



# RÉPUBLIQUE FRANÇAISE

*Liberté  
Égalité  
Fraternité*

Institut national de l'information  
géographique et forestière



CHANGER  
D'ÉCHELLE



RÉPUBLIQUE  
FRANÇAISE

Liberté  
Égalité  
Fraternité

**IGN**  
INSTITUT NATIONAL  
DE L'INFORMATION  
GÉOGRAPHIQUE  
ET FORESTIÈRE

CHANGER  
D'ÉCHELLE

# LIDAR HD ET IA

Données d'apprentissage, réentraînement de modèle & passage en production.

**Webinaire MAGIS du 5/12/2024 – Floryne Roche**

[Floryne.Roche@ign.fr](mailto:Floryne.Roche@ign.fr)



# Sommaire

1. Données d'apprentissage

2. Concentrer la donnée

3. Insérer le modèle dans la chaîne de production

4. Un processus dans le processus

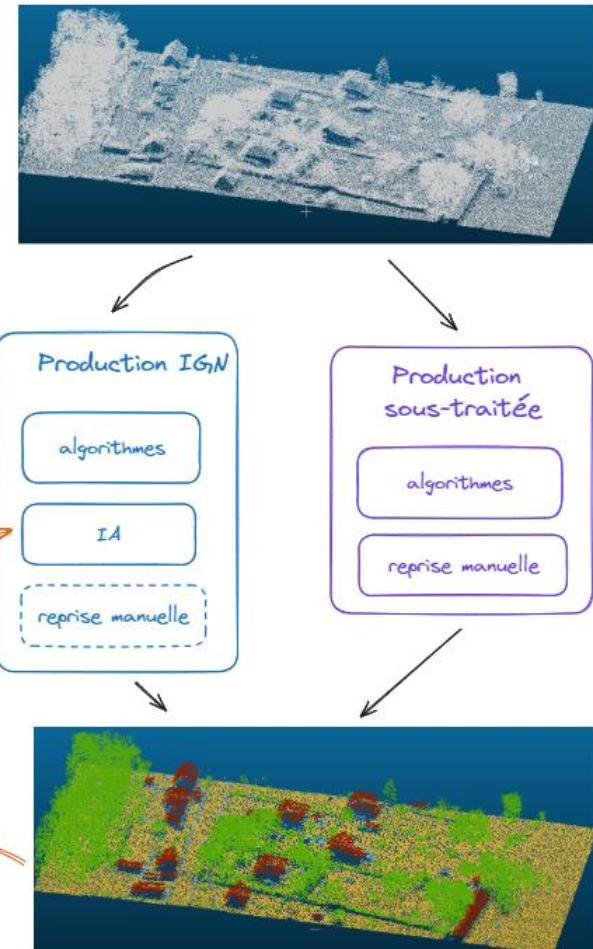
5. Bilan & Perspectives

6. Et encore... quelques liens utiles

# 1. Le nerf de la guerre en IA : la donnée

# Données d'apprentissage

- La donnée finale est une donnée annotée, qui peut servir de donnée d'apprentissage
- Constitution de la donnée d'apprentissage:
  - Effort initial sur la constitution en interne IGN
  - Utilisation possible de données de sous-traitance obtenues à partir d'un processus automatique et de reprises manuelles.



