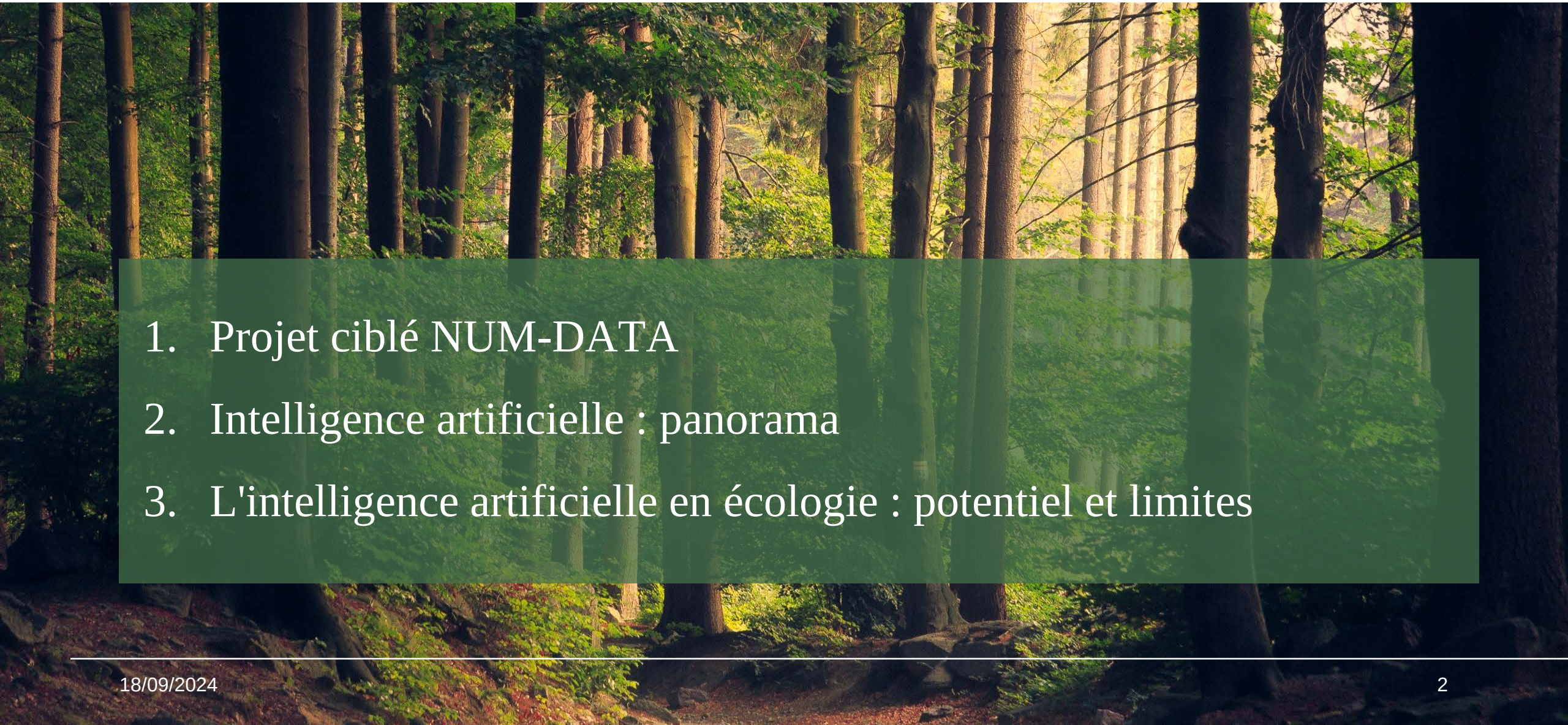
Partage, accessibilité et services d'exploitation des données

Projet Ciblé NUM-DATA

-

Focus sur l'Intelligence Artificielle



1. Projet ciblé NUM-DATA
2. Intelligence artificielle : panorama
3. L'intelligence artificielle en écologie : potentiel et limites

1. Projet Ciblé NUM-DATA

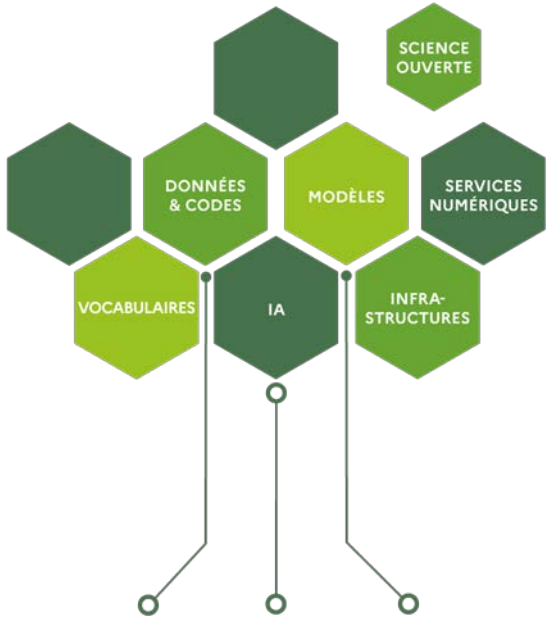
Projet ciblé NUM-DATA (1 M€)

Pilotes : Sophie Fortuno (CIRAD), Christian Pichot (INRAE), Fabrice Bénédet (CIRAD)

Le projet NUM-DATA répond à un besoin transverse dans le domaine du numérique et des données en soutien :

- ✓ aux projets ciblés du programme FORESTT
- ✓ et aux futurs projets lauréats de l'appel à projets

Pour développer le partage, l'accessibilité et l'exploitation des données des socio-écosystèmes forestiers tempérés et tropicaux



PARTENAIRES NUM-DATA

UMR TETIS - Territoires, Environnement, Télédétection et Information Spatiale, CIRAD
Sophie FORTUNO

UMR ECOFOG - Ecologie des Forêts de Guyane, AGROPARISTECH
Gaëlle JAOUEN

UR Forêts et Sociétés, CIRAD | Fabrice BENEDET

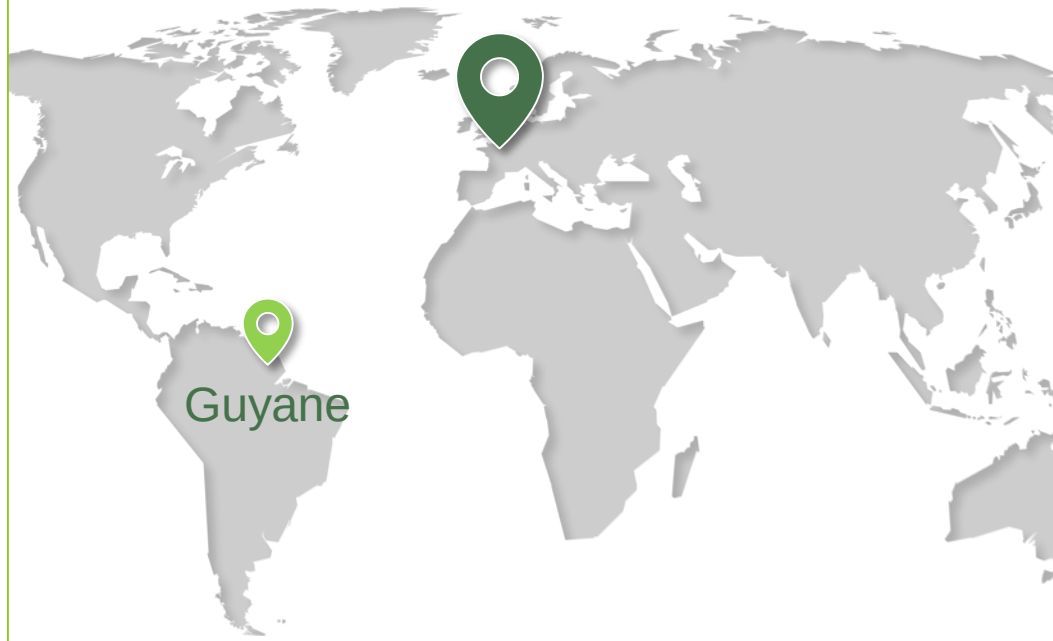
UMR AMAP - botAnique et Modélisation de l'Architecture des Plantes et des végétations, IRD | Paul TRESSON, Philippe VERLEY

URFM - Unité de Recherche écologie des Forêts Méditerranéennes, INRAE Christian PICHOT, Philippe CLASTRE

UMR Silva, INRAE | Alain BENARD, Damien MAURICE

UMR Biogeco - Biodiversité Gènes et Communautés, INRAE | François Ehrenmann

UR ETTIS - Environnement, Territoires en Transition, Infrastructures, Sociétés, INRAE | David CARAYON



CONTEXTE

- Place centrale et grandissante du numérique
- Politique de science ouverte, contextes juridique&éthique
  - Plan National pour la Science Ouverte, loi Lemaire, RGPD
- Gestion et partage des données FAIR
- Ecosystème numérique (inter)national de la Recherche
  - Recherche Data Gouv, Méso centres, Infrastructures de Recherche & Pôles de données et de services thématiques et numériques
  - EOSC Cloud Européen, …

