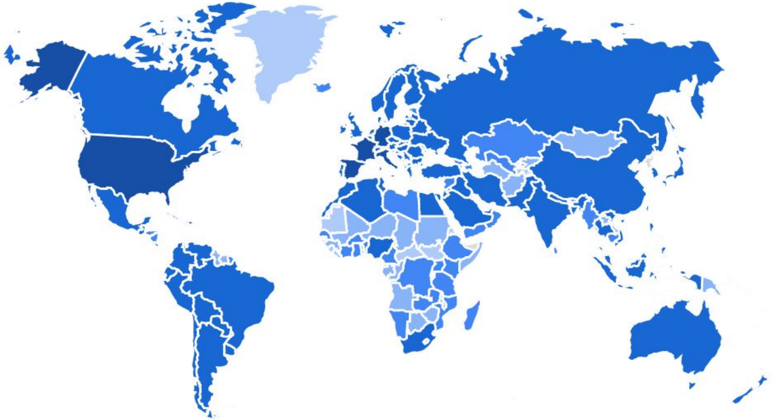
Alexis Joly, Pierre Bonnet, Hervé Goëau, Antoine Affouard, J.C. Lombardo, Mathias Chouet, Hugo Gresse, Christophe Botella, Titouan Lorieul, Benjamin Deneu, Joaquim Estopinan, Cesar Leblanc, Camille Garcin, Diego Marcos, Maximilien Servajean, François Munoz, Joseph Salmon

PART I
Pl@ntNet overview

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

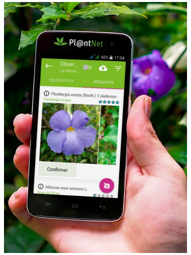
Pl@ntNet

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



Personal Usage


Nature, walks


Gardening


Phytotherapy

Professional Usage


Agro-ecology

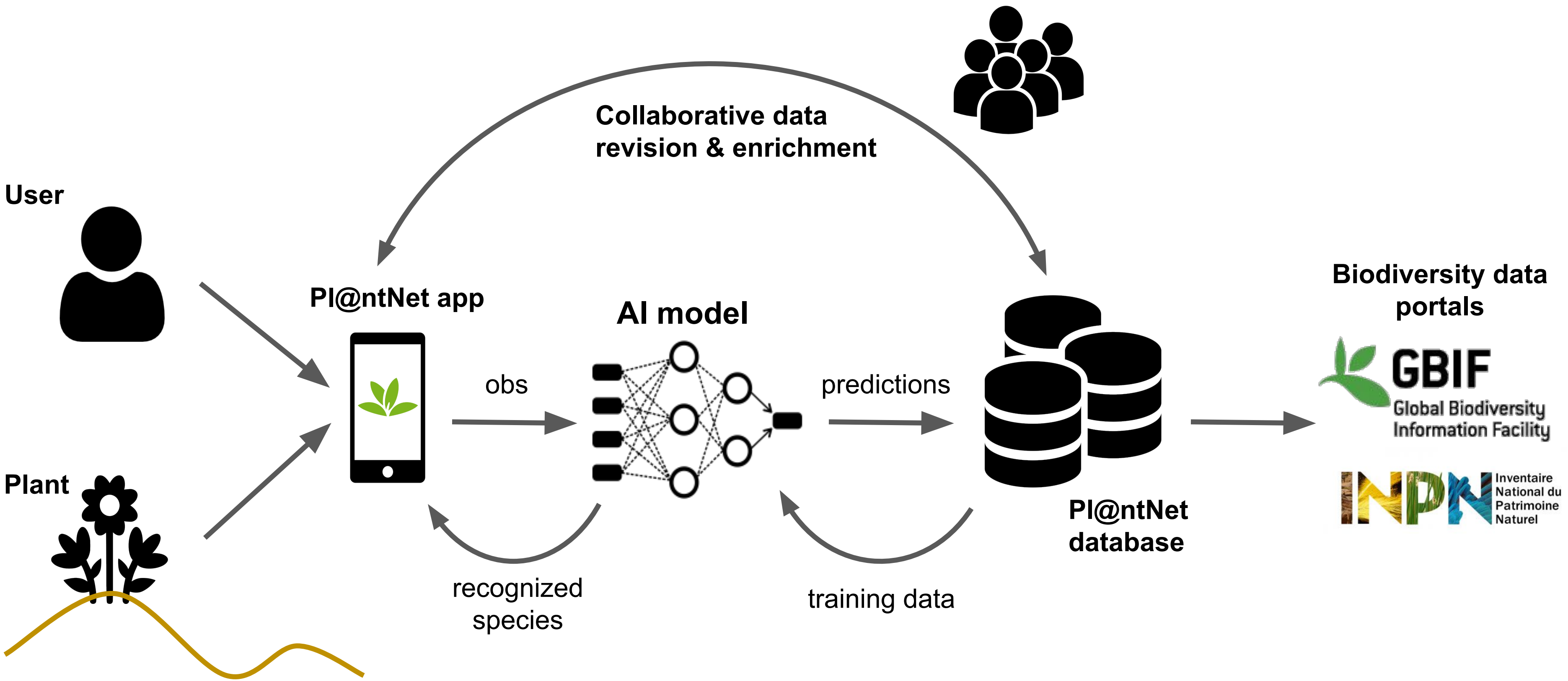

Natural Areas Management


Education, animation


Tourism

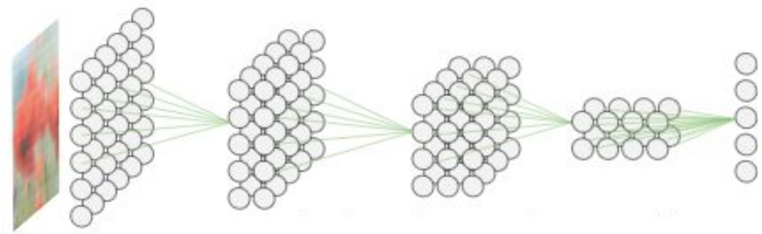

Trade

Key concept of Pl@ntNet: Cooperative Learning



Pl@ntNet AI model

Model trained with the cross-entropy loss on the set of valid observations (Jean Zay, a few days of training)



Softmax output (46K-dimensional)

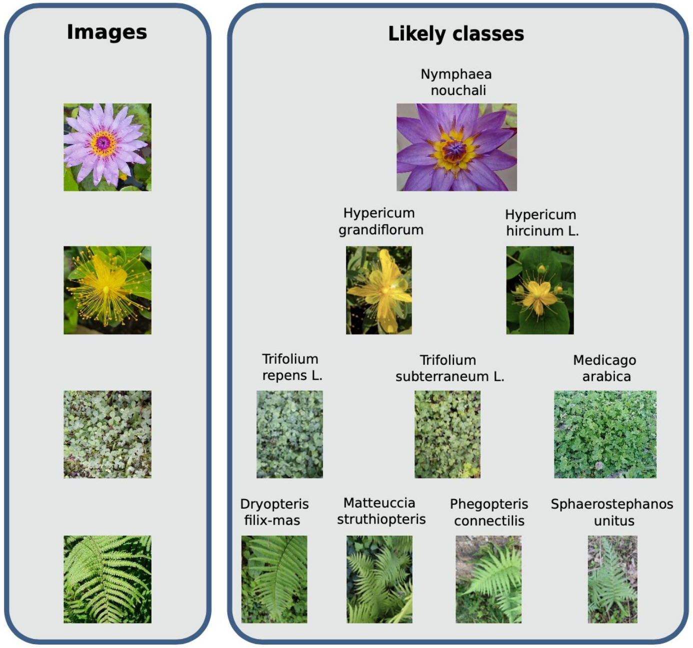
$$\sigma(f(x))$$

Production version: Convolutional Neural Network (IV3) → **Top1 accuracy = 0.70**
Beta version: Vision transformer (BEIT) → **Top1 accuracy = 0.73**

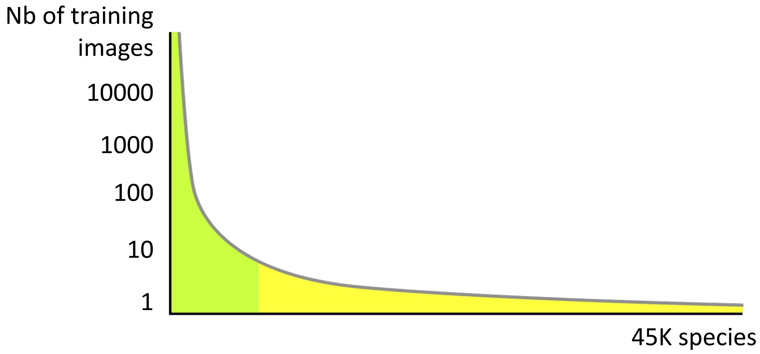
46K species (+ reject classes)
5M training images (undersampling for classes > 1000 images)

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**
Glaucium flavum L.	**0.98**

0.95

Average set size control

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

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

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

0.1

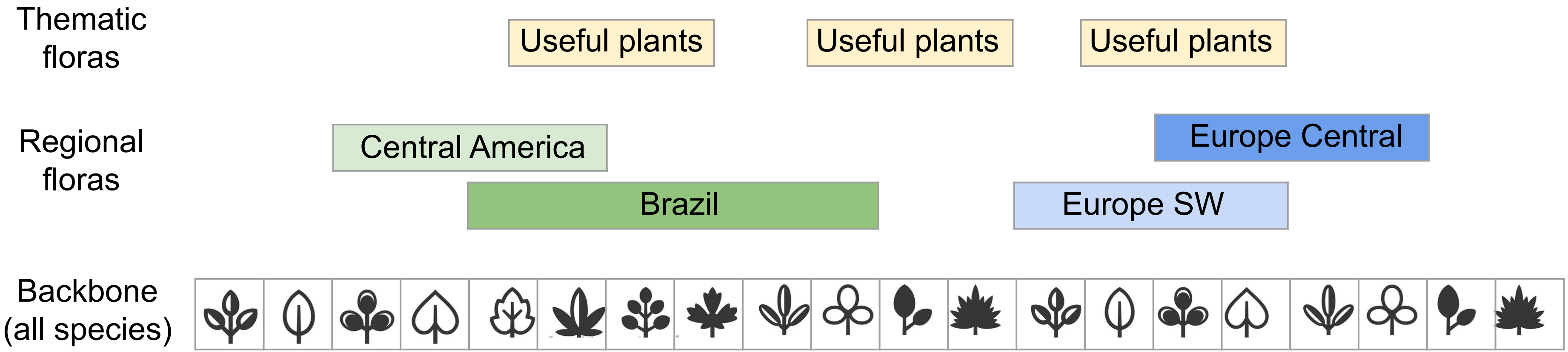
→ Average-K classification

(proof of consistency)

Use of regional or thematic floras

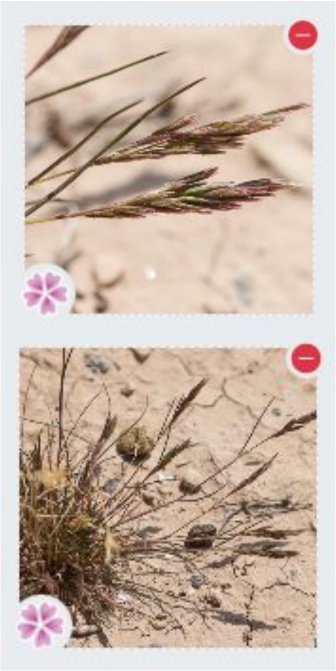
Restricting the hypothesis space to a particular flora allows improving the identification accuracy

