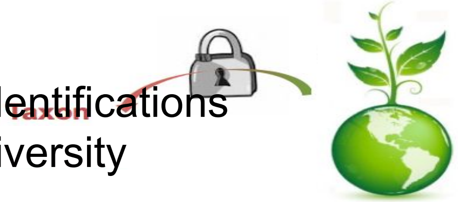
Pl@ntNet: towards the recognition of the world's flora

Alexis Joly *et al.*

Challenge

- More than **369K species** of flowering plants in the world
- Increasing our knowledge of them is of crucial importance
  - Health
  - Food crisis
  - Biodiversity crisis
- However, the **taxonomic gap** is penalizing the aggregation of new data and knowledge
  - Only specialists can identify plants
  - Specialists cannot carry the burden of all routine identifications
  - Particularly in south countries with the richest biodiversity



Pl@ntNet

An innovative **citizen science** platform making use of **machine learning** to help people **identify plants** through their mobile phone

Image Recognition Technology: Convolutional Neural Networks

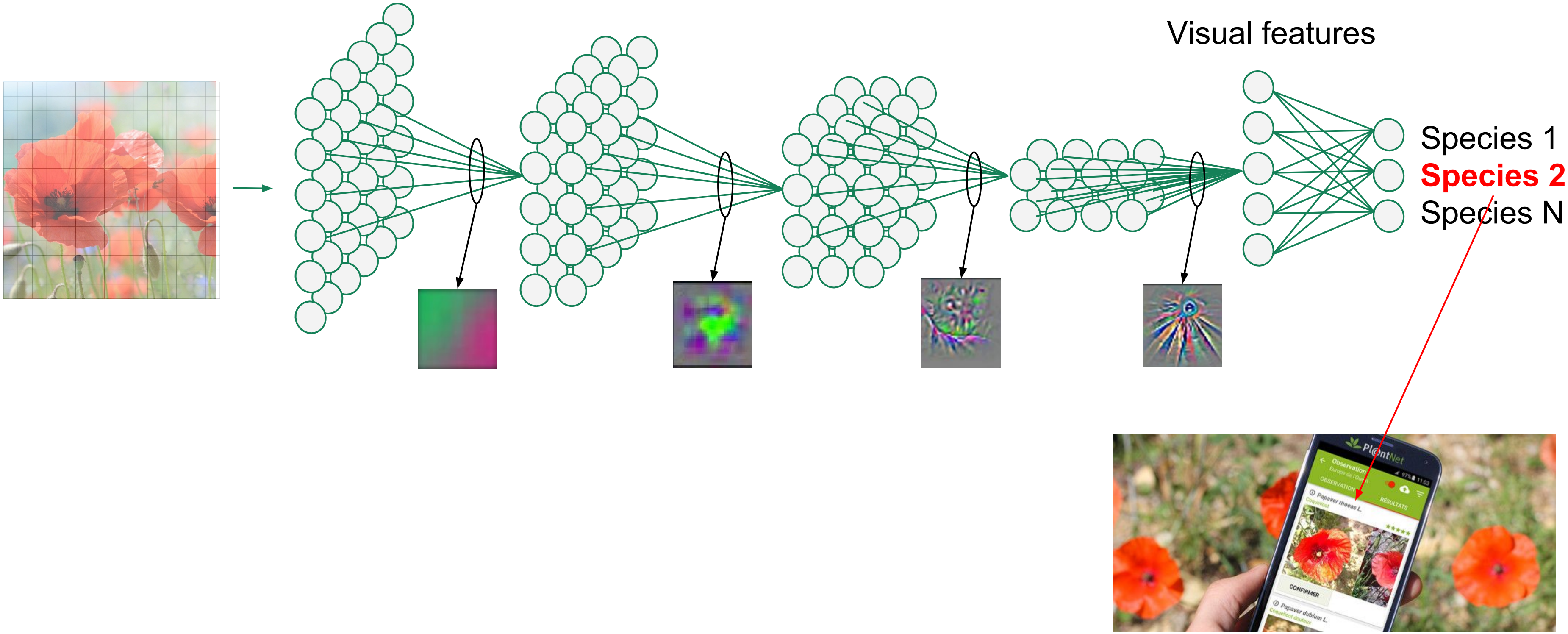
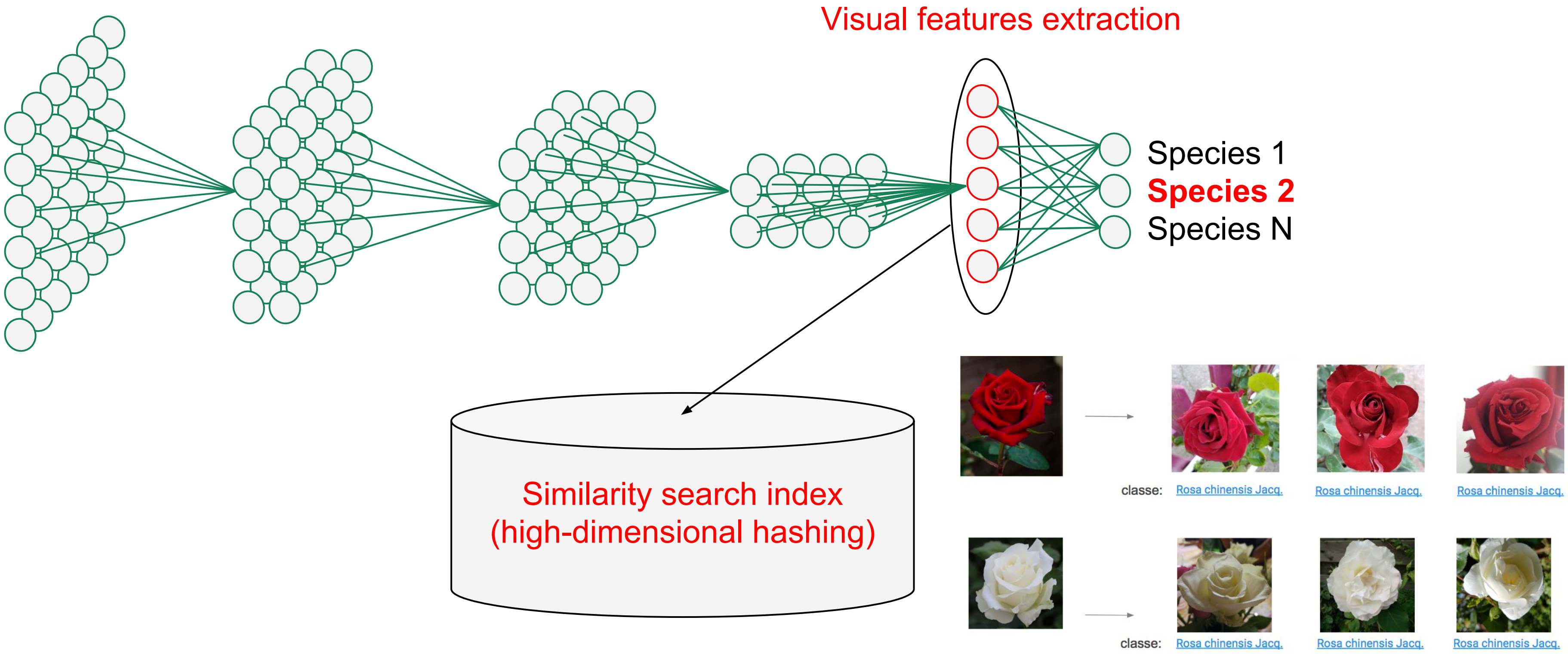
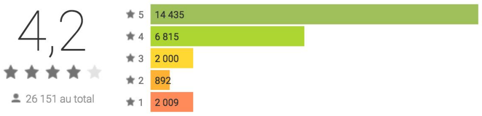
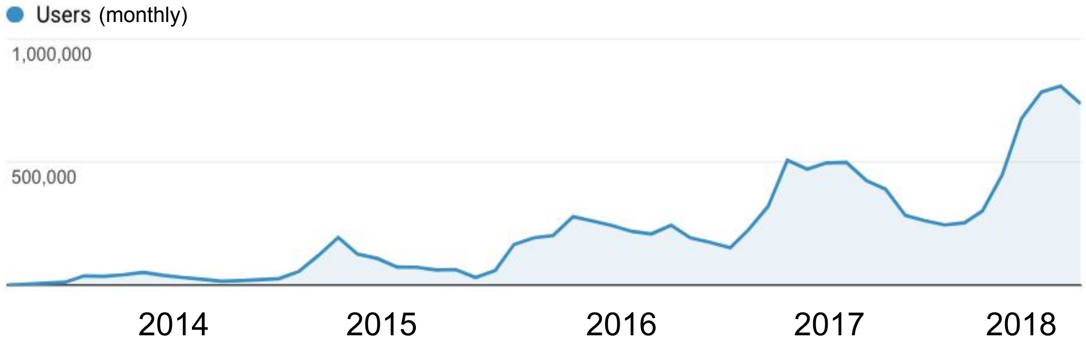