POSITIONNEMENT

NUM-DATA est dédié aux transversalités numériques du PEPR FORESTT

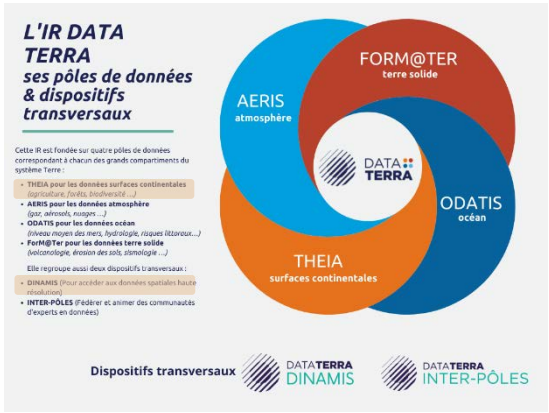
En soutien et **en synergie au sein et inter PC** sur expertise numérique

PROJETS CIBLES

&

PROJETS LAUREATS DE L'AAP

En **interaction avec les acteurs amont et aval du numérique scientifique** et les **institutions partenaires**



In-Sylva FRANCE

AnaEE

Huma-Num

AgroParisTech

cirad

INRAE

IRD

Partage, accessibilité et services d'exploitation des données NUM-DATA

ENJEUX

- **Traduire l'écosystème numérique et la politique nationale de science ouverte**
- **Identifier/relever les défis numériques**
- **Science des données / IA**
- **Porter à connaissance**
- **Fédérer**
- **Renforcer les capacités**
- **Traduire les besoins de la communauté,**
- **Les rapprocher des services**
- **En créer de nouveaux**

- Données **FAIR** *Facile à trouver, Accessible, Interopérable, Réutilisable*

- PC1-REGE ADAPT
- PC2-XRISKS
- PC3-MONITOR
- PC4-FORESTT HUB
- PC5 NUM-DATA

- Orienter vers les services supports
- Sensibiliser, former

Fournir des recommandations et services caractérisés <u>avec et pour</u> la communauté

ORGANISATION DU PROJET - 5 WORK PACKAGES

WP1 - Gouvernance, coordination et dissémination

WP2 - Pratiques et Ressources numériques FAIR

WP3 - Partage et accessibilité des ressources numériques & outils

NUM-DATA

WP5 - Communs numériques

WP4 - Data Science **IA**

Première action de soutien

Premier atelier Plan de Gestion des Données (PGD) organisé par le projet ciblé NUM-DATA

Ce 10 juin s'est tenu au CIRAD de Montpellier un atelier Plan de Gestion Données (PGD), réunissant des responsables des Projets Ciblés (PC) du programme de recherche PEPR FORESTT sur la résilience des forêts.

RESULTATS ET SERVICES ATTENDUS

1 - Promouvoir de meilleures pratiques de gestion et de partage des données

- Données FAIR Facile à trouver, Accessible Interopérable, Réutilisable
- Lignes directrices communautaires

2 - Renforcer les capacités de recherche

- Infrastructures de Recherche & services régionaux et institutionnels
- Référentiels de vocabulaires, interopérabilité des actifs numériques
- Science des données / IA - VRE et travaux exploratoires

3 - Socialiser, porter à connaissance avec un accès aux données, informations et services NUM-DATA

- Capitaliser les acquis. Mettre en visibilité sur un portail web
- Renforcer les capacités : webinaires, ateliers, formations…
- Identifier /mettre en réseau les experts
- Contribuer aux communs numériques FAIR (inter)national
- Contribuer à des évènements (RDA, CODATA...)

2. Intelligence artificielle : panorama

Intervention de Richard Moreno

Directeur Technique
de l’Infrastructure de Recherche Data Terra

visio

3. L'intelligence artificielle en écologie : potentiel et limites

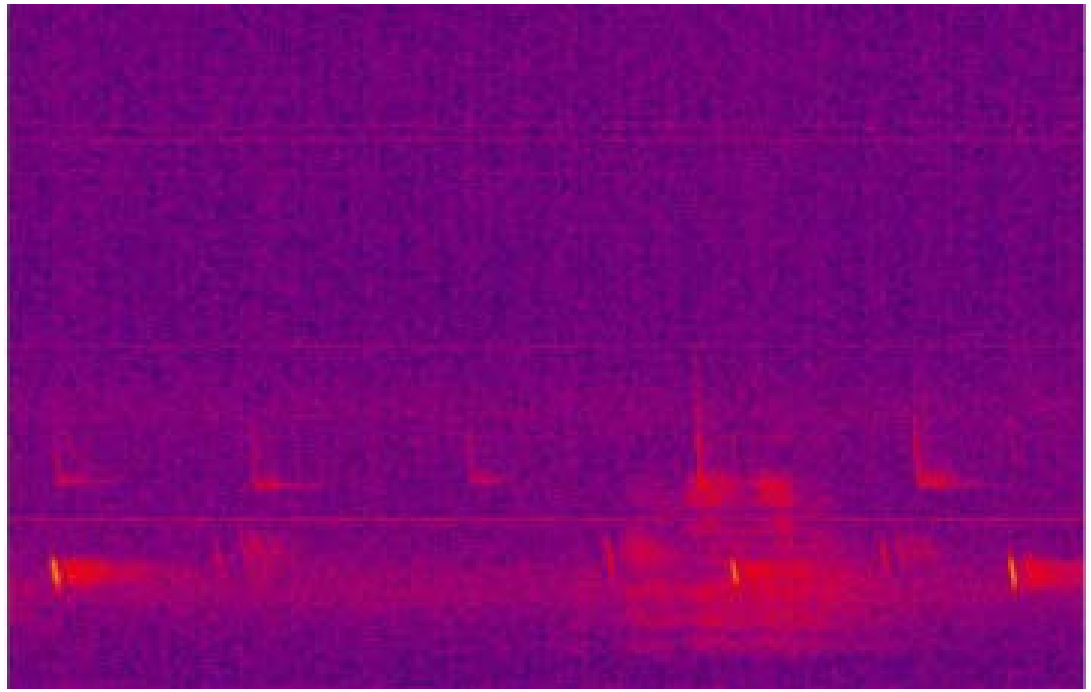
Pourquoi utiliser le Deep Learning?

