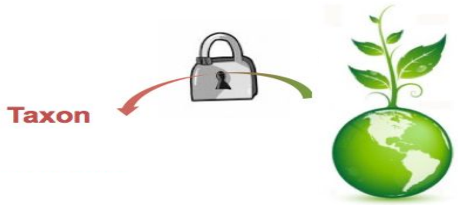
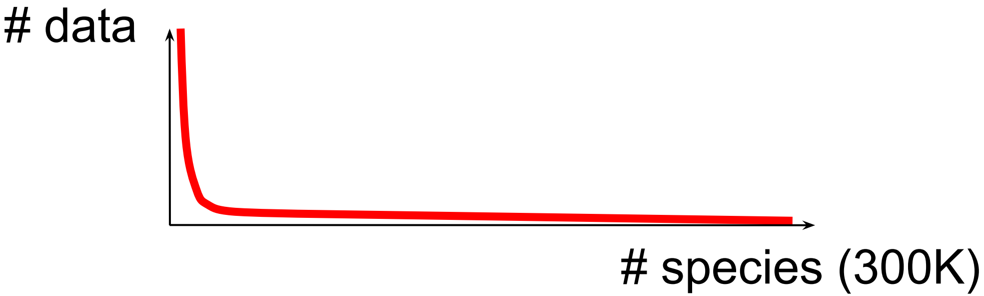
Towards the recognition of the world's flora

Alexis Joly, Hervé Goëau, Valeriu Codreanu, Jean-Christophe Lombardo

Taxonomic gap

- Plant identification is crucial for sharing and accessing knowledge about plants
  - Food crisis
  - Biodiversity crisis
- But the taxonomic gap is a tricky problem
  - Traditional tools only suitable for specialists
  - Less and less specialists
- Particularly in south countries with the richest bioc



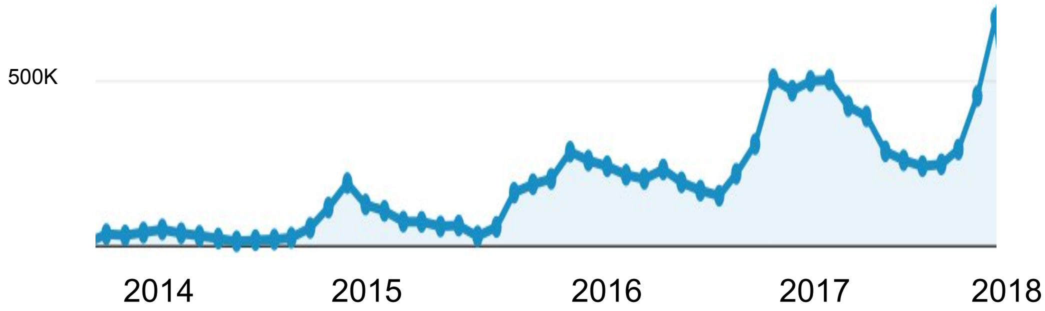


Pl@ntNet

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

Pl@ntNet app

Users / month

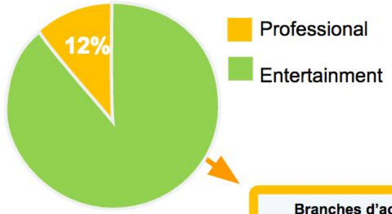


Last 12 months: 3,352,788 users in 235 countries	
More than 5 sessions	1,469,423
More than 10 sessions	876,698
More than 25 sessions	172,666
More than 100 sessions	14,167
1. United States	412,062 (17.63%)
2. France	357,001 (15.28%)
3. Germany	191,180 (8.18%)
4. Italy	160,388 (6.86%)
5. Spain	126,940 (5.43%)
6. Brazil	115,821 (4.96%)
7. United Kingdom	77,864 (3.33%)
8. India	64,488 (2.76%)
9. Netherlands	61,830 (2.65%)
10. Canada	57,159 (2.45%)

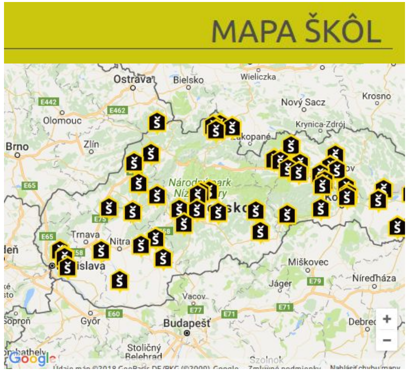
- 11 languages
- 17K species *(illustrated by 800K revised images)*
- 23 projects & micro-projects *(e.g. asian plants, trees of South Africa, etc.)*
- 30M raw plant images
- 55M sessions / 192M screen views
- 12K followers on social networks

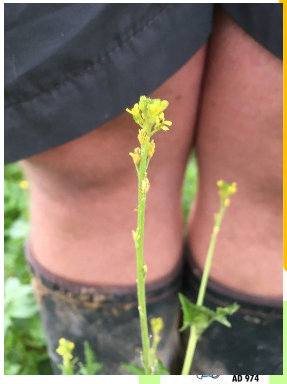
Pl@ntNet

Usage



Branches d'activités	Eff.	%
Agriculture et agroalimentaire	34	40,0%
Industrie	0	0,0%
Energie	0	0,0%
Commerce et artisanat	8	9,4%
Tourisme	2	2,4%
Télécoms et internet	0	0,0%
Recherche	5	5,9%
Enseignement	15	17,6%
Finance et assurance	0	0,0%
Autre service	21	24,7%
Total	**85**	**100,0%**