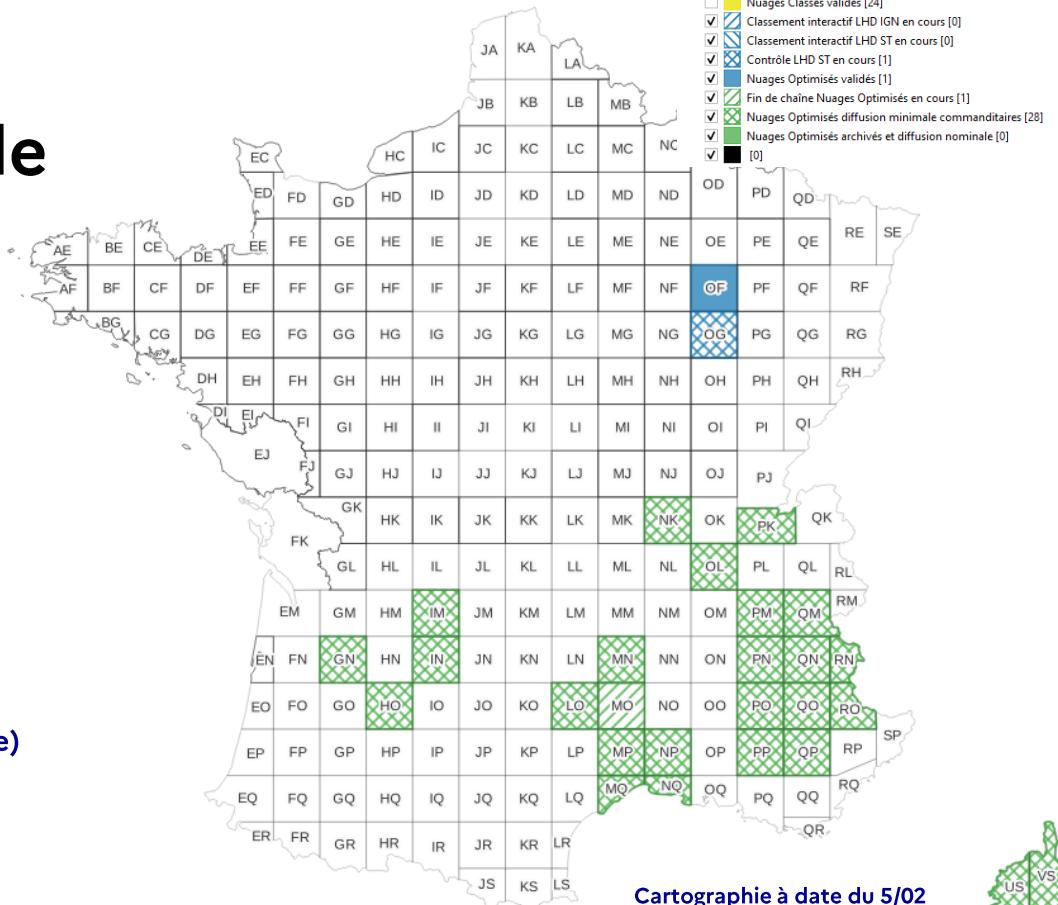
# De la donnée disponible

## Données produite:

- Donnée classée (= automatique)
- Donnée optimisée (avec reprise manuelle)

29 blocs de données sous-traitée, soit 40 000 km<sup>2</sup> de données

=> c'est plus de km<sup>2</sup> que nécessaire! (150 km<sup>2</sup> nécessaire)



Cartographie à date du 5/02



## 2. Concentrer la donnée, comment?

Des outils pour concentrer la données

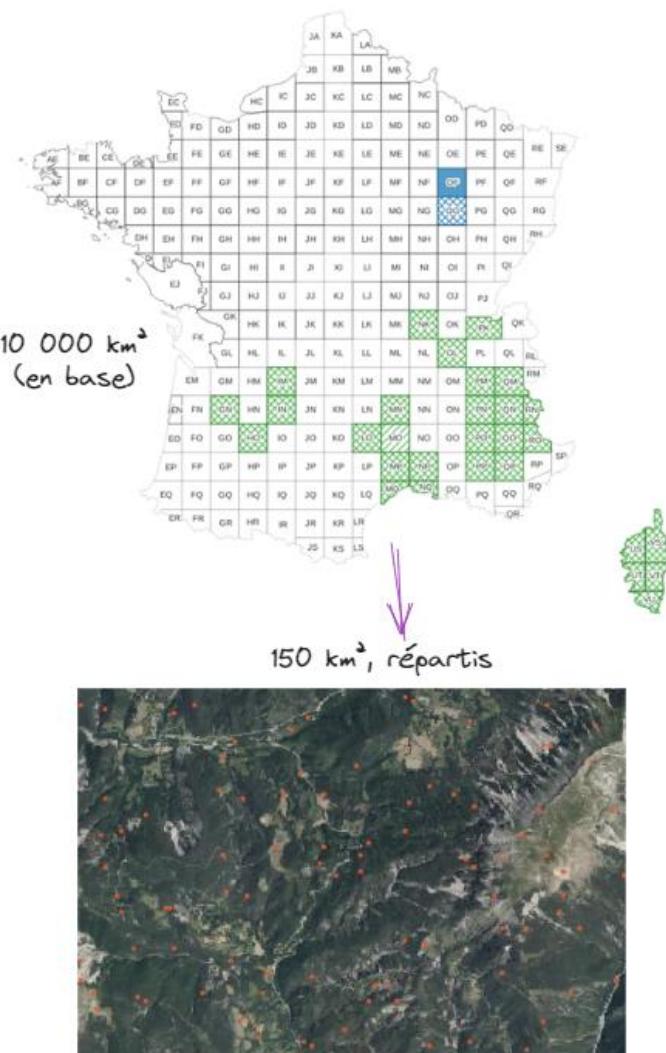
# Concentrer la donnée

**Le besoin :** obtenir 150 km<sup>2</sup> soit 60 000 patchs de 50 m x 50m, intelligemment répartis

- Sélectionner les patchs (répartition spatiale, critères)
- « Construire » les patchs (extraction, colorisation)