Paul Tresson, UMR AMAP, IRD

De plus en plus de données

- Drones, satellites

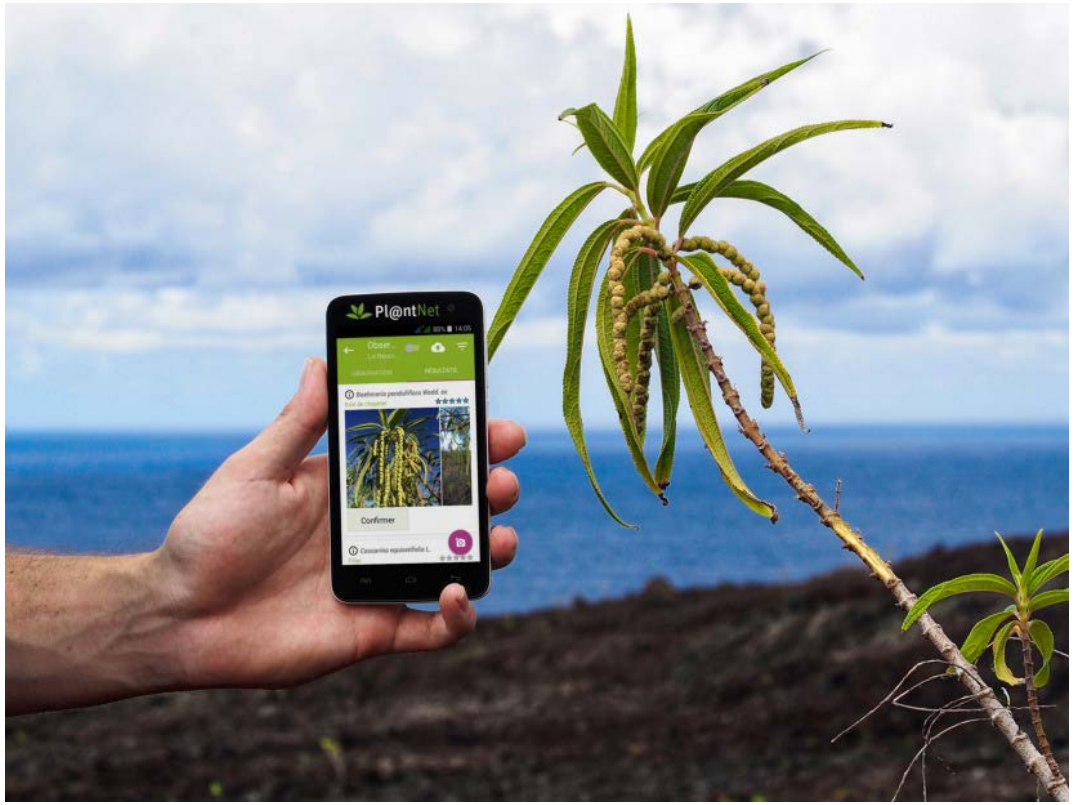
De plus en plus de données



Mac Aodha et al. 2022

- Drones, satellites
- Pièges photos, suivi acoustique

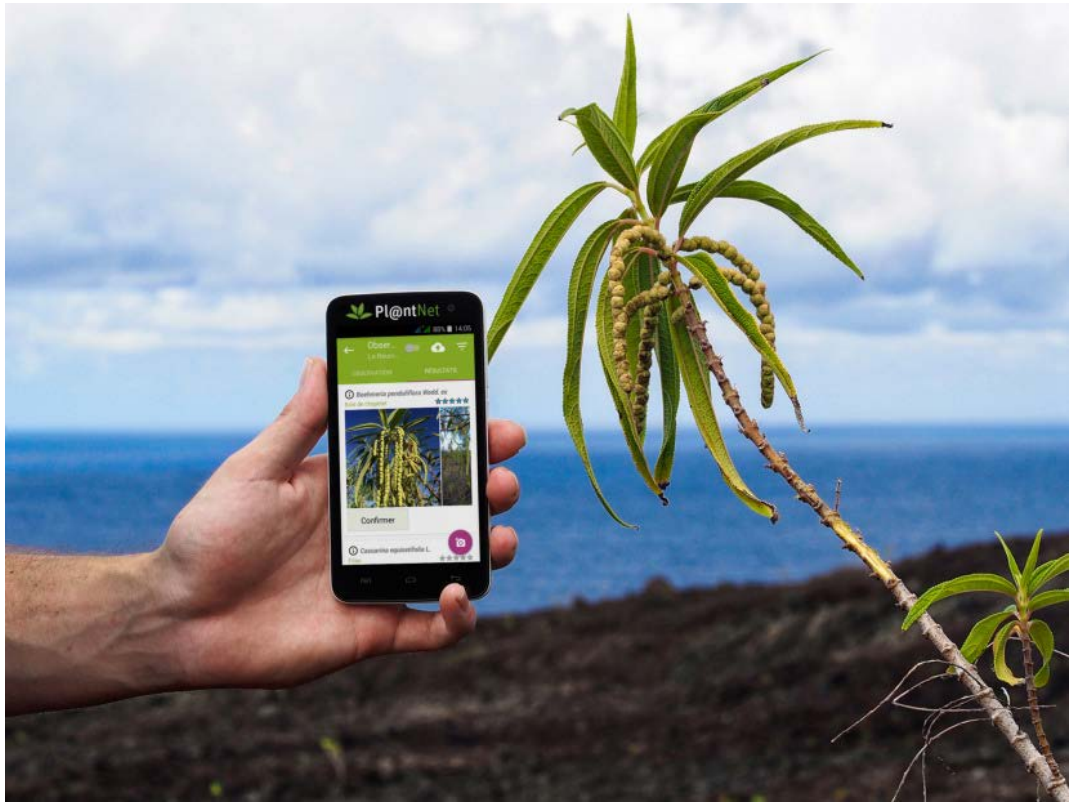
De plus en plus de données



- Drones, satellites
- Pièges photos, suivi acoustique
- Science participative

plantnet.org

De plus en plus de données



plantnet.org

- Drones, satellites
- Pièges photos, suivi acoustique
- Science participative

➡ **Meilleure couverture, meilleur suivi**

L’analyse de données et l’interprétation sont chronophages

- Un ordinateur permet un suivi constant
- Un ordinateur est moins prône à des erreurs de fatigue

Grieshop *et al.* 2012

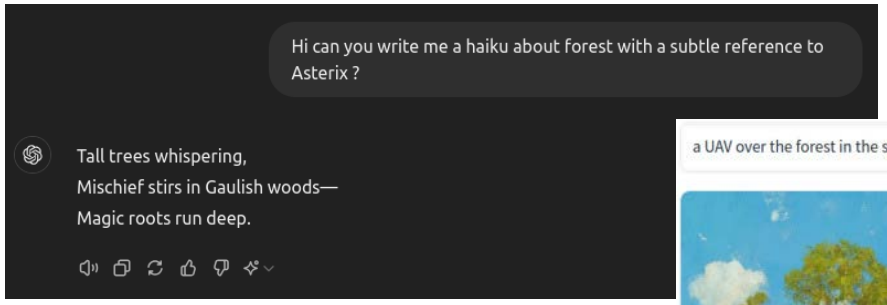
L’automatisation de certaines tâches était impossible jusqu’à il y a quelques années

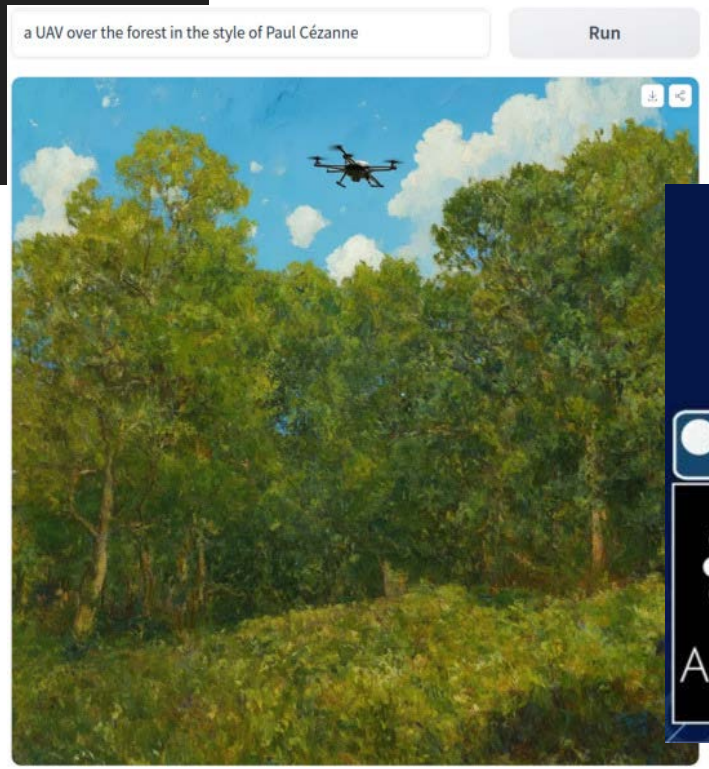
xkcd 1425, 2014

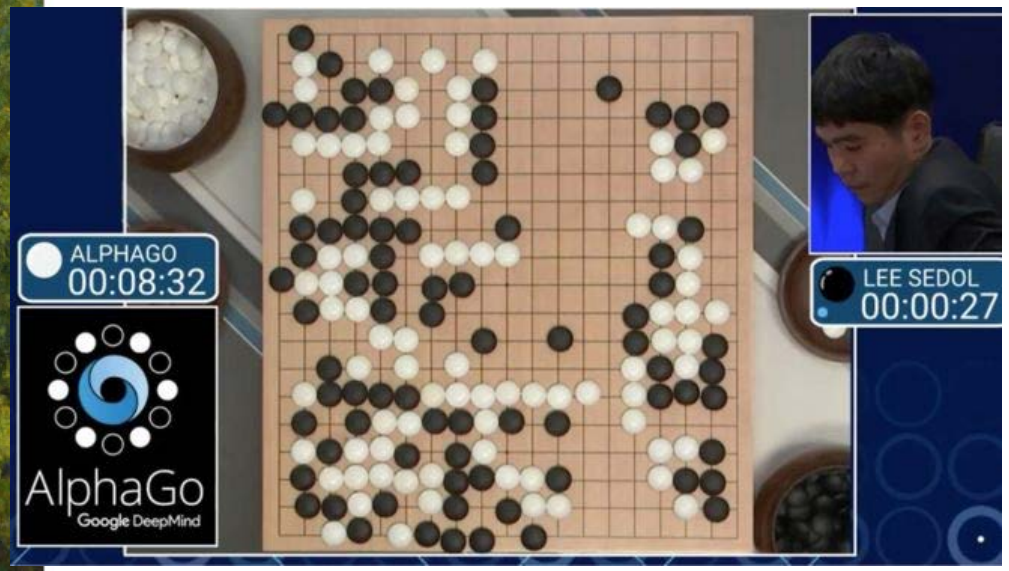
Quels sont les cas où le Deep Learning fonctionne bien ?

(et les autres modèles non)