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

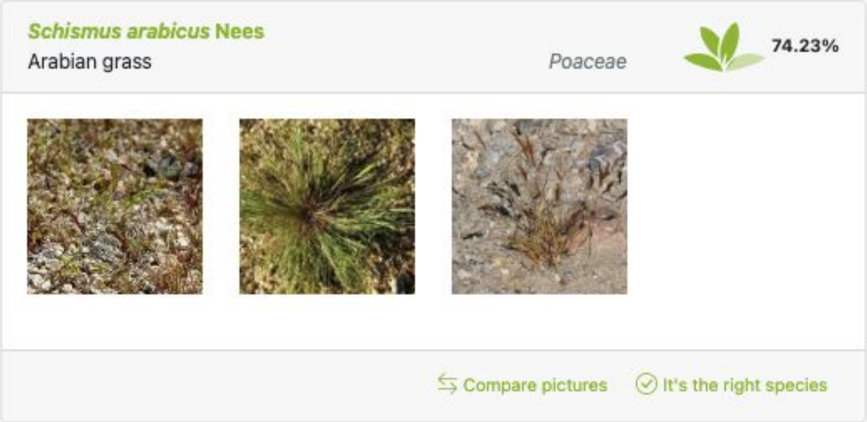


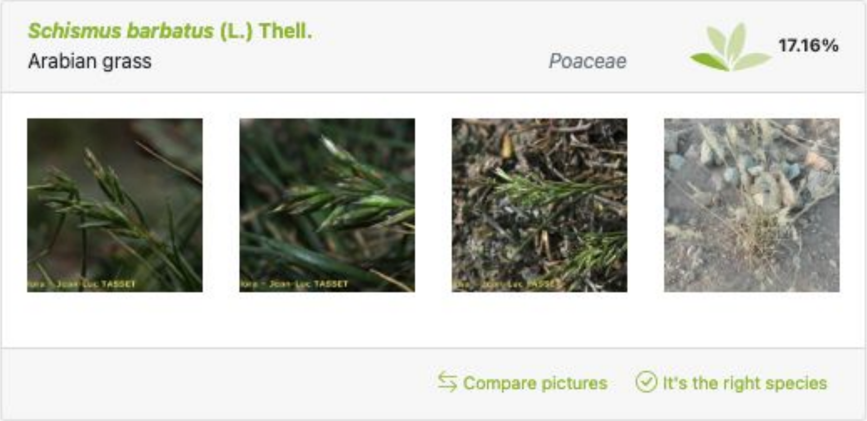
Use of regional or thematic floras

Query



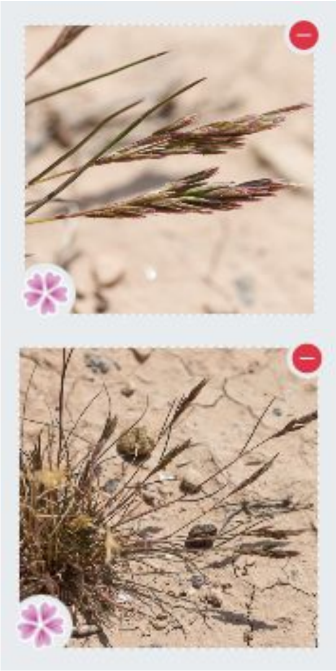
Identify in World flora



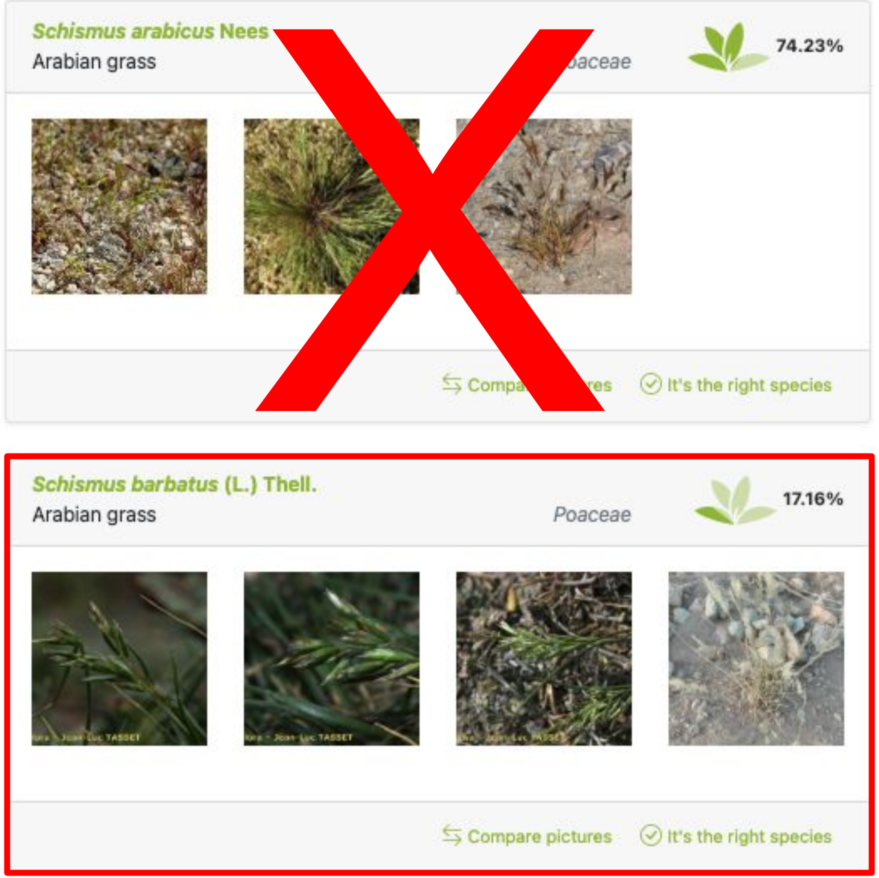


Use of regional or thematic floras

Query

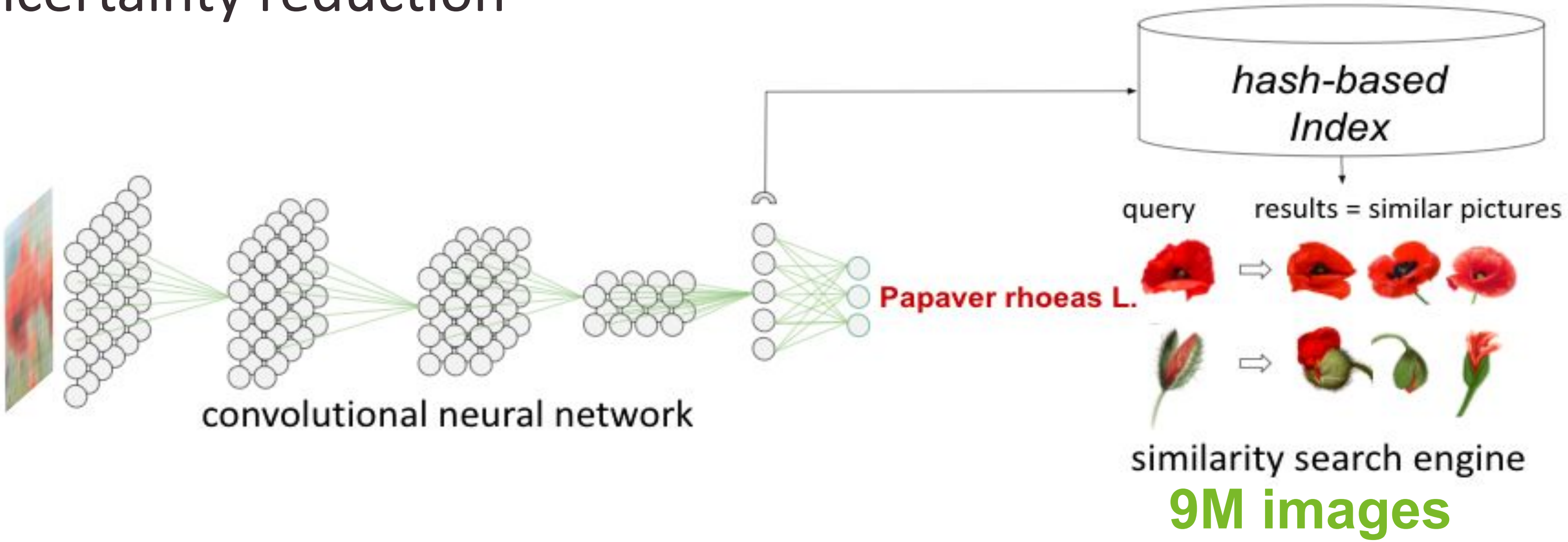


Identify in | **West Europe**



Pl@ntNet Similarity search

User's visual control =
uncertainty reduction

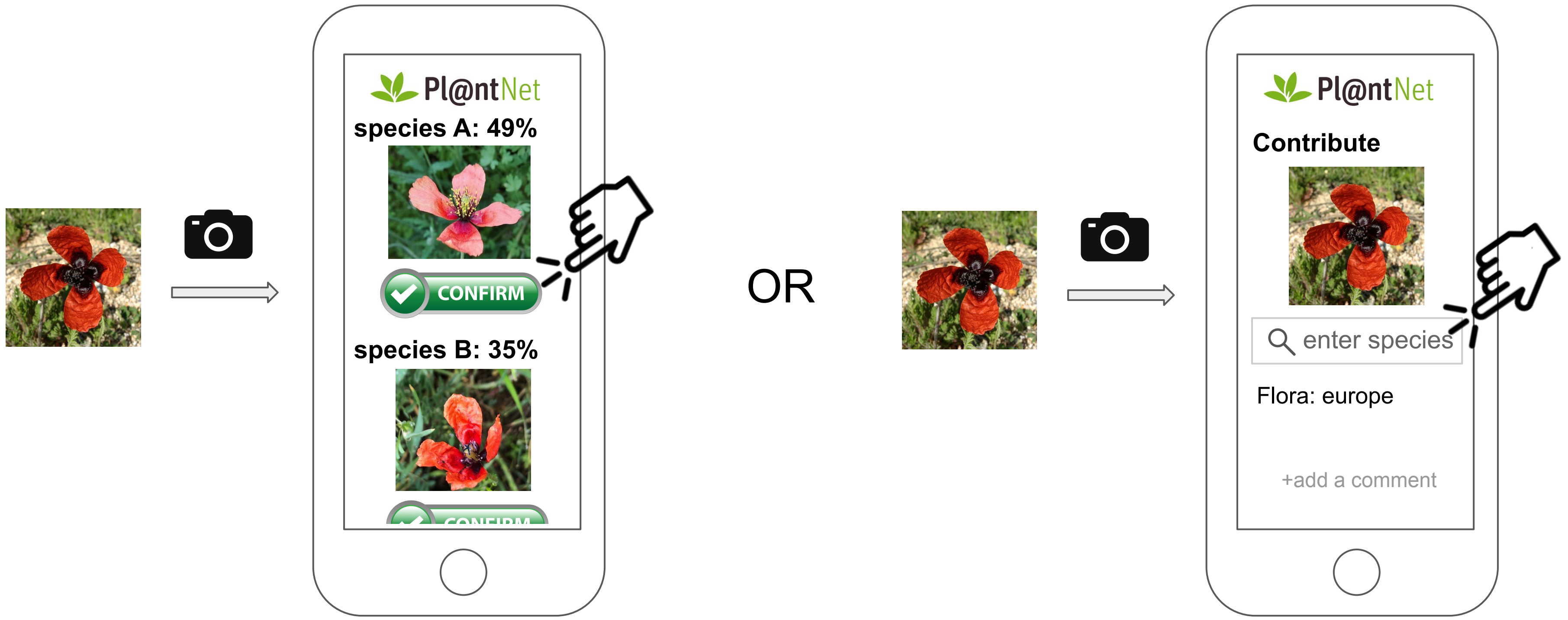


→ Sub-linear algorithm based on locality sensitive hashing

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

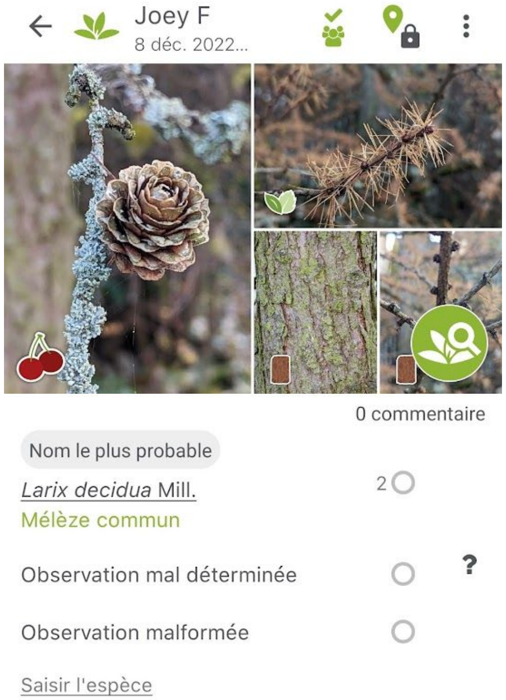
Contribution

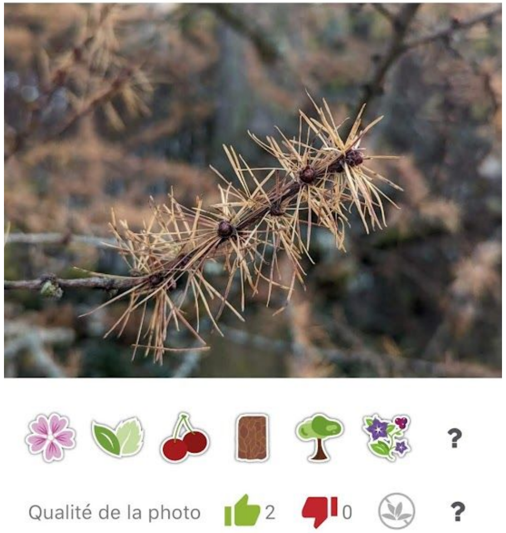
Users can contribute their observations



Revision

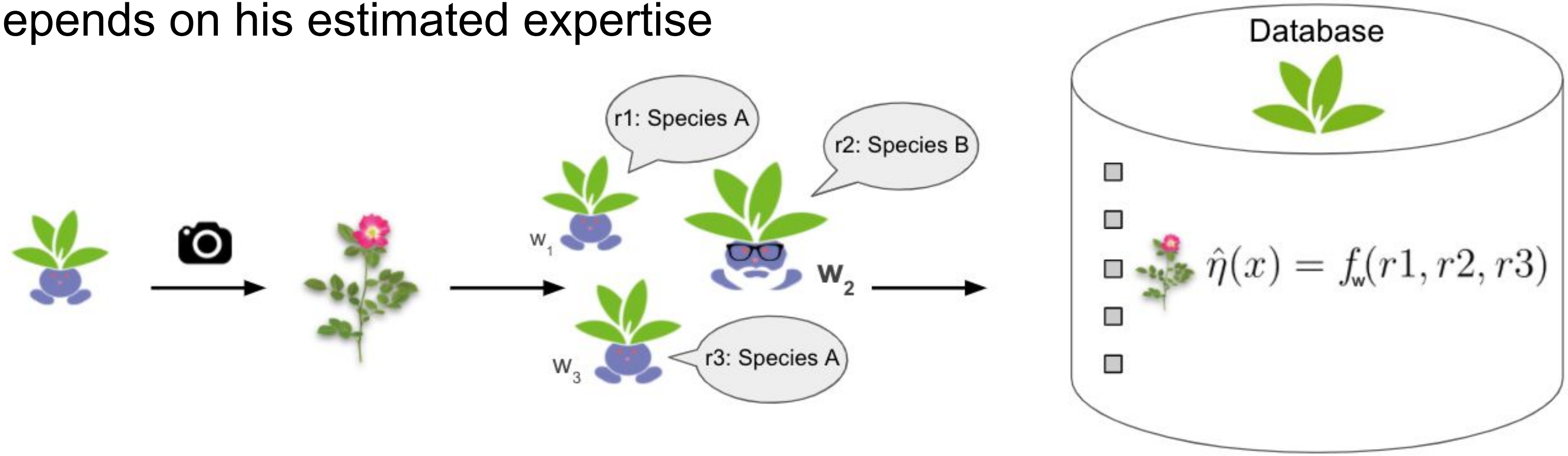
Users can revise observations of other users.





Cooperative learning

The weight of a user in the decision process depends on his estimated expertise

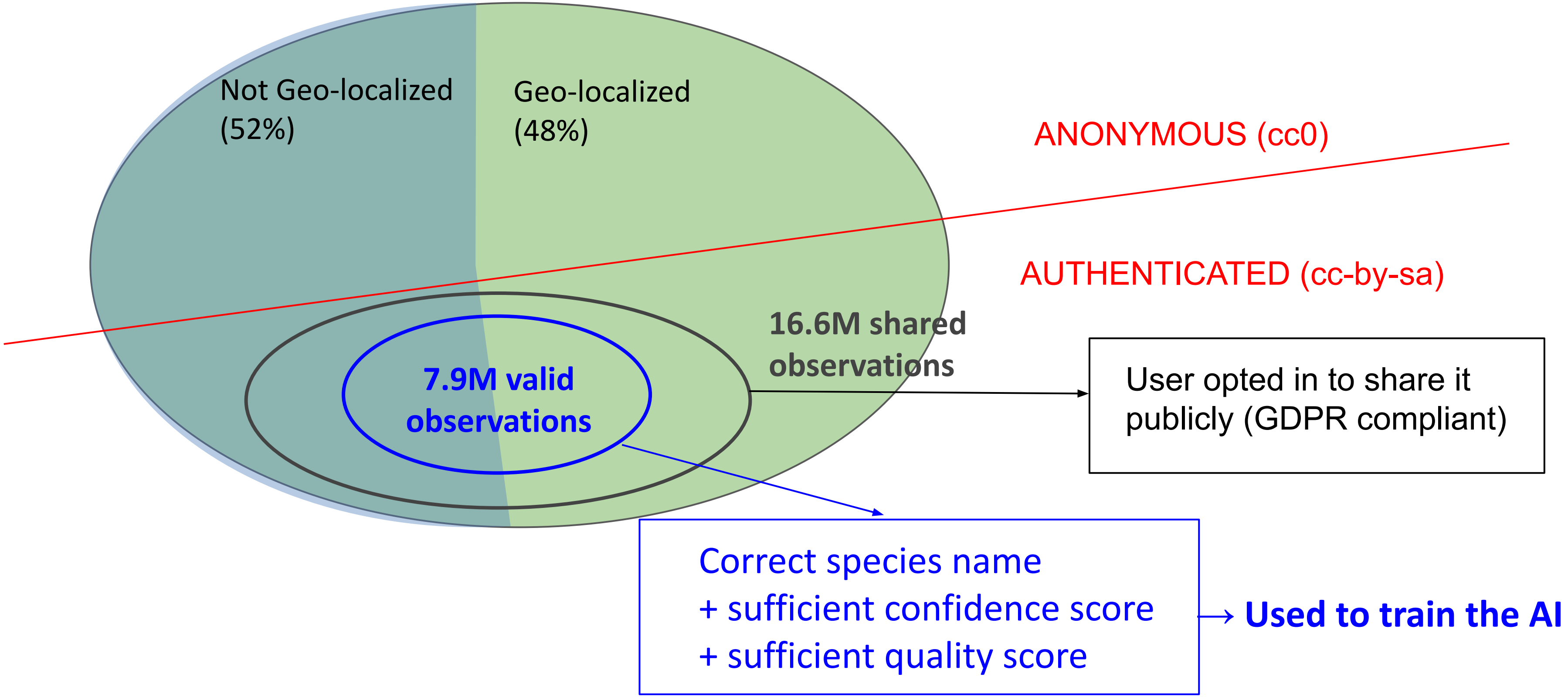


Most probable species $y = \arg\max_j \hat{\eta}_j(x)$

Validation decision (valid → used by AI) $\hat{\eta}_y(x) > \theta$

Pl@ntNet Data

750M raw observations (=queries)



Pl@ntNet Data shared in GBIF