Image Recognition Technology: Similarity search



Pl@ntNet Statistics





In 2018 : 3,352,788 users in 235 countries			
	More than 5 sessions		1,866,423
	More than 10 sessions		1,293,698
	More than 25 sessions		735,666
	More than 100 sessions		96,167
1.	France	641,569	(19.19%)
2.	Germany	345,933	(10.35%)
3.	United States	345,880	(10.34%)
4.	Italy	282,842	(8.46%)
5.	Spain	180,291	(5.39%)
6.	Brazil	172,949	(5.17%)
7.	Netherlands	101,057	(3.02%)
8.	India	96,576	(2.89%)
9.	United Kingdom	86,670	(2.59%)
10.	Belgium	79,050	(2.36%)

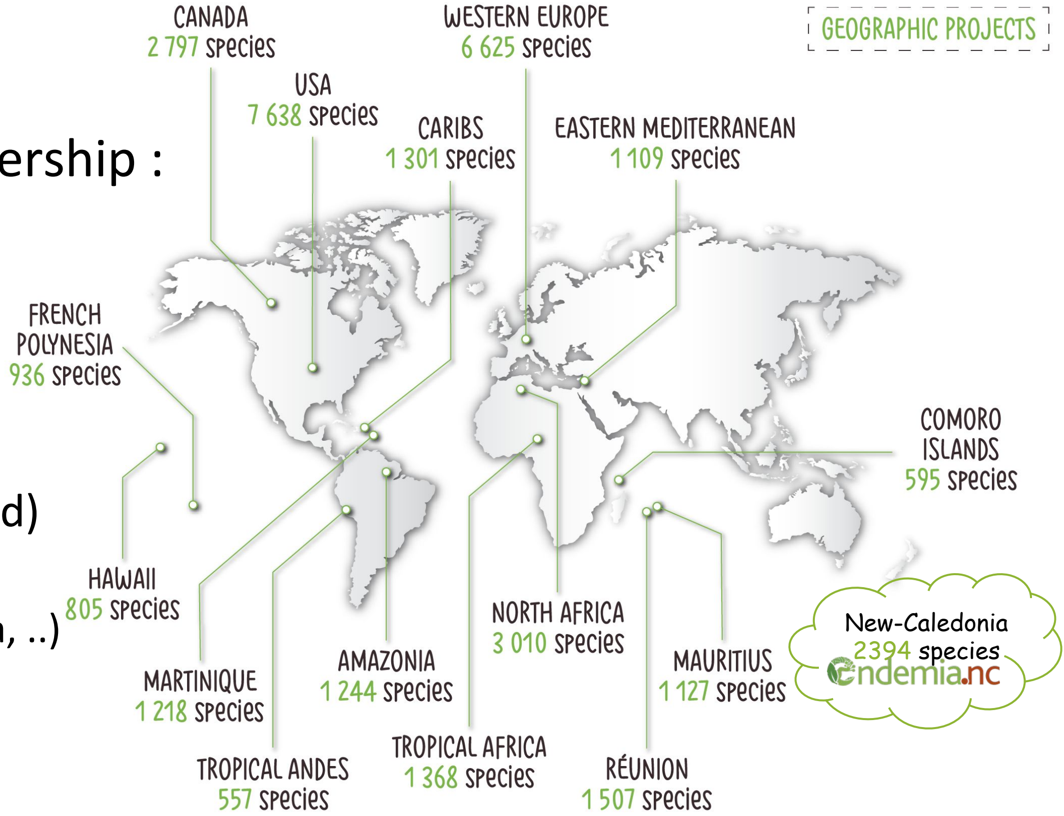
- More than **8M downloads**
- Between **60k - 100K users / day**
- **11 languages**
- **17K species** *(illustrated by 1M revised images)*
- **22** projects & micro-projects
- 35M raw plant images / 55M users sessions
- 12K followers on social networks

Pl@ntNet

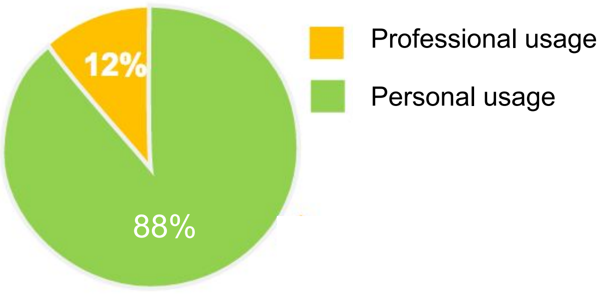
22 projects around the world

Based on a wide international partnership :

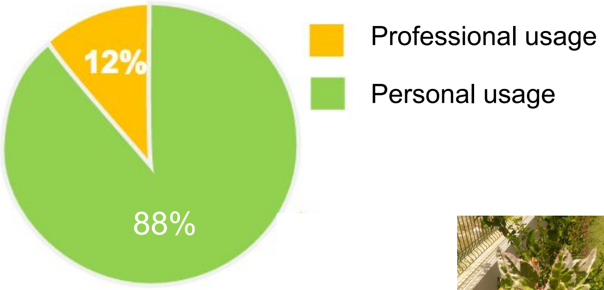
- Univ. TEC (Costa-Rica),
- Univ. Los Andes (Bolivie),
- Univ. Bobo-Dioulasso (Burkina F.),
- Univ. Nat. Maurice (Maurice),
- National herbarium of Comores,
- Botanical Garden Geneva (Switzerland)
- National parks
- NGOs (Tela Botanica, iScanTree, Endémia, ..)



Pl@ntNet Mobile App Usage



Pl@ntNet Mobile App Usage



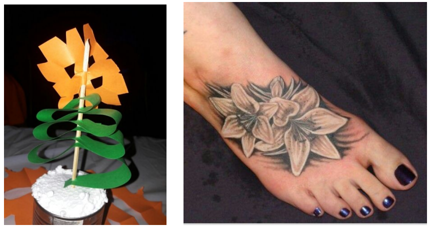
Personal usage (88%)

Houseplants

Gardening

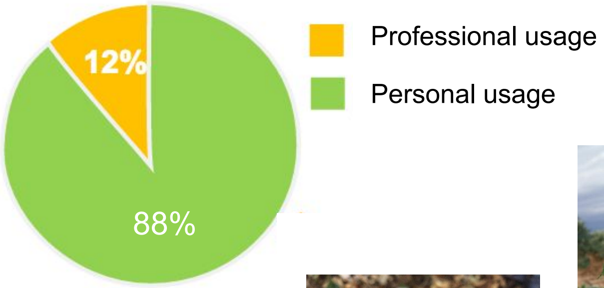
Walk, jannie, trekking

Phytotherapy, eatable



Fun, delusional

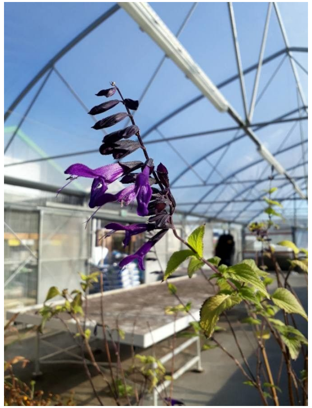
Pl@ntNet Mobile App Usage



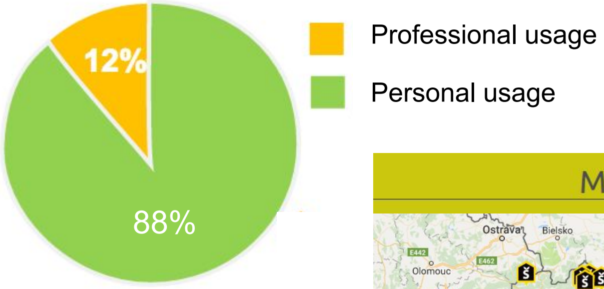
Agriculture & Agri-food industry (4.8%)