Modéliser des relations complexes, abstraites et non linéaires







Interpolation et généralisation

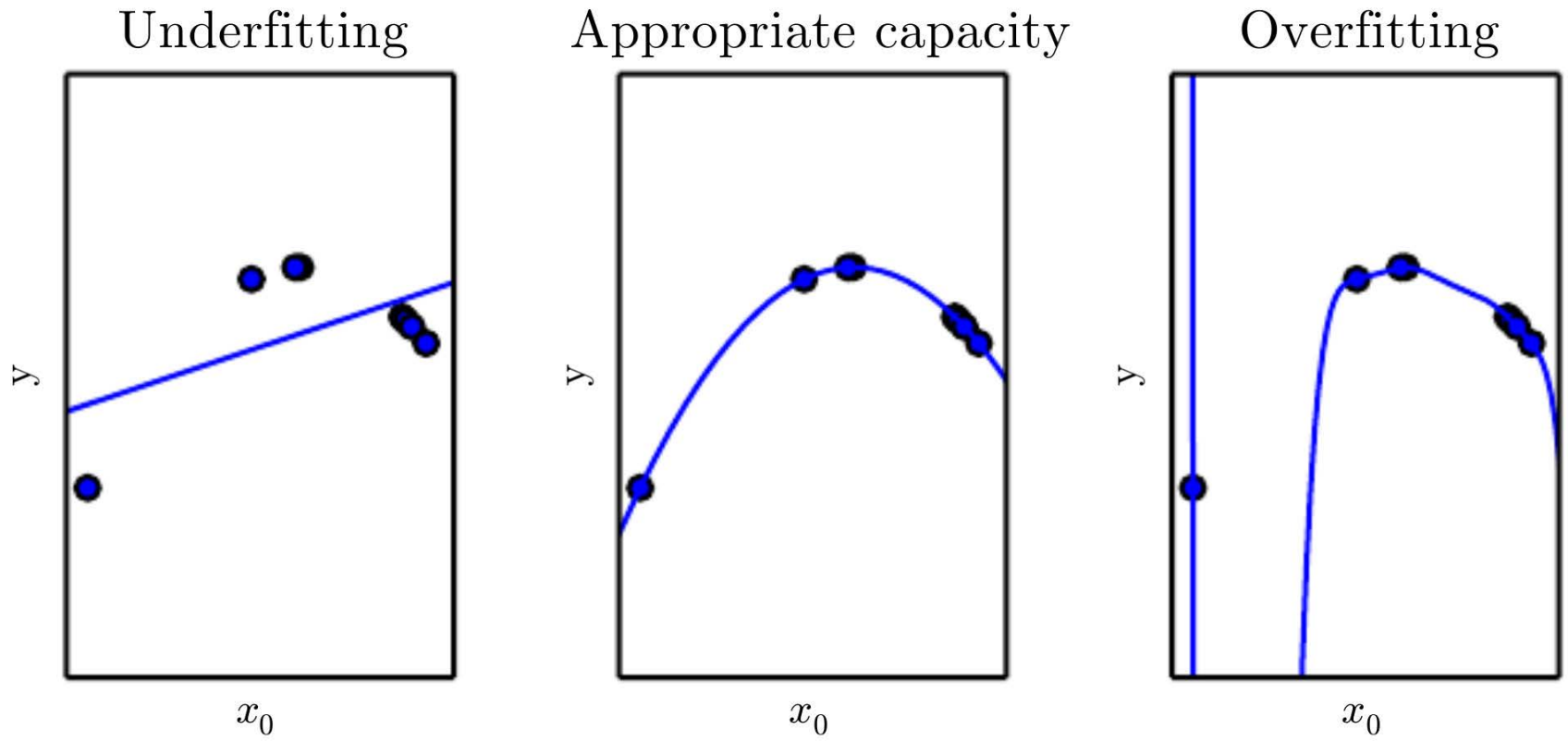
Unonopsis stipitata Diels

PlantClef 2020

Quels sont les cas où le Deep learning ne fonctionne pas bien ?

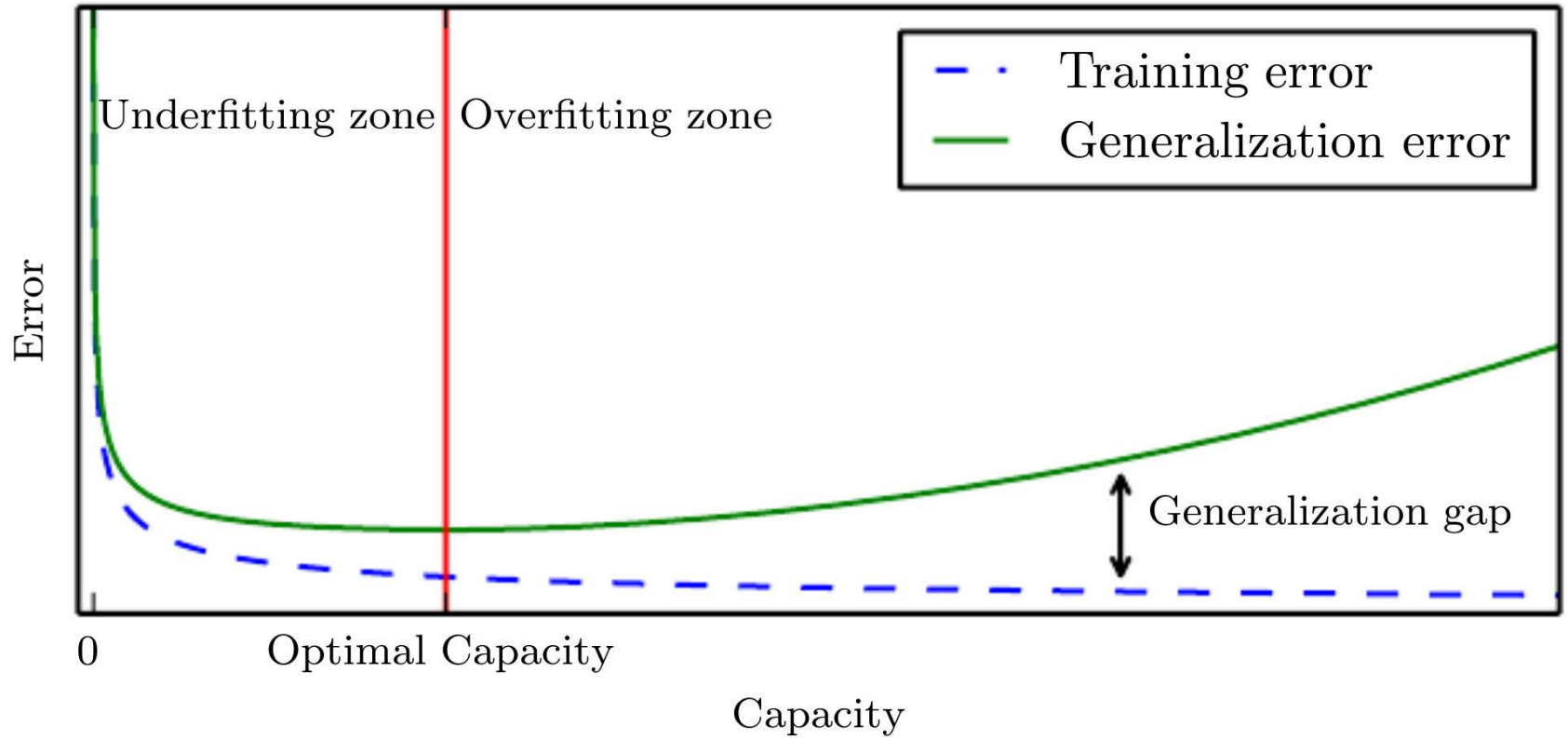
(exemples en écologie)

Pas assez de données



Godfellow *et al.* 2016

Pas assez de données



Godfellow *et al.* 2016

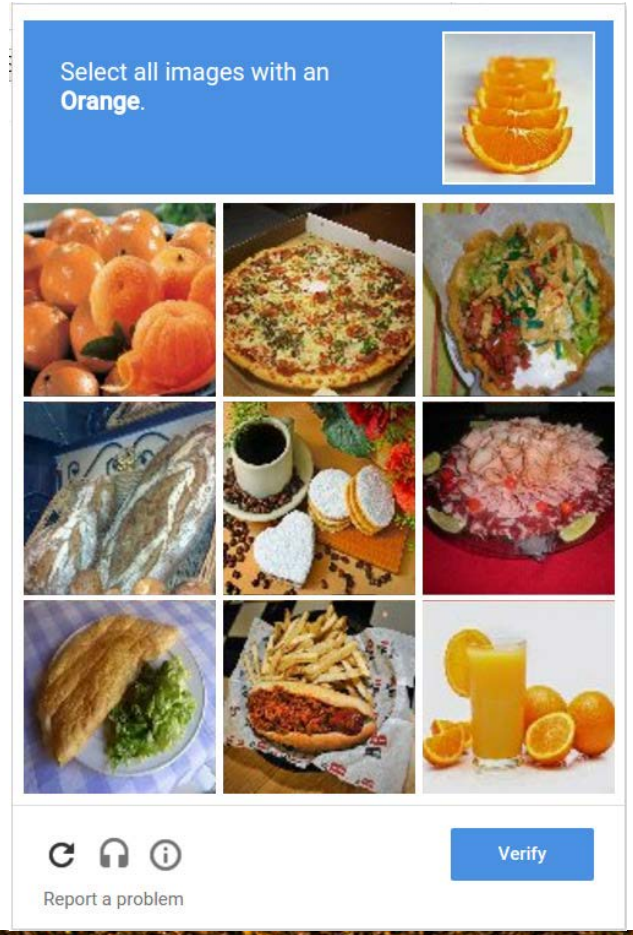
C’est quoi, « pas assez de données » ?

2. What is the small data problem?

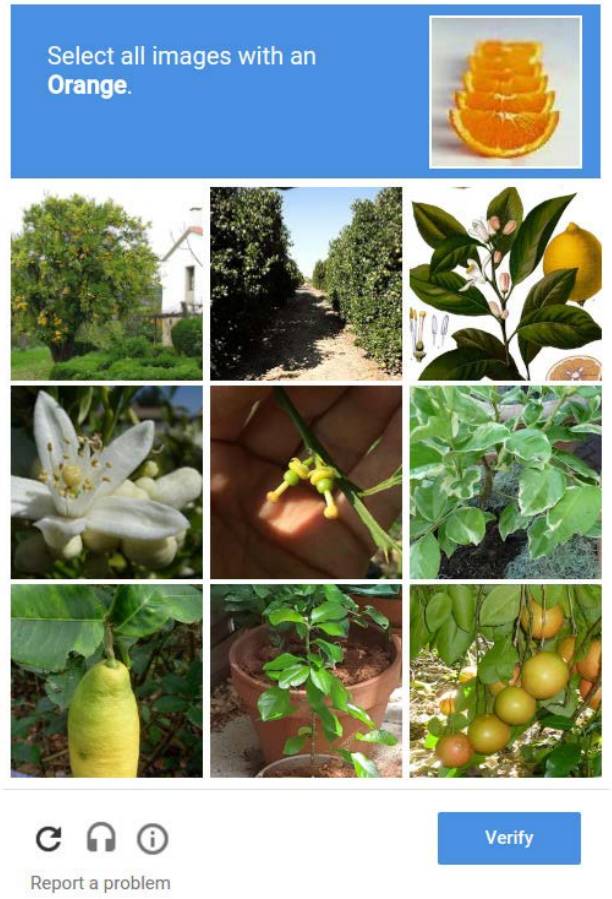
We argue that a dataset can be considered large (not small) when the dataset consists of>100,000 annotated samples, or when it covers the entire probability distribution in a high-dimensional space. For example, there are several free large datasets that can be used for DL: the ImageNet dataset, containing over 14 million annotated images (Russakovsky et al., 2015), the Common Objects in Context (COCO) dataset, containing 330K images, 1.5 million object instances, and 80 object categories (Lin et al., 2015), and the OpenImages dataset, containing over 9 million images (Kuznetsova et al., 2020).