- **Top-4 data provider to GBIF** (world's largest infrastructure for biodiversity data)
- **Valid observations + trusted queries identified by the AI** (AI score>0.9)
- **Additional quality filters**: potted & cultivated plants removal, region-based filtering (Kew POWO)

GBIF

13 856 500 OCCURRENCES

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

421 CITATIONS

RED LIST

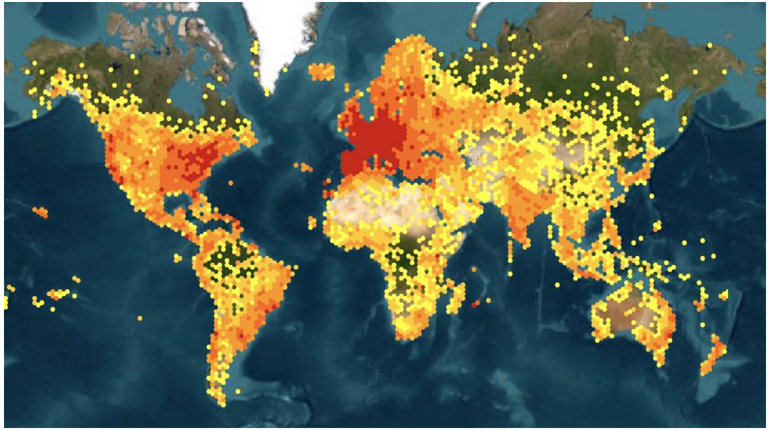
nature

PLOS ONE

ANNALS OF BOTANY Founded 1887

WILEY Publishers Since 1807

ELSEVIER



https://doi.org/10.15468/mma2ec

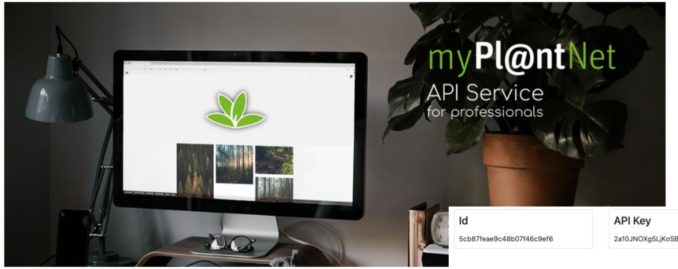
Pl@ntNet Latest major developments

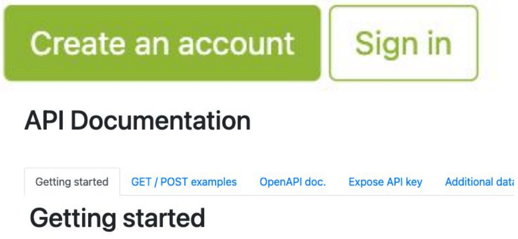
Cos4Cloud

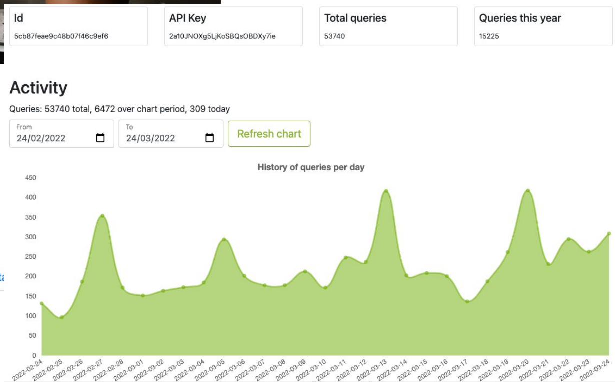
EUROPEAN OPEN SCIENCE CLOUD

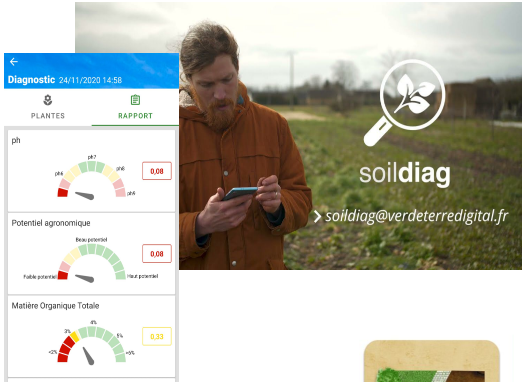
⚙ API

- A secured API providing developers programmatic access to Pl@ntNet engine
- **6K developer accounts** (researchers, companies, citizen observatories)
- Integrated in European Open Science Cloud (EOSC)



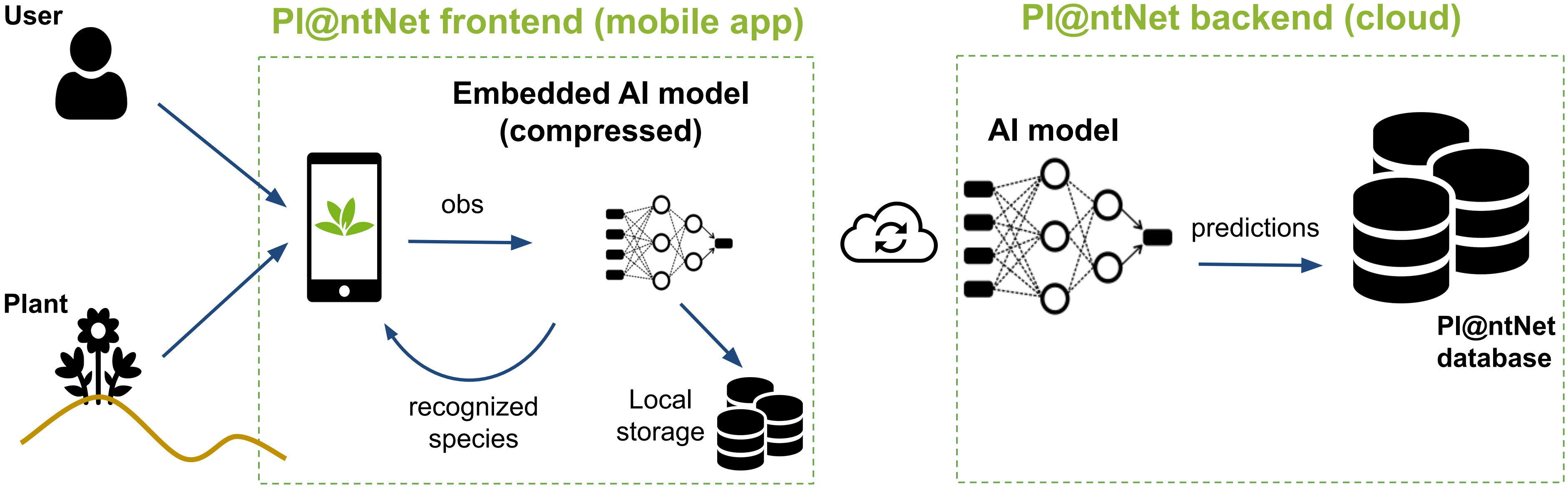






Pl@ntNet Latest major developments

Pl@ntNet offline: identify plants without connection



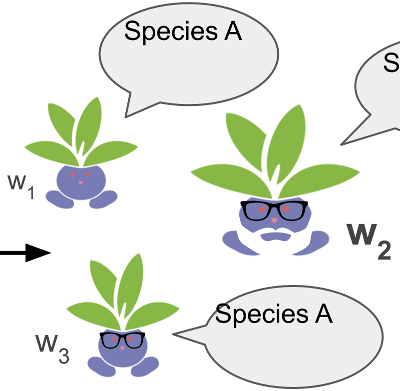
PART II
Latest cooperative learning algorithm