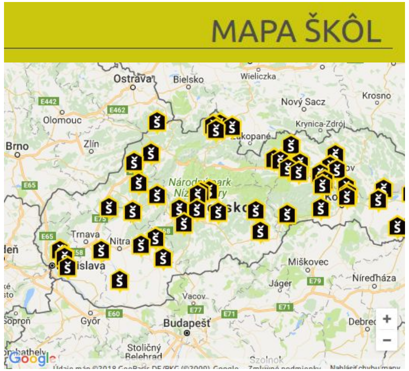


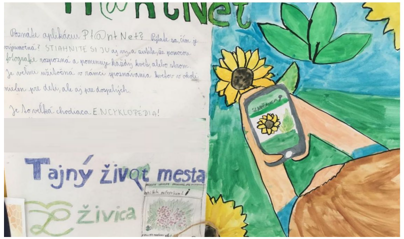



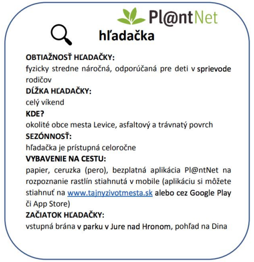

Pl@ntNet Mobile App Usage



Education & animation (3.2%)






Pl@ntNet

hľadačka

OBTIAŽNOSŤ HĽADAČKY:
fyzicky stredne náročná, odporúčaná pre deti v sprievode rodičov

DĹŽKA HĽADAČKY:
celý víkend

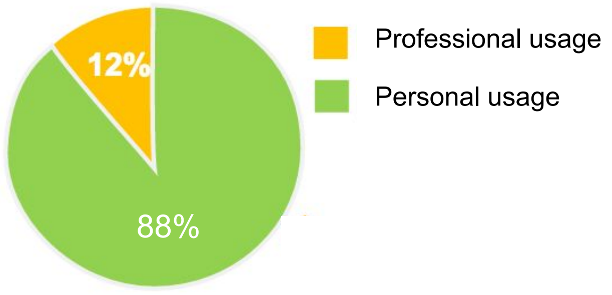
KDE?
okolité obce mesta Levice, asfaltový a trávnatý povrch

SEZÓNNOSŤ:
hľadačka je prístupná celoročne

VYBAVENIE NA CESTU:
papier, ceruzka (pero), bezplatná aplikácia Pl@ntNet na rozpoznanie rastlín stiahnutá v mobile (aplikáciu si môžete stiahnuť na www.tajnyzivotmesta.sk alebo cez Google Play či App Store)

ZAČIATOK HĽADAČKY:
vstupná brána v parku v Jure nad Hronom, pohľad na Dina

Pl@ntNet Mobile App Usage



Other professional usage (4%)



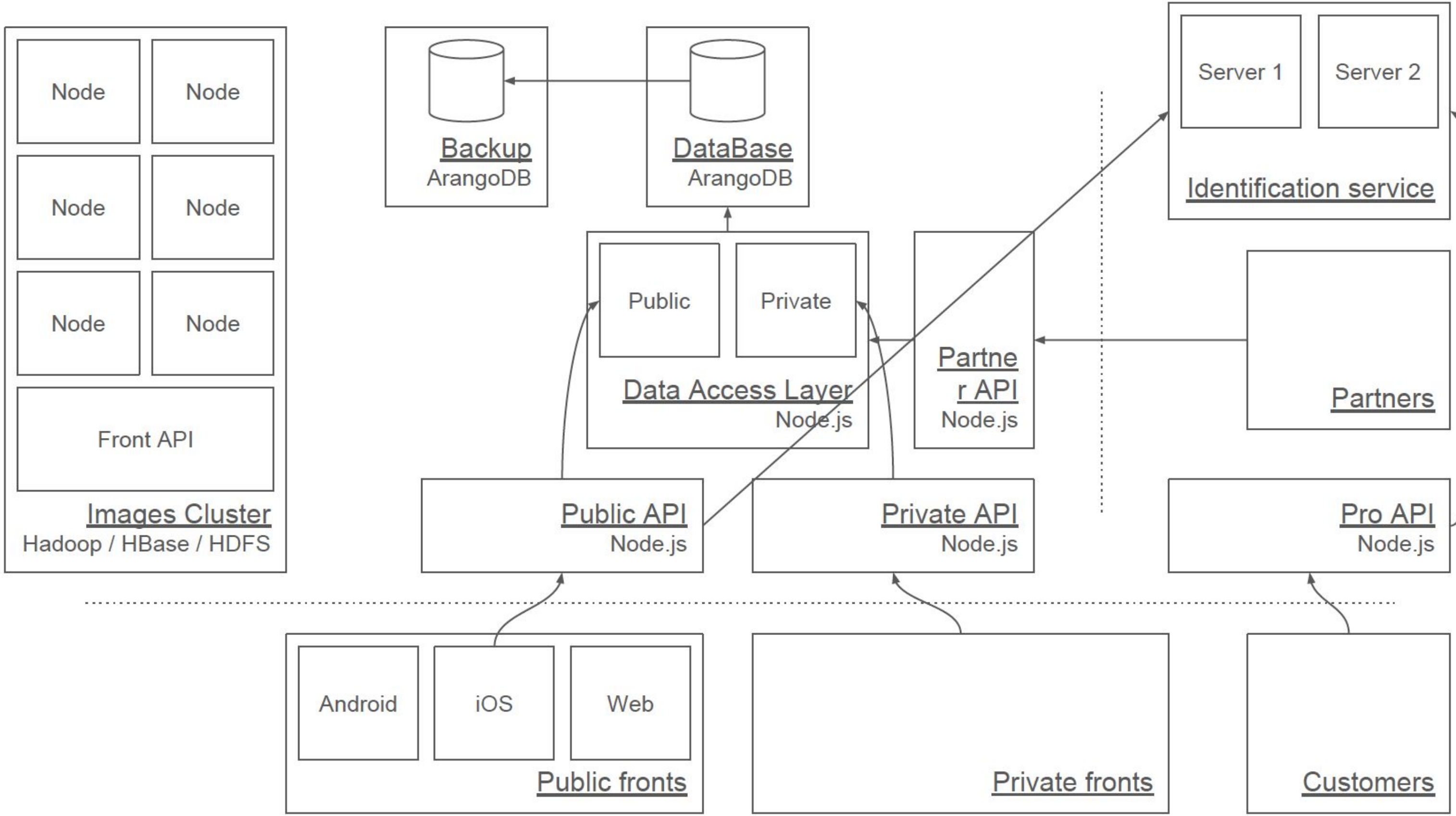
Professional botanists, consulting, expertise