Safanova *et al.* 2023

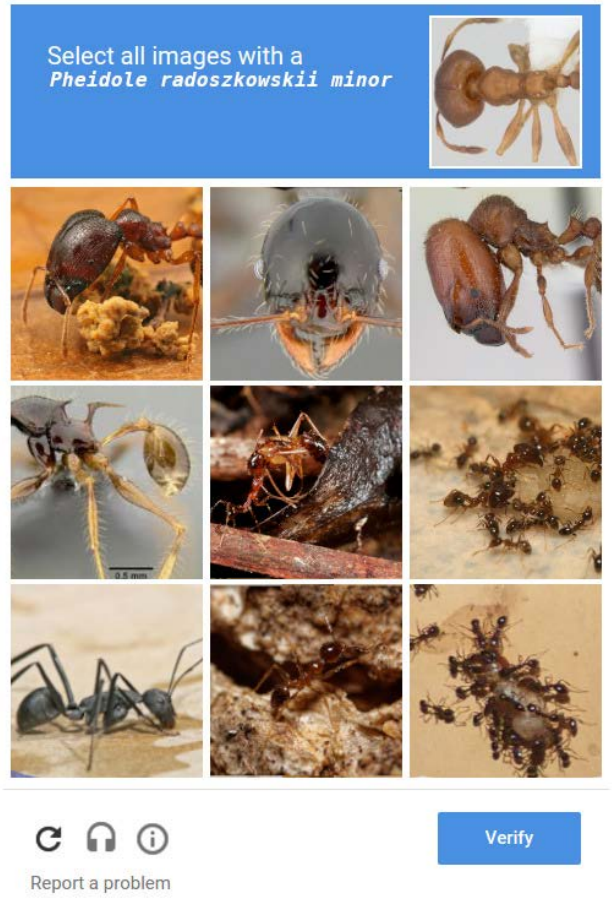
Pas assez de données : besoin d'experts



Pas assez de données : besoin d'experts



Pas assez de données : besoin d’experts



Pas assez de données : acquisition

- Les données peuvent être difficiles à acquérir
- Les métadonnées également
- Qualité des données ?

➡ **Petits jeux de données dans les standards du Deep Learning**

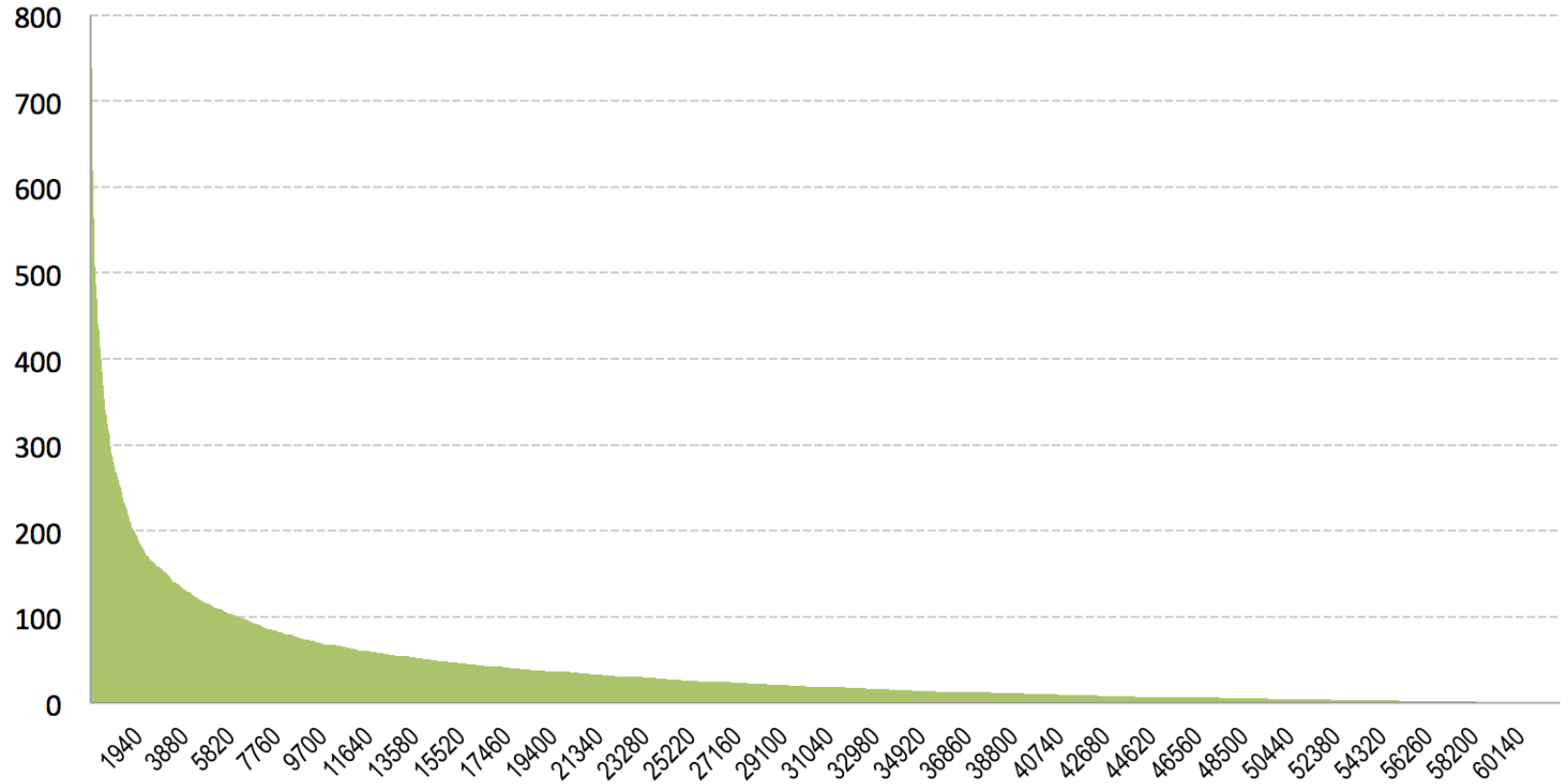
Jeux de données déséquilibrés

Image classification annotations (1000 object classes)

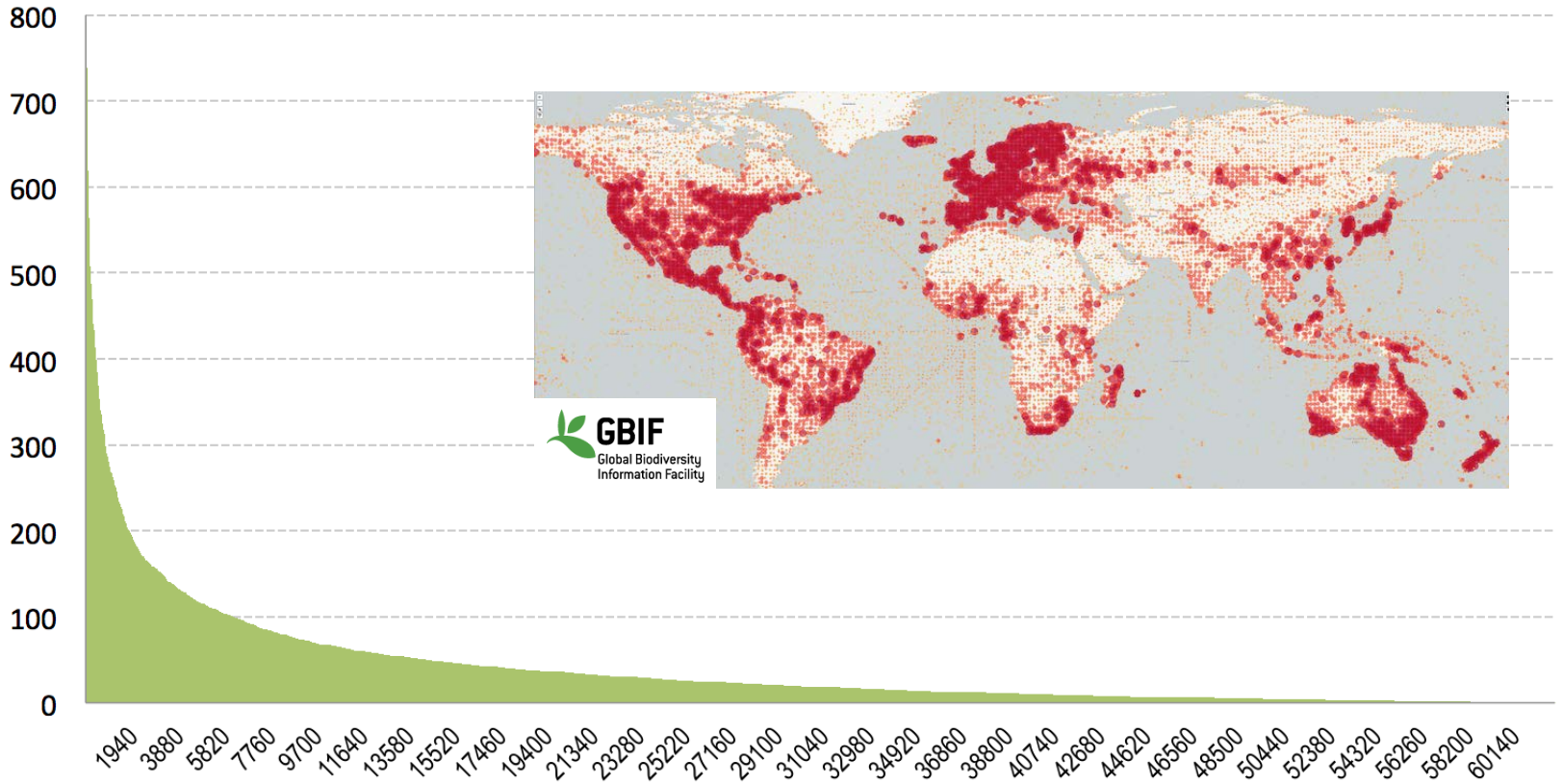
Year	Train images (per class)	Val images (per class)	Test images (per class)
ILSVRC2010	1,261,406 (668-3047)	50,000 (50)	150,000 (150)
ILSVRC2011	1,229,413 (384-1300)	50,000 (50)	100,000 (100)
ILSVRC2012-14	1,281,167 (732-1300)	50,000 (50)	100,000 (100)