Cooperative Learning algorithm in detail

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 in detail

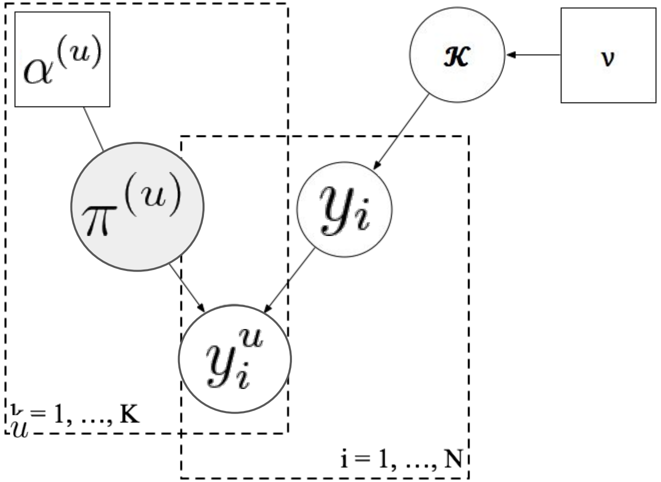
Unlike most state-of-the-art crowdsourcing approaches, the weight of a user is not determined by his estimated probability of success

Inferred confusion matrix of a user u

$\pi^{(u)}$ =

0.8	0.1	0.1
0.2	**0.6**	0.1
0.1	0.1	**0.7**

$$w_u = Tr(\pi^{(u)})$$



Problems:

- Not tractable for 45K classes
- Very sparse data for most users and species
- A user might be highly successful but only on a few very common species
- User scores interpretability (people love leaderboards)

Cooperative Learning algorithm in detail

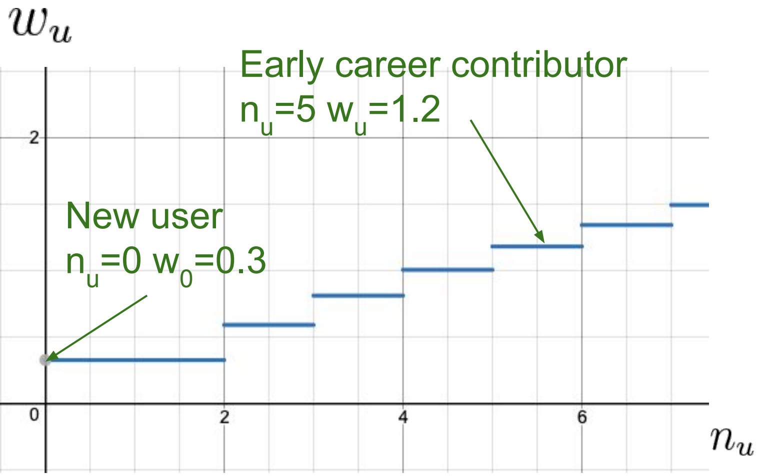
Rather, the weight of a user in Pl@ntNet is a function of the **estimated number of species** he is able to identify

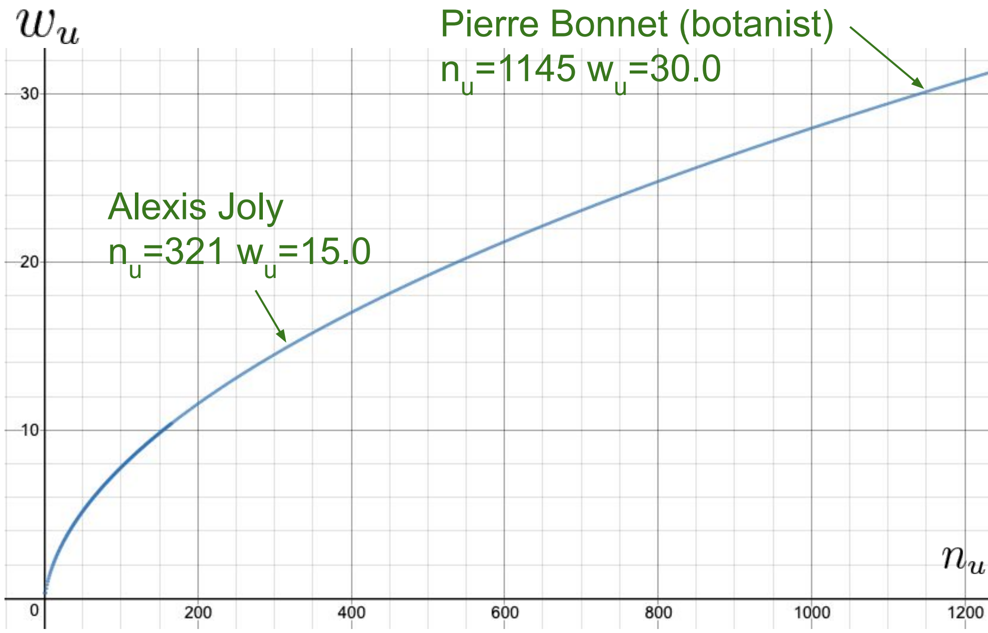
$$w_u = g(n_u) \qquad n_u = \left|\{j : \exists i \; y_i^u = y_i\}\right|$$

Cooperative Learning algorithm in detail

Rather, 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 in detail

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)

$$\eta_{y_i}(x_i) = \frac{w_{y_i}(x_i)}{\sum_k w_k(x_i)}$$

Cooperative Learning algorithm in detail

Parameters are estimated through an iterative algorithm similar to expectation-maximisation :

Initialization:

$w_u = w_0$ for all users

Repeat until convergence:

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

$$s_{y_i}(x_i) = \sum_{u \in U_i} w_u \mathbf{1}(y_i^u = y_i) \qquad \eta_{y_i}(x_i) = \frac{w_{y_i}(x_i)}{\sum_k w_k(x_i)}$$ Confidence and agreement scores

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

$$n_u = |\{j : \exists i\ y_i^u = \hat{y}_i | v(x_i) = 1\}| \qquad w_u = g(n_u)$$ Update user weights

Cooperative Learning algorithm in detail

A **new iteration** is ran **each night** but only on **new incremental data**:

1 - **Update user weights** for

- users who voted since last iteration
- users who created new observation(s) since last iteration
- users whose observations received a vote since last iteration

2 - **Compute validity score** for

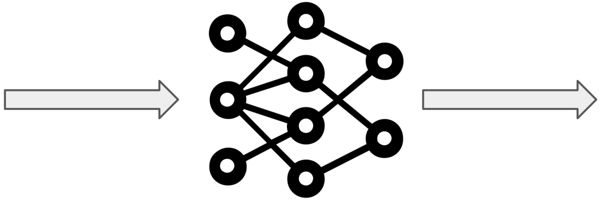
- new observations created since last iteration
- updated observations since last iteration (including the ones with new votes)
- observations having a vote whose author has had its weight modified since last iteration

Computation time: from **2 to 3 hours** depending on the volume of new data (e.g. longer the week-end)

Cooperative Learning algorithm in detail

Valid observations (i.e. $v(x_i) = 1$) are the only ones:
- used for training the AI
- appearing in Pl@ntNet galleries
- appearing in the identification results (visual similarity search)

Papaver argemone



A valid observation can be revised at any time within the application so that the label noise can be reduced afterwards

Cooperative Learning algorithm in detail

New observations

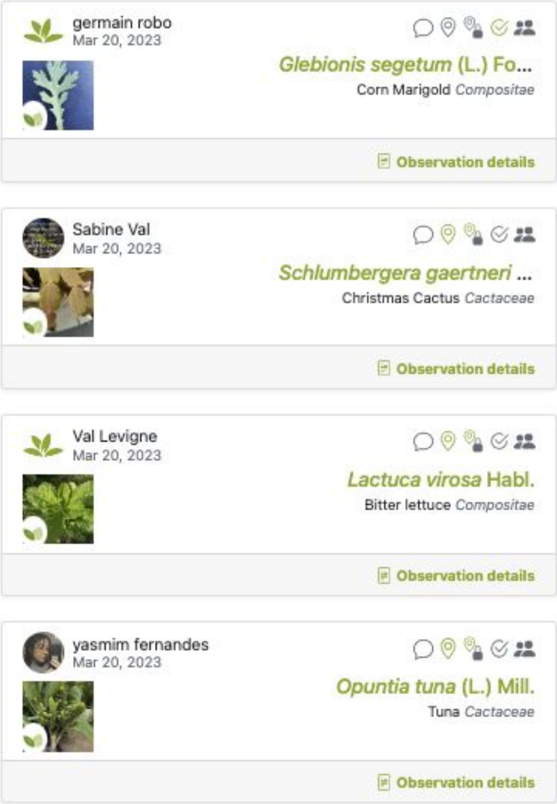
Appear only once in the contribution stream
$\longrightarrow$ they can be revised/confirmed on the fly (low rate)

They can be directly *valid* if the author has a sufficient weight

$$w_u > \theta \longrightarrow \quad s_{y_i}(x_i) = \sum w_u \mathbf{1}(y_i^u = y_i) > \theta$$

Such users are said *self-validating* ($\theta = 2.0$)

Obs of self-validating users can be unvalidated by a user with similar weight: $\dfrac{w_u}{w_u + w_{u'}} < \theta' \quad (\theta' = 0.7)$



Pl@ntNet Contributors

4M users accounts, 1M active contributors

Top 10 contributors

#	Weight	Species count	Observations	User
1	78.14	6932	17627	Diego Alex
2	65.43	4923	16408	Daniel Barthelemy
3	60.76	4269	15868	Liliane Roubaudi
4	53.81	3381	13653	Maarten Vanhove
5	52.45	3219	11567	Yoan Martin
6	51.35	3091	11209	Dieter Albrecht
7	49.3	2859	10463	Michal Svit
8	49.06	2832	9964	William Coville
9	46.46	2552	9210	Martin Bishop
10	46.25	2530	8757	Sylvain Piry