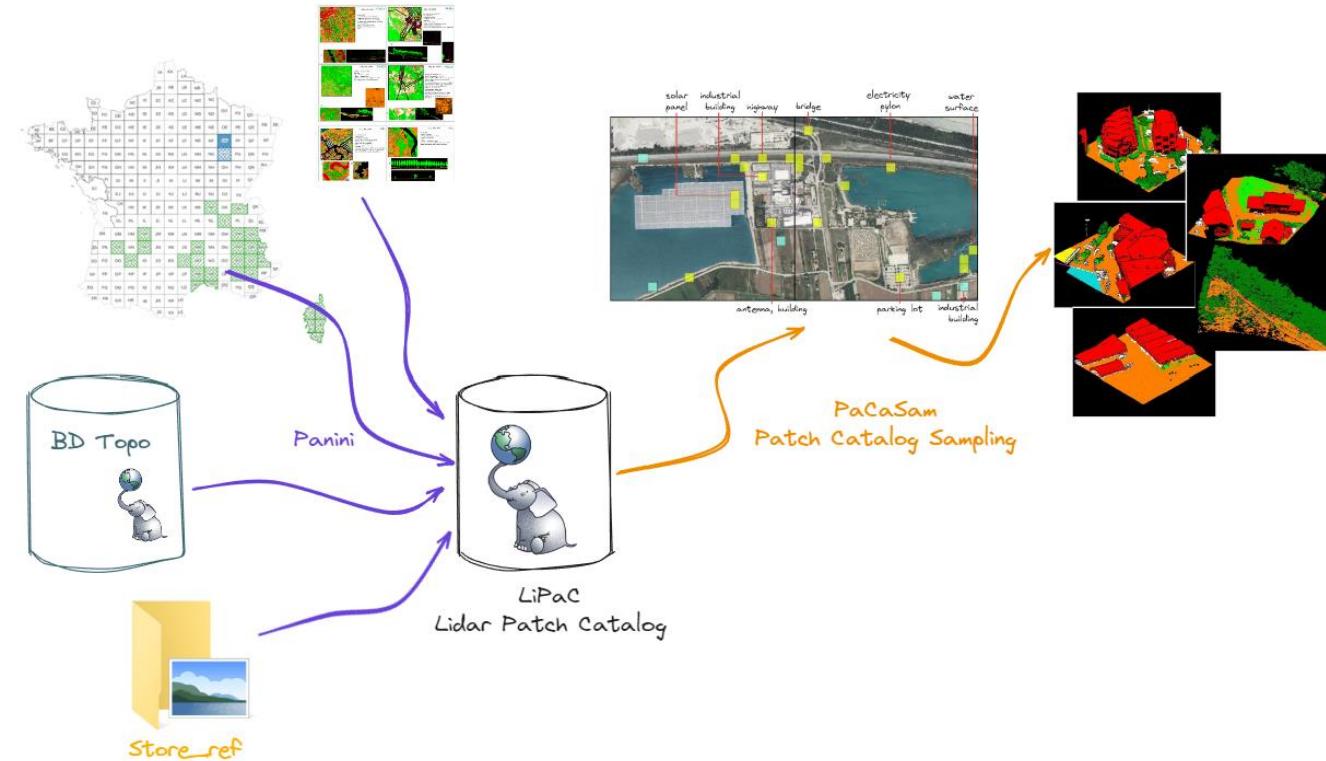
**La réponse :** 3 outils

- Une base de données (Lipac)
- Un outil pour la peupler (Panini)
- Un outil d'extraction (PaCaSam)



# Échantillonner

- Sur le territoire disponible/ utile
- Sélectionner selon des critères
  - Issus de la données (nombre de points par classe)
  - Par connaissance du territoire (croisement BD Topo)
  - + quelques autres contraintes (test, contrôle, ...)