Merchants

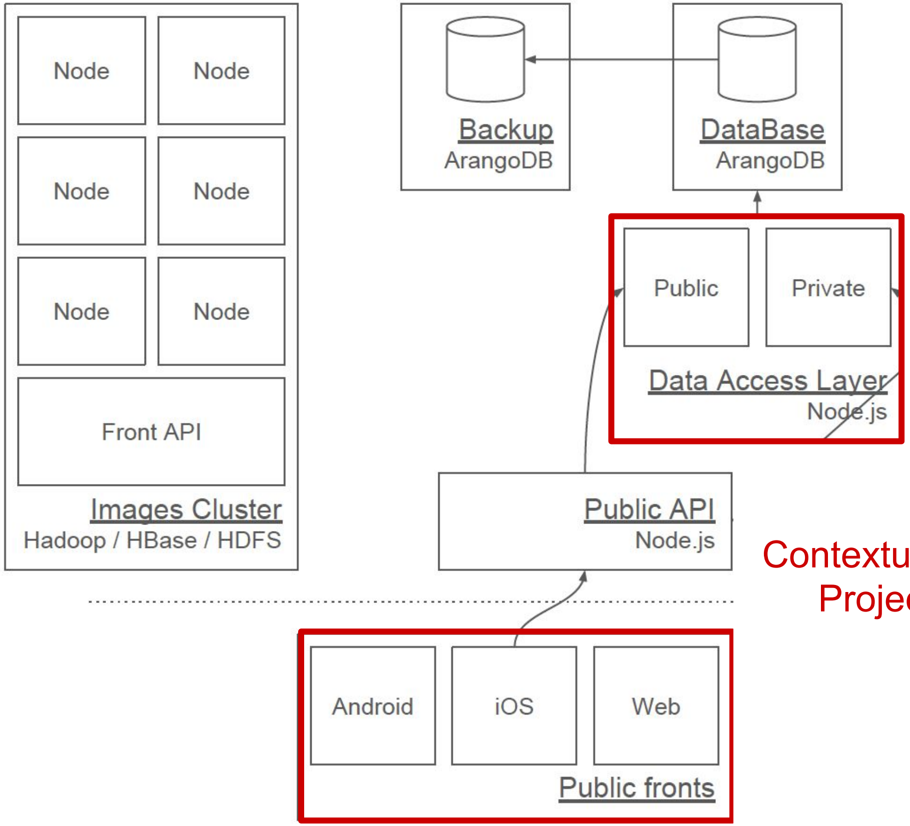

Natural area management

Tourism

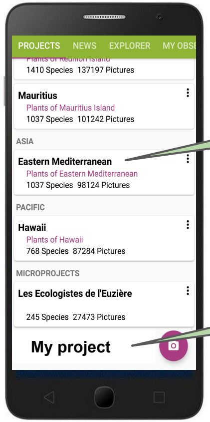
Infrastructure



Infrastructure



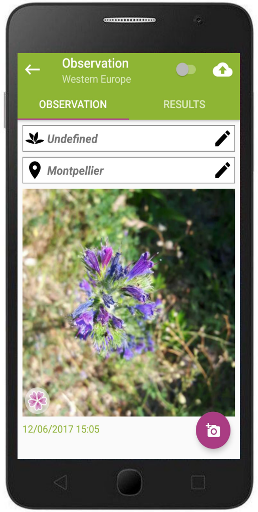
Contextualized Projects

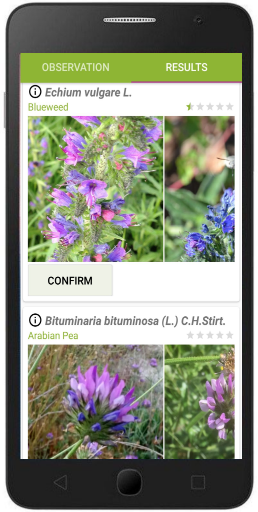


Large regional or thematic projects: *e.g.* “western europe”, “Hawai”, “Useful plants”

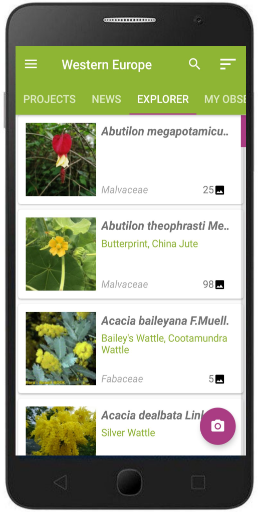
Micro-project: a very specific project dedicated to a small region, or a dedicated flora