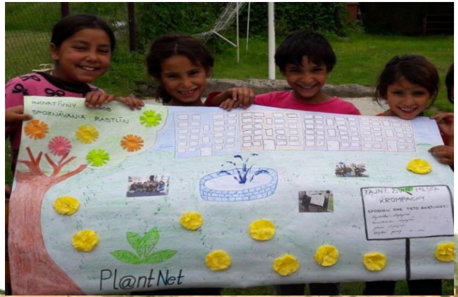




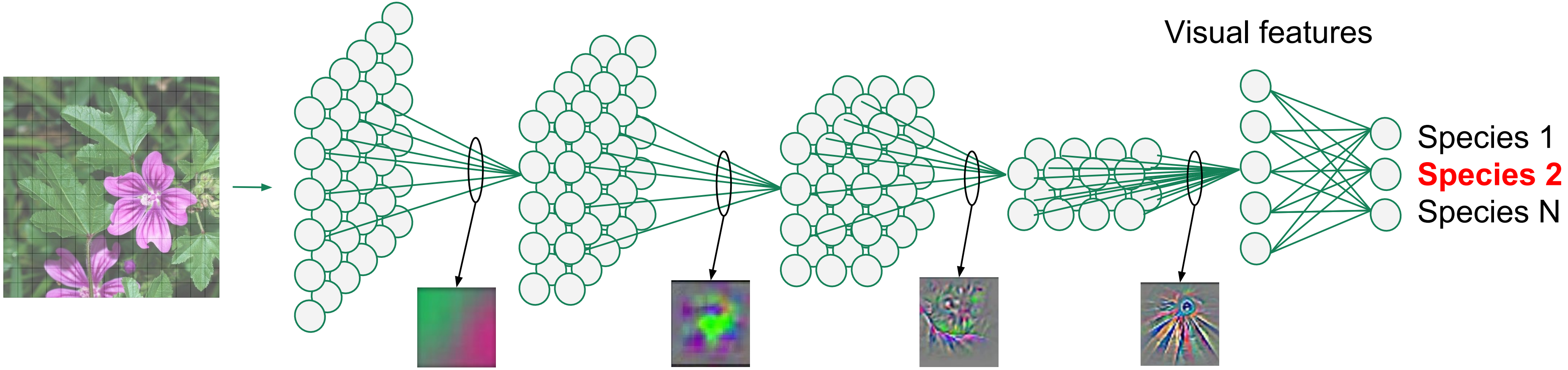
Énigmes végétales

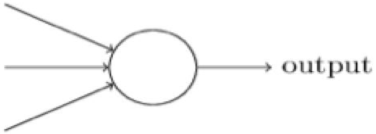
OCCE AUTONOMES & SOLIDAIRES AD 974 Pédagogie coopérative

Association départementale OCCE de la Réunion
18 rue de la Gare - BP 70043 - 97803 SAINT-DENIS-CEDEX 9
Tel : 02 62 21 54 50 – 06 92 34 54 50 - Courriel : ad974@occe.coop
www.occe.coop/ad974



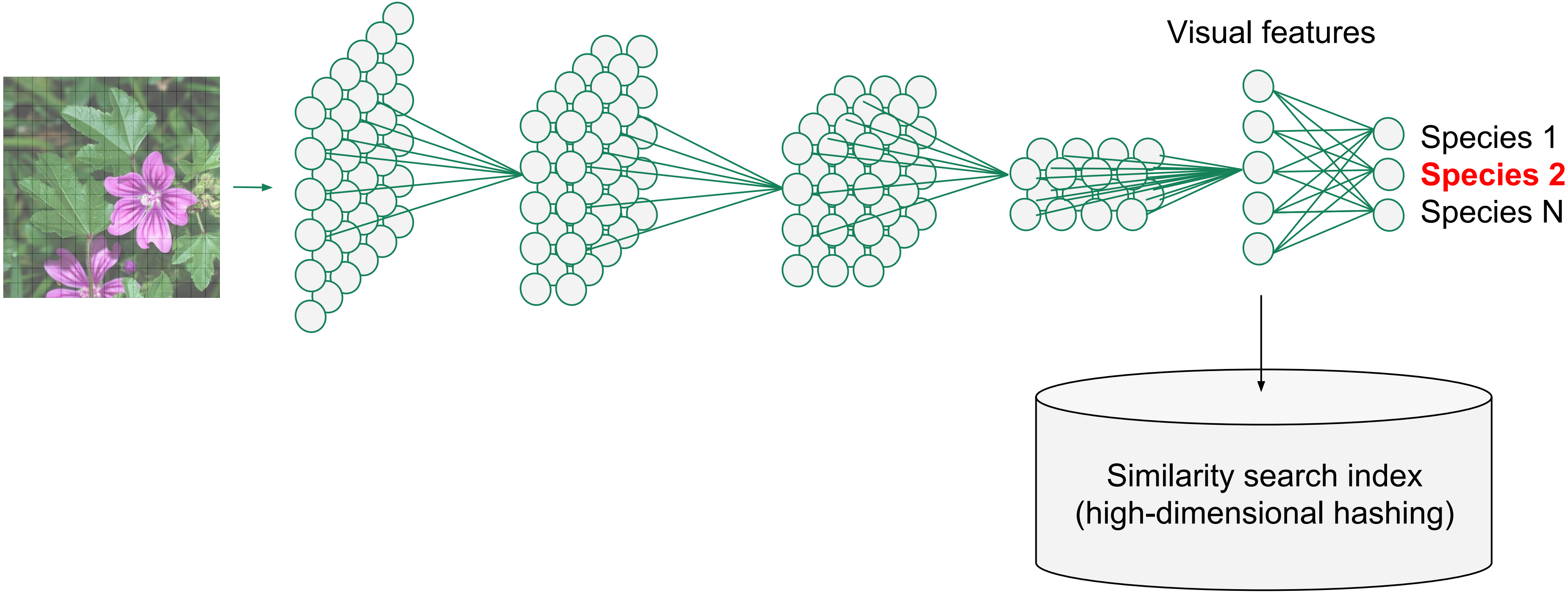
Technology:
Convolutional Neural Networks