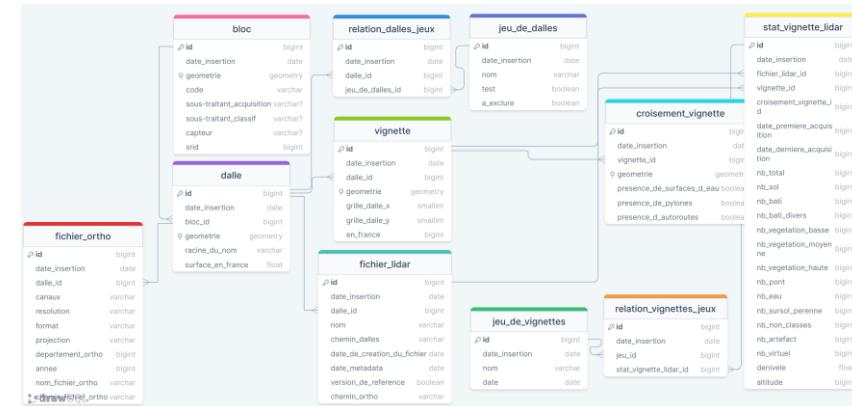
# Cataloguer la donnée : Lidar-Patch-Catalogue (LiPaC)

Une base de données PostGIS extensible

- ✓ 5000 à 20 000 km<sup>2</sup>
- ✓ 2M à 8M patchs (50m x 50m)



- Attributs liés à la dalle
  - Donnée d'acquisition: capteur, date d'acquisition ...
  - Info de localisation du nuage de points & de l'ortho associée
- Attributs intrinsèques au nuage
  - nb de points dans chaque classe,
  - Altitude
  - ...
- Attributs calculés
  - Données calculées (pente, ...)
  - Présence d'objets par croisement (éolienne, autoroutes, ...) avec la BD Topo ®



# Cataloguer les vignettes : Panini

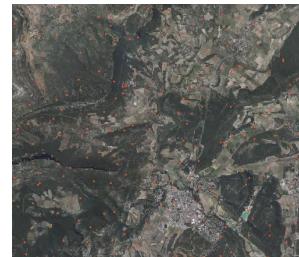
**Objectif:** peupler Lipac

- Insérer des nouvelles données
- Déetecter les éventuelles incohérences (chemin d'accès)
- Associer les orthophotos associées
- Peupler la base pour assurer la cohérence avec l'IA (données réservées aux test, ne devant être utilisées ni pour l'entraînement, ni pour la validation)

# Extraire la donnée d'apprentissage (PaCaSam)

- Patch-Catalogue-Sampling  un outil python pour
  1. **Se connecter** au catalogue Lipac
  2. **Echantillonner** les patchs de données (via des critères, les histogrammes ou des contraintes spatiales)
  3. **Extraire** les données et les préparer pour l'entraînement.
- Adapté à nos besoins et disponible sur [Github](#).





# Les résultats

- Données prototype: 150 x 1km<sup>2</sup>
- 1ère version d'un jeu de données simple 150 km<sup>2</sup> en 60 000 vignettes de 50m x 50m
  - Amélioration globale avec un volume de données identique (~150km<sup>2</sup>)
  - Meilleures performances sur les classes rares

Critère	Prévalence minimale désirée
Nombre de points pont $\geq 50$	2%
Nombre de points sursol pérenne $\geq 50$	2%
Nombre de points bâti $\geq 500$	10%
Dénivelé entre 30m et 45m	1%
Dénivelé $\geq 45m$	1%
Altitude entre 1000m et 2000m	0.5%
Altitude $\geq 2000m$	0.5%

Classes rares dans le jeu de données précédent

Model benchmarking on eval67								
Model ID	Unclassified	Ground	Vegetation	Building	Water	Bridge	Permanent structures	Mean IoU
proto151_V2.0_epoch_100_Myria3DV3.1.0	16%	90%	88%	86%	41%	27%	0%	50%
20230930_60k_basic_targetted-epoch_037	19%	90%	91%	84%	62%	39%	46%	62%

# Les résultats

- Données prototype: 150 x 1km<sup>2</sup>
- 1ère version d'un jeu de données simple 150 km<sup>2</sup> en 60 000 vignettes de 50m x 50m
- FRACTAL (FFrench ALS Clouds from TArgetted Landscapes) : 250 km<sup>2</sup> en 100 000 vignettes de 50m x 50m
  - Meilleures performances sur les classes rares
  - Amélioration sur les camions et sur les serres