Identification screens

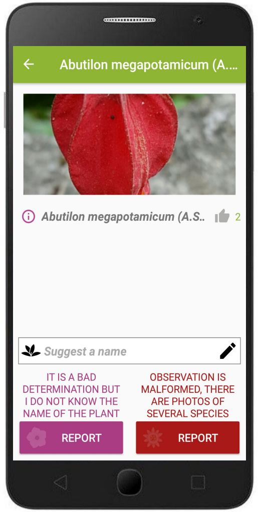


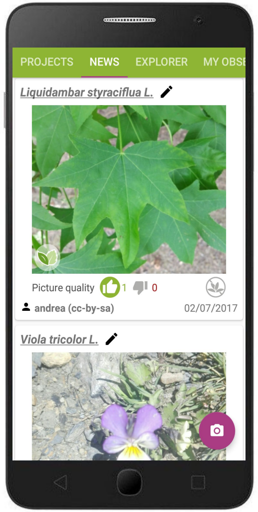


Exploration screens



Collaborative revision screens





All screens are **contextualized** with the project's **species of interest.**

Android	iOS	Web

Public fronts

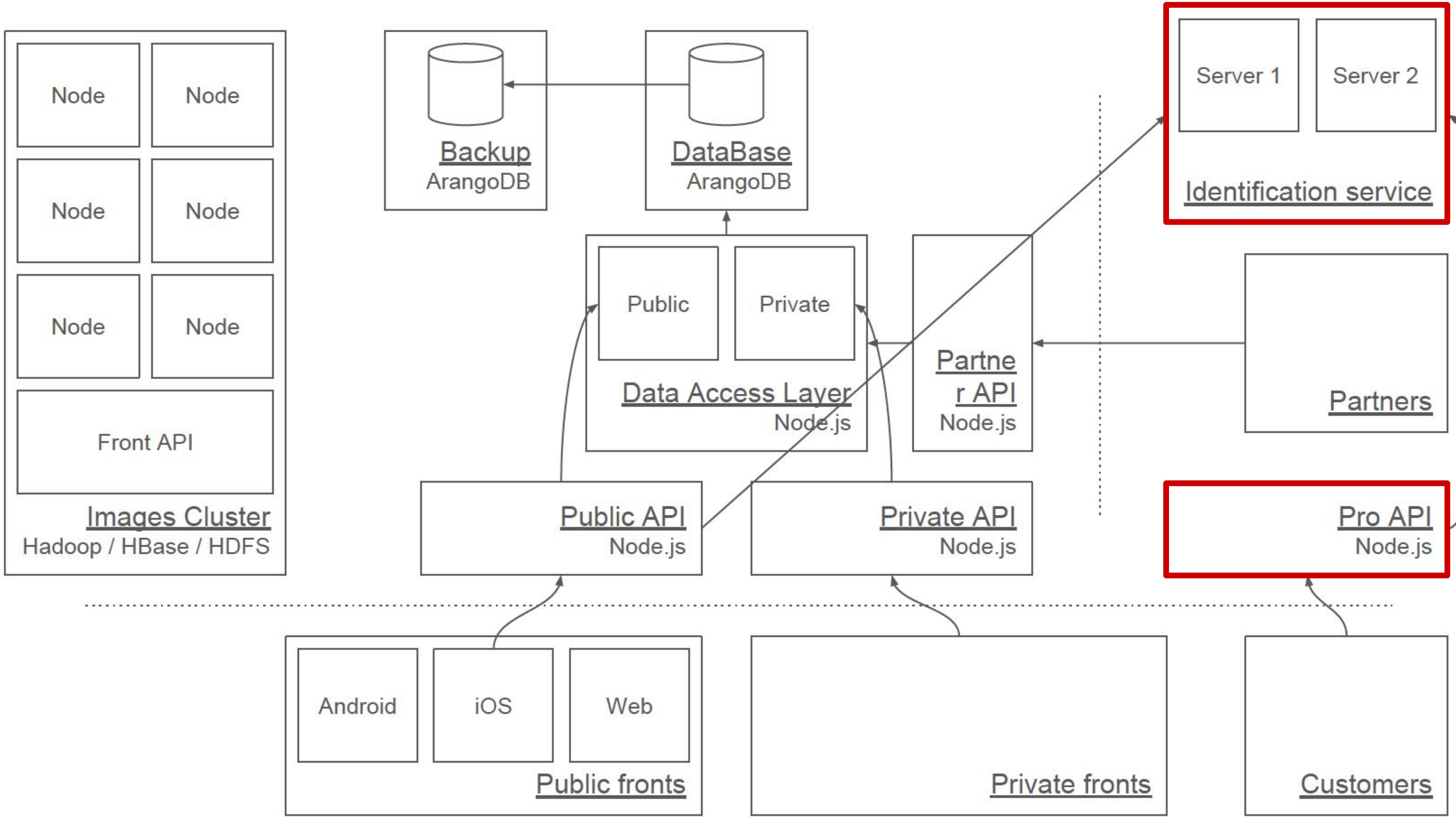
Infrastructure: Micro-projects

- Currently 3 micro-projects, several others in discussion

Wild salads

Infrastructure

Plant identification as a service

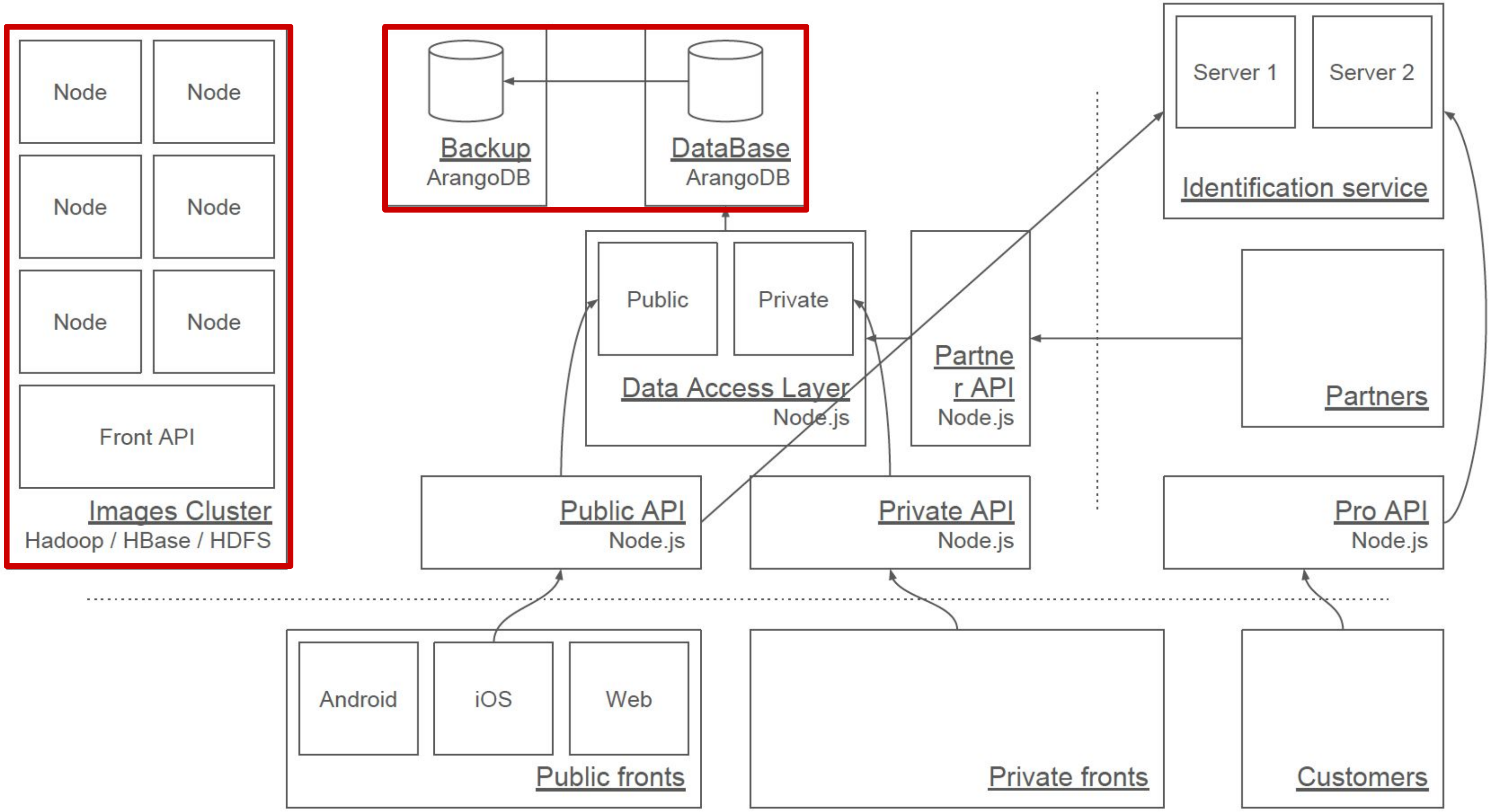


Infrastructure: Pro API

- Currently experimented by 15 beta-testers (app developers)
  - Start-ups: BiodivGo, NaturalSolutions, ecoBalade, Garden-answers, Jardin Imaginaire, etc.
  - Universities, public bodies, associations, student projects

Infrastructure

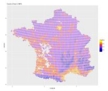
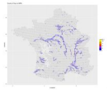
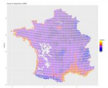
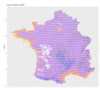
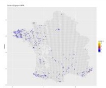
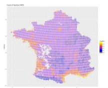
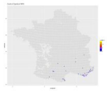
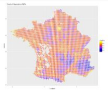
Research projects driven by Pl@ntNet data



Research projects in plant sciences

- Two examples of projects centrally driven by Pl@ntNet data

Invasive species distribution models

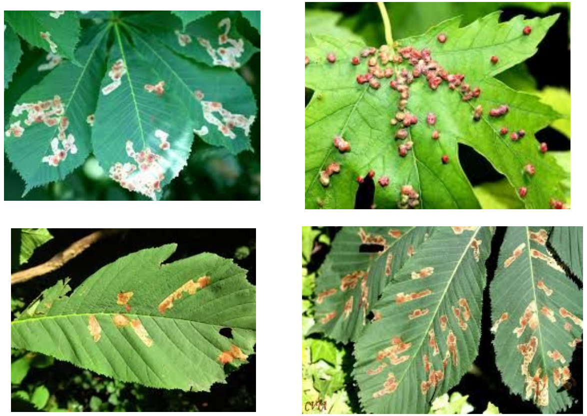
Species names	Species distribution computed from Pl@ntNet data	Maps from expert count data of INPN
Acer negundo L. Sensibility : 0,66 Specificity: 0,72		
Carpobrotus edulis (L.) N.E.Br. Sensibility : 0,94 Specificity: 0,89		
Erigeron karvinskianus DC. Sensibility : 0,64 Specificity: 0,72		
Opuntia ficus-indica (L.) Mill. Sensibility : 0,82 Specificity: 0,95		
Reynoutria japonica Houtt. Sensibility : 0,90 Specificity: 0,22		