Russakovsky *et al.* 2015

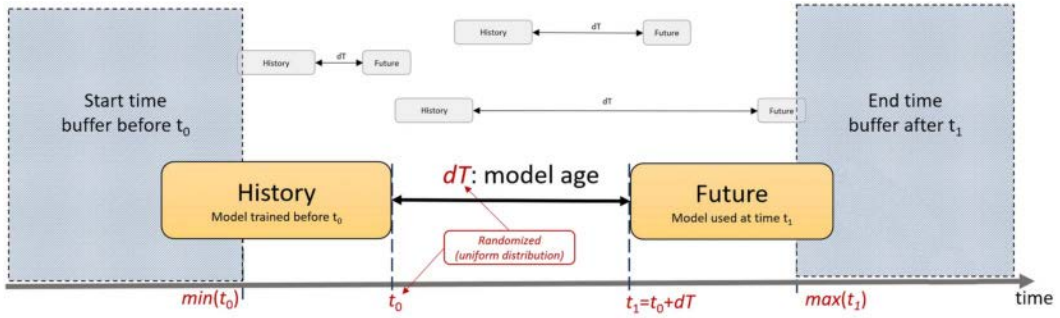
Jeux de données déséquilibrés

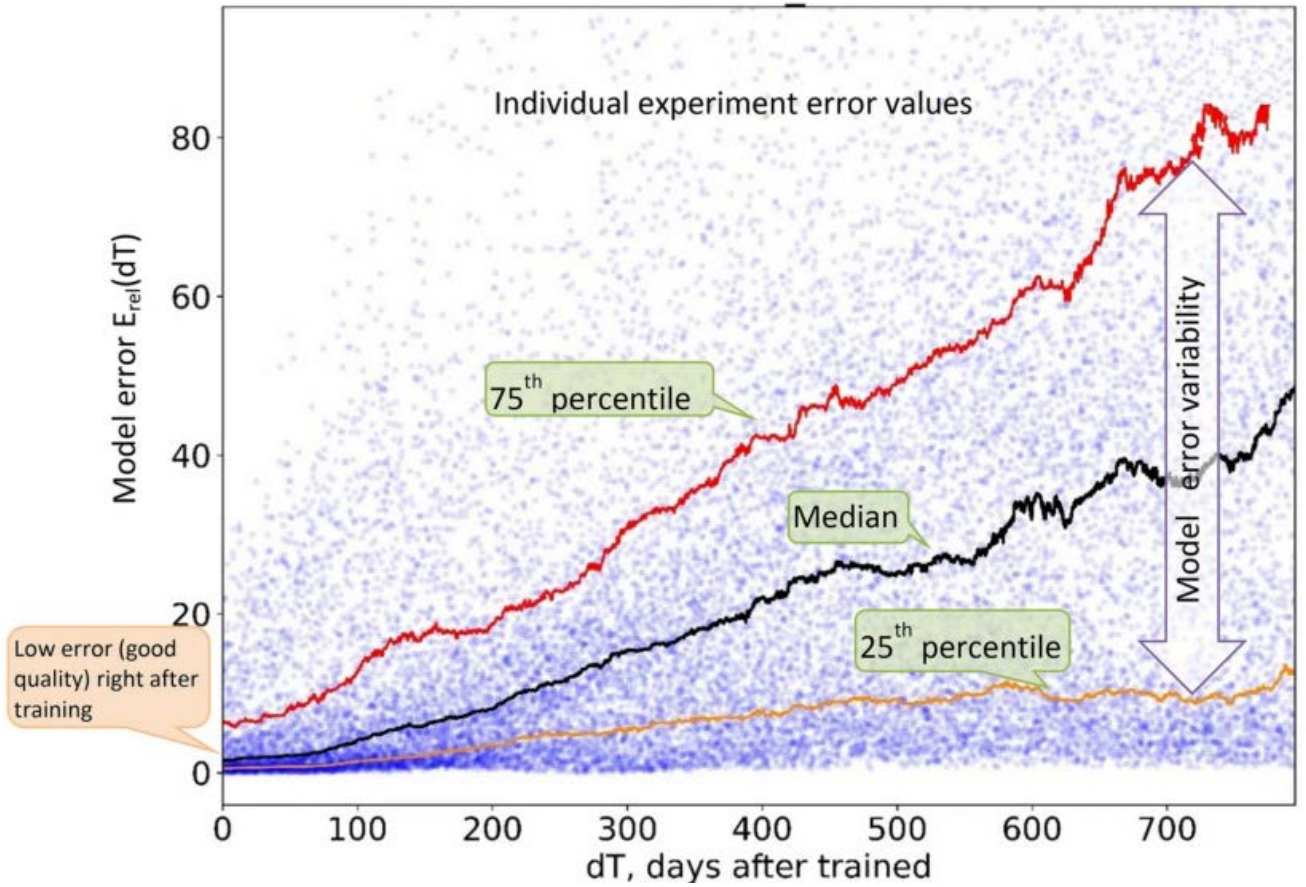


Jeux de données déséquilibrés



Out of distribution : Évolution des modèles





Adapté de Vela *et al.* 2022

Out of distribution : changements globaux

Les écosystèmes vont connaître des conditions sans précédent :

– Conditions climatiques inconnues jusqu’alors
– Espèces invasives
– ...

Arrivée d’une espèce invasive inconnue

Faux Positif

Confusion avec une espèce connue par le modèle

Faux Négatif

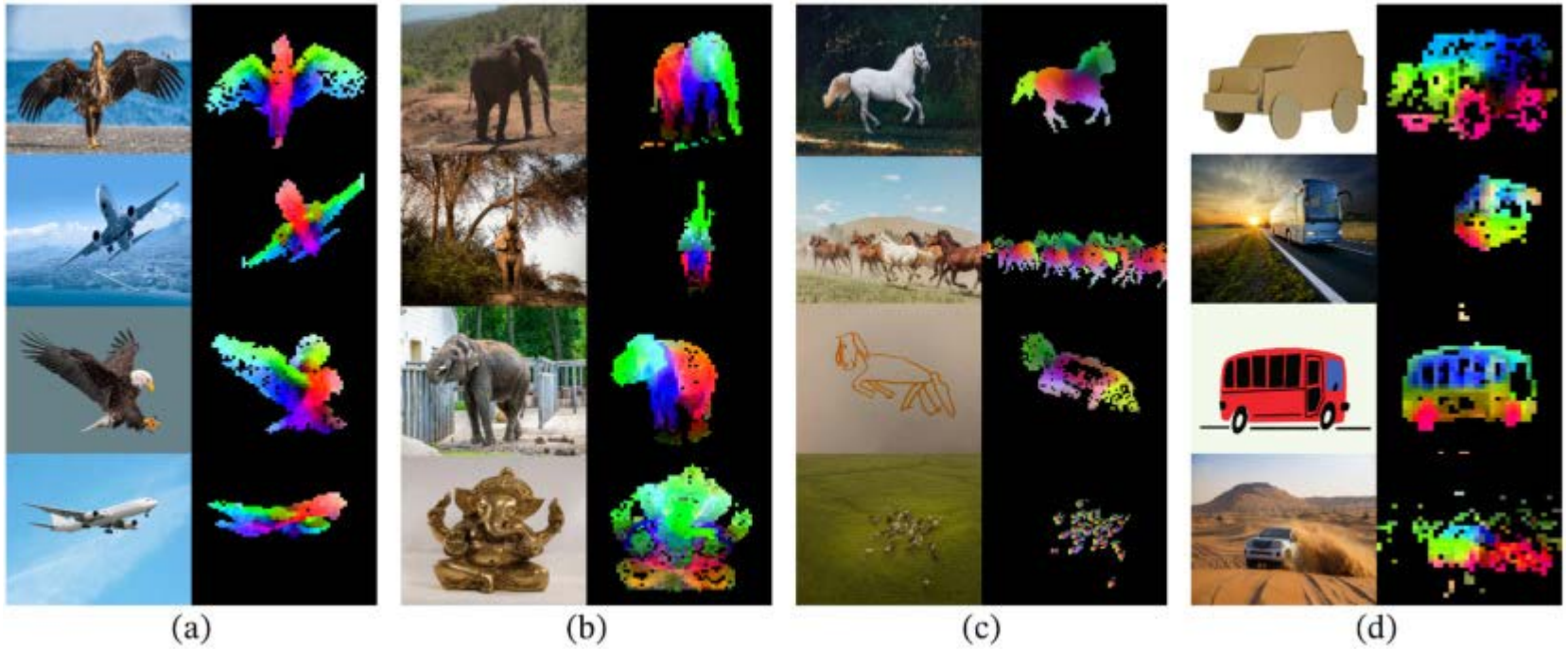
Le modèle ne détecte pas l’espèce invasive

Vérification

Implémentation de vérifications manuelles sur les prédictions à faible confiance

Quelles tendances ?