Typical contributor

Weight = 9.0

Rossen Vassilev

Stats

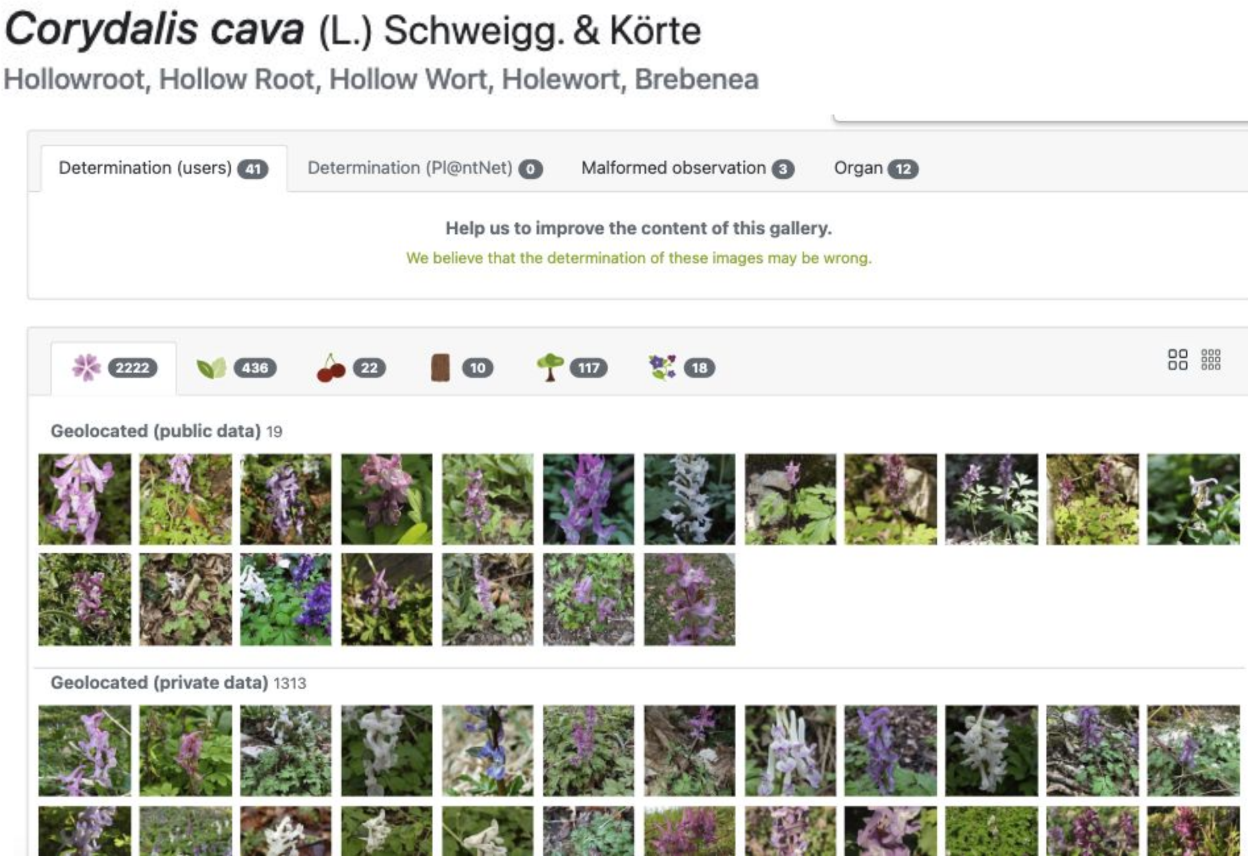
Rank 14 062

Observations
- Observed species 134
- Contributions 143
- Images 463

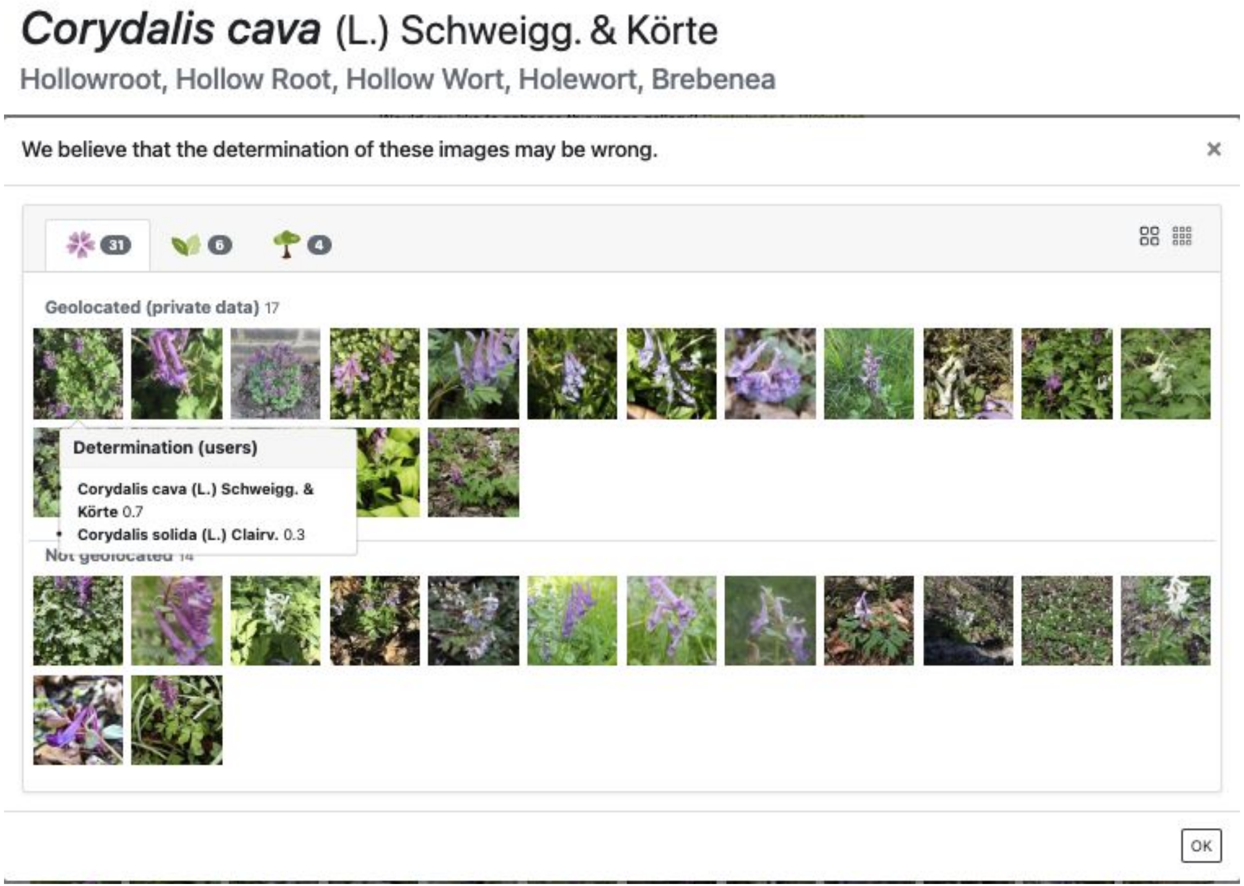
Votes
- Votes 54

Queries
- Identification requests 520
- Images 1005

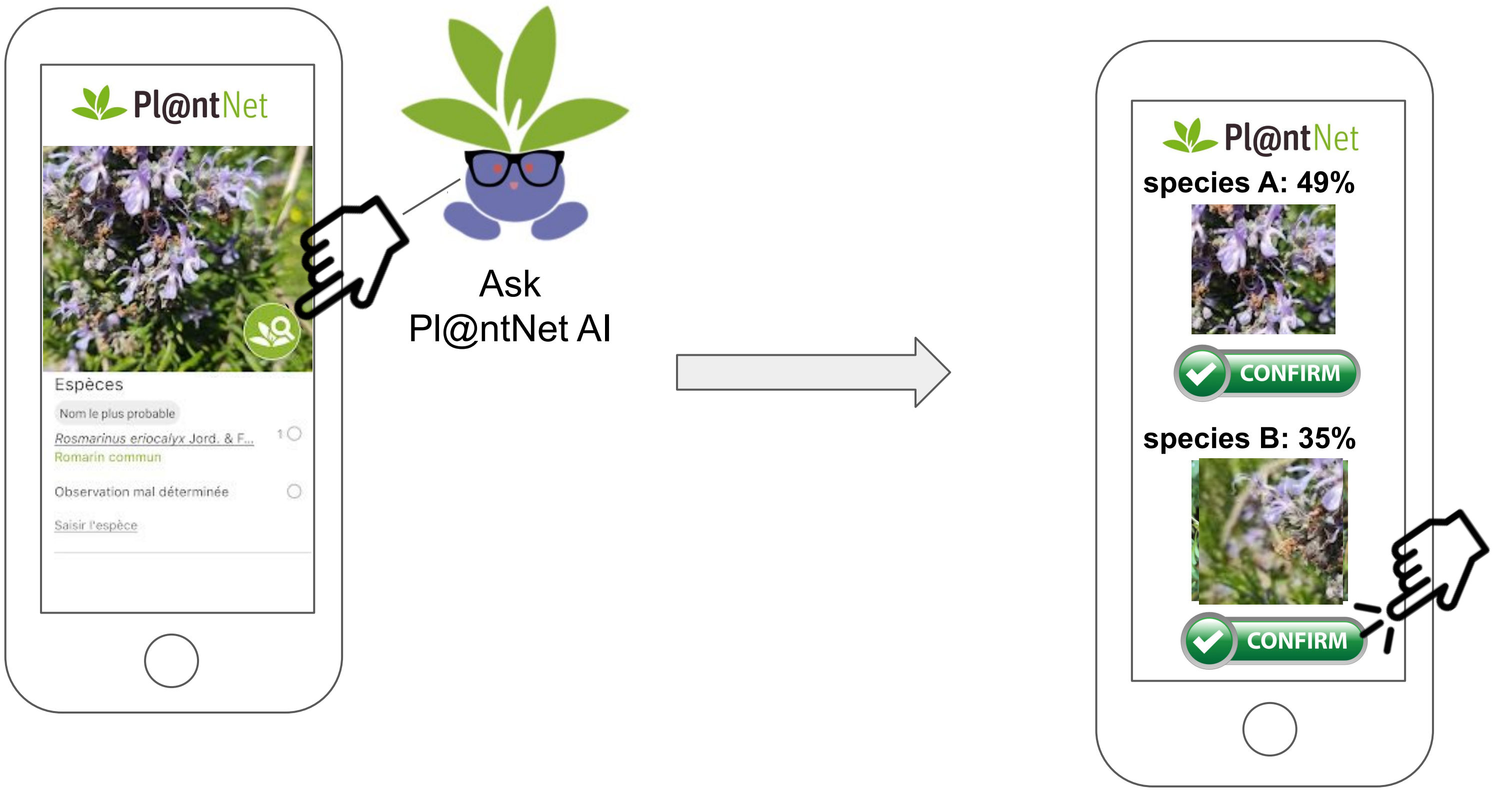
Active learning



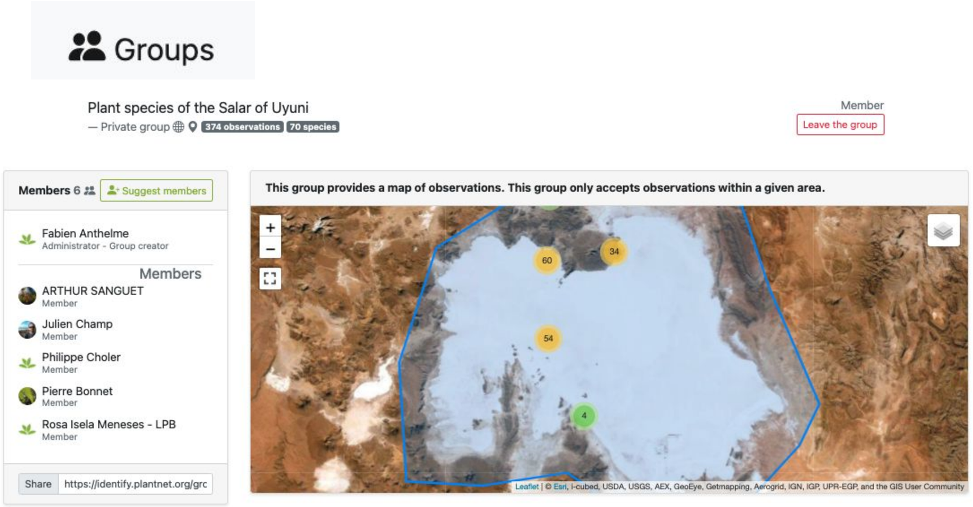
Active learning

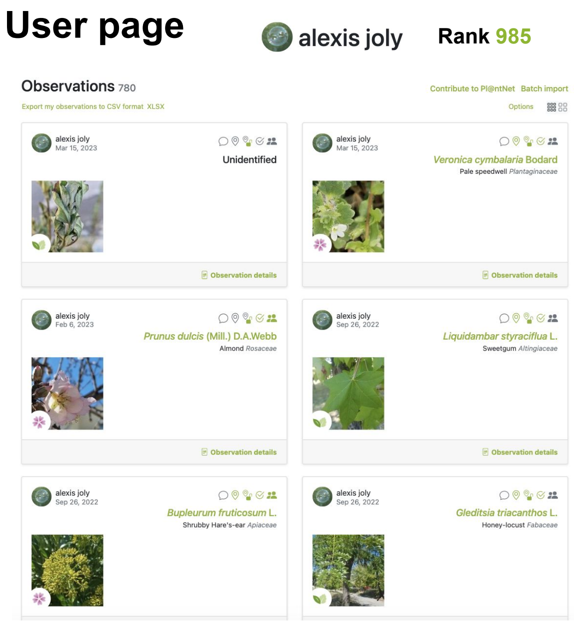


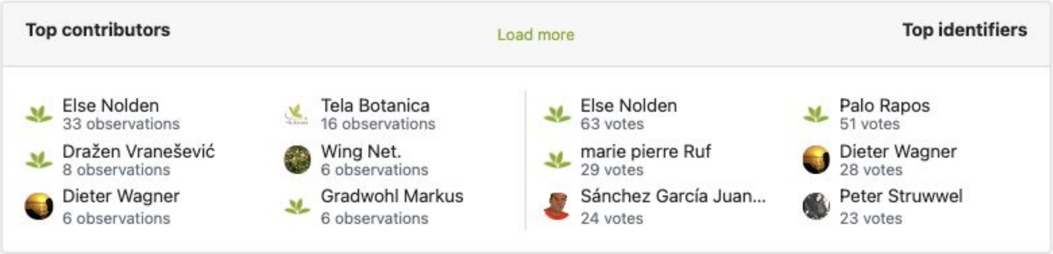
Active learning



Other collaborative tools





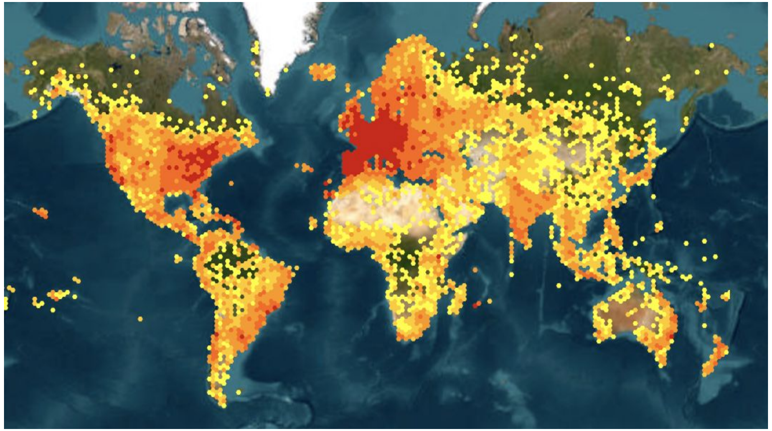


PART III
Deep Species Distribution Modeling

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