$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right)$$

Technology:
Convolutional Neural Networks



Technology:
Convolutional Neural Networks

Billions of adjustable parameters (weights)

Requires high computing resources (GPUs or large clusters of CPUs)

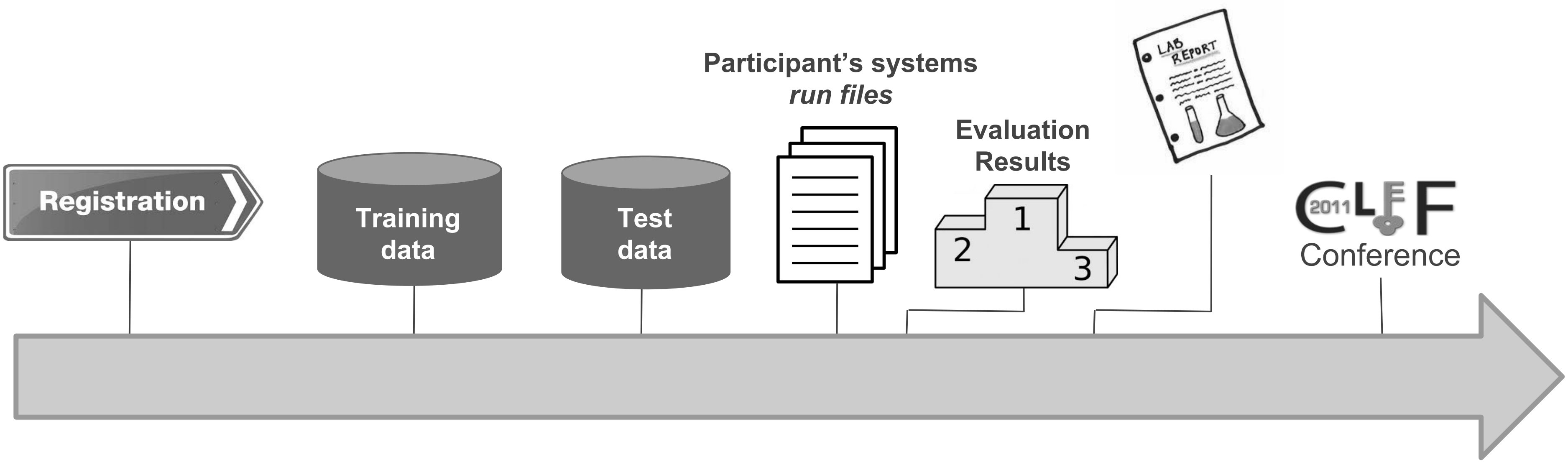
Transfer learning (fine-tuning)

Problem: CNNs require huge training data to learn the billions of parameters
Solution: Learn domain specific features by transfer learning