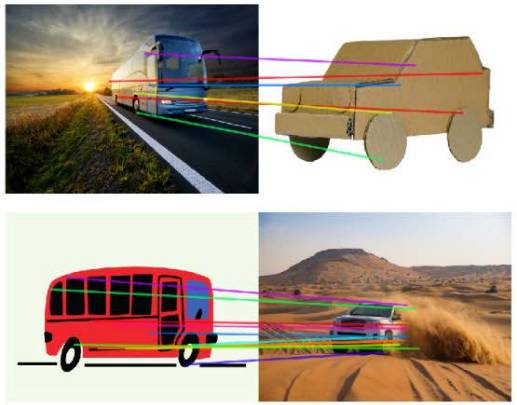
Modèles de plus en plus robustes



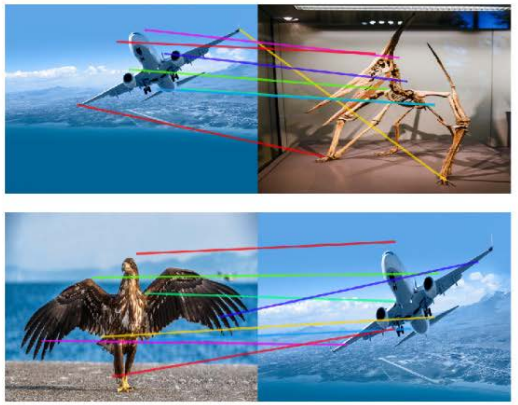
Pré-entraînement des modèles

Oquab *et al.* 2024

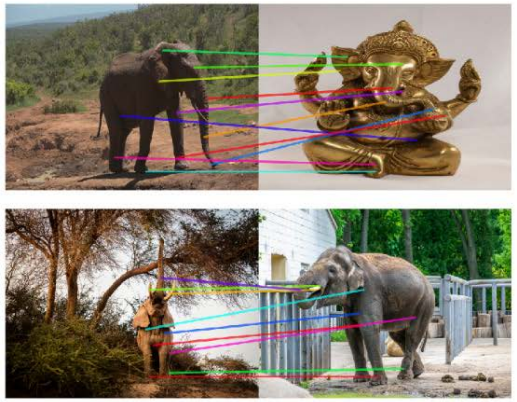
Modèles de plus en plus robustes



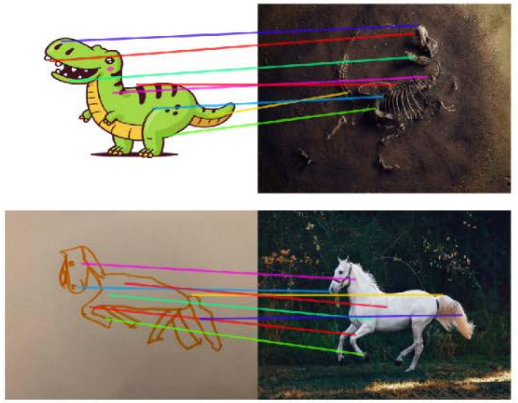
(Vehicles)



(Birds / Airplanes)

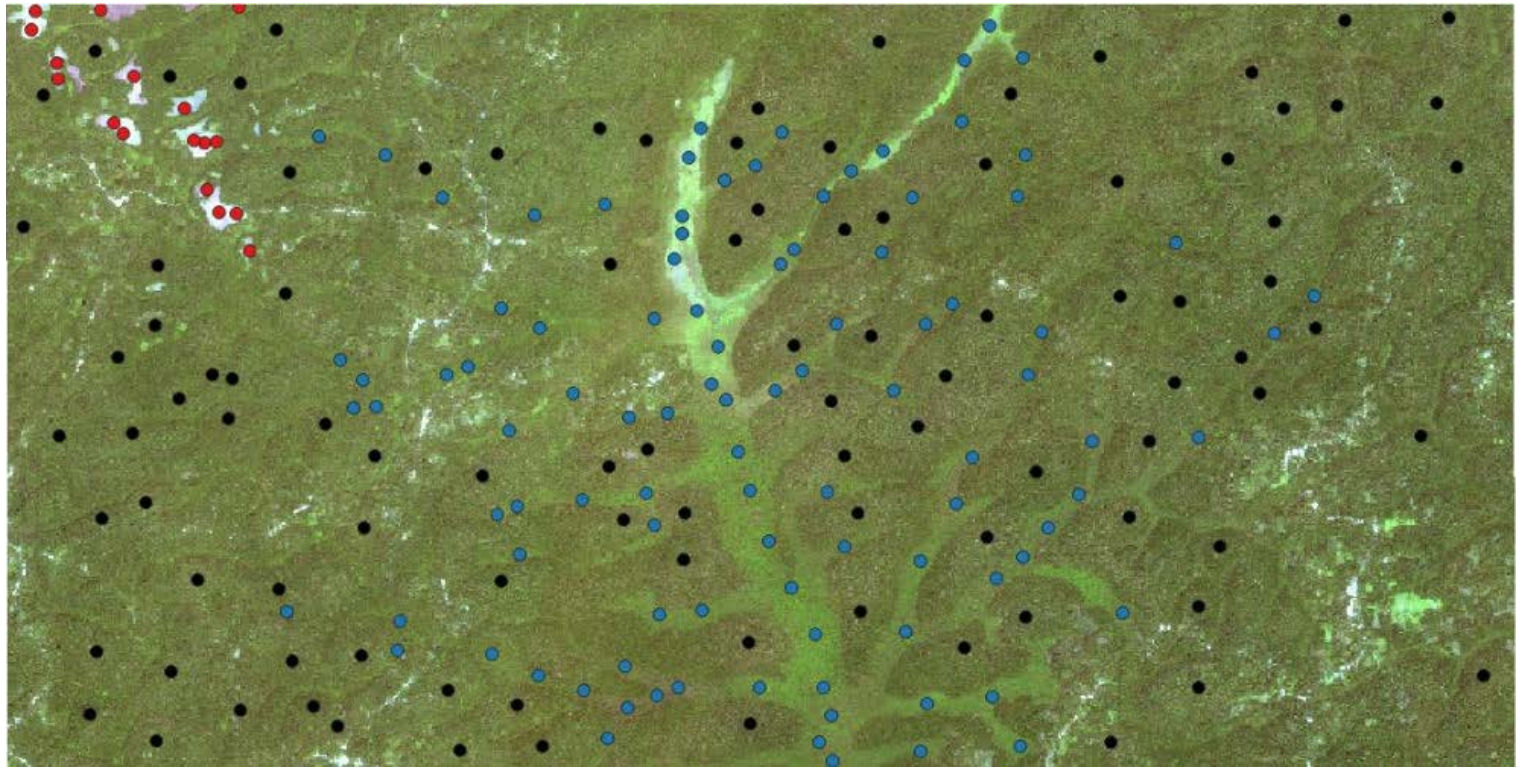


(Elephants)