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

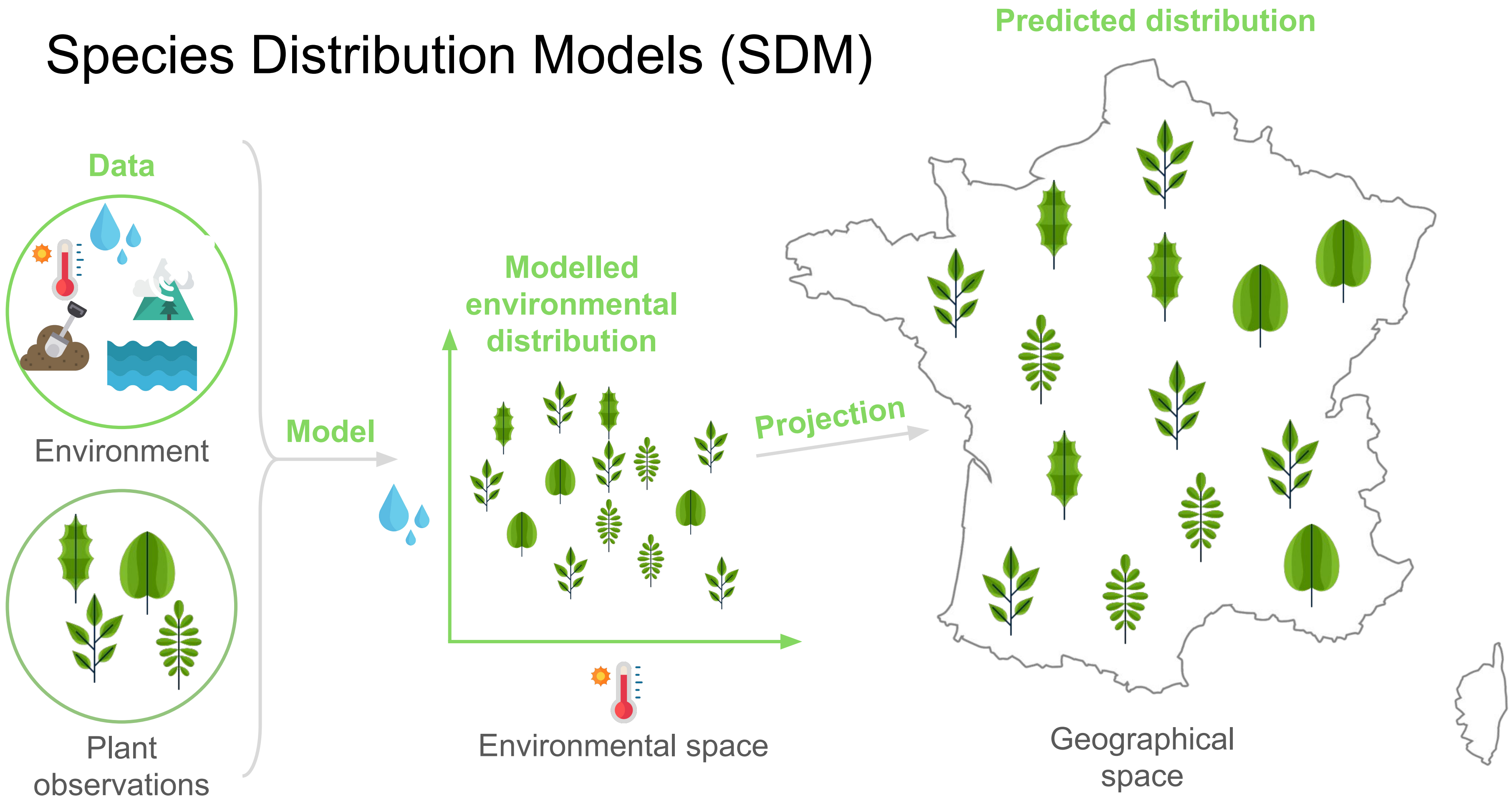
Many plant occurrences at world scale



GBIF

But very few locally for most species

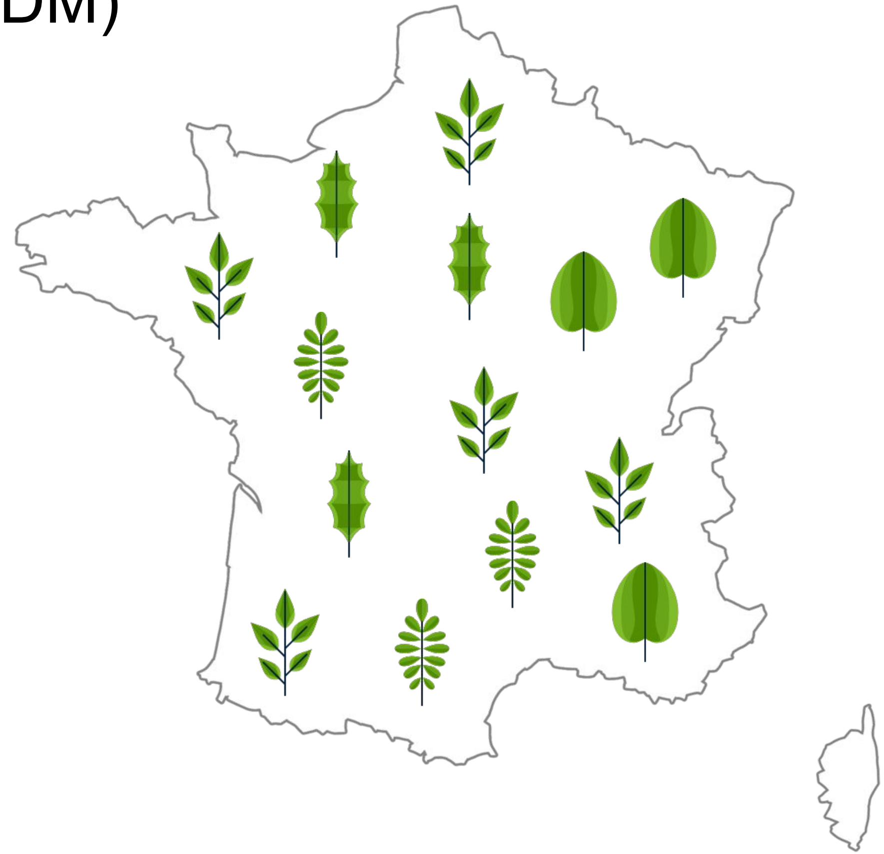
Species Distribution Models (SDM)



Species Distribution Models (SDM)

Motivations

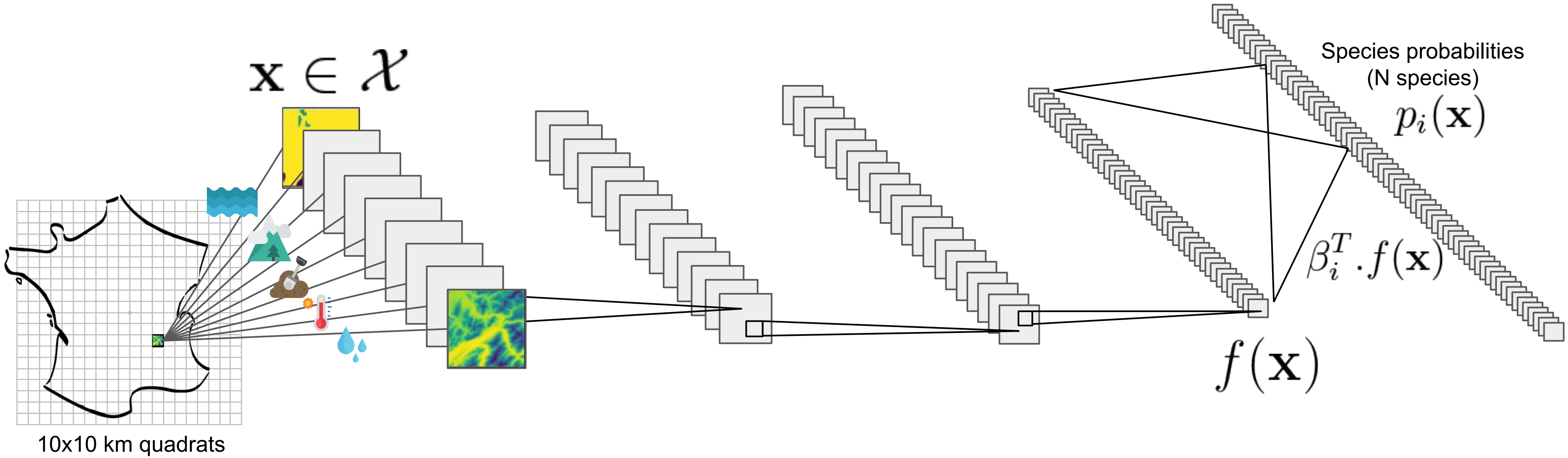
- Help conservation plans
- Invasive plant monitoring
- Learn about species preferences
- Simulation under climate change



A deep learning approach to species distribution modelling

Christophe Botella *et al.*, "A deep learning approach to species distribution modelling." *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*. Springer, 2018. 169-199.