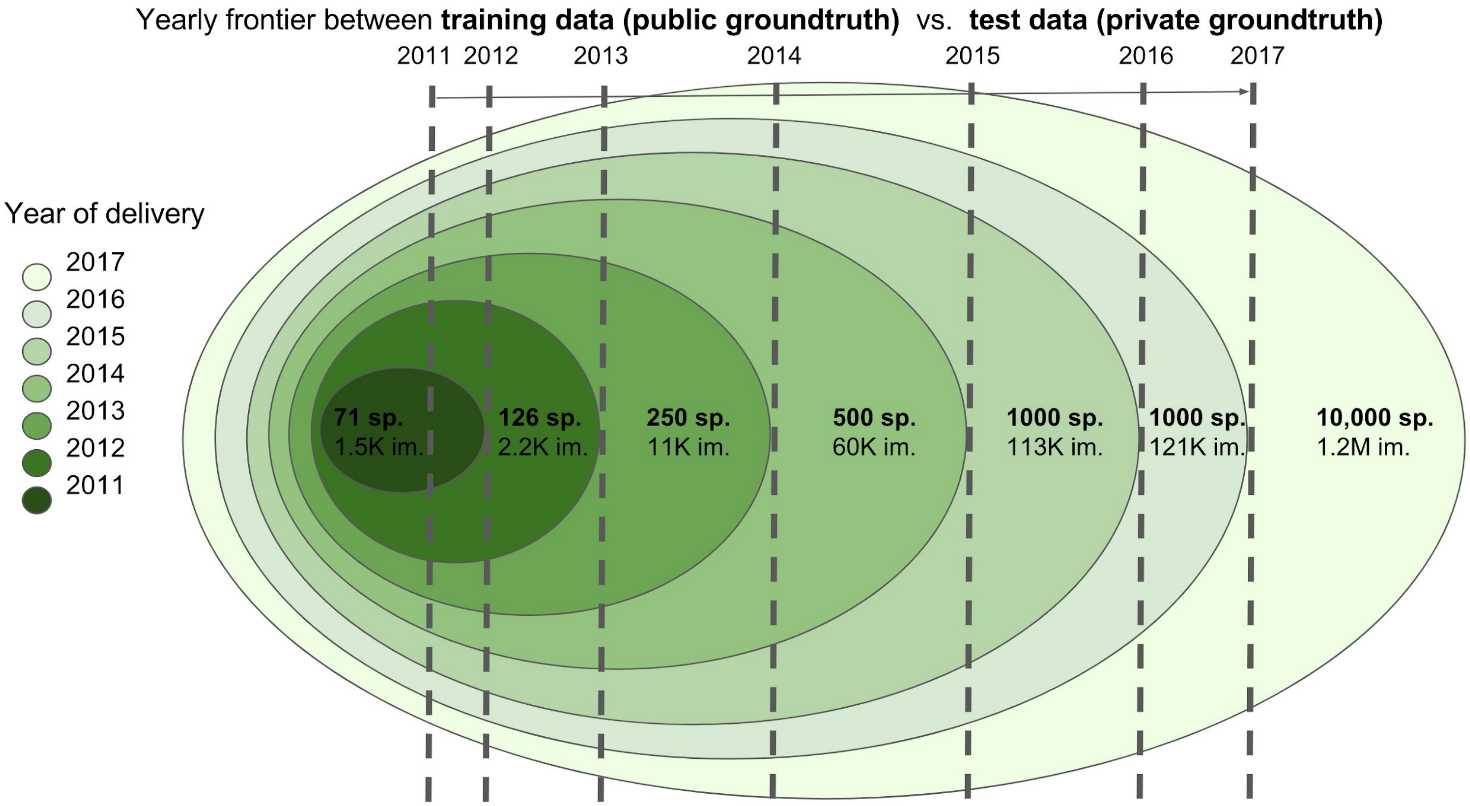
1. Train CNN on a generalist image dataset with millions of images
2. Keep the weights of the lowest layers but remove/reset the top layers
3. Feed forward and back-propagate new domain specific images

Evaluation

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



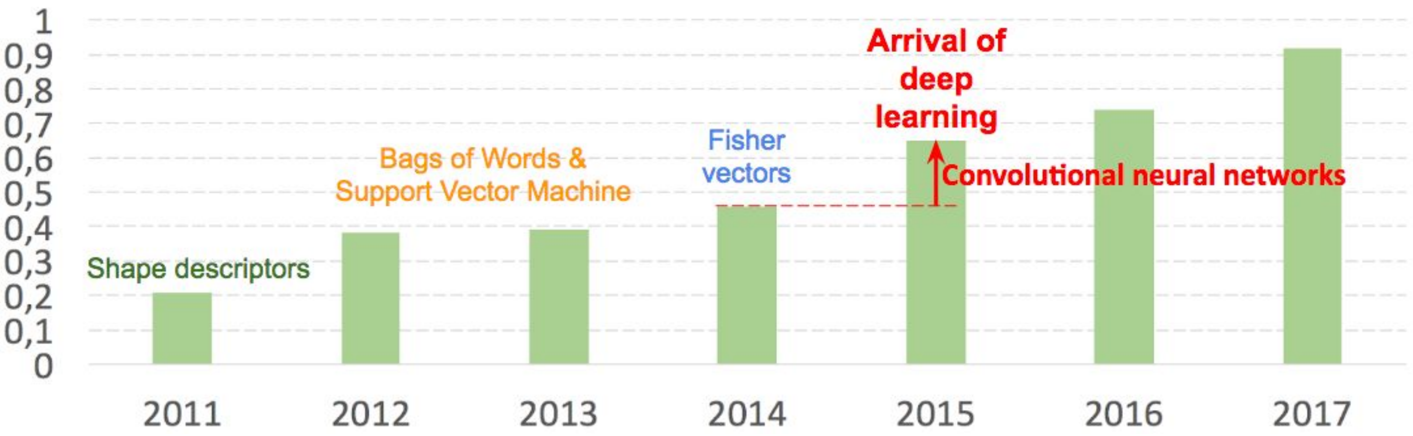
PlantCLEF



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.