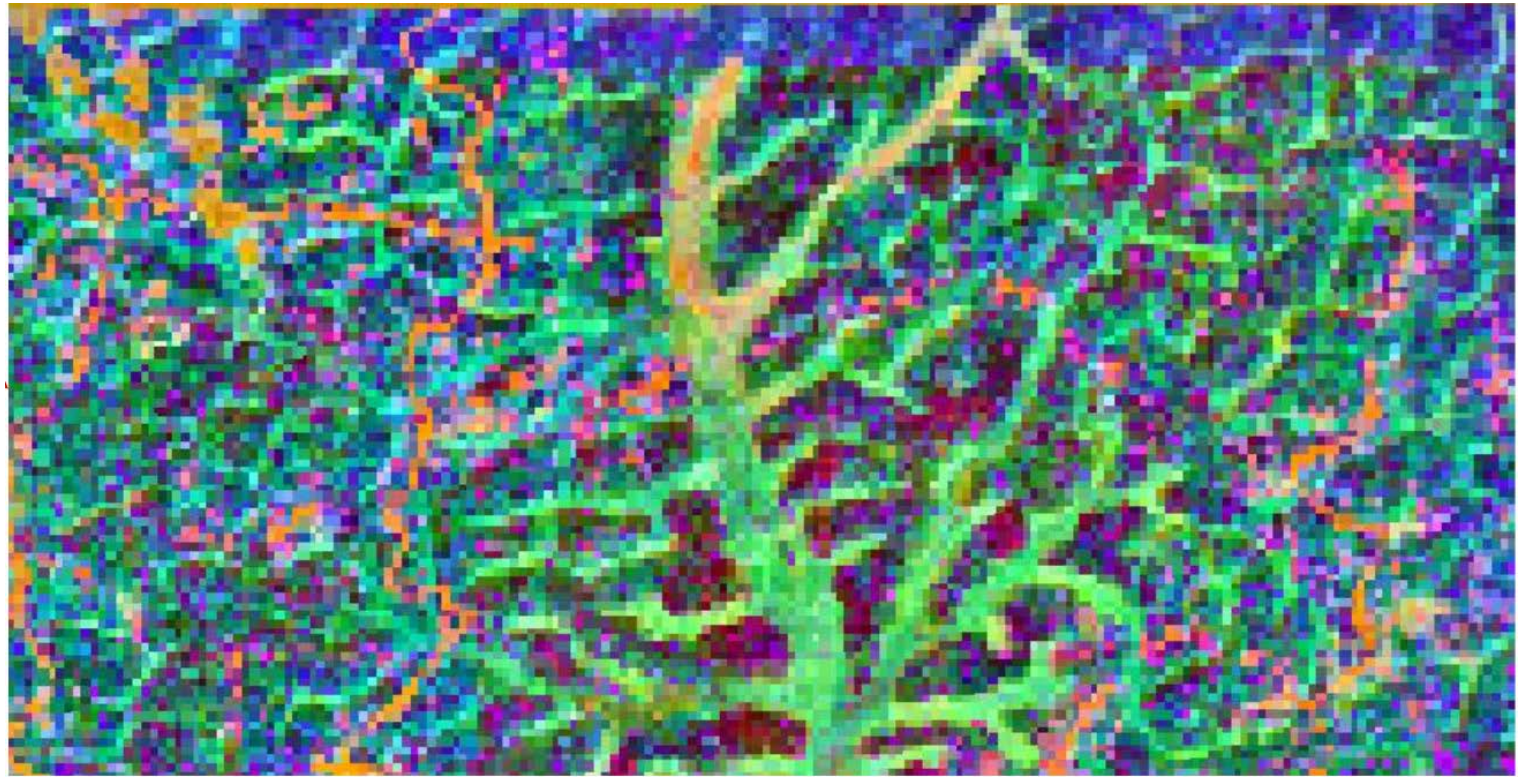


(Drawings / Animals)

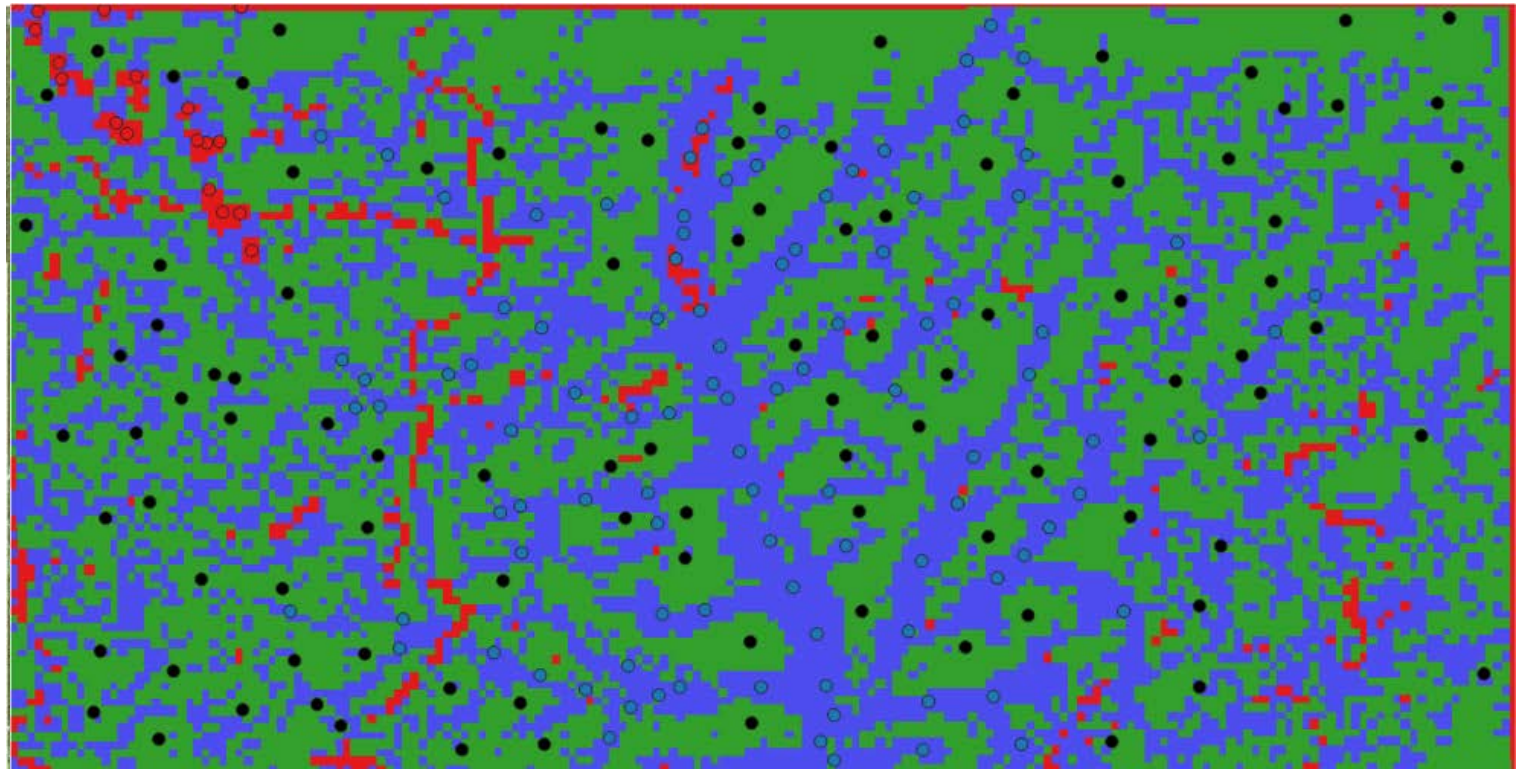
Approches hybrides ?



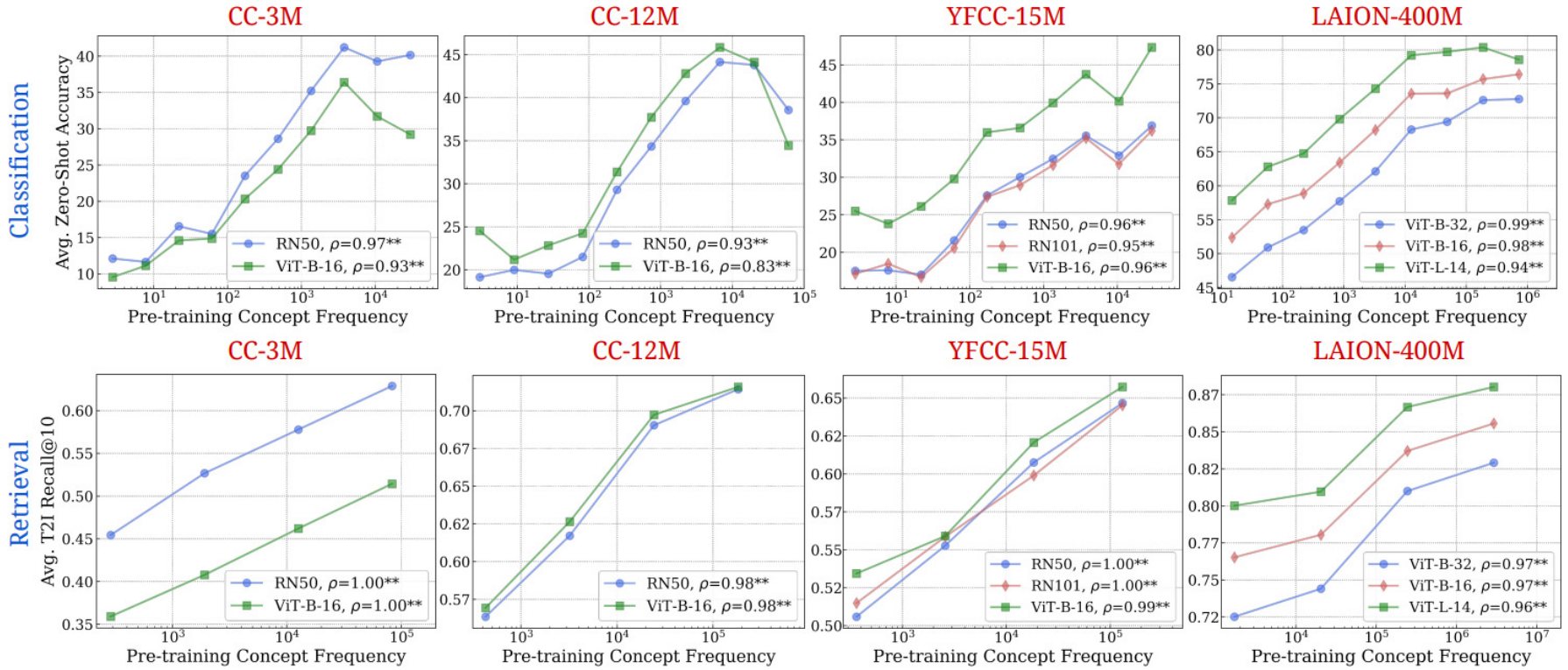
Approches hybrides ?



Approches hybrides ?



Le « zero-shot » a un coût



Udandarao *et al.* 2024

Conclusion

- Beaucoup de potentiel
- Permet de passer à des échelles non-atteintes jusqu’à présent

Mais

Beaucoup de pièges de par la nature des données en écologie

Merci pour votre attention

NUM-DATA

sophie.fortuno@cirad.fr WP2
christian.pichot@inrae.fr WP3
paul.tresson@cirad.fr WP4
fabrice.benedet@cirad.fr WP5

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

Partage, accessibilité et services d'exploitation des données NUM-DATA