New invasion occurrences

Risks of invasion

Pl@ntHealth: automated plant epidemiology

Chesnut gall



#DigitAg

Labex NUMEV

agropolis fondation

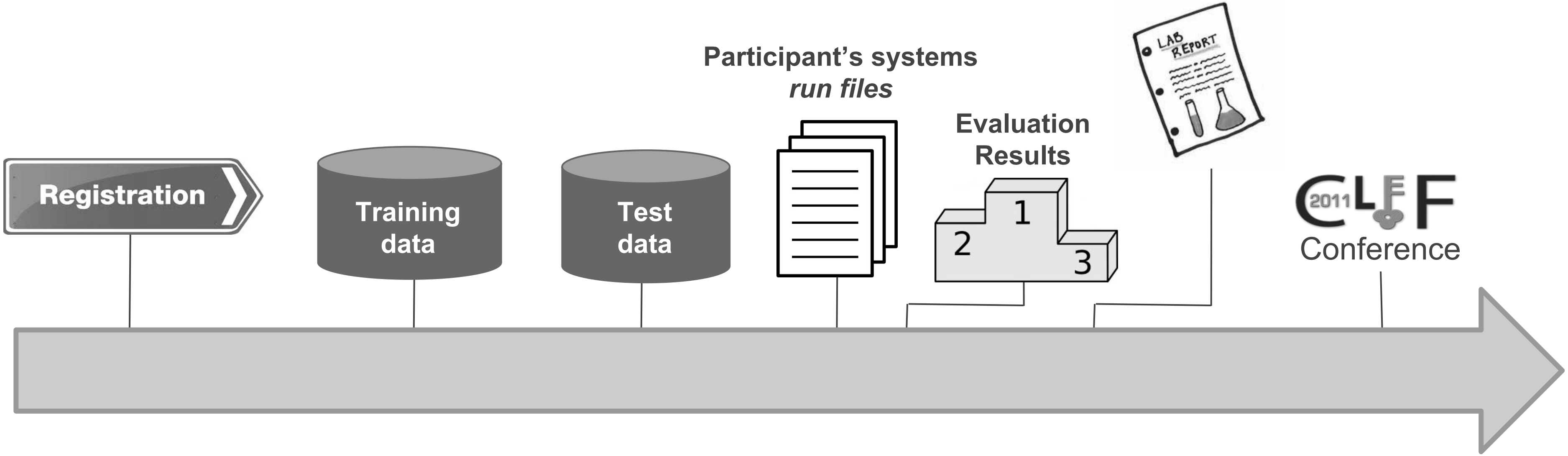
INRA SCIENCE & IMPACT

cirad

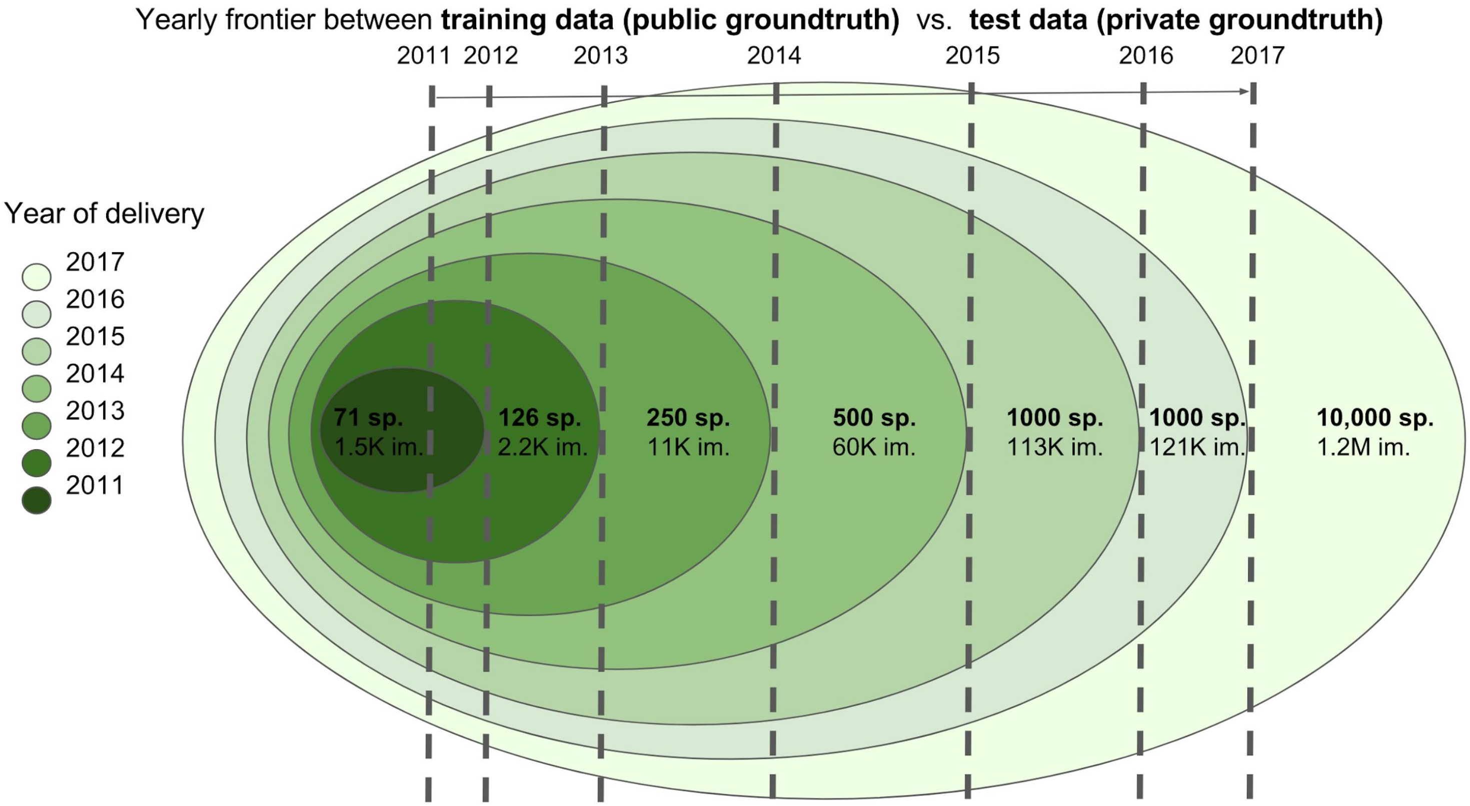
Inria inventeurs du monde numérique

Biodiversity informatics research within CLEF forum

- Pl@ntNet organizes a world-wide challenge since 2011
- Tens of research teams working on Pl@ntNet data
- **System-oriented** benchmarks/competitions



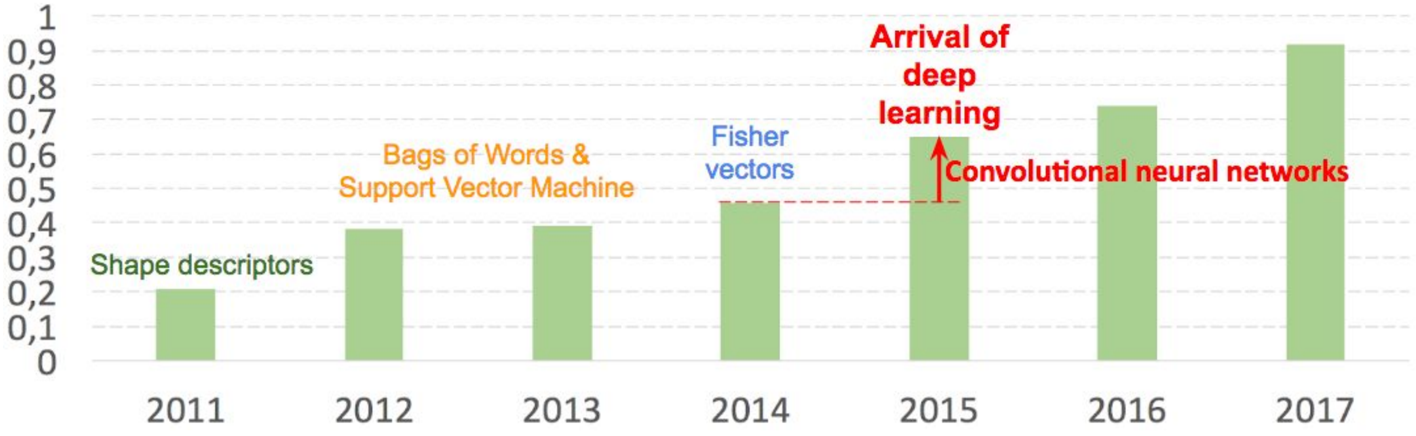
PlantCLEF



Yearly frontier between training data (public groundtruth) vs. test data (private groundtruth)
2011 2012 2013 2014 2015 2016 2017
Year of delivery
2017
2016
2015
2014
2013
2012
2011
71 sp.
1.5K im.
126 sp.
2.2K im.
250 sp.
11K im.
500 sp.
60K im.
1000 sp.
113K im.
1000 sp.
121K im.
10,000 sp.
1.2M im.