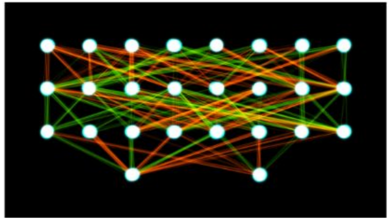
PlantCLEF

	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 !**



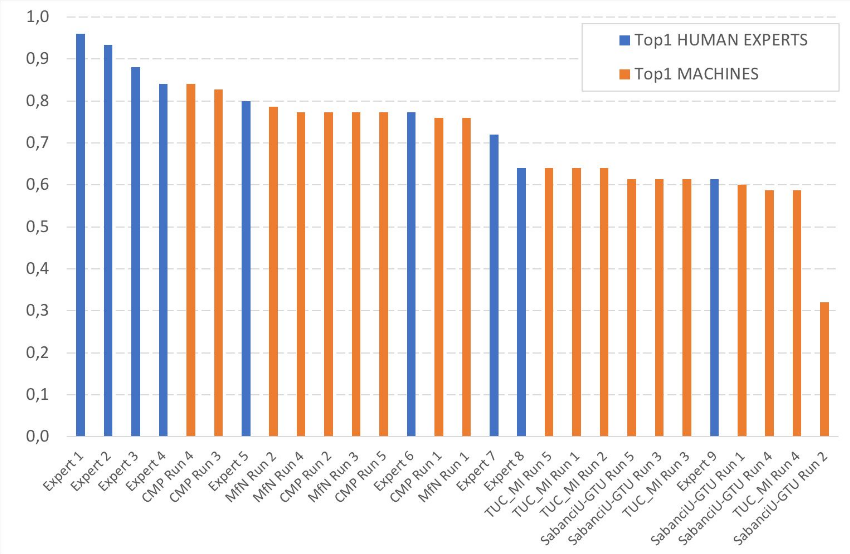
PlantCLEF

vs.



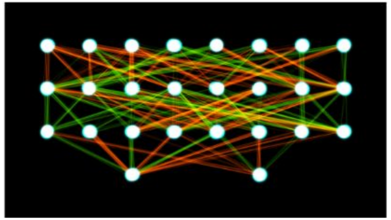
PlantCLEF 2018: Experts vs. Machines plant images identification

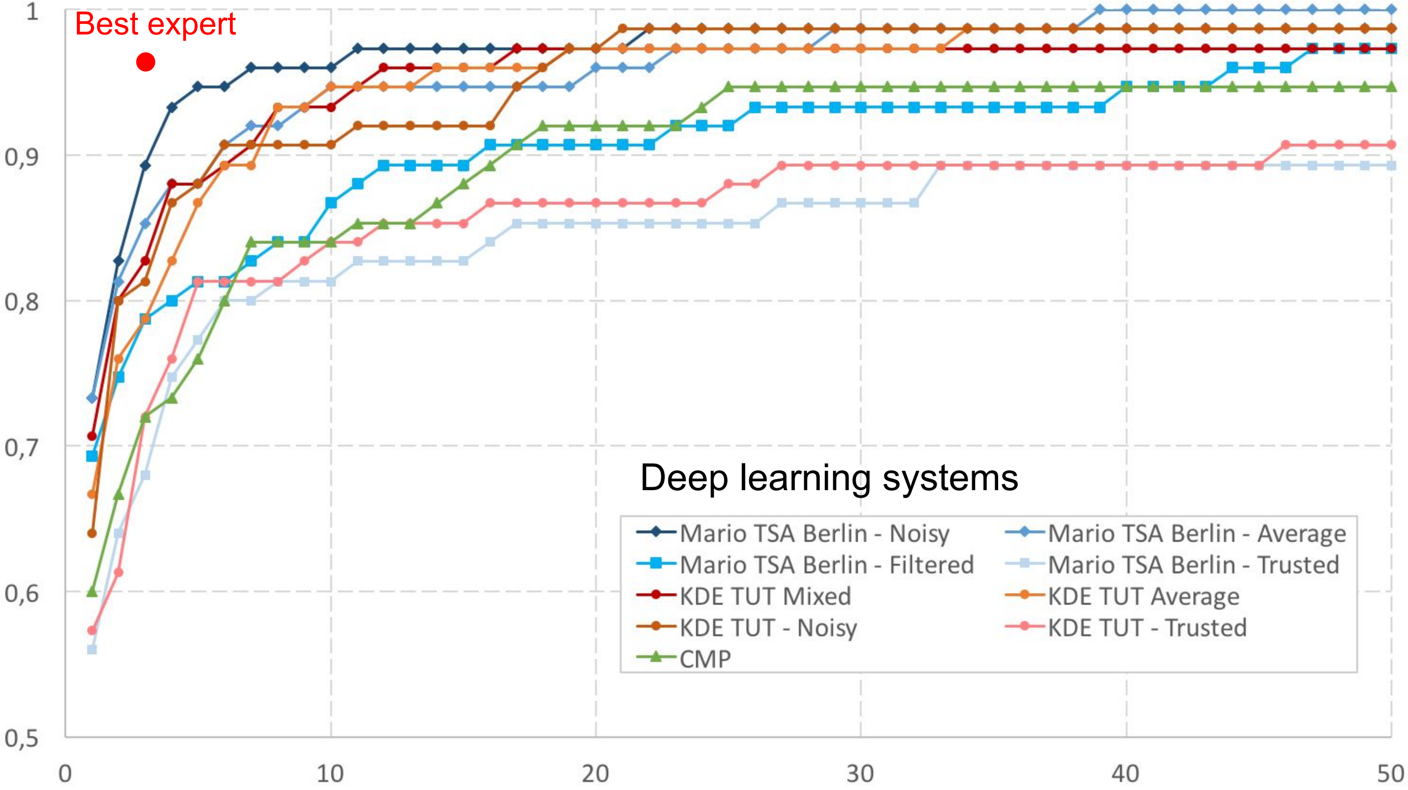
- **9 of the best of the best experts of the French flora**
- 100 obs. including very difficult taxonomic groups



PlantCLEF

vs.