Motivation	Scene Type	Definition	Target (%)
Classes	BUILD	building ≥ 500 pts	8
	BUILD_GREENHOUSE	greenhouse (BD TOPO)	1 10
	BUILD_BIG	non-residential building (BD TOPO)	1
	BRIDGE	bridge ≥ 50 pts	5 5
	WATER	eau ≥ 50 pts	4
	WATER_SURFACE	water area (BD TOPO) & eau ≥ 50 pts	1 5
	PERMSTRUCT	permanent structure ≥ 50 pts	3
	PERMSTRUCT_PYLON	pylon (BD TOPO) & permanent structure ≥ 50 pts	1 5
	PERMSTRUCT_ANTENNA	antenna (BD TOPO) & permanent structure ≥ 50 pts	1
	OTHER	unclassified ≥ 250 pts	3
Landscapes	OTHER_PARKING	parking lot (BD TOPO) & unclassified ≥ 250 pts	1 5
	OTHER_HIGHWAY	highway (BD TOPO) & unclassified ≥ 400 pts	1
	FOREST	high vegetation ≥ 90% of points	5
	HIGHSLOPE1	35 m ≤ elevation gain < 45 m	2
	HIGHSLOPE2	elevation gain ≥ 45 m	2
	MOUNTAIN	elevation ≥ 1000 m	4
	WATER_ONLY	water ≥ 50 pts & ground = 0 pts	1
	SEASHORE	-10 ≤ elevation < 10 & water ≥ 50 & ground ≥ 100 pts	1
	URBAN	building ≥ 25% of points	5

Classes rares dans le jeu de données précédent

Model benchmarking on eval67								
Model ID	Unclassified	Ground	Vegetation	Building	Water	Bridge	Permanent structures	Mean IoU
proto151_V2.0_epoch_100_Myria3DV3.1.0	16%	90%	88%	86%	41%	27%	0%	50%
20230930_60k_basic_targetted-epoch_037	19%	90%	91%	84%	62%	39%	46%	62%
FRACTAL	52%	92%	94%	90%	91%	57%	62%	78%

### 3. Insérer le modèle dans la chaîne de production

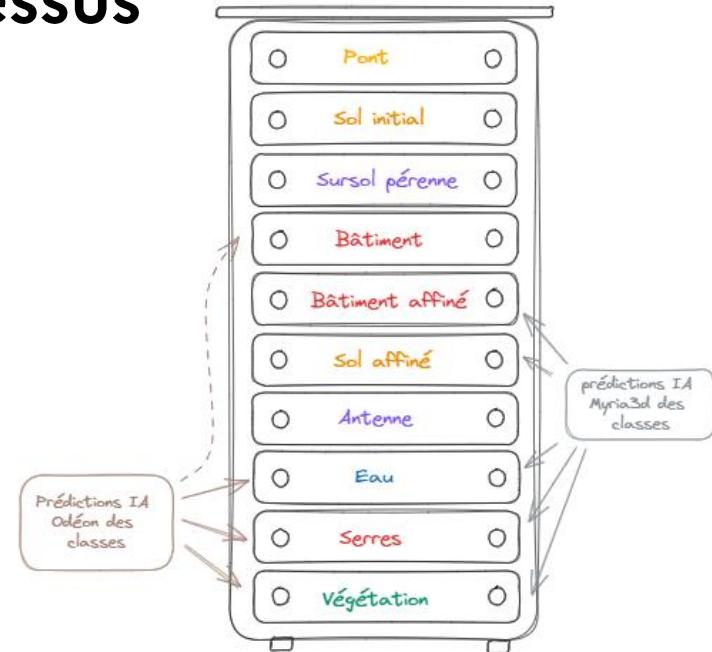
# Un modèle imbriqué dans le processus

Un **processus hybride** (algorithme, données vecteur, IA 2D, IA 3D)

=> des **résultats interdépendants**

Des évolutions – et des analyses d'évolution – au sein de chaque brique.

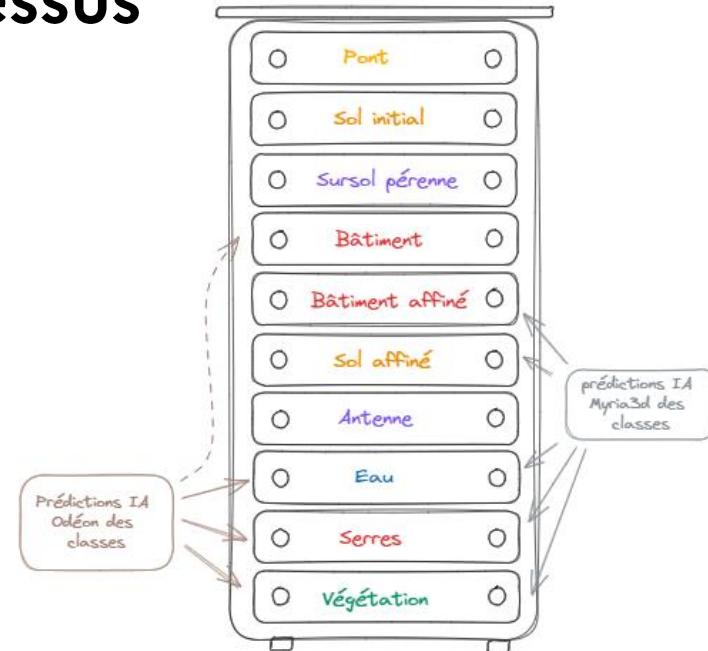
=> Comment analyser que le **processus complet** évolue dans le bon sens quand on modifie une brique et en particulier le modèle IA?



# Un modèle imbriqué dans le processus

Objectif d'une évolution de processus: améliorer **le résultat final**

- Analyser les métriques du modèle ne suffit pas, c'est le résultat final qui doit être analysé et amélioré
- Les métriques du modèle pourraient s'améliorer sans que le processus final ne le soit, voire même en le dégradant.
- Algorithmes conçus localement : comment s'assurer qu'il n'y a pas dégradation ailleurs?

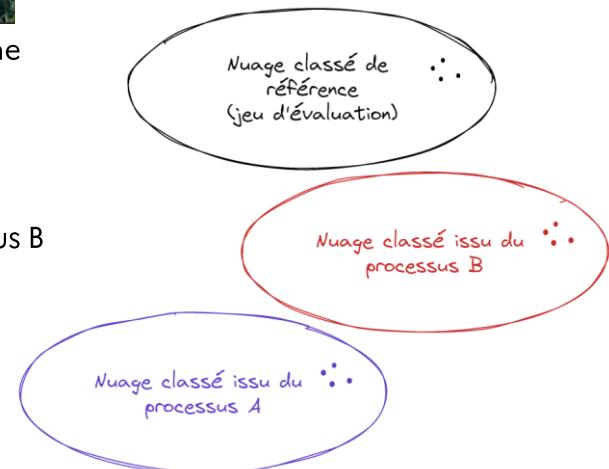


# Quelle version du processus?

⇒ Besoin d'un processus d'évaluation des évolutions:

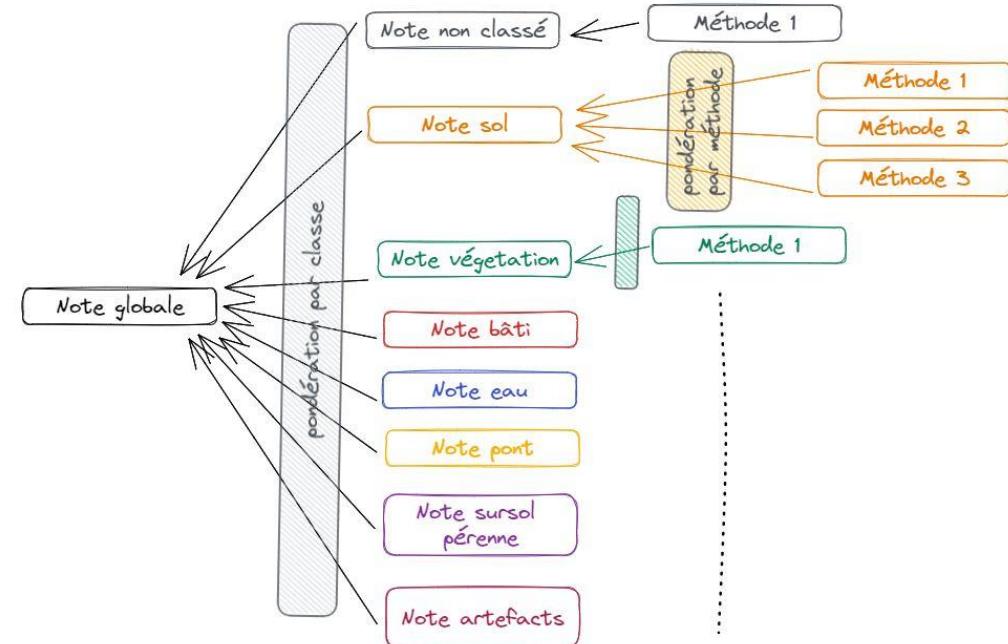