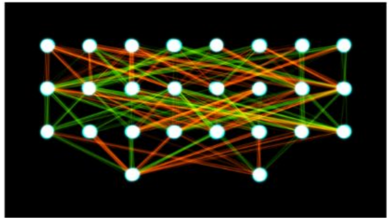
PlantCLEF

	2011	2012	2013	2014	2015	2016	2017
Espèces	71	126	250	500	1,000	1,000	10,000
Images	5,400	11,500	26,077	60,962	113,205	121,205	1.2 M
Nb. of particip.	8	11	12	22	15	16	17
Best perf.	**0,209**	**0,38**	**0,393**	**0,456**	**0,652**	**0,742**	**0,92 !**



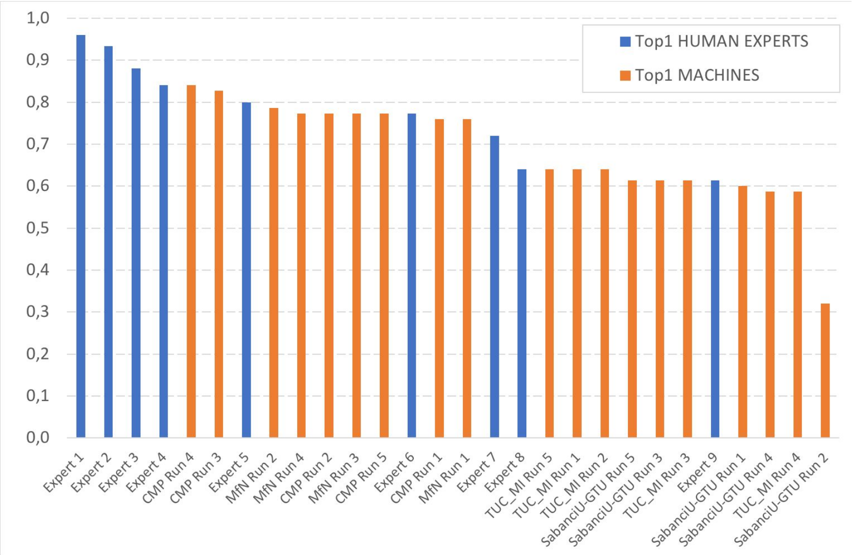
PlantCLEF

vs.



PlantCLEF 2018: Experts vs. Machines plant images identification

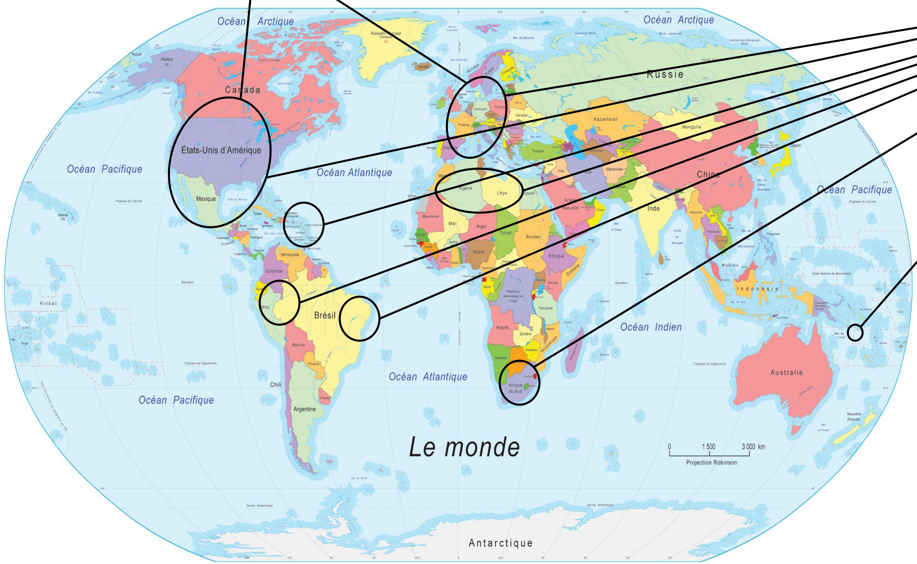
- **9 of the best of the best experts of the French flora**
- 100 obs. including very difficult taxonomic groups



Is the problem solved ? Not really...

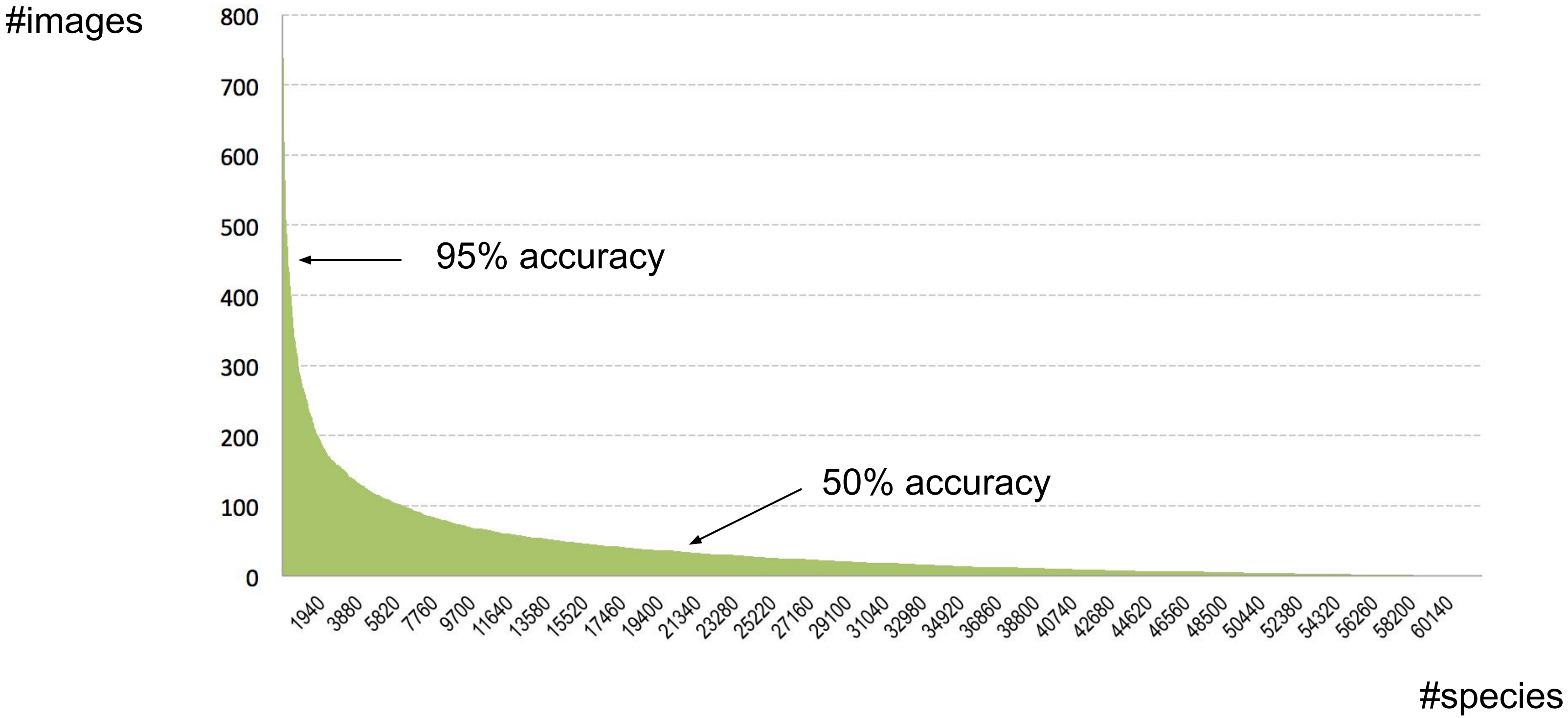
LifeCLEF: 10K species

Pl@ntNet: 17K species



World: 369K species

Is the problem solved ? Not really...