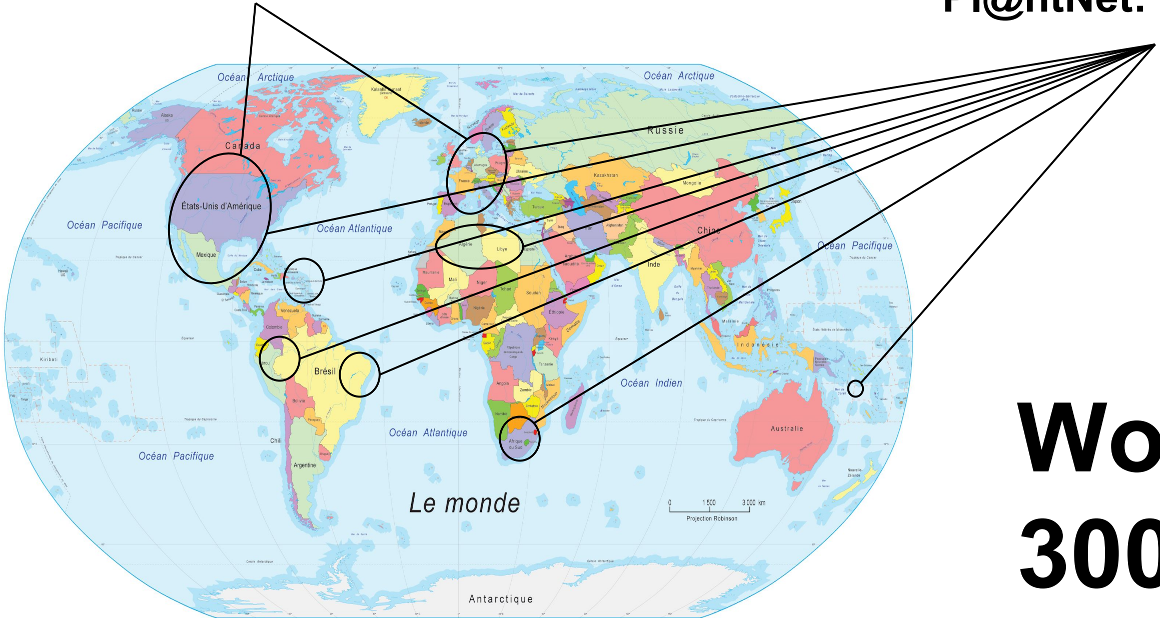




Is the problem solved ? Not really...

LifeCLEF: 10K species

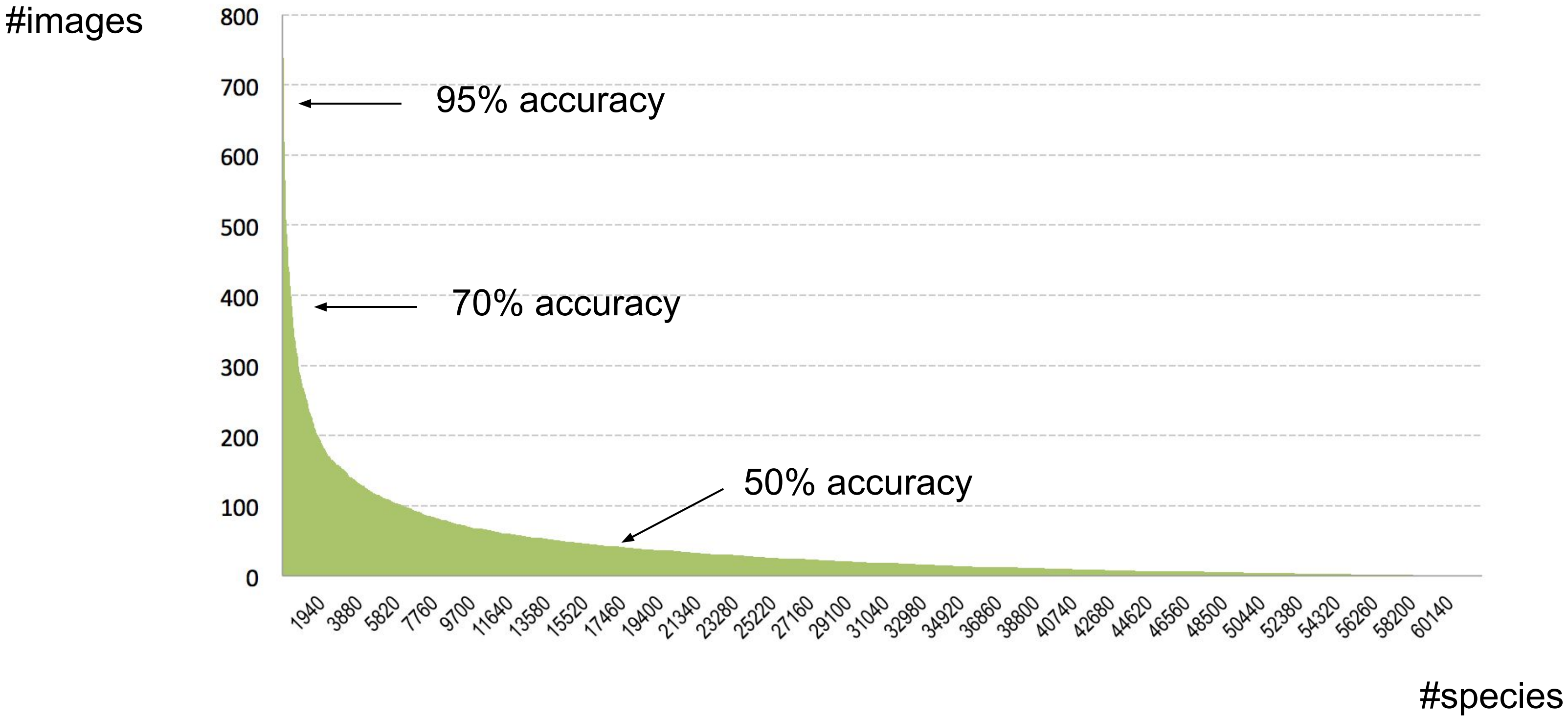
Pl@ntNet: 17K species



Encyclopedia of Life: 50K species

World: 300K 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 ?