- NN can model complex relationships from heterogeneous data sources
- Learn a joint representation space $f(\mathbf{x})$ of the environment for all species (≈ latent variables)
- Capturing multi-scale spatial information thanks to convolutional layers (CNN)



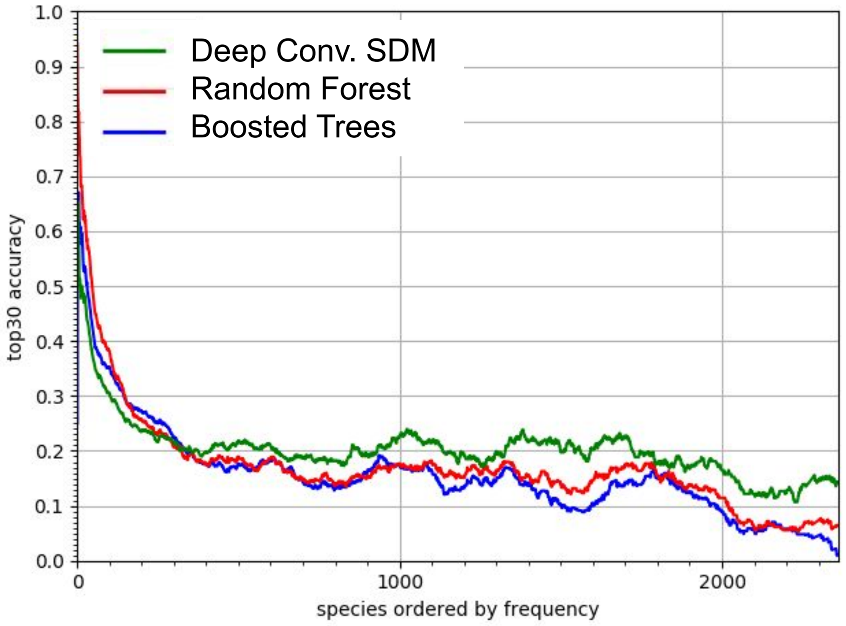
Understanding Deep Convolutional SDMs

Benjamin Deneu *et al.*, "Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment", *PLOS Computational Biology*

- Better knowledge transfer to least frequent species

<u>Model</u>
Architecture: Inception v3
Loss: categorical loss

<u>Data</u>
Source: GBIF
Type: occurrences
Nb of occurrences: 97 683
Nb of species: 4520
Environmental data:
33 geographic rasters (19 bioclimatic, 1 evapotranspiration, 10 pedologic, altitude, 1 hydro, Corine Land Cover)



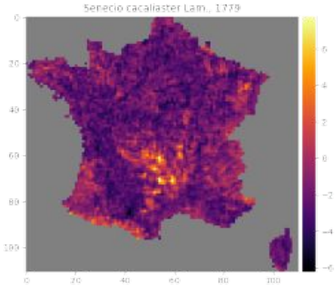
Understanding Deep Convolutional SDMs

Benjamin Deneu *et al.*, "Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment", *PLOS Computational Biology*

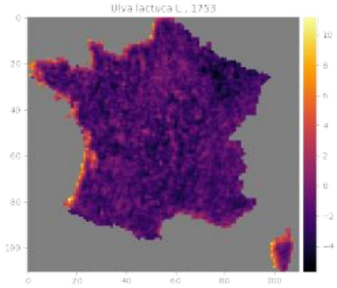
- Better knowledge transfer to least frequent species

Occurrences in training set | Predicted distribution | Comparison with another data source (INPN)

Senecio cacaliaster Lam.



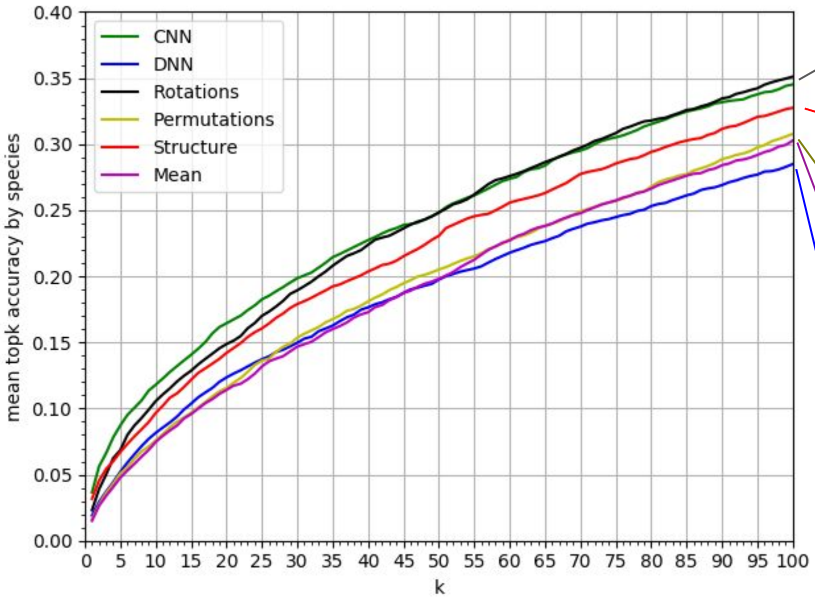

Ulva lactuca L.



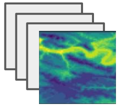

Deriving knowledge from Deep SDMs

Benjamin Deneu *et al.*, "Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment", *PLOS Computational Biology*

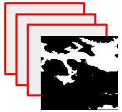
- Spatial structure of the local environment plays an important role in species distribution (landscape, barriers, relief, etc.)



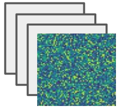
Complete information (structure+values)



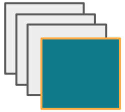
Only structural information (standardized values)



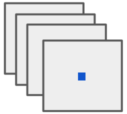
Only values (no structure)



Mean value (no structure)



No spatial information (central value)



How to train Deep SDM models ?

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:

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

If we have an **estimator** : $\hat{\eta}_k(x)$ of $\mathbb{P}_{X,Y}(Y = k|X = x)$

We can define the following *plug-in* estimator of the assemblage:

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

Predicting species assemblages from presence only data

How to get a good estimator $\hat{\eta}_k(x)$ of the conditional probability ?

→ Train a model using the **negative log-likelihood** = a **strictly proper loss**, i.e. it is minimized only when the model predicts the true $\eta_k(x) = \mathbb{P}_{X,Y}(Y = k|X = x)$

$$\arg\min_{\theta} \sum_{i} -log\ \hat{\eta}_{y_i}(x_i) \text{ e.g. with } \hat{\eta}_k(x) = \frac{exp(f_\theta^k(x))}{\sum_j exp(f_\theta^j(x))} =$$

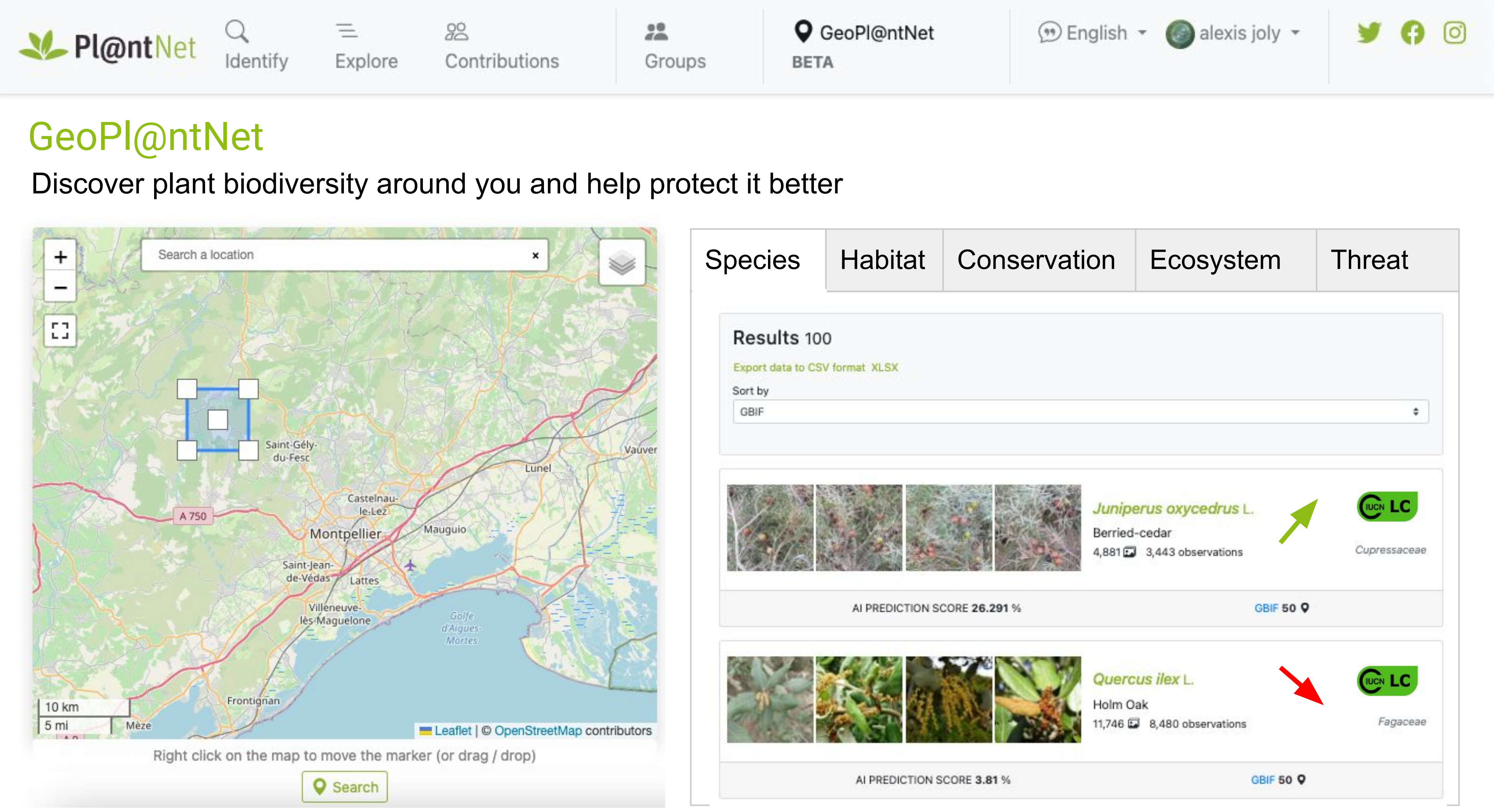
neural network output

In brief:

- Our plug-in predictor simply consists in **thresholding the softmax output** of a neural network trained with the so-called ***cross-entropy*** loss

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

- It is proved that $S_\lambda(x)$ assymptotically converges towards $S^*_\lambda(x)$


Pl@ntNet
Identify
Explore
Contributions
Groups
GeoPl@ntNet
BETA
English
alexis joly
GeoPl@ntNet
Discover plant biodiversity around you and help protect it better
Search a location
Saint-Gély-du-Fesc
Vauvert
Lunel
Castelnau-le-Lez
A 750
Montpellier
Mauguio
Saint-Jean-de-Védas
Lattes
Villeneuve-lès-Maguelone
Golfe d'Aigues-Mortes
Frontignan
Mèze
10 km
5 mi
Leaflet | © OpenStreetMap contributors
Right click on the map to move the marker (or drag / drop)
Search
Species
Habitat
Conservation
Ecosystem
Threat
Results 100
Export data to CSV format XLSX
Sort by
GBIF
Juniperus oxycedrus L.
Berried-cedar
4,881 3,443 observations
IUCN LC
Cupressaceae
AI PREDICTION SCORE 26.291 %
GBIF 50
Quercus ilex L.
Holm Oak
11,746 8,480 observations
IUCN LC
Fagaceae
AI PREDICTION SCORE 3.81 %
GBIF 50

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:

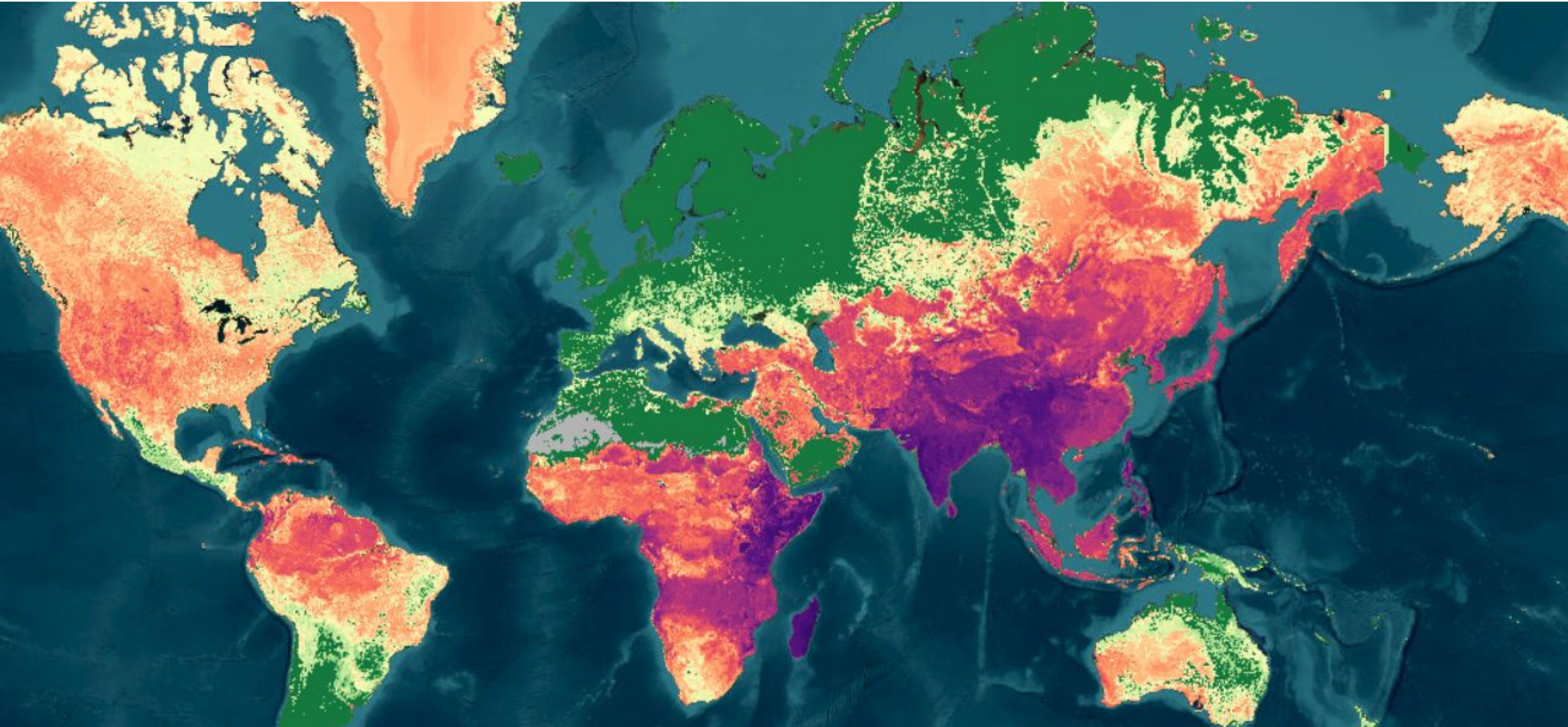
- The proportion of endangered species (e.g. on IUCN red list)
- The proportion is woody species
- The diversity of species (e.g. Shanon index)