The Big One

- **We did query Bing and Google image with 300K species names**
  - Using ThePlantList: the first effort to list all plants on earth
- **We collected 12 million images** of **294K plant species (1.5 Tb)**:
  - Expert data (Encyclopedia of Life, 350K images) + Citizen science data (Pl@ntNet data, 400K images) + Web data (11 M images)
- **Highly imbalanced distribution**: only 50K species with more than 10 images, 50% with 1 images)
- **Noise:** depends on the species

"Arnica montana"

Challenges/questions

Scalability to hundreds of thousands of classes

- Which hardware ?
  - Memory usage: last layer is 300 times larger than state-of-the-art models
  - To distribute or not to distribute ?: communication cost, large batch size
  - CPU vs GPU ?
- Which network architecture ?
  - Convergence of state-of-the-art models ? No guaranty
  - Do we need a new dedicated architecture?
  - Acceptable training time ?
- Quality of the learned models ?
  - Top-1, top-5, top-30 accuracy ? On average ? In the long tail ?
  - Robustness to noise in the training data ?

Evaluation methodology: test set

- **30K never published images** of expert botanists
    - Stored on their local disks or on slides
    - Complex groups in the long tail of the distribution
        - 342 Orchids species
        - 1K Guyana species
        - 469 Alpine species
        - 75 Grass species

- **PlantCLEF 2017 test set (25K Pl@ntNet images)**
    - 1K species living in America and Europe (including common ones)
    - Never published labels

Platforms & frameworks

GENCI proposed us to be **beta-tester** of prototype platforms

- **Ouessant: GPU cluster** hosted by IDRIS (IBM OpenPOWER platform)
    - **12 nodes** IBM Power Systems **x 4 GPU** Nvidia P100 + Infiniband
    - **IBM powerAI** framework v4:
        - Caffe-DLL & TensorFlow-DLL
        - Stochastic gradient

- **Irene: CPU cluster** hosted by CEA (Intel skylakes platform)
    - **1600 nodes** x **48 Intel Skylakes**
    - Intel-CAFFE library

Ouessant/GPU experiments (1/3)

By Hervé Goëau, data scientist Pl@ntNet (CIRAD / Inria)

- **Encountered difficulties**: feedback from a data scientist without experience in HPC or distributed deep learning
  - **File systems / inodes issues**: quota exceeded notifications, file creation errors, etc.
  - **No internet access:** no wget, no curl to download pre-trained models, tests, etc.
  - **Lack of documentation**
  - **Limitation of the installed frameworks:** old versions, no data augmentation, no shuffling, etc.
  - **Jobs limitation** (20h00 & 4 nodes)
  - **Within the allocated time: No efficiency gain observed in multi-nodes**

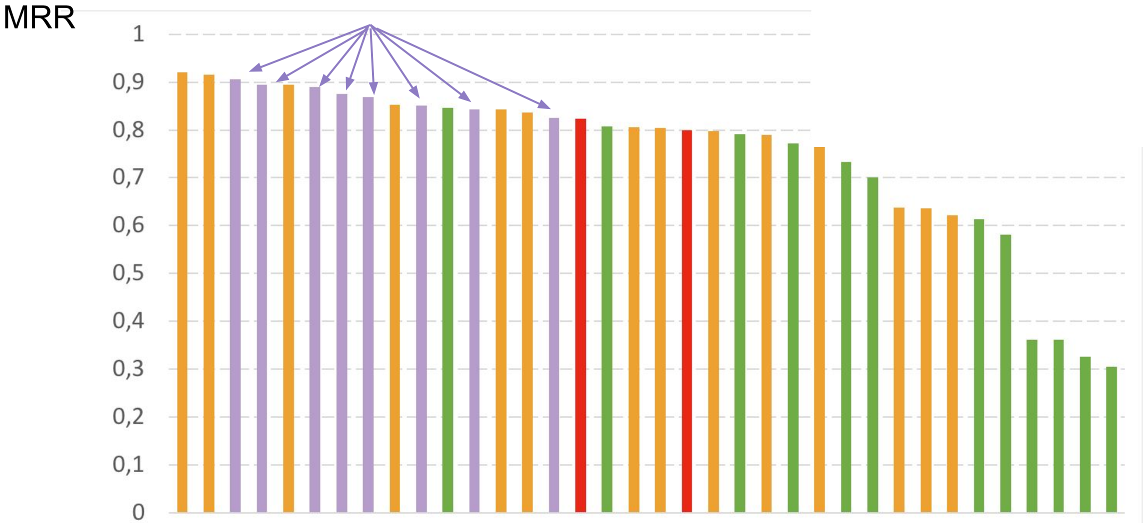
- **Succeeded in training a model at the scale of the world's flora using transfer learning**
  - Inception v2 model pre-trained on ImageNet and fine-tuned on 294K species during about 7 epochs
  - About **60h of training** on 1 node with 4 P100 GPUs

Ouessant/GPU experiments (2/3)

By Hervé Goëau, data scientist Pl@ntNet (CIRAD / Inria)

- **The model works ! state-of-the-art performance on PlantCLEF 2017 dataset** (without using ensembles)

Our world's flora model (with different testing configurations: data augmentation, post-filtering, duplicates removal, multi-image)



Ouessant/GPU experiments (3/3)

By Hervé Goëau, data scientist Pl@ntNet (CIRAD / Inria)

Performance in the long tail is low but fair with regard to 294K classes

Dataset	**Top1 accuracy (single image)**	**Top1 accuracy + multi-image**	**Top5 accuracy + multi-image**
Orchids	0.04	0.12	**0.22**
Alpine	0.19	0.25	**0.40**
Guyana	0.07	0.07	**0.12**
Grasses	0.37	0.57	**0.71**
Random	0.000003	0.000003	0.000015

Irene/CPU experiments (1/3)

- **Team**
  - Valeriu Codreanu & Damian Podareanu (Research engineers at **SURFsara**, state-of-the-art results on1K Intel Skylake)
  - Jean-Christophe Lombardo (Research engineer at **Inria - Pl@ntNet**)
  - Gabriel Hautreux (HPC engineer, **CINES/GENCI**)
  - Vikram A Saletore (Principal Engineer for Artificial Intelligence Products at **Intel**)

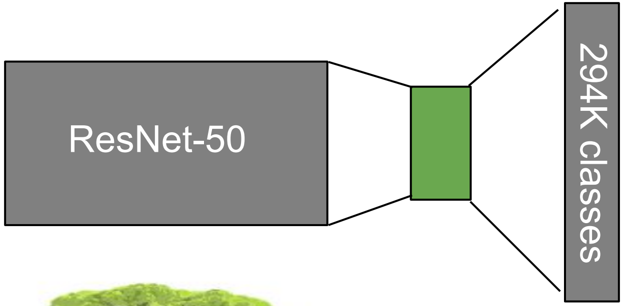
- **Preparatory phase on Occigen & Frioul CPU cluster from CINES**
  - Occigen: 3306 nodes x 2 Intel processors (12-14 cores)
  - Frioul: 48 nodes x Intel KNL processor (68 cores)

Irene/CPU experiments (2/3)

Encountered difficulties

- Intel-CAFFE (MLSL library) requires a password less ssh connexion for initialization (only possible to run in interactive mode)
- Protobuf library is limited to 2Gb files: impossible to serialize ResNet-50 model with 275K classes → dimensionality reduction trick

Last layer size: 2.3GB



Last layer size: 1.8GB

Irene/CPU experiments (3/3)

	Top1 accuracy (all world flora test sets)	Top5 accuracy (all world flora test sets)	Training time
Ouessant: 1 node - 4 x P100 Inception v2 fine-tuned 10 epochs	0.356	0.454	60 hours 6 hours/epoch
Irene: 512 skylake nodes ResNet-50 from scratch 50 epochs	**0.375**	**0.463**	**10 hours 12 minutes/epoch**
Irene: 1320 skylake nodes ResNet-50 from scratch 82 epochs	0.362	0.451	9 hours 9 minutes/epoch

Conclusions

- Data deficiency in the long tail remains the core problem: 50% of species illustrated by only 1 image on the web
- State-of-the-art CNNs scale to 300K classes (without much modifications)
- Synchronous SGD on hundreds of nodes provides high scaling efficiency but this requires significant know-how

Perspectives

- Integrate The Big One in Pl@ntNet platform
- Sustain the platform to continue the aggregation of data and knowledge about plants

Thank you