Platforms & frameworks

GENCI proposed us to be **beta-tester** of 2 prototype platforms

- **Oussant: 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
- **Irene: CPU cluster** hosted by CEA (Intel skylakes platform)
  - **1600 nodes** x **48 Intel Skylakes**
  - Intel-CAFFE library
- Preparatory phase on **CINES CPU clusters (**Intel-CAFFE library)
  - **Occigen:** 3306 nodes x 2 Intel processors (12-14 cores)
  - **Frioul**: 48 nodes x Intel KNL processor (68 cores)

Evaluation methodology: set up

- Use state-of-the-art ConvNet with good **size/performance tradeoff**
  - Inception v2, ResNet-50
  - Extend them to **294K classes**

- Distribution with **synchronous Stochastic Gradient Descent**

$$\vec{R} = \sum_{j=1}^{N} \vec{V}_j$$

  - synchronize the gradients of N learners through a **collaborative reduction/communication** (such as **allreduce**)
  - High-dimensional = **millions of values**
  - Best **performances** in literature
    - FaceBook, ImageNet in 1 hour, 32x8 GPU P100, All reduce, batch size=8192
    - IBM powerAI, ImageNet in 50 min, 64x4 GPU P100, Multi-ring, batch size=8192
    - Intel-CAFFE , ImageNet in 28 min, 1.5K Intel KNL (100K cores), MLSL, batch size=48K
    - Preferred Networks, ImageNet in 15 min, 1K GPU P100, All reduce, batch size=32K
    -

Evaluation methodology: test set

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

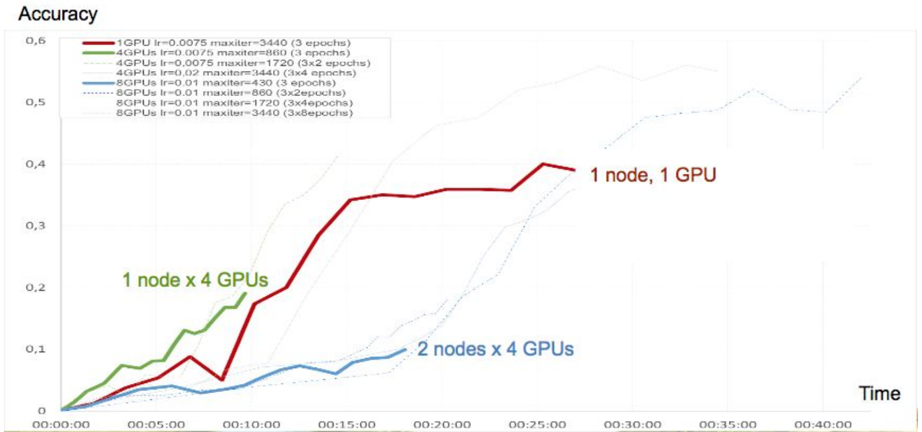
- **PlantCLEF 2017 test set (25K images)**
  - 1K species living in America and Europe (including common ones)
  - Never published labels

Ouessant experiments (1/4)

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**



Ouessant experiments (2/4)

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

- **Training models “from scratch” was not possible**
  - Several weeks on a single node
  - Without guaranty of (good) convergence
  - With a 20h jobs limitation

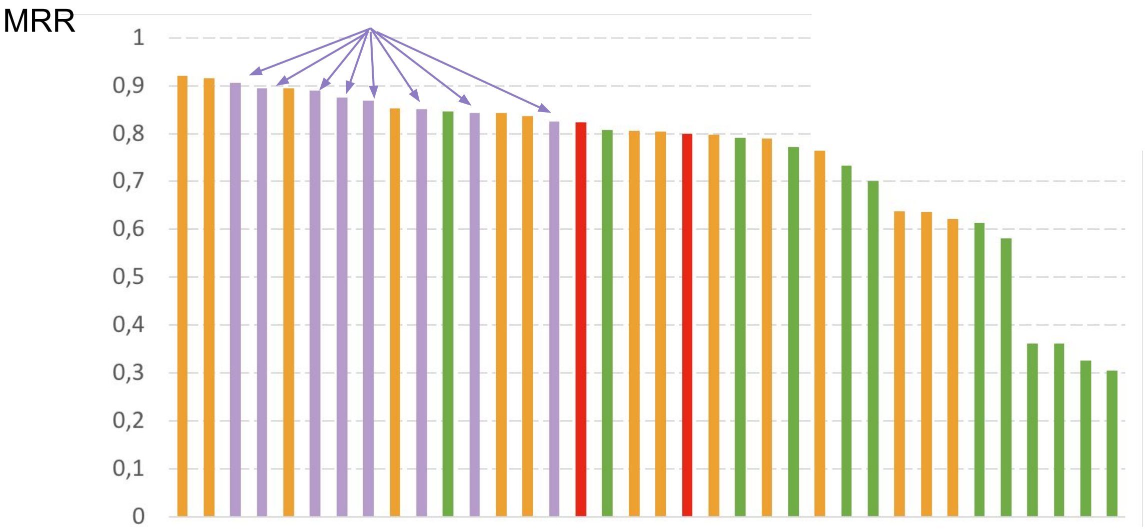
- **Succeeded in training a model at the scale of the world’s flora using transfer learning**
  - Inception v2 pre-trained on ImageNet and fine-tuned on 294K species in 2 steps
    1. Freeze the model except the last layer
    2. Fine-tune all layers
  - About **60h of training** on 1 node with 4 P100 GPUs