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)

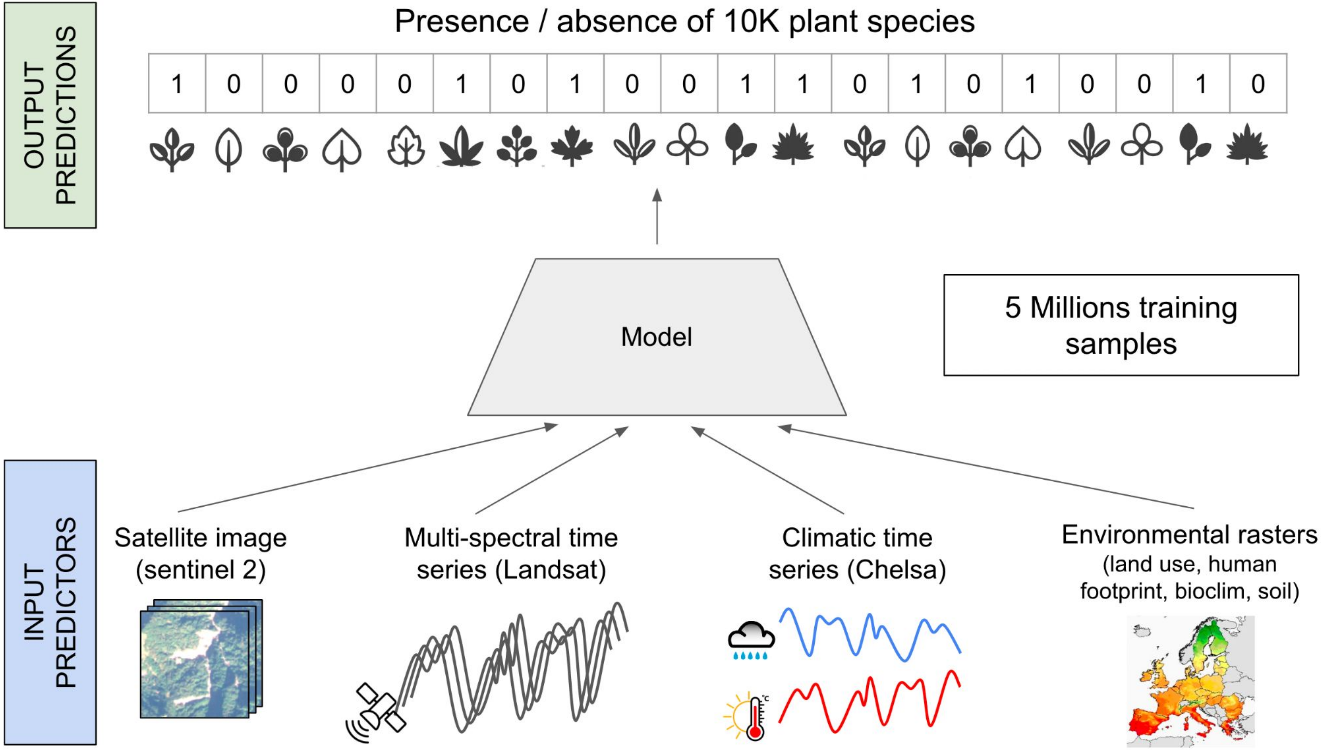
1x1 km resolution (view online)

PhD of Joaquim Estopinan



PART IV
Other ongoing stuff

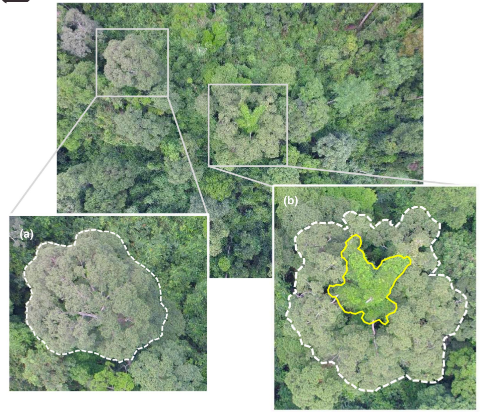
GeoLifeCLEF challenge 2023

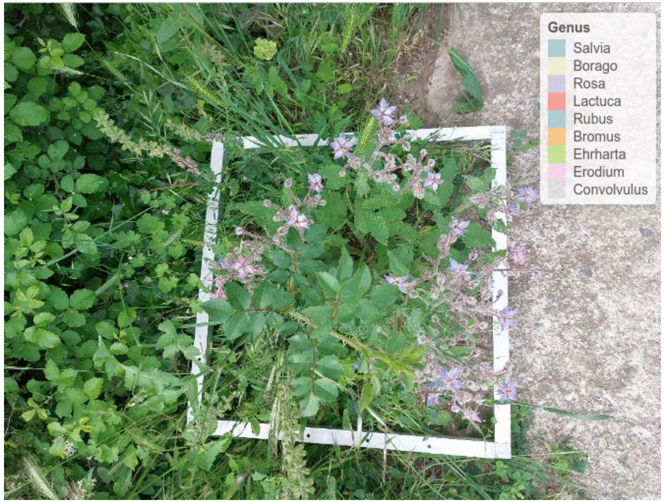


JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

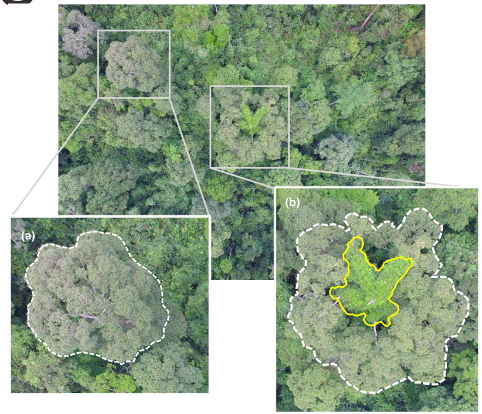
New biodiversity monitoring approaches

- Car views for the monitoring of invasive species (human vector)
- Quadrat images for the monitoring of vulnerable habitats or fields biodiversity
- Drones for the monitoring of forest canopies

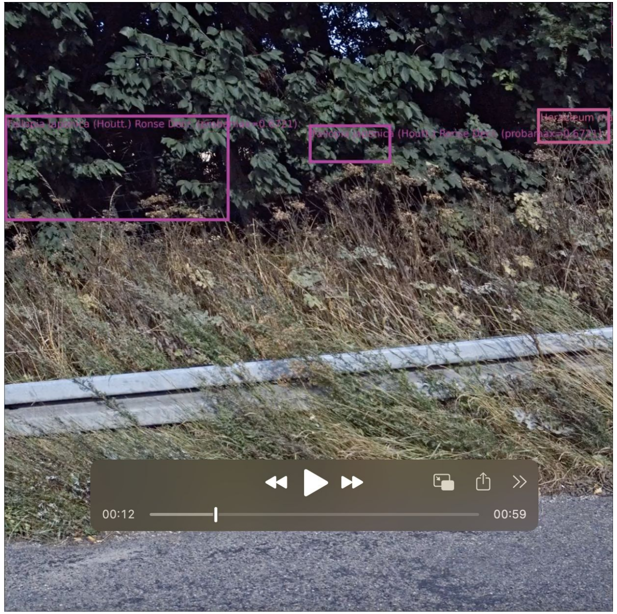


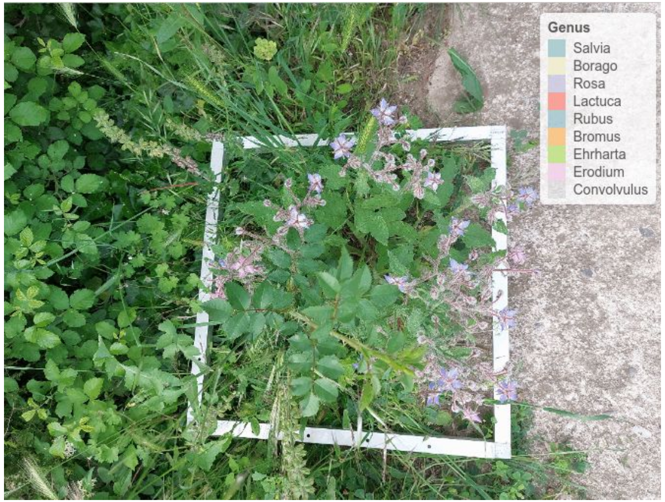


New biodiversity monitoring approaches



- Car views for the monitoring of invasive species (human vector)
- Quadrat images for the monitoring of vulnerable habitats or fields biodiversity
- Drones for the monitoring of forest canopies





Habitats mapping and future trajectories prediction

MAMBO GUARDEN

PhD thesis of Cesar Leblanc

Deep SDM → **Plants community** → **ML model** → **Habitat type**

3
0
1
0
0
0
14

Input data = tabular data
- abundance
- presence/absence

Species-to-habitat classifier →

Habitat N14

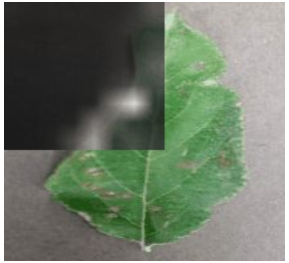
Mediterranean shifting coastal dune

Pl@ntAgroEco

Designing new services for agroecology within the Pl@ntNet platform

Plant **disease** identification

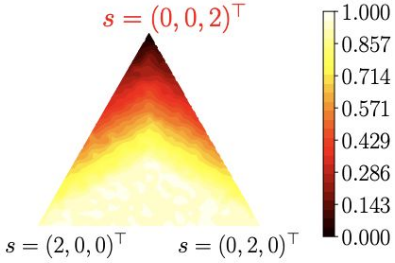
- Collaborative epidemiology surveillance
- Reduction of phytosanitary products
- Jointly with ephytia



Identification of **infra-specific** taxa

- Crop varieties, horticol varieties, cultivar, hybrids, etc.
- Towards a selection more respectful of the environment

Handling uncertainty and bias of species identification



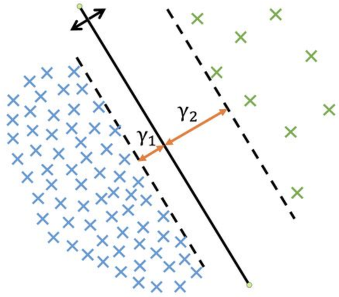
Advanced optimization techniques

- Uncertainty: top-K loss function
- Imbalance: shifting of the decision frontier

K	ℓ_{CE}
1	36.3±0.3 (12.6/42.9/71.7)
3	58.8±0.4 (32.4/**75.3**/92.0)
5	68.7±0.2 (45.1/**86.3**/95.4)

→

$\ell^{K,0.01,5,\max m_y=0.2}_{\text{Noised imbal.}}$
42.4±0.3 (**23.9**/**46.3**/72.1)
64.9±0.4 (**44.8**/74.5/92.1)
73.2±0.5 (**55.3**/84.2/95.3)



Statlearn poster today:

<u>Camille Garcin</u>, M. Servajean, A. Joly, J. Salmon. *Stochastic smoothing of the top-K calibrated hinge loss for deep imbalanced classification*. ICML 2022.

Thank you