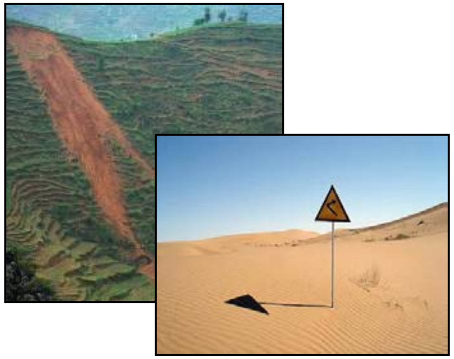
Plant identification in an open-world

The LifeCLEF 2016 Plant Identification Task

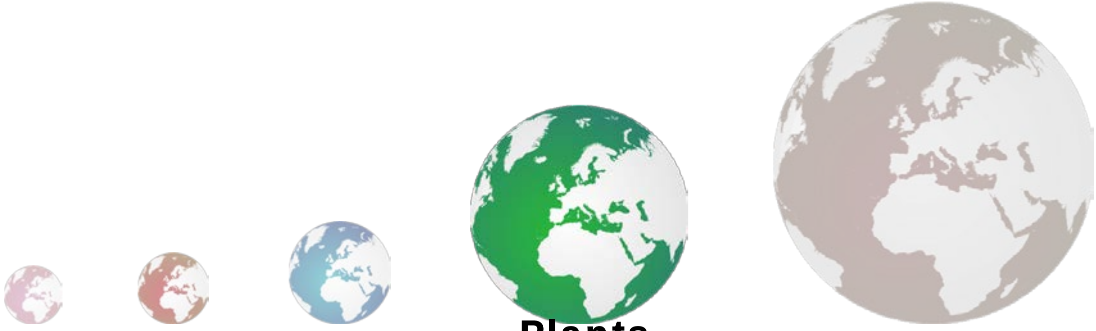
Hervé Goëau, Alexis Joly, Pierre Bonnet

Context & challenges



Plant identification is the **key** for gathering and sharing information in order to have a better knowledge about plants

Taxonomic gap: a huge and unknown number of species



Mammals
5416 sp.

Birds
9956 sp.

Fishes
30 000 sp.

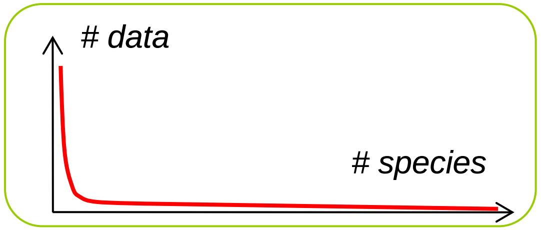
Plants
400 000 sp.!

Insects
950 000 sp.

http://www.factmonster.com/ipka/A0934288.html

Botanical data is:

- sparse and incomplete ("long tail distribution")
- decentralized and heterogeneous
- complex (un-structured tags, empirical measures...)

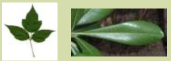
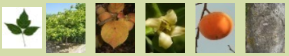
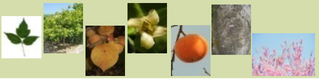
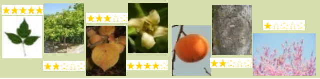


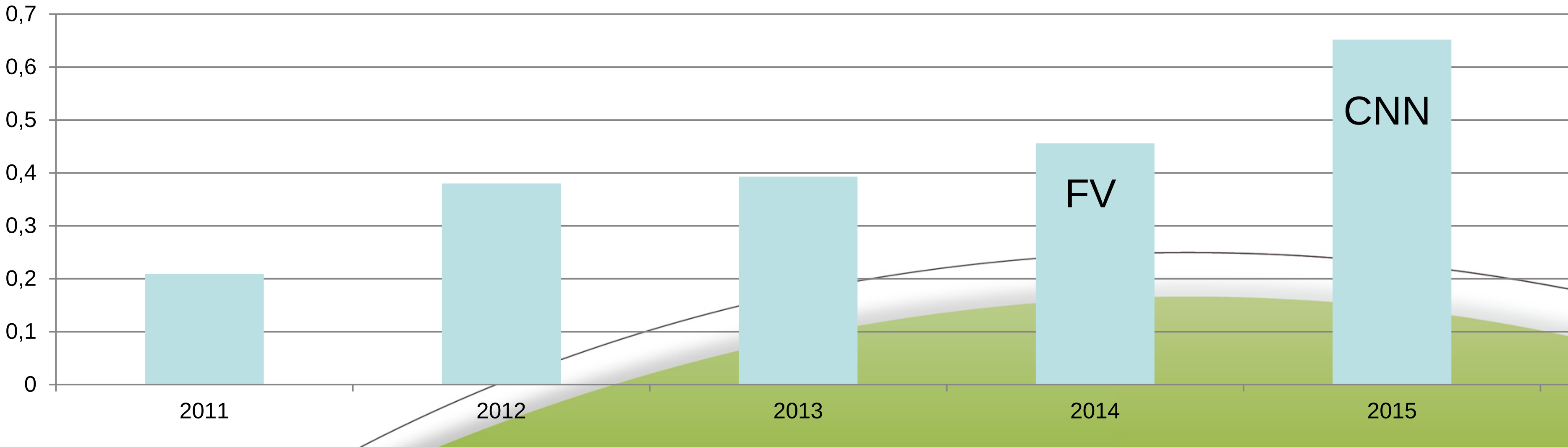
Multimedia identification tools is considered as one of the most **promising solution** to help bridging the taxonomic gap

5 years of Plant Identification Task

A lot of work, a lot of progress…

- From single scans of leaf to multi-organ plant identification
- From few dozens of species to 1000 species
- from scientific protocols (scans in lab), to mobile crowsourcing data

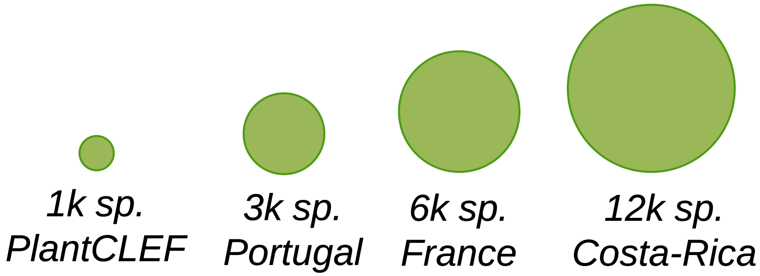
	2011	2012	2013	2014	2015
Species	71	126	250	500	1,000
Images	5,400	11,500	26,077	60,962	113,205
Views					
Perf.	0,209	0,38	0,393	0,456	0,652



5 years of Plant Identification Task

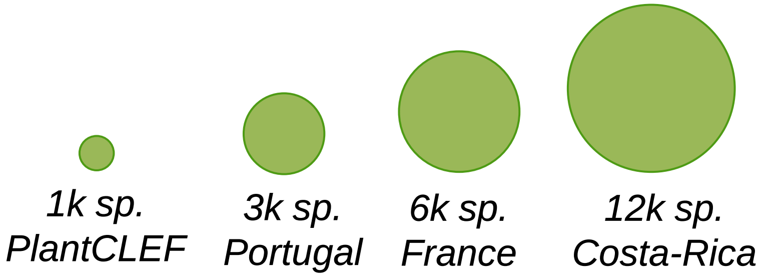
5 years of Plant Identification Task

However, measured performances are still far from that it can be expected in a real scenario



5 years of Plant Identification Task

However, measured performances are still far from that it can be expected in a real scenario



1k sp.
PlantCLEF

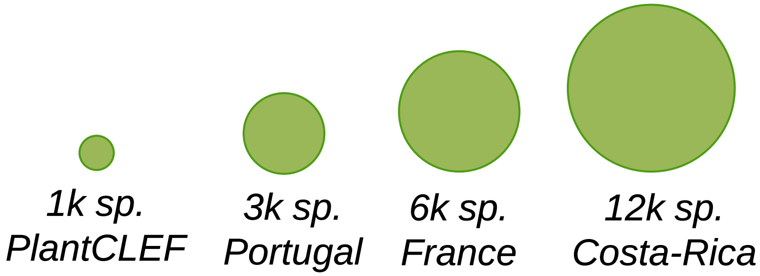
3k sp.
Portugal

6k sp.
France

12k sp.
Costa-Rica

5 years of Plant Identification Task

However, measured performances are still far from that it can be expected in a real scenario



1k sp.
PlantCLEF

3k sp.
Portugal

6k sp.
France

12k sp.
Costa-Rica

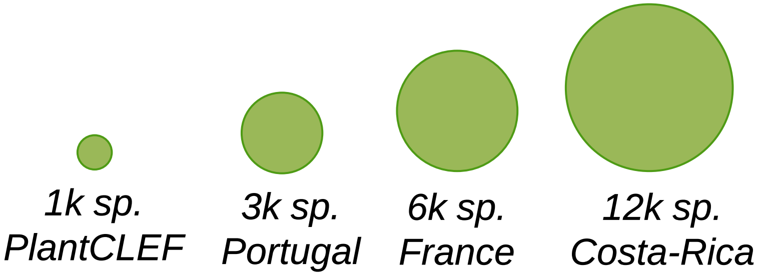
At the same time popular apps appeared …

- *with image based automatic identification*
- *or / and with collaborative identification*

FlowerChecker

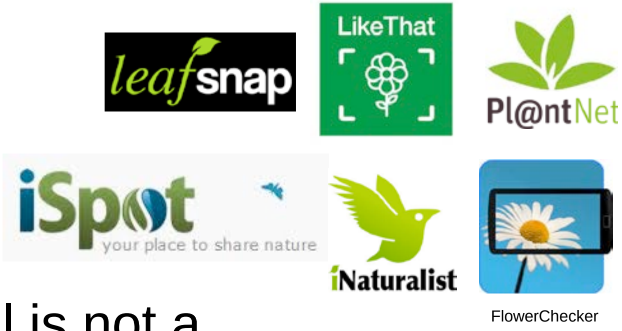
5 years of Plant Identification Task

However, measured performances are still far from that it can be expected in a real scenario



At the same time popular apps appeared …

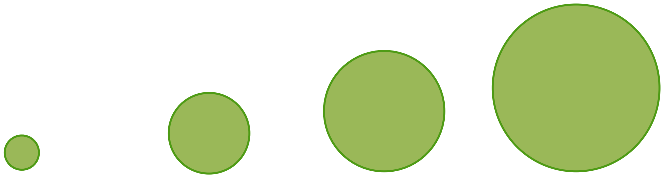
- *with image based automatic identification*
- *or / and with collaborative identification*



… expressing that biodiversity information retrieval is not a narrow topic and does interest people as much as other entertainments ...

5 years of Plant Identification Task

However, measured performances are still far from that it can be expected in a real scenario



1k sp.
PlantCLEF

3k sp.
Portugal

6k sp.
France

12k sp.
Costa-Rica

At the same time popular apps appeared …

- *with image based automatic identification*
- *or / and with collaborative identification*

FlowerChecker

… expressing that biodiversity information retrieval is not a narrow topic and does interest people as much as other entertainments ...

… and finally creating a huge number of new plant observations and images:

- explicitly shared with the communities
- non shared but recorded as raw observations for future usages (the queries)

Pl@ntNet app

Pl@ntNet

iOS

android

Launched in February 2013 on 800 species from French flora

It is now actually working on about 10k species from France, French Guyana, Reunion Island, North Africa

- **2.4 M** users cumulating **11,5 M** sessions from **150 Countries**
- **Between 10k and 50 K users per day during the year 2016**
- **Generated more than 7 M of pictures through the queries !**

Countries	Downloads
France	750 000
US	500 000
Italy	140 000
Spain	125 000
Germany	125 000
Brazil	100 000
Canada	90 000
Belgium	80 000
UK	78 000
Australia	44 000

Pl@ntNet app: queries and shared observations for validation

0. select a flora

Tela Botanica

Back office

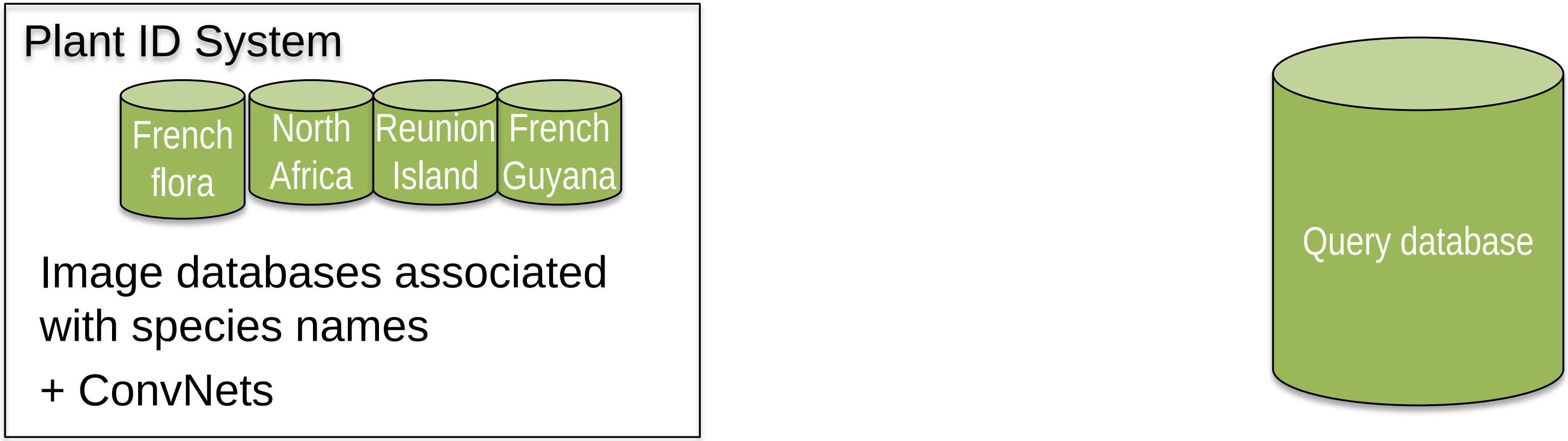
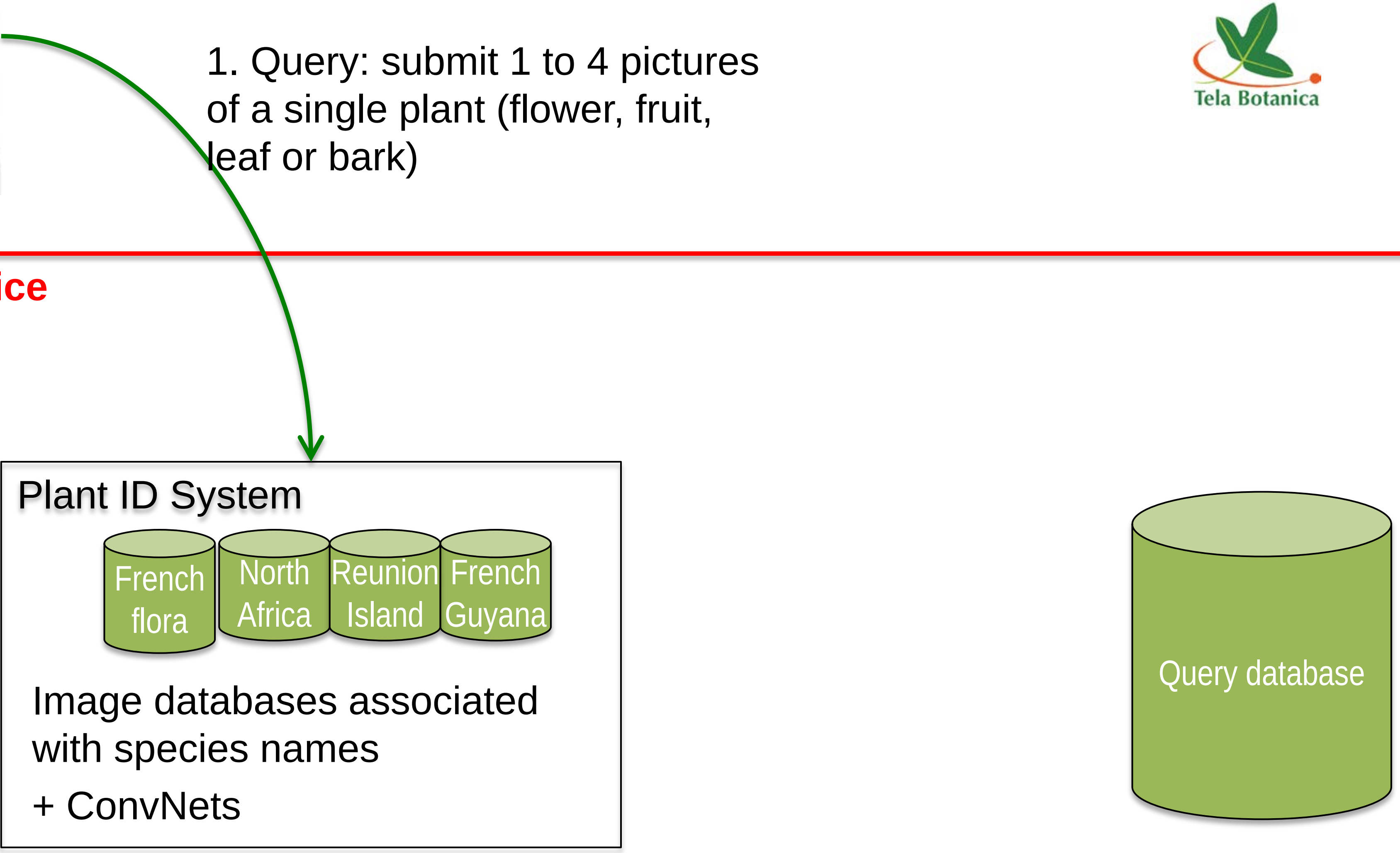