Ouessant experiments (3/4)

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 experiments (4/4)

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

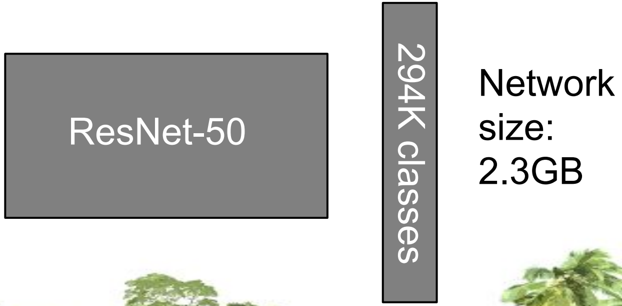
CINES experiments (1/5)

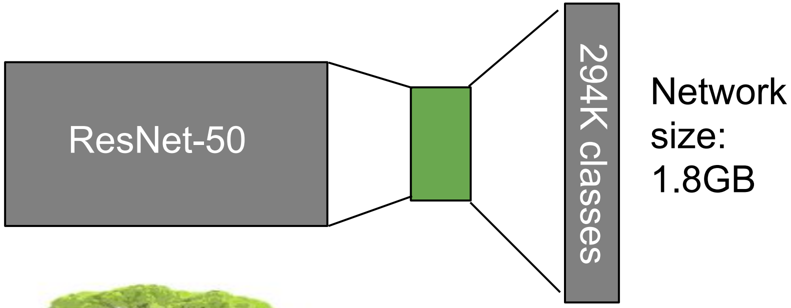
- **Team**
  - Valeriu Codreanu & Damian Podareanu (Research engineers at **SurfSara**, state-of-the-art results on1K Intel Skylake)
  - Jean-Christophe Lombardo (Research engineer at **Inria - Pl@ntNet**)
  - Gabriel Hautreux (HPC engineer, **CINES/GENCI**)
  - Vikram A Saletore (Principal Engineer for Artificial Intelligence Products at **Intel**)
- **Objective**
  - Scaling Deep learning on CPUs using INTEL-CAFFE (optimized for skylakes CPUs)
  - CEA Irene cluster (1600x48 Skylake hearts) in July (machine delivery)
- **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)

CINES experiments (2/5)

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





CINES experiments (3/5)

Scaling efficiency experiments

- **Lustre striping makes a big difference**

	2 nodes	32 nodes	**Scaling efficiency**
No striping	**47.5 img/s** 23.7 img/s/node	**303 img/s** 9.5 img/s/node	**40.1%**
Lustre striping stripe count: 64 stripe size: 32M	**47.5 img/s** 23.7 img/s/node	**688 img/s** 21.5 img/s/node	**90.7%**

CINES experiments (4/5)

Scaling efficiency experiments

Broadwell (BDW28) scaling results:

2 nodes; global batch size: 128; Throughput: 28.45 img/s/node. Aggregate throughput: 56.9 img/s

32 nodes; global batch size: 2048; Throughput: 25.6 img/s/node. Aggregate throughput: 819 img/s

64 nodes; global batch size: 4096; Throughput: 25.1 img/s/node. Aggregate throughput: 1606 img/s

128 nodes; global batch size: 8192; Throughput: 24.6 img/s/node. **Aggregate throughput: 3150 img/s**

→ **86.5% scaling efficiency when going from 2 to 128 BDW nodes**

Haswell (HSW24) scaling results :

128 nodes; global batch size: 8192; Throughput: 20.15 img/s/node. Aggregate throughput: 2580 img/s

→ **82.2% scaling efficiency when going from 2 to 128 HSW nodes**

CINES experiments (5/5)

Succeeded to learn two new models on Frioul and Occigen CPU clusters

	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	0.336	0.437	60 hours 6 hours/epoch
Frioul: 40 KNL nodes ResNet-50 fine-tuned 37 epochs	**0.355**	**0.440**	**37 hours 1 hour/epoch**
Occigen: 128 nodes (BDW) ResNet-50 from scratch 100 epochs	**0.363**	**0.449**	**28 hours 17 minutes/epoch**

Conclusion/perspectives

Conclusions

- State-of-the-art CNNs scale to 300K classes (without much modifications)
- Synchronous SGD on hundreds of CPU nodes provides high scaling efficiency but this requires significant know-how
- Training data remains a core problem

Perspectives

- Irene cluster: 1600 skylake nodes
- Inria Project Lab HPC-Deep Learning: PhD on the joint optimization of network architecture and resource allocation

Thank you