Plant ID System

Image databases associated with species names

+ ConvNets

Pl@ntNet app: queries and shared observations for validation

0. select a flora



1. Query: submit 1 to 4 pictures of a single plant (flower, fruit, leaf or bark)

Tela Botanica

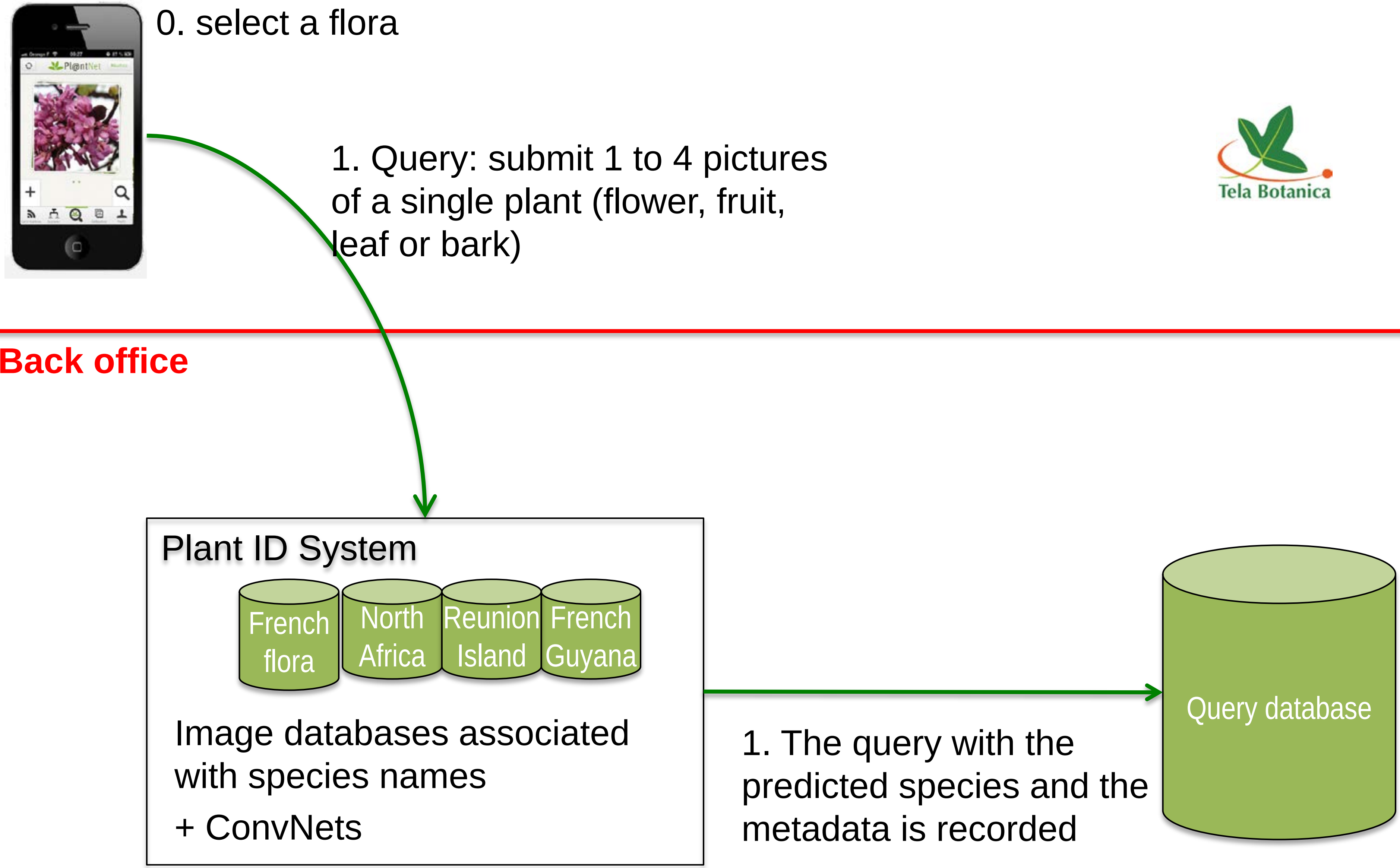
Back office

Plant ID System

Image databases associated with species names

+ ConvNets

Pl@ntNet app: queries and shared observations for validation



0. select a flora

1. Query: submit 1 to 4 pictures of a single plant (flower, fruit, leaf or bark)

Tela Botanica

Back office

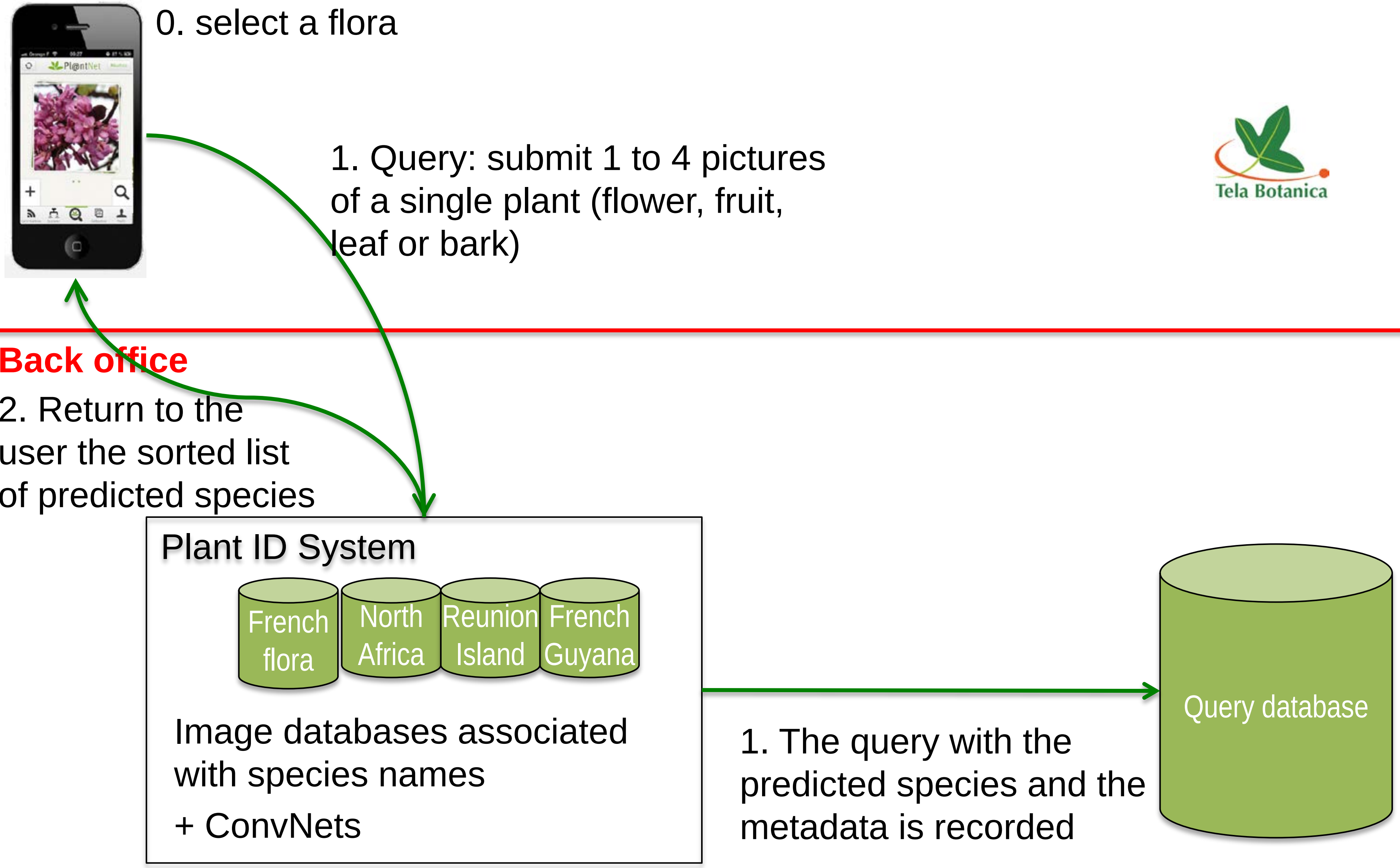
Plant ID System

Image databases associated with species names

+ ConvNets

1. The query with the predicted species and the metadata is recorded

Pl@ntNet app: queries and shared observations for validation



0. select a flora

1. Query: submit 1 to 4 pictures of a single plant (flower, fruit, leaf or bark)

Back office

2. Return to the user the sorted list of predicted species

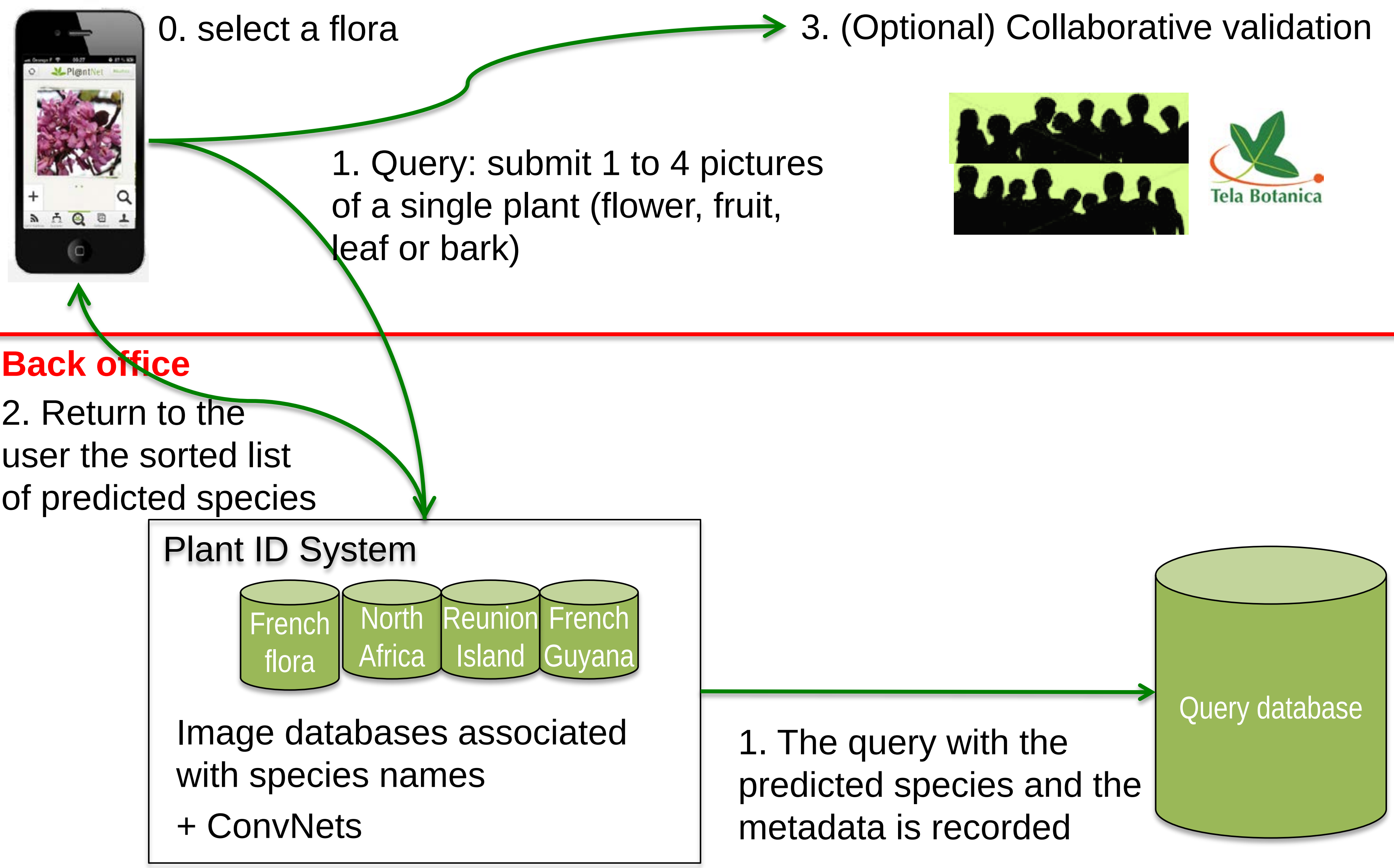
Plant ID System

Image databases associated with species names

+ ConvNets

1. The query with the predicted species and the metadata is recorded

Pl@ntNet app: queries and shared observations for validation



0. select a flora

1. Query: submit 1 to 4 pictures of a single plant (flower, fruit, leaf or bark)

3. (Optional) Collaborative validation

Back office

2. Return to the user the sorted list of predicted species

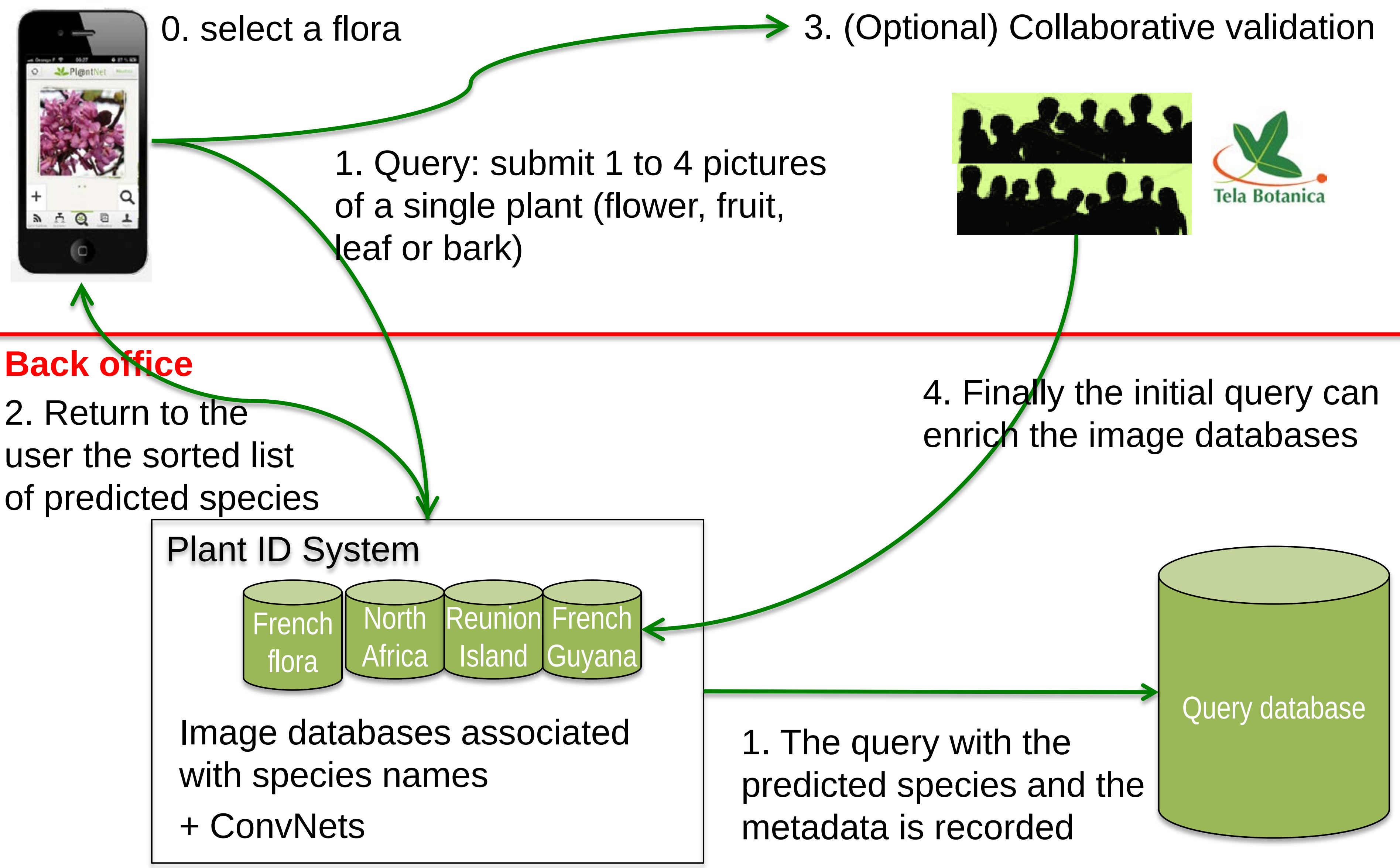
Plant ID System

Image databases associated with species names

+ ConvNets

1. The query with the predicted species and the metadata is recorded

Pl@ntNet app: queries and shared observations for validation



0. select a flora

3. (Optional) Collaborative validation

1. Query: submit 1 to 4 pictures of a single plant (flower, fruit, leaf or bark)

Tela Botanica

Back office

2. Return to the user the sorted list of predicted species

4. Finally the initial query can enrich the image databases

Plant ID System

French flora

North Africa

Reunion Island

French Guyana

Image databases associated with species names

+ ConvNets

1. The query with the predicted species and the metadata is recorded

Query database

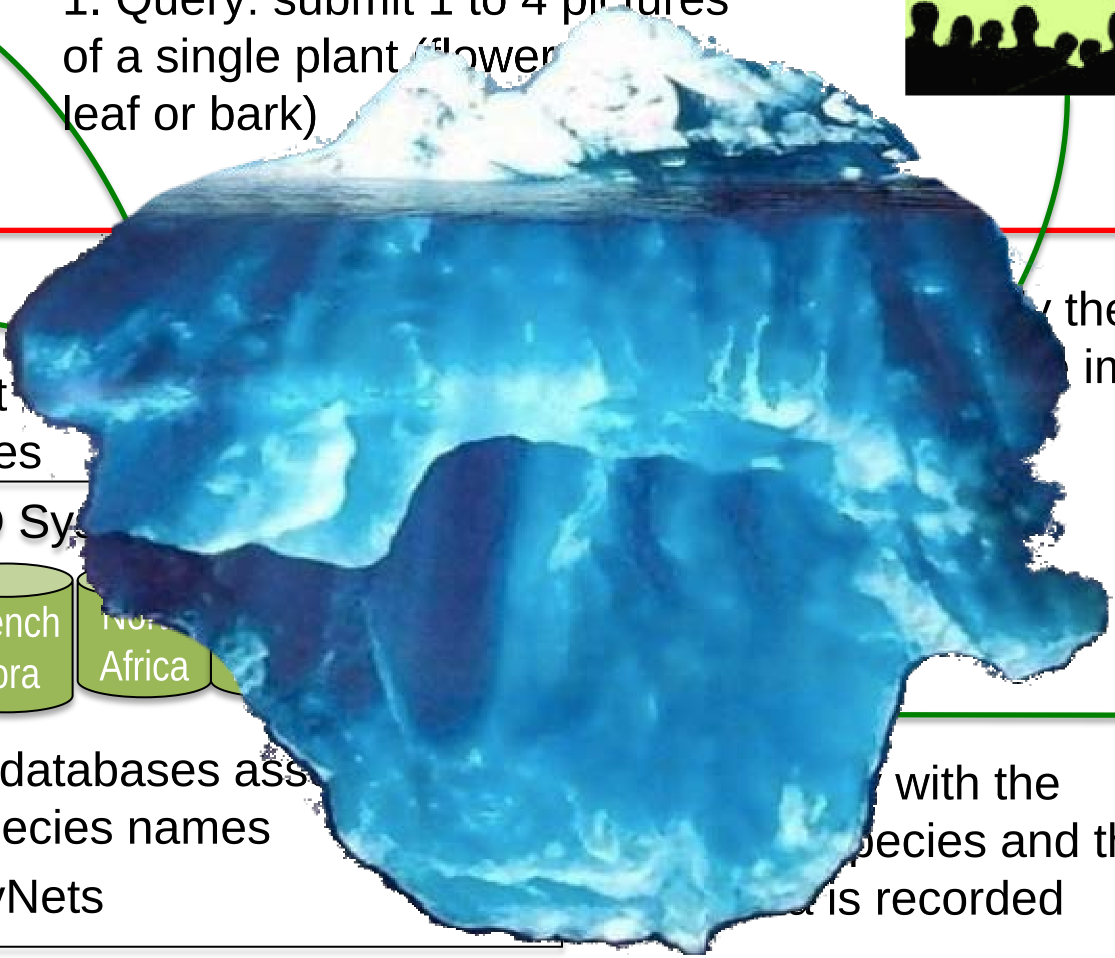
Pl@ntNet app: queries and shared observations for validation

0. select a flora

3. (Optional) Collaborative validation

Tela Botanica

1. Query: submit 1 to 4 pictures
of a single plant (flower
leaf or bark)



Back office

2. Return to the
user the sorted list
of predicted species

Plant ID Sys

French flora

No
Africa

Image databases ass
with species names

+ ConvNets

the initial query can
image databases

with the
pecies and the
is recorded

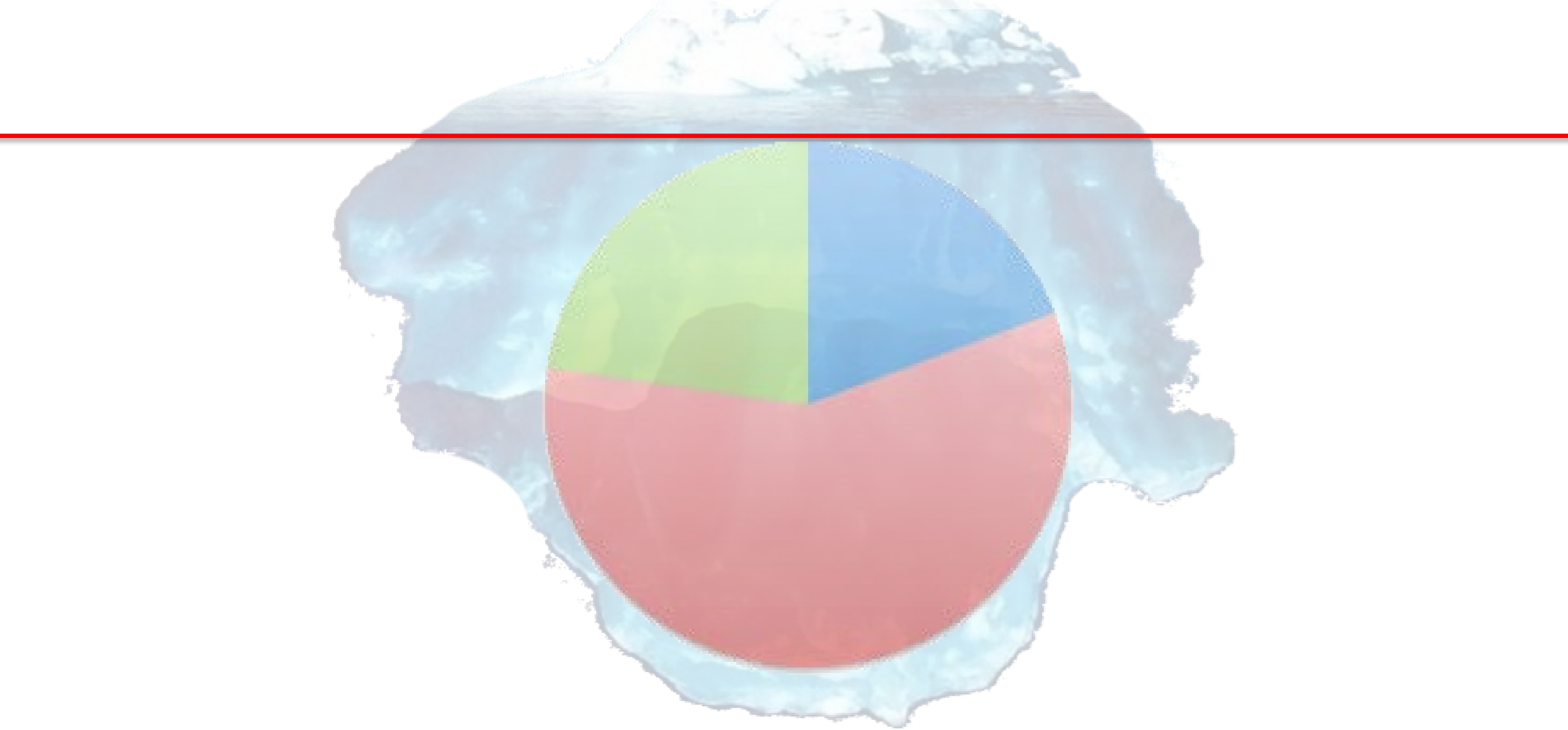
Query database

Pl@ntNet app: queries and shared observations for validation

A lot of unlabeled data through the queries: 7M pictures generated!

How much the raw unlabeled data is valuable for the biodiversity ?

Ex: a valuable resource for biodiversity issues such as the ecological monitoring of invasive plants ?



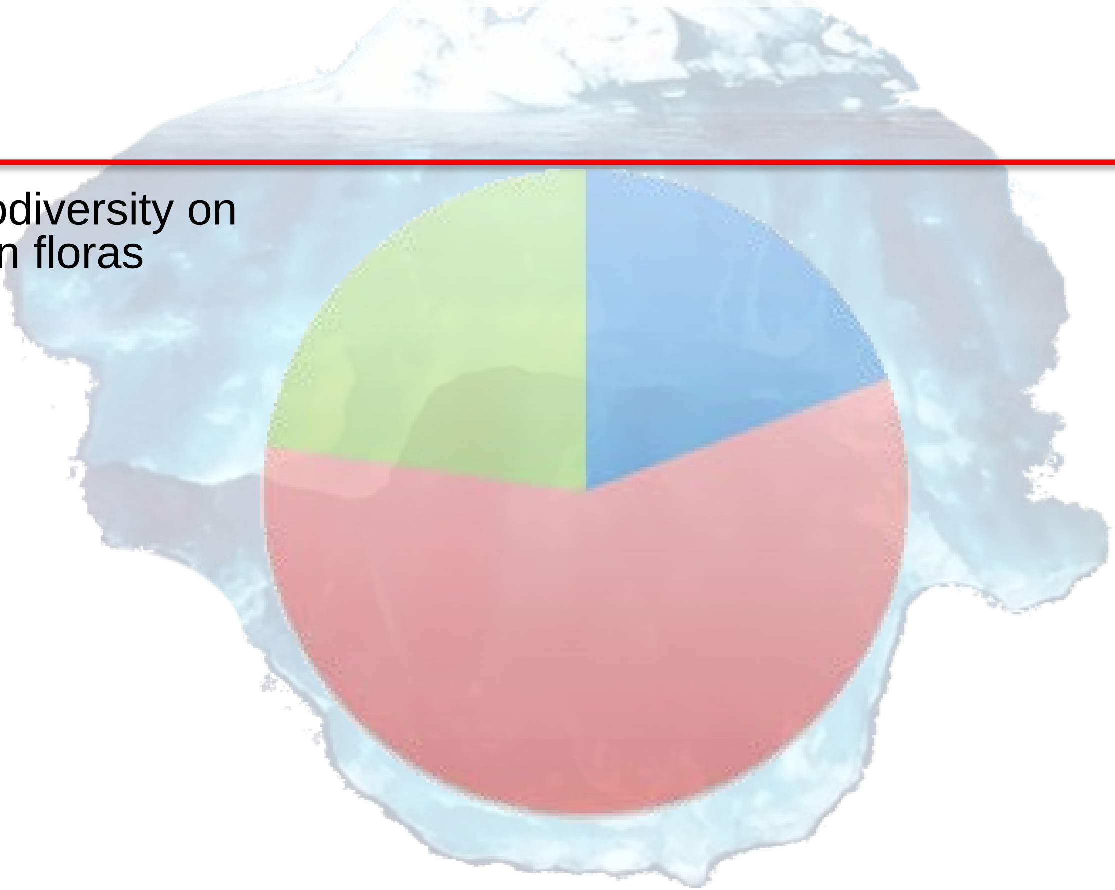
Pl@ntNet app: queries and shared observations for validation

A lot of unlabeled data through the queries: 7M pictures generated!

How much the raw unlabeled data is valuable for the biodiversity ?

Ex: a valuable resource for biodiversity issues such as the ecological monitoring of invasive plants ?

% Biodiversity on known floras



Pl@ntNet app: queries and shared observations for validation

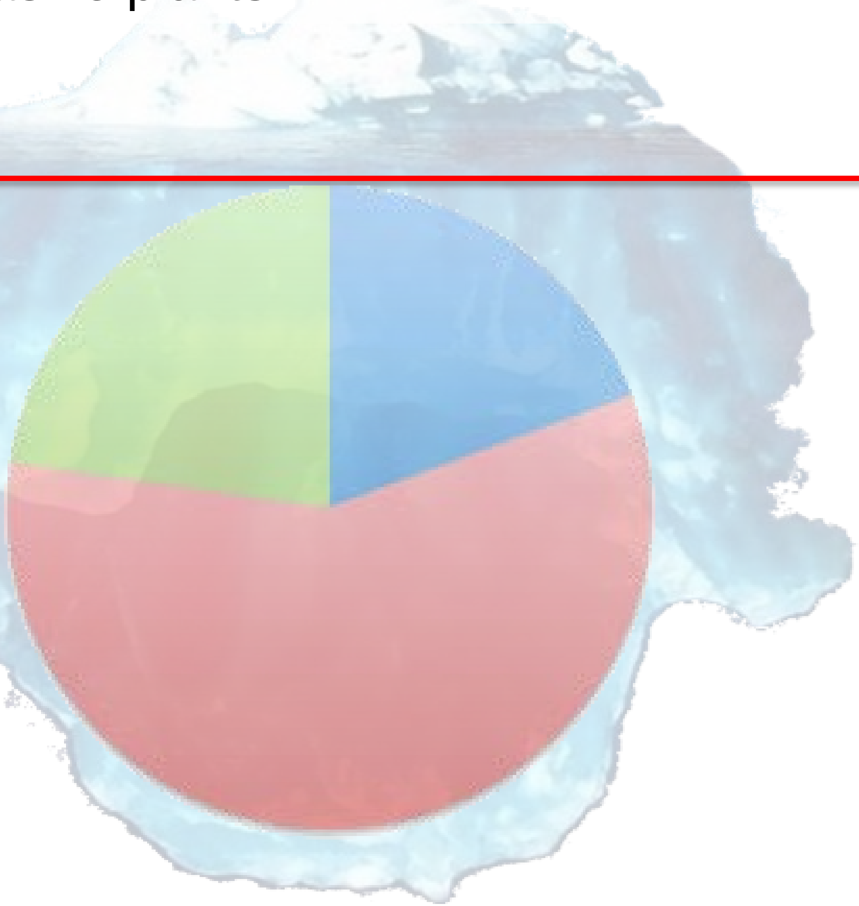
A lot of unlabeled data through the queries: 7M pictures generated!

How much the raw unlabeled data is valuable for the biodiversity ?

Ex: a valuable resource for biodiversity issues such as the ecological monitoring of invasive plants ?

% Biodiversity on known floras

% Biodiversity on other floras

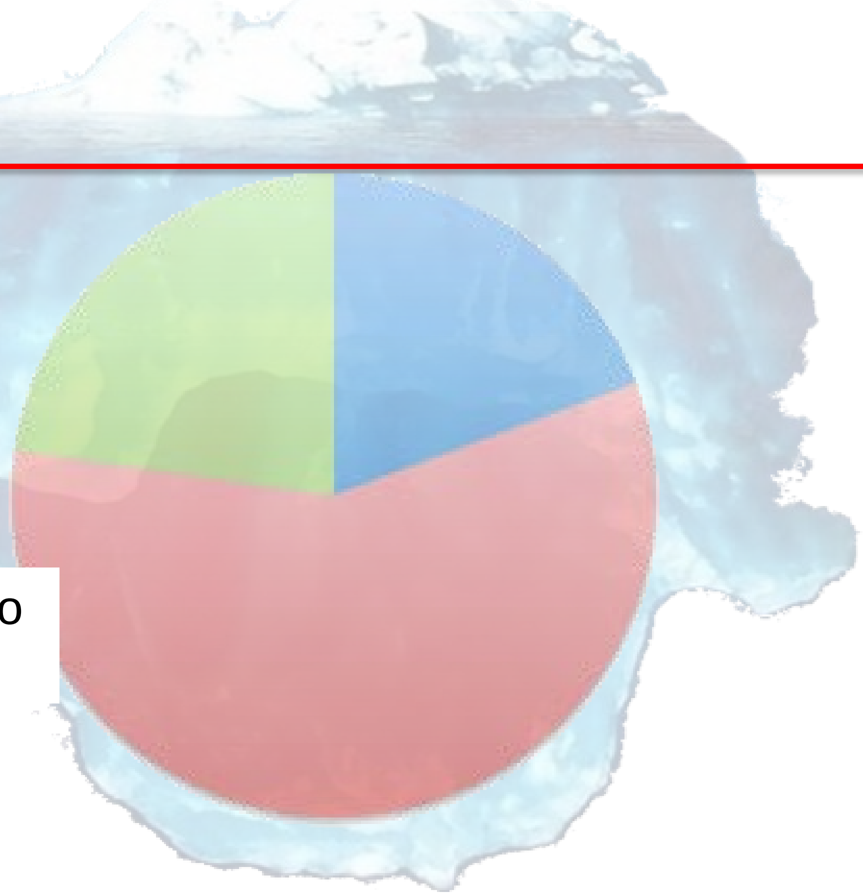


Pl@ntNet app: queries and shared observations for validation

A lot of unlabeled data through the queries: 7M pictures generated!

How much the raw unlabeled data is valuable for the biodiversity ?

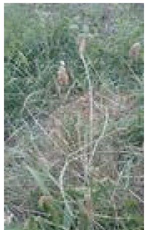
Ex: a valuable resource for biodiversity issues such as the ecological monitoring of invasive plants ?



% Biodiversity on known floras

% Biodiversity on other floras

% wild plants but impossible to identify at species level



Pl@ntNet app: queries and shared observations for validation

A lot of unlabeled data through the queries: 7M pictures generated!

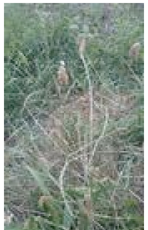
How much the raw unlabeled data is valuable for the biodiversity ?

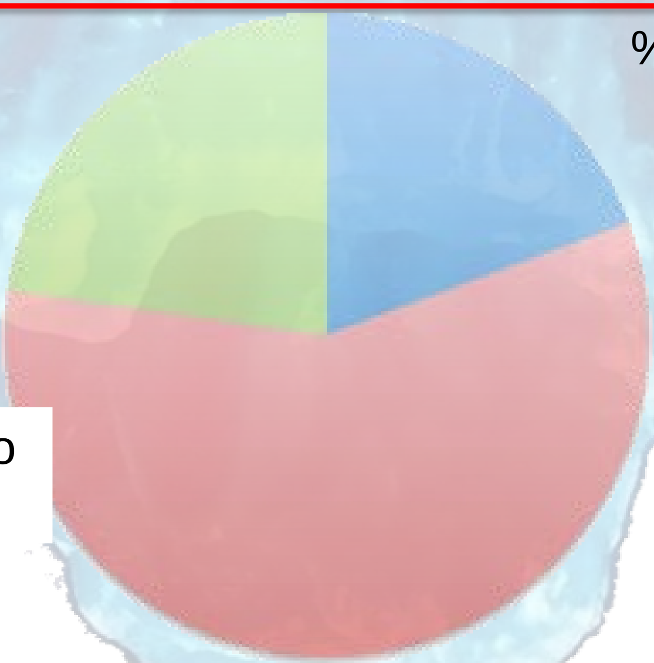
Ex: a valuable resource for biodiversity issues such as the ecological monitoring of invasive plants ?

% Biodiversity on known floras

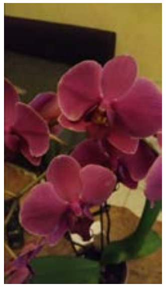
% Biodiversity on other floras

% wild plants but impossible to identify at species level





% Garden and house plants?

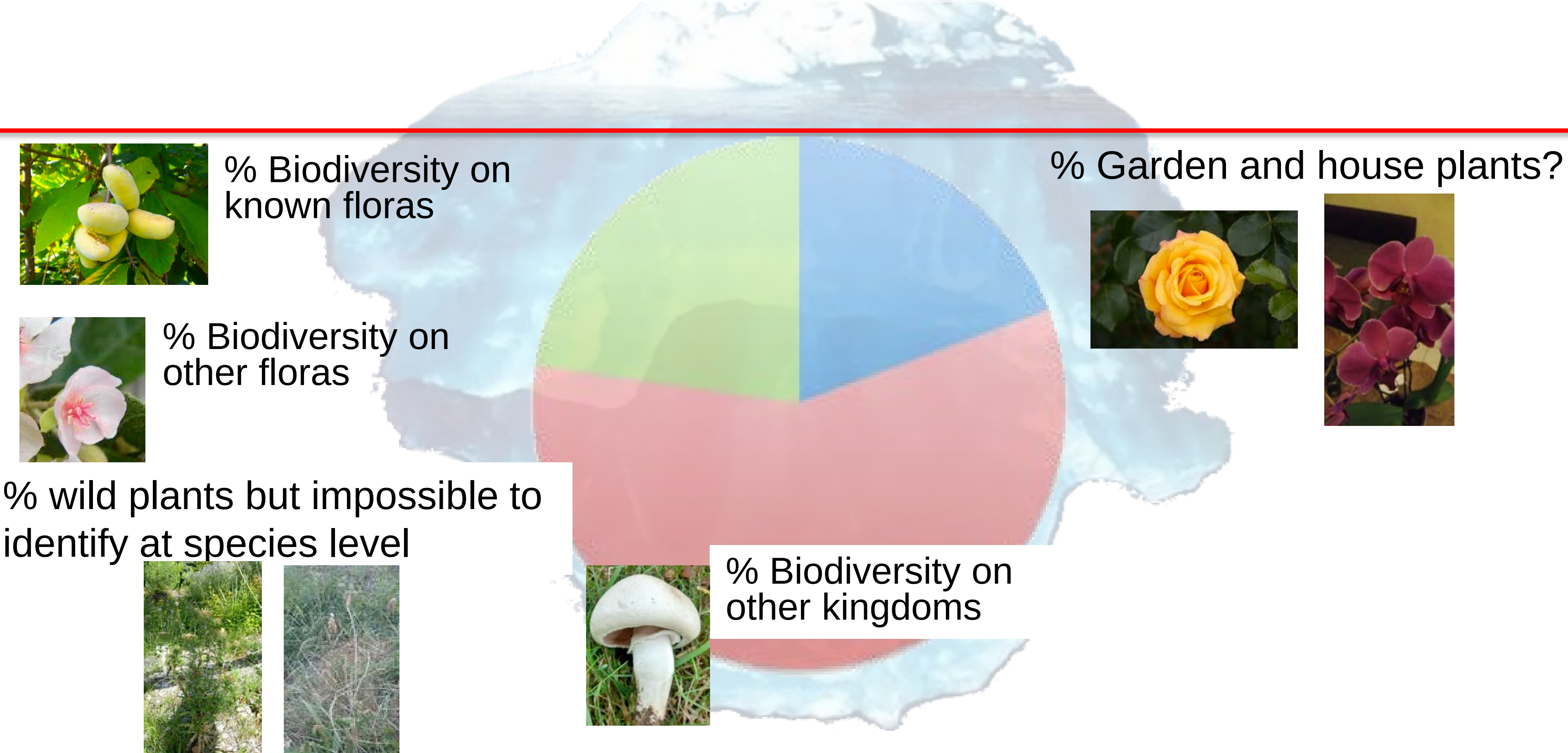


Pl@ntNet app: queries and shared observations for validation

A lot of unlabeled data through the queries: 7M pictures generated!

How much the raw unlabeled data is valuable for the biodiversity ?

Ex: a valuable resource for biodiversity issues such as the ecological monitoring of invasive plants ?



% Biodiversity on known floras

% Garden and house plants?

% Biodiversity on other floras

% wild plants but impossible to identify at species level

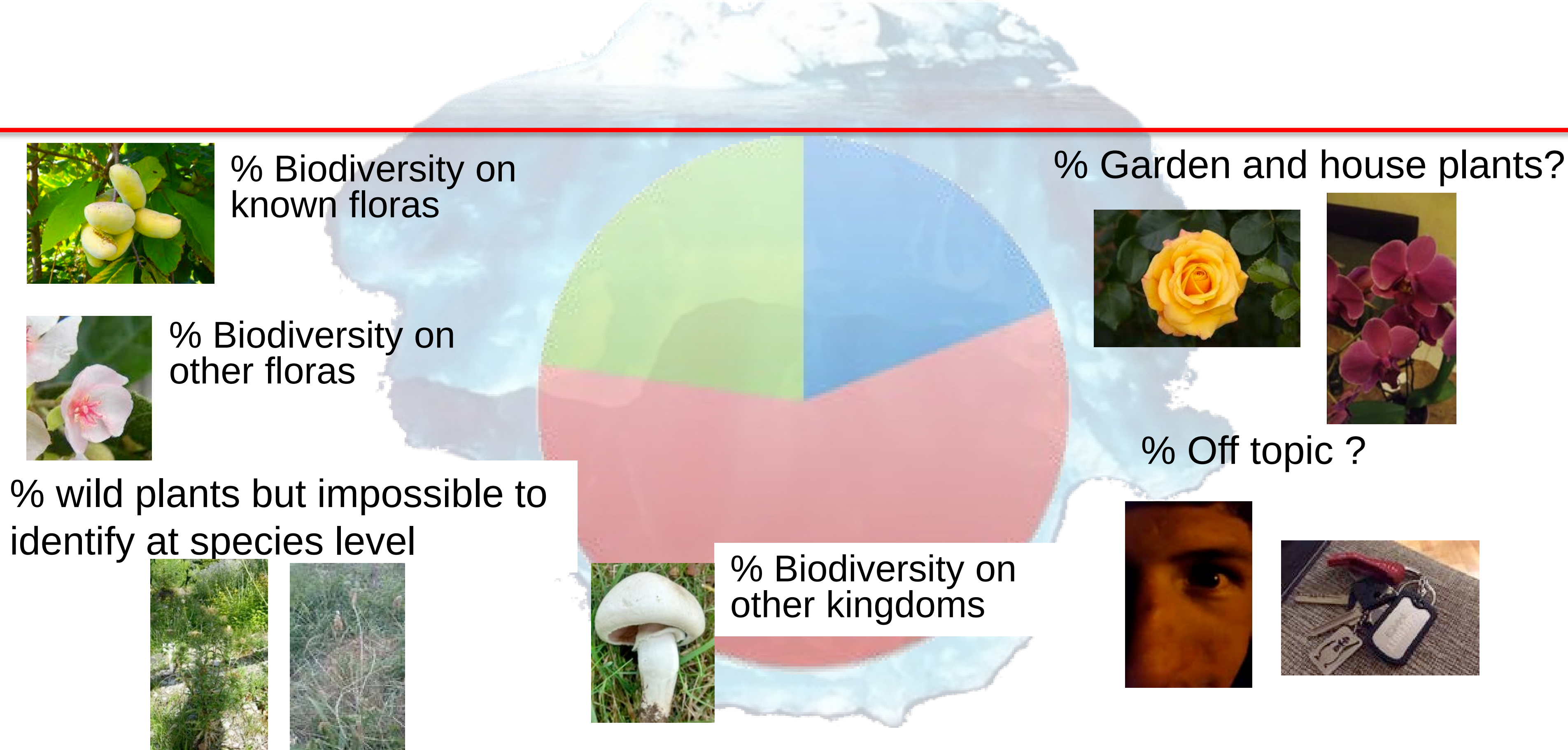
% Biodiversity on other kingdoms

Pl@ntNet app: queries and shared observations for validation

A lot of unlabeled data through the queries: 7M pictures generated!

How much the raw unlabeled data is valuable for the biodiversity ?

Ex: a valuable resource for biodiversity issues such as the ecological monitoring of invasive plants ?



% Biodiversity on known floras

% Biodiversity on other floras

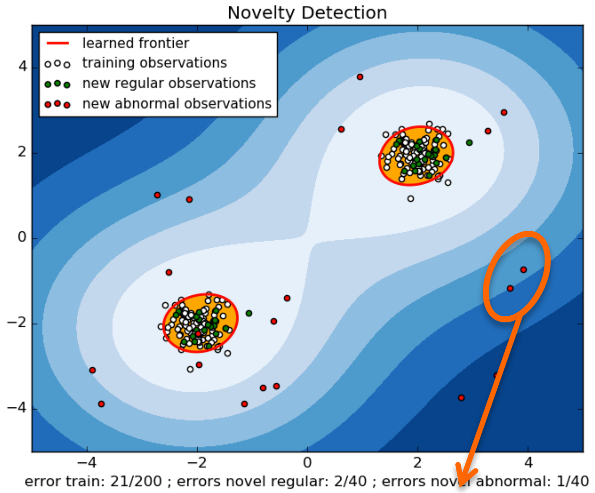
% wild plants but impossible to identify at species level

% Biodiversity on other kingdoms

% Garden and house plants?

% Off topic ?

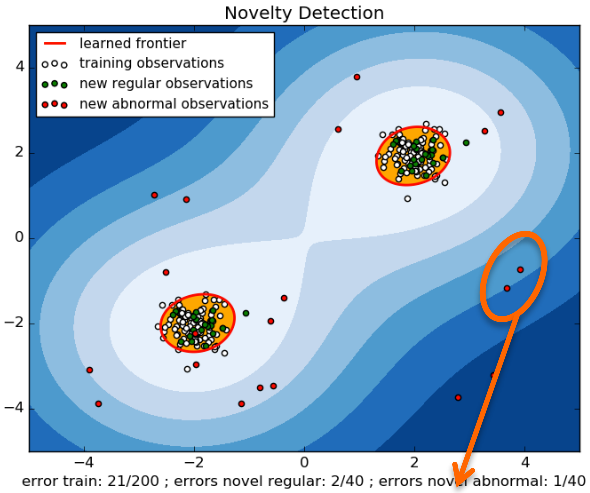
Plant identification in an open-world



New class ?

Plant identification in an open-world

Instead of proposing a wrong plant species
Be able to detect a unknown class
And give an appropriate response to the user

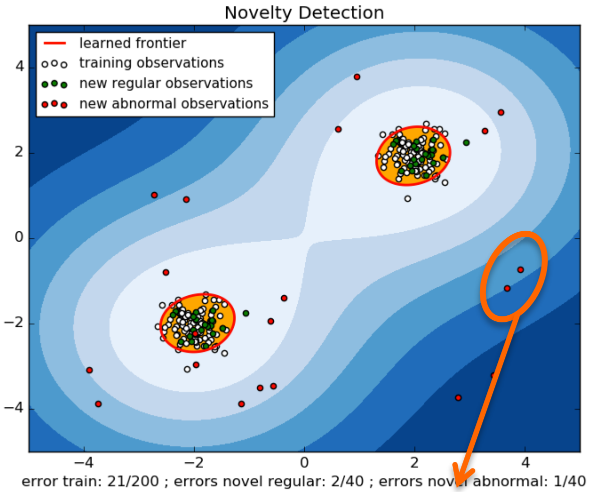


New class ?

Plant identification in an open-world

Instead of proposing a wrong plant species
Be able to detect a unknown class
And give an appropriate response to the user

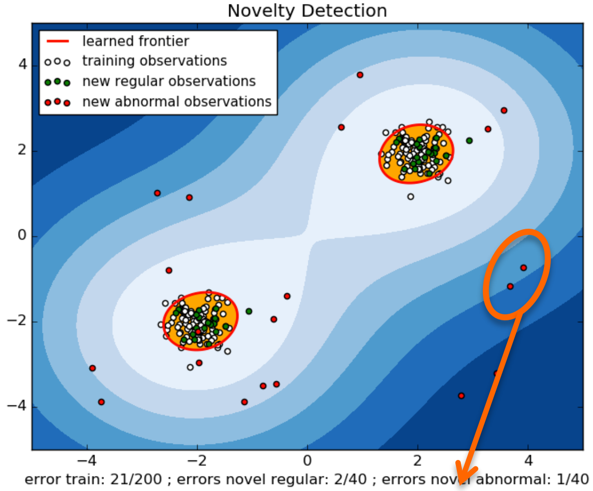
« Not related to the Western Europe flora (try with the Indian Ocean flora.) »



New class ?

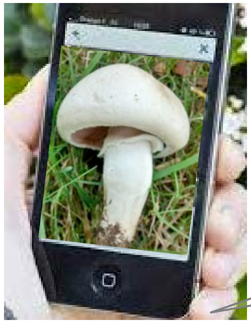
Plant identification in an open-world

Instead of proposing a wrong plant species
Be able to detect a unknown class
And give an appropriate response to the user



New class ?

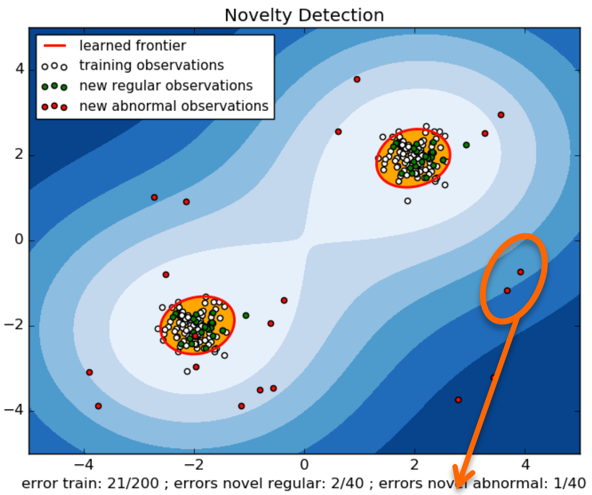
« Not related to the Western Europe flora
(try with the Indian Ocean flora.) »



« Not related to the Western
Europe flora
(it is a mushroom.) »

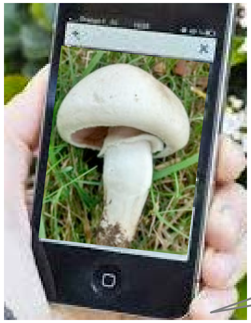
Plant identification in an open-world

Instead of proposing a wrong plant species
Be able to detect a unknown class
And give an appropriate response to the user



New class ?

« *Not related to the Western Europe flora (try with the Indian Ocean flora.)* »



« *Not related to the Western Europe flora (it is a mushroom.)* »

« *Not related to the Western Europe flora (it is a Homo sapiens L.)* »

Metric: Mean Average Precision

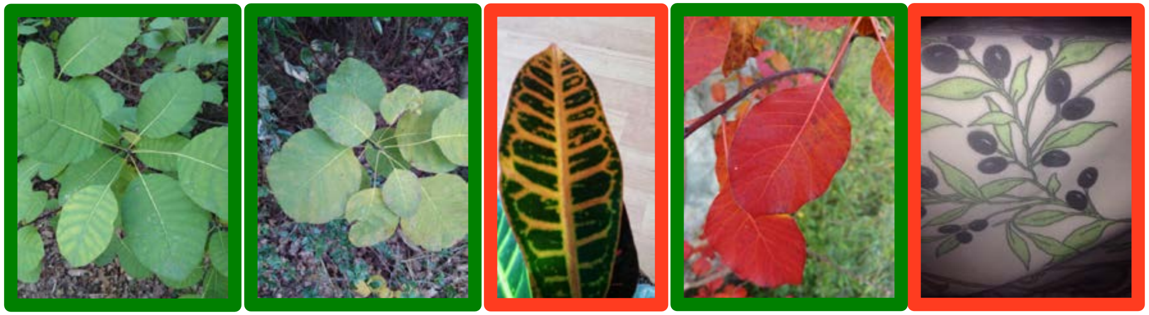
Each known species is seen as a query

Test images are ranked by descending probabilities on this species if it appear as a proposition in the run file

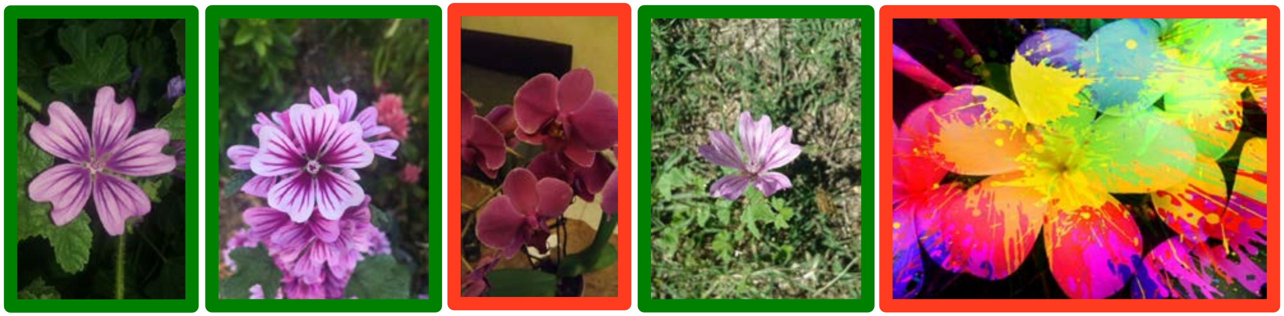
Compute the Av. P.

$$\mathrm{AveP} = \frac{\sum_{k=1}^{n}(P(k) \times \mathrm{rel}(k))}{\text{number of relevant images}}$$

Cotinus coggygria ?



Malva sylvestris ?



Irrelevant test images can be unfortunately predicted as a plant species, degrading thus the Average Precision

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

Training set

= LifeCLEF Plant Identification task 2015
- training set
- test set with the groundtruth

1 000 sp.

113 204 pictures

Collected and shared by thousands of contributors involved in various citizen projects

Tela Botanica

Different organs & views

Various metadata

USEFULNESS
« Quality »

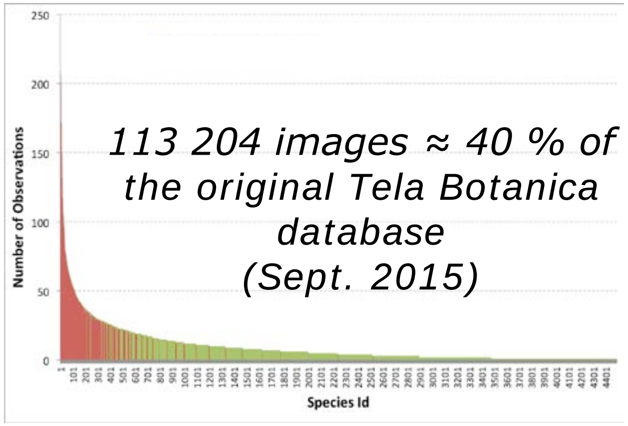
AUTHOR

DATE

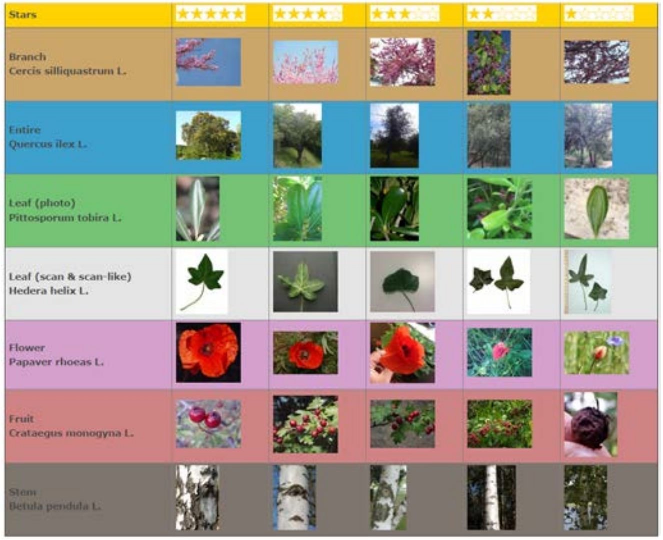
LOCALIZATION

TAXONOMY

ORGAN TAG

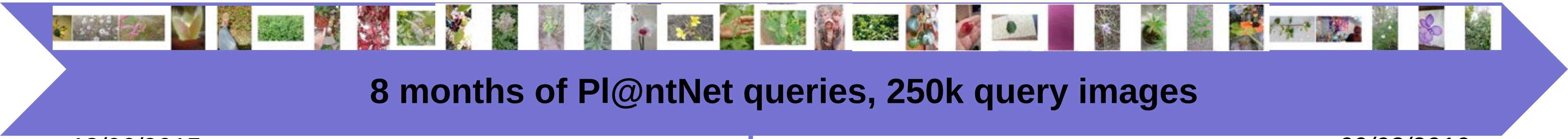


113 204 images ≈ 40 % of the original Tela Botanica database (Sept. 2015)



Types of views	**Pictures**
Branch	10 218
Entire plant	22 348
Flower	36 552
Fruit	9 143
Leaf	16 057
Stem	6 060
Scans of leaf	12826
	113 204

Test dataset creation



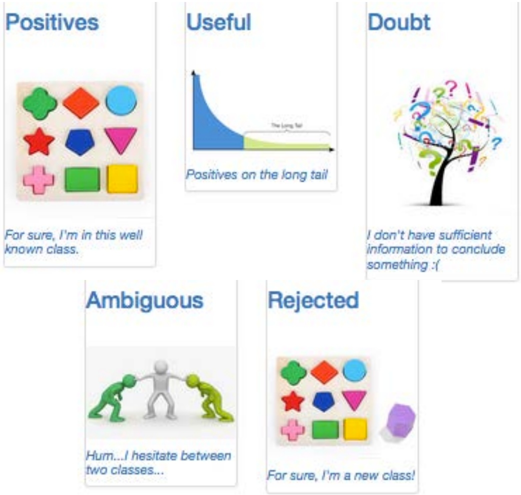
8 months of Pl@ntNet queries, 250k query images

18/06/2015

09/03/2016

30k query images from authenticated users

Interactive navigation and annotation



Positives

For sure, I'm in this well known class.

Useful

Positives on the long tail

Doubt

I don't have sufficient information to conclude something :(

Ambiguous

Hum...I hesitate between two classes...

Rejected

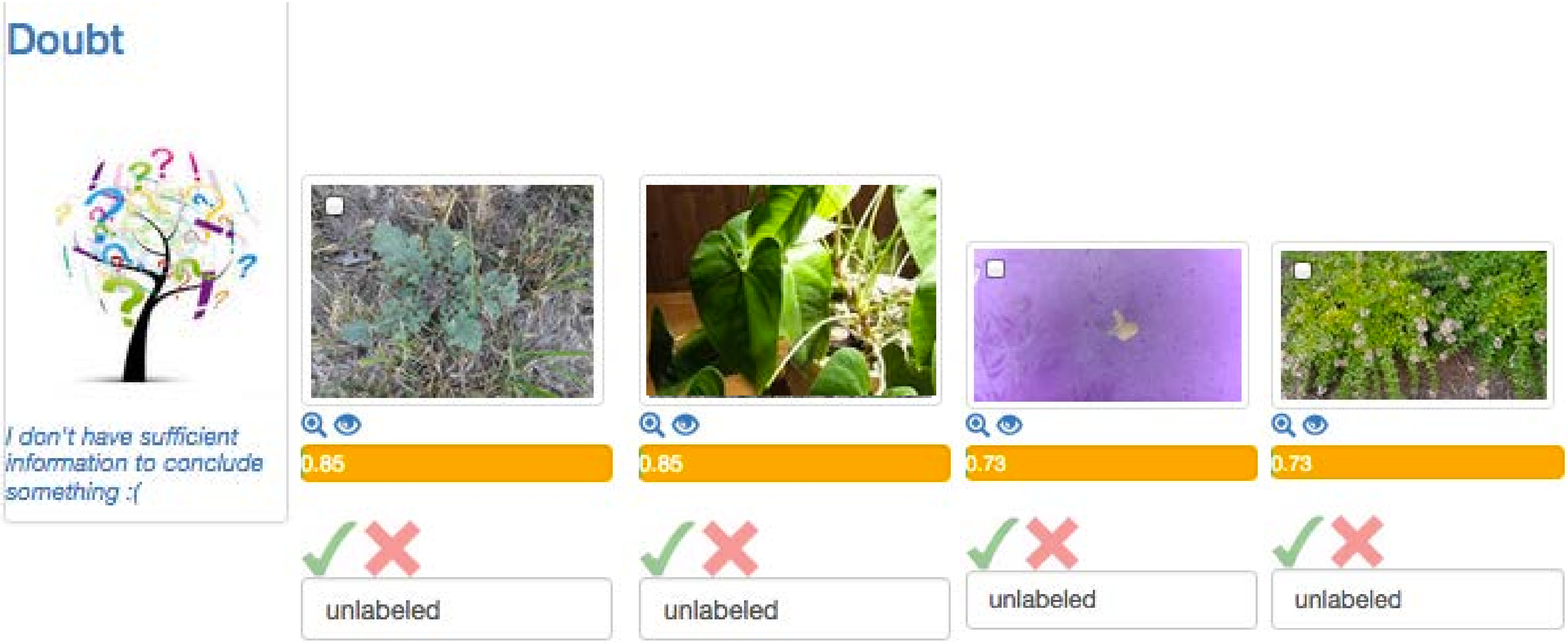
For sure, I'm a new class!

Based on a classifier (ConvNet) continuously updating the label predictions on still unlabeled data

Learn progressively new classes

Test dataset creation

Interactive navigation and annotation



Doubt

I don't have sufficient information to conclude something :(

0.85 ✓ ✗ unlabeled

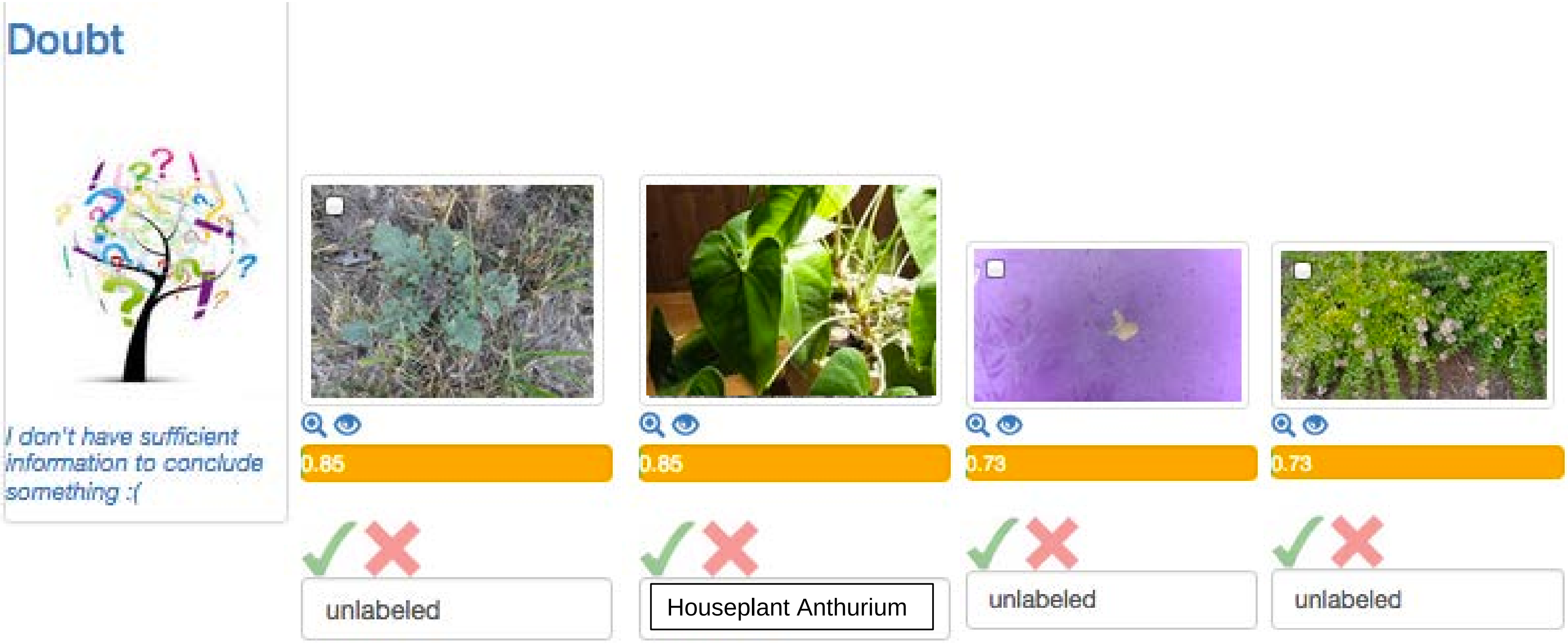
0.85 ✓ ✗ unlabeled

0.73 ✓ ✗ unlabeled

0.73 ✓ ✗ unlabeled

Test dataset creation

Interactive navigation and annotation



Doubt

I don't have sufficient information to conclude something :(

0.85 — unlabeled

0.85 — Houseplant Anthurium

0.73 — unlabeled

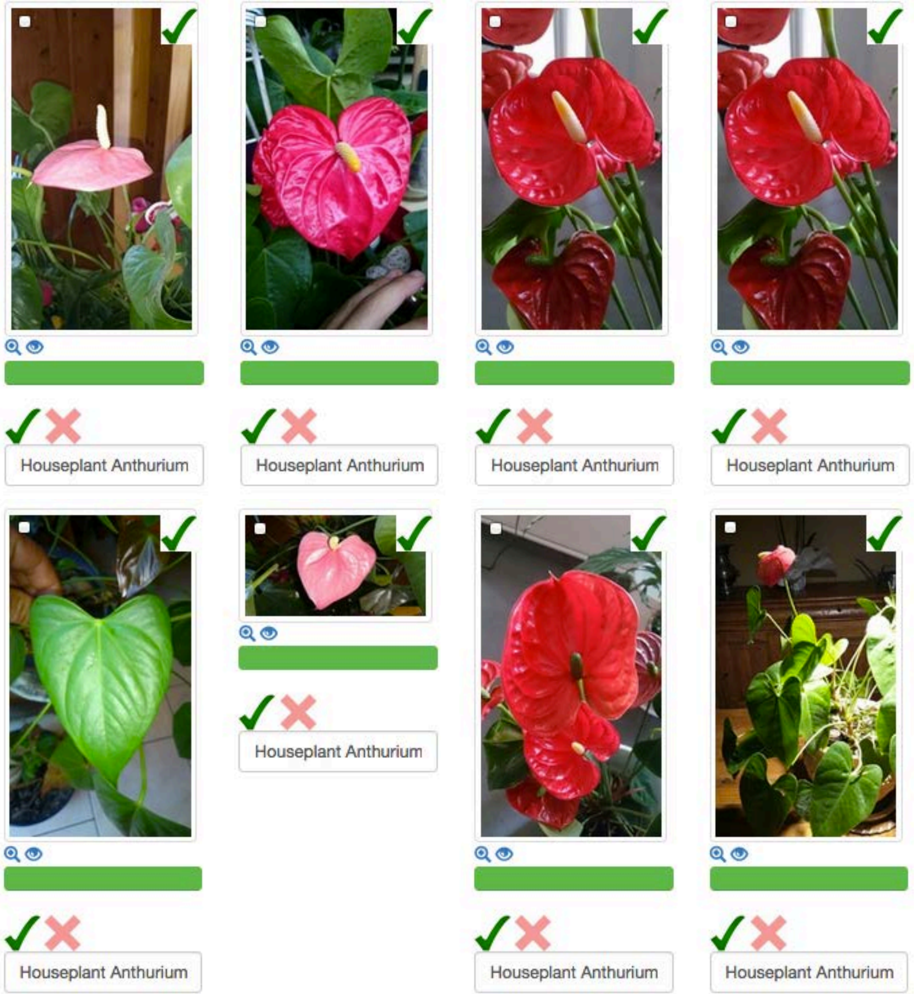
0.73 — unlabeled

Test dataset creation

Interactive navigation and annotation

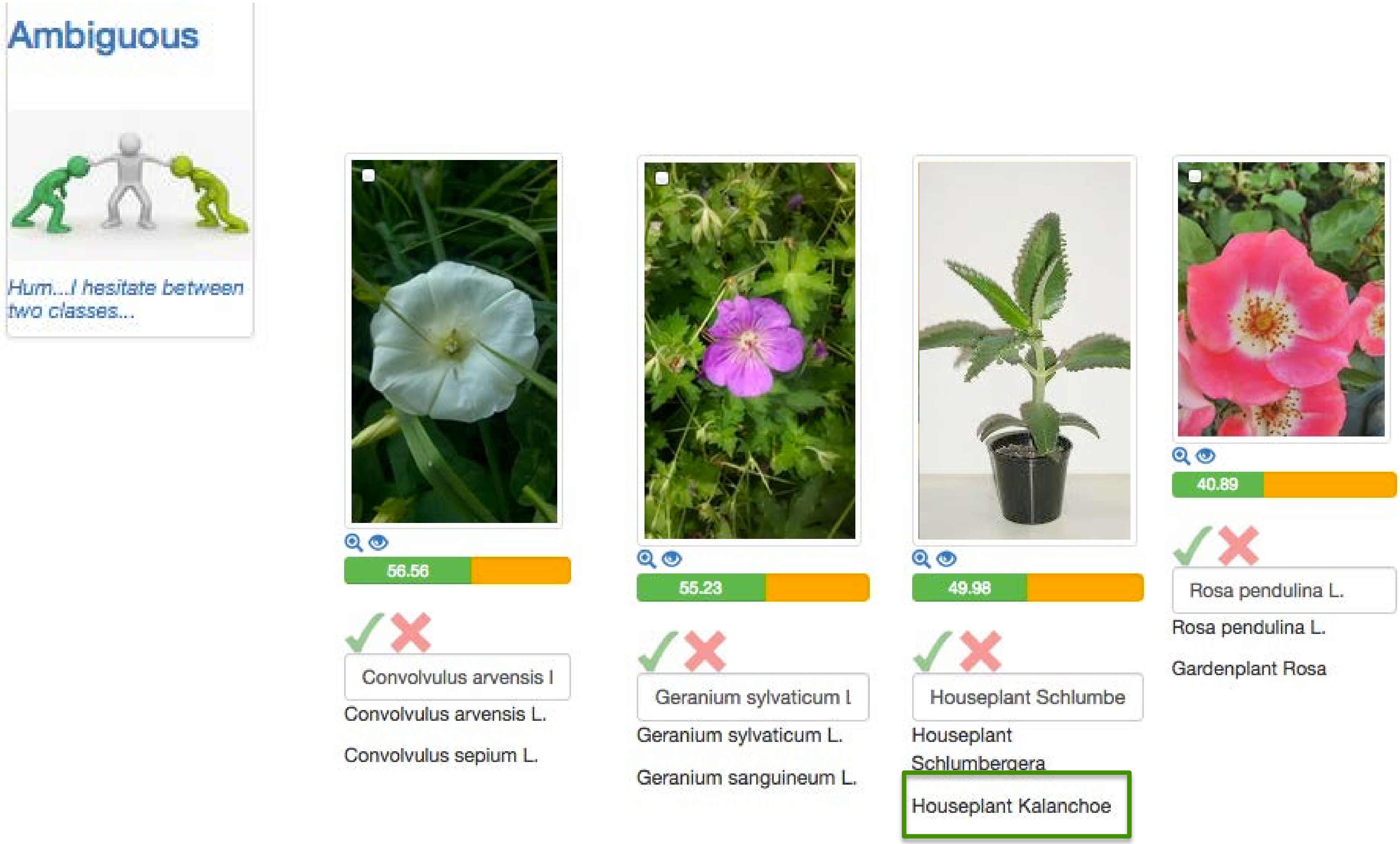
Positives

For sure, I'm in this well known class.



Test dataset creation

Interactive navigation and annotation



Ambiguous

Hum...I hesitate between two classes...

56.56

✓ ✗

Convolvulus arvensis l

Convolvulus arvensis L.

Convolvulus sepium L.

55.23

✓ ✗

Geranium sylvaticum l

Geranium sylvaticum L.

Geranium sanguineum L.

49.98

✓ ✗

Houseplant Schlumbe

Houseplant Schlumbergera

Houseplant Kalanchoe

40.89

✓ ✗

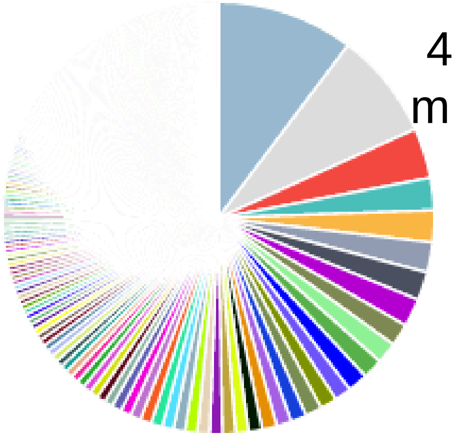
Rosa pendulina L.

Rosa pendulina L.

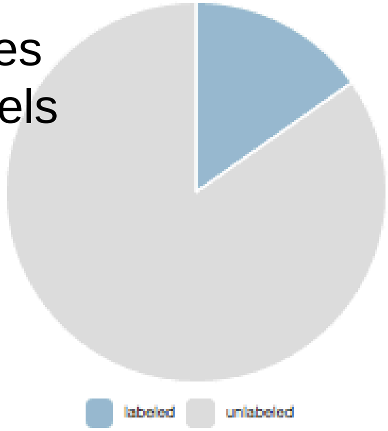
Gardenplant Rosa

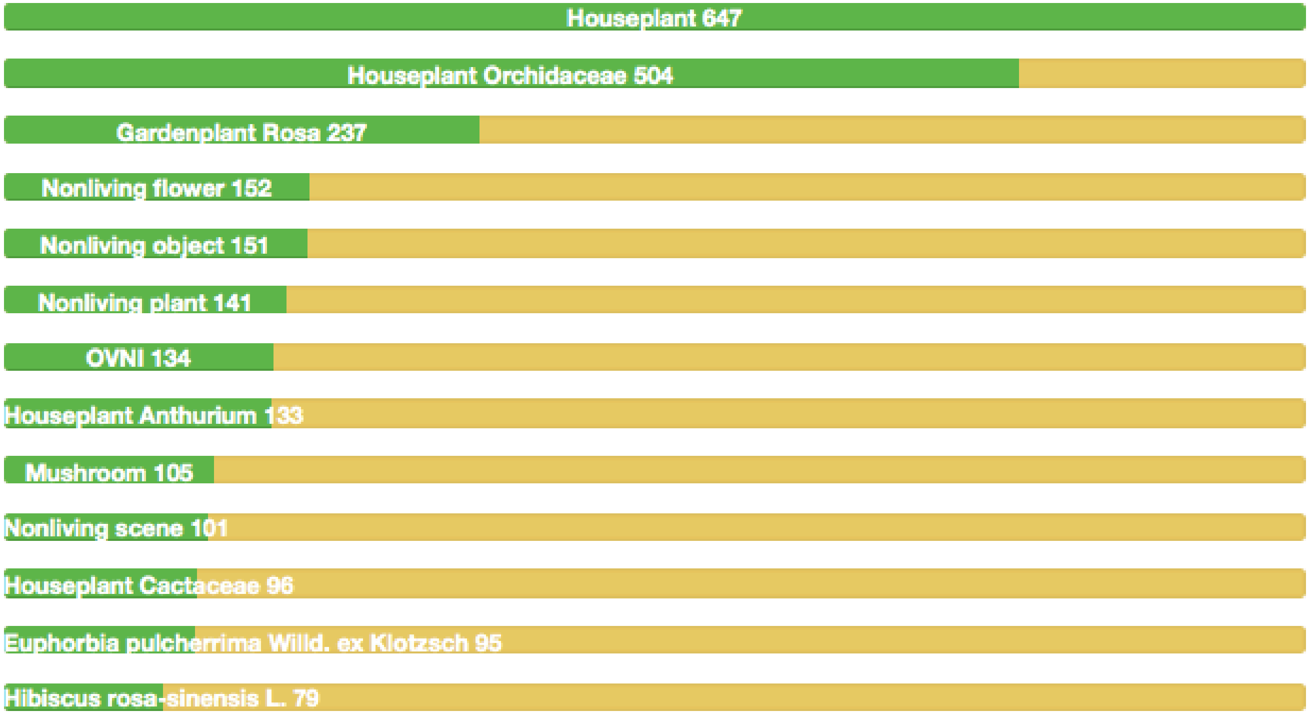
Test dataset creation

Interactive navigation and annotation

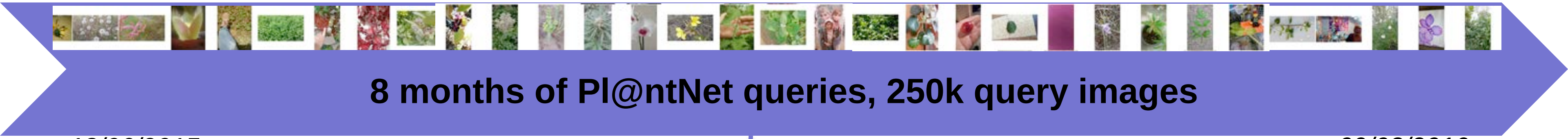


4k annotated images
mainly with new labels





Test dataset creation



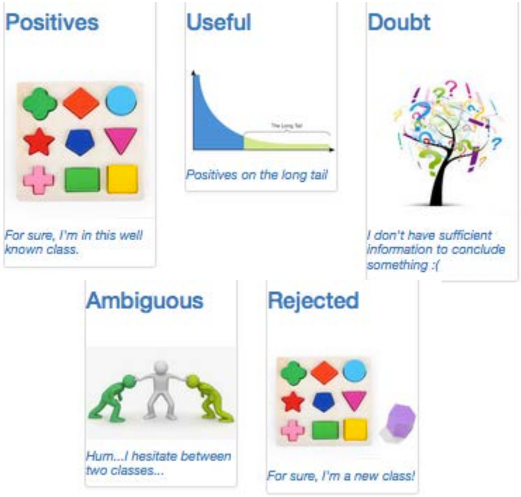
8 months of Pl@ntNet queries, 250k query images

18/06/2015

09/03/2016

30k query images from authenticated users

Interactive navigation and annotation



Positives

For sure, I'm in this well known class.

Useful

Positives on the long tail

Doubt

I don't have sufficient information to conclude something :(

Ambiguous

Hum...I hesitate between two classes...

Rejected

For sure, I'm a new class!

4k annotated images,
2k removed (near duplicates &
over-represented classes

1821 pictures related
to off-topic pictures
or cultivated plants

Test dataset creation

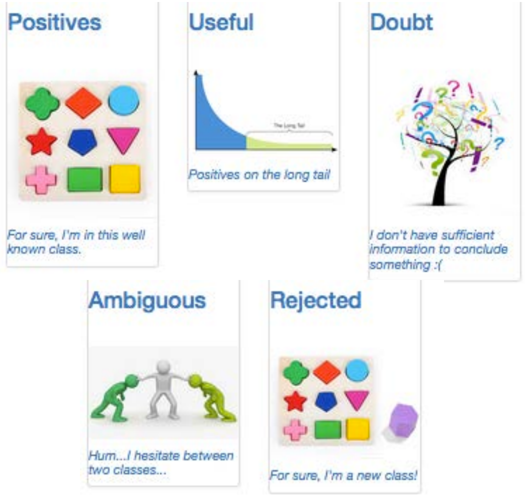
8 months of Pl@ntNet queries, 250k query images

18/06/2015

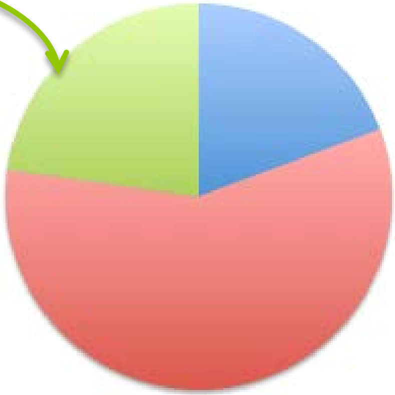
09/03/2016

30k query images from authenticated users

Interactive navigation and annotation


Positives
For sure, I'm in this well known class.
Useful
The Long Tail
Positives on the long tail
Doubt
I don't have sufficient information to conclude something :(
Ambiguous
Hum...I hesitate between two classes...
Rejected
For sure, I'm a new class!

4k annotated images,
2k removed (near duplicates &
over-represented classes



1821 pictures related to off-topic pictures or cultivated plants

4633 (collaboratively revised) pictures related to known species

Test dataset creation

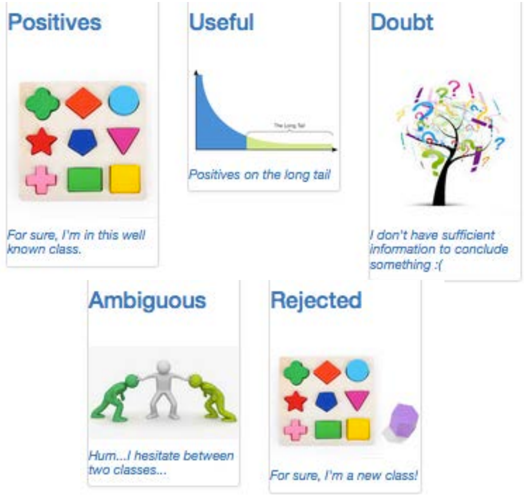
8 months of Pl@ntNet queries, 250k query images

18/06/2015

09/03/2016

30k query images from authenticated users

Interactive navigation and annotation



Positives

For sure, I'm in this well known class.

Useful

Positives on the long tail

Doubt

I don't have sufficient information to conclude something :(

Ambiguous

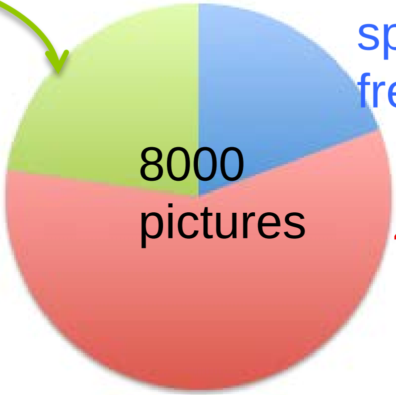
Hum...I hesitate between two classes...

Rejected

For sure, I'm a new class!

4k annotated images,
2k removed (near duplicates & over-represented classes

1821 pictures related to off-topic pictures or cultivated plants



1546 (collaboratively revised) pictures related to wild unknown plant species outside the french flora

4633 (collaboratively revised) pictures related to known species

Ratio = 58 % of known species in the test dataset

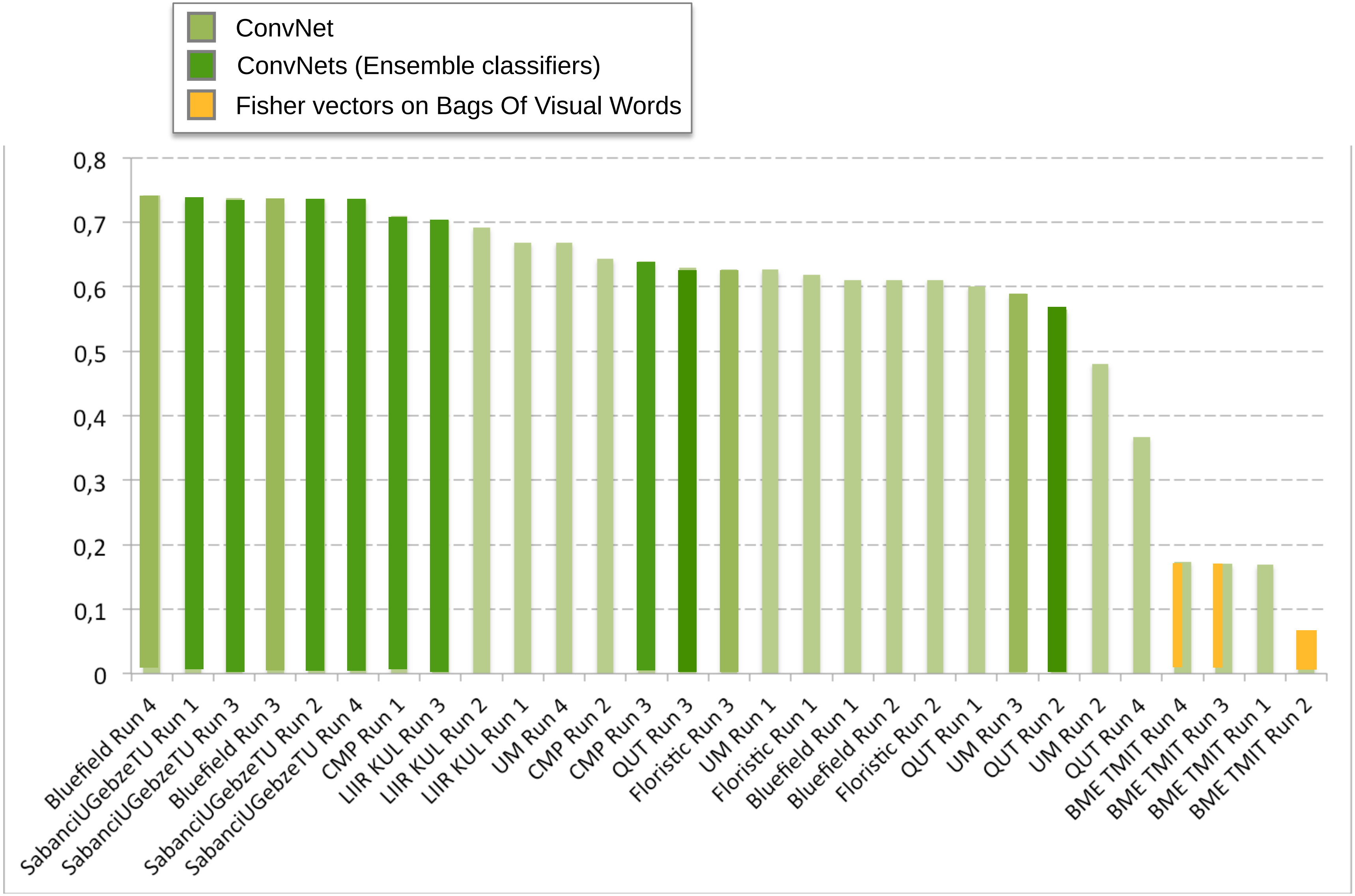
Participation and Methods

94 teams registered, including 34 teams exclusively registered to the plant task

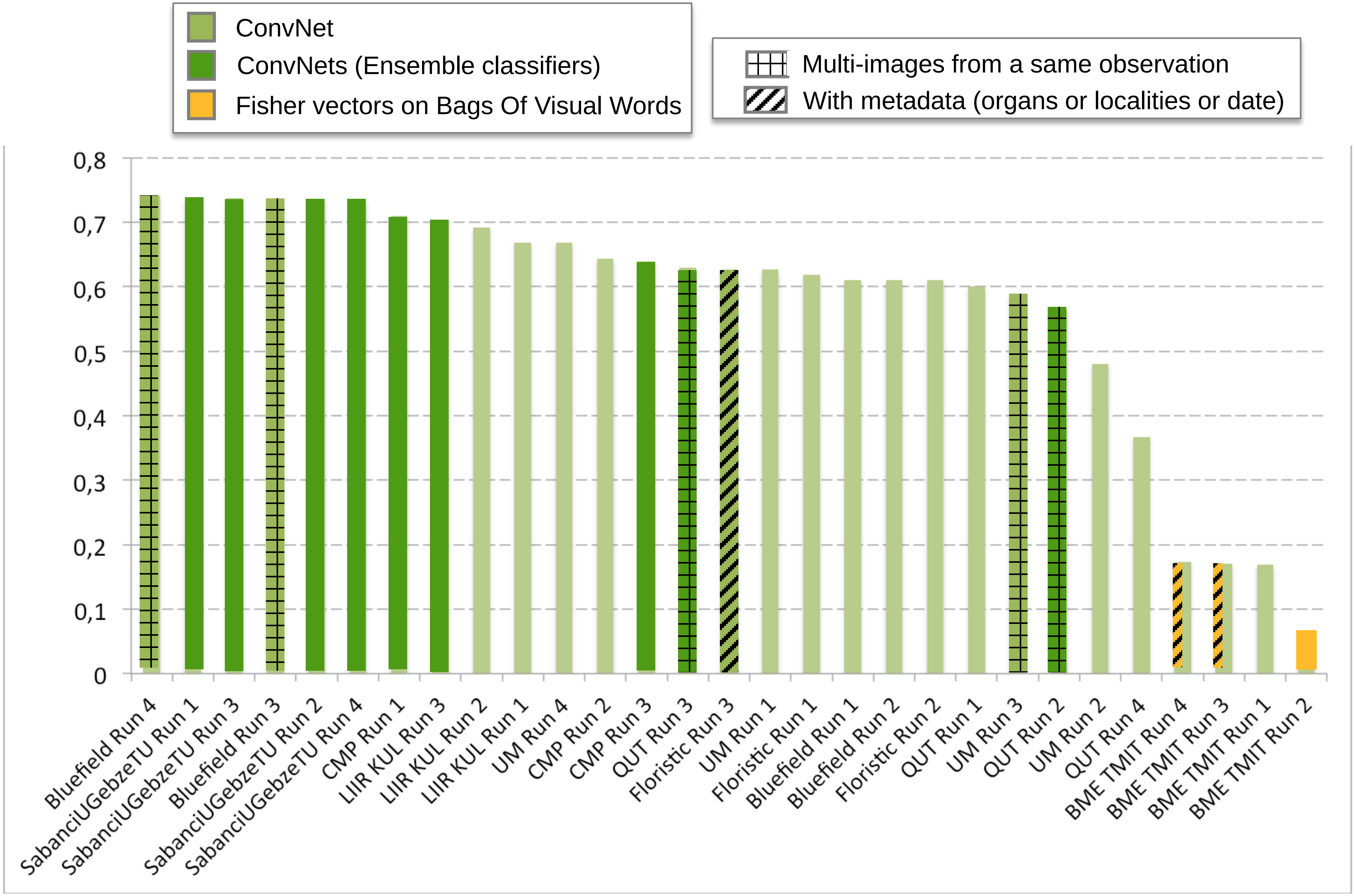
2011: 8 teams / 20 methods
2012: 10 teams / 30 methods
2013: 12 teams / 33 methods
2014: 10 teams / 28 methods
2015: 7 teams / 18 methods
2016: 8 teams / 29 methods

Team	Methods (key-words)	Rejection ?	BestMAP
Bluefield	VGGNet Multi-images by observation	Adaptive thresholds by species	0.742
BME TMIT	AlexNet & BVWs & Fisher vectors Metadata	Adaptive thresholds by species	0.174
CMP	Bagging of 3xResNet-152	_	0.710
Floristic	GoogleNet, metadata	Adaptive thresholds by species	0.627
LIIR KUL	CaffeNet, VGGNet16, 3xGoogleNet + external plant images	Global threshold	0.703
QUT	GoogleNet + 6xGoogleNet/organs Multi-images by observation	_	0.629
UM	VGGNet16, organ and species layers	_	0.627
Sabanci/ Gebze	VGGNet, GoogleNet	GoogleNet tuned on plants and imagenet no plants pictures	0.738

Results: Mean Average Precision



Results: Mean Average Precision



Results: MAP on the black list of potential invasive species

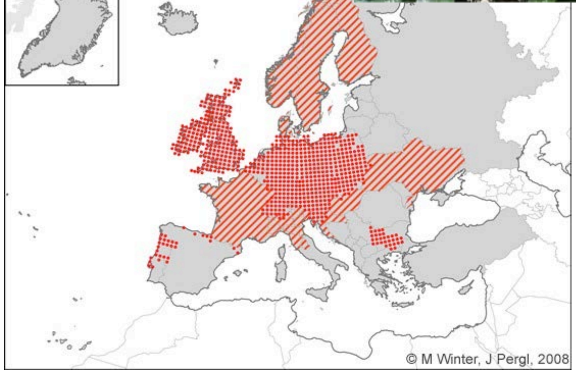
A valuable resource for biodiversity issues such as the ecological monitoring of invasive plants ?

Ludwigia grandiflora

Before…

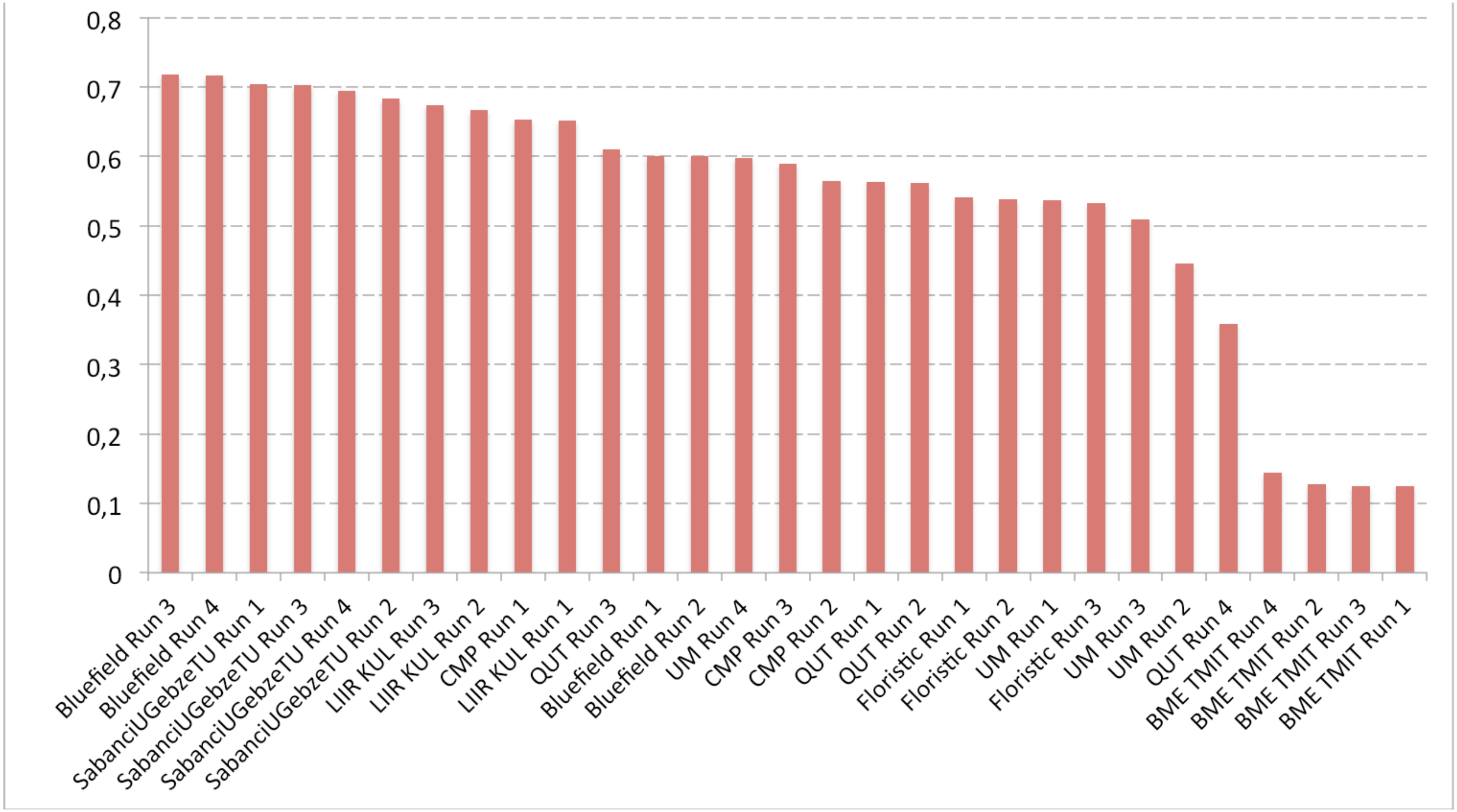
After…

Reynoutria japonica



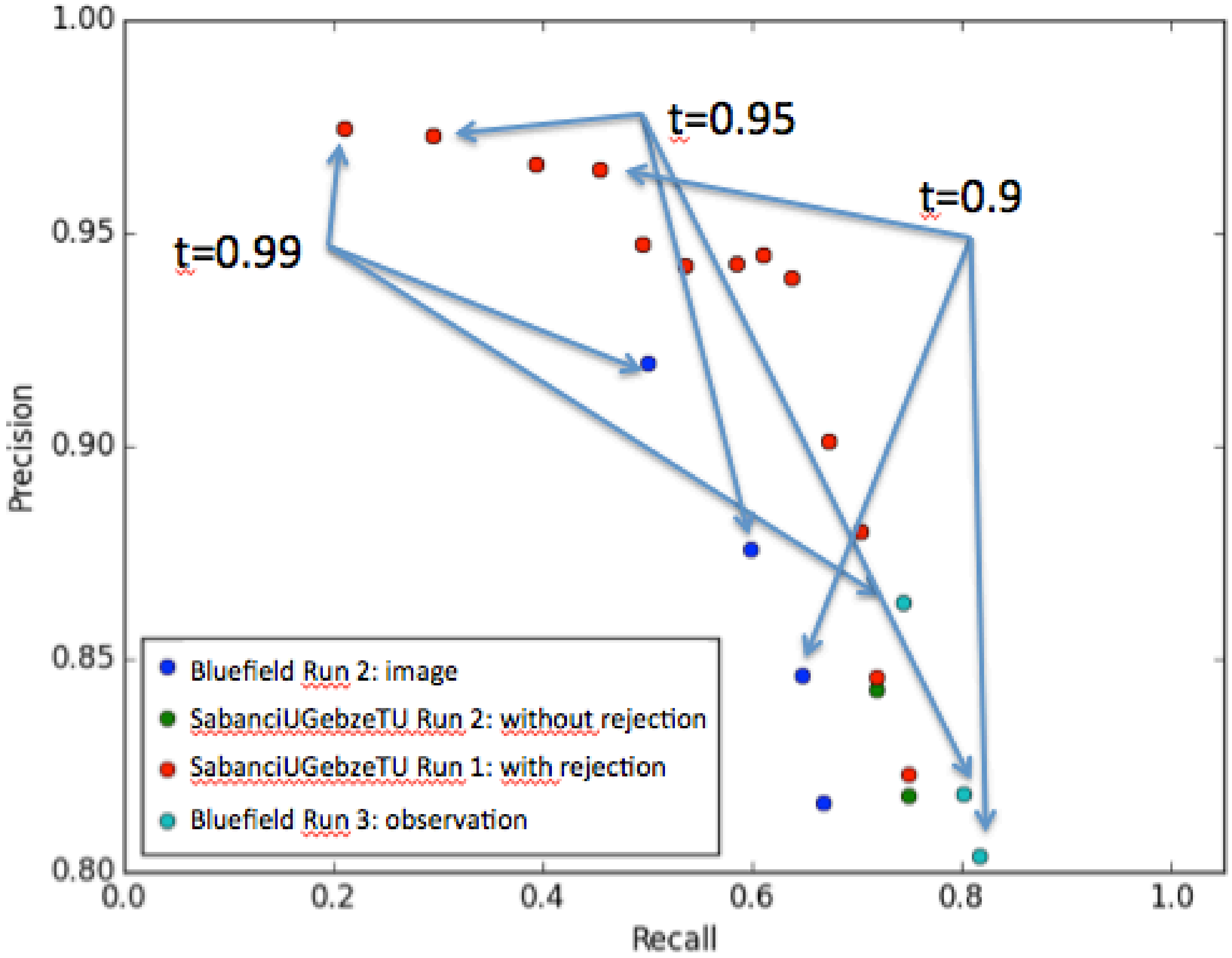
© M Winter, J Pergl, 2008

Results: MAP on the black list of potential invasive species



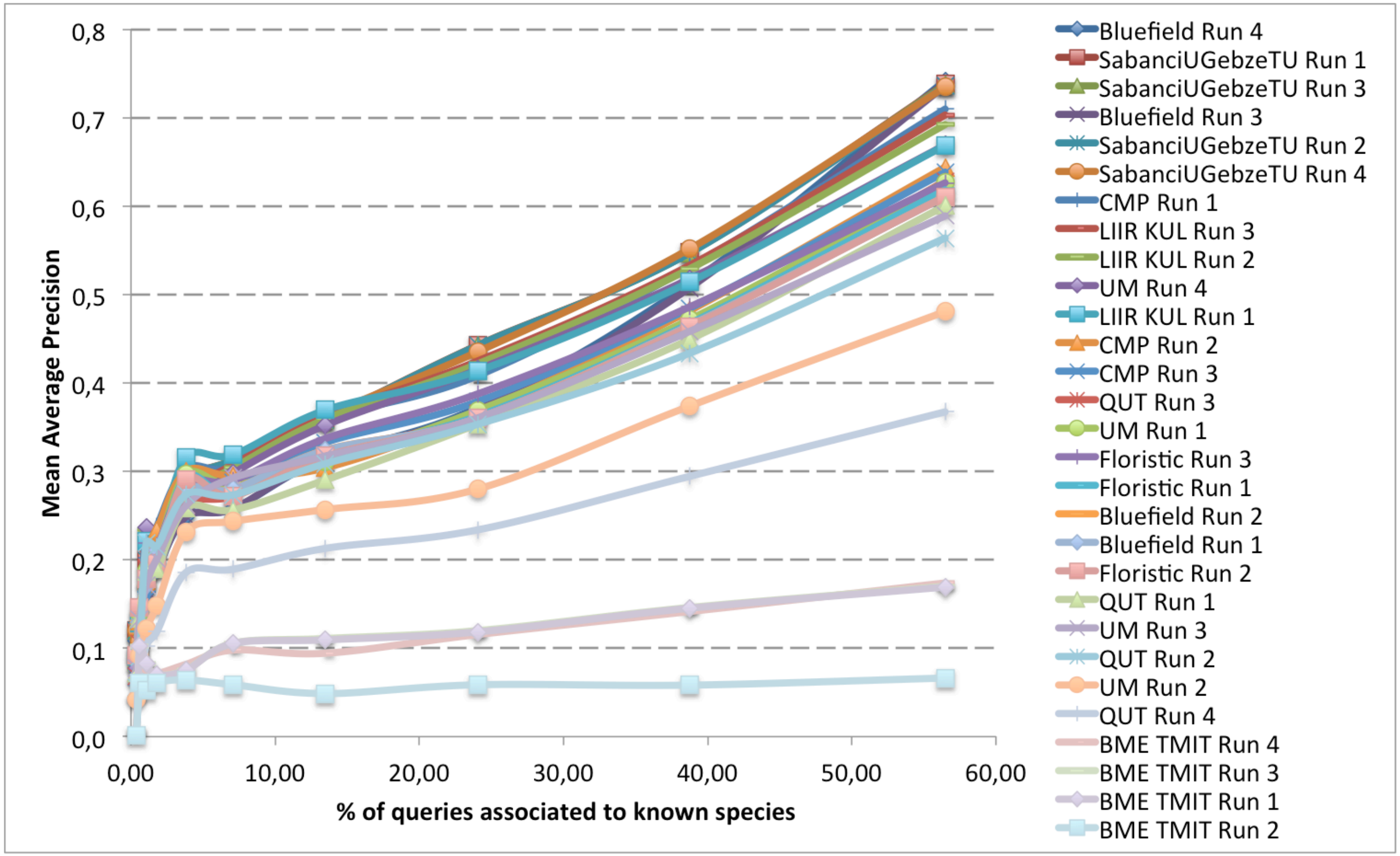
Results: Precision/Recall on the black list of potential invasive species

Ready to be shared with network of ecologists working on invasive species ?



A too easy task ?

How performances decrease when we reduce the number of test images related to known species



In order to be closer to a real stream of user queries

Conclusion

- Supremacy of ConvNet approaches, naturally robust to novelty until a certain level

- Multi-image combination from one test observation lead to better performances than single images

- Metadata under-exploited ? while it is so essential for a botanist...

JULY
28
USEFULNESS
DATE
LOCALIZATION
TAXONOMY

- Known & (obviously) Unknown species... But how to deal with the huge number of difficult pictures to identify?

Be ready for the next year Plant Identification Task ?

1000 x 10 species ?

more unknown species and off-topic pictures

Thank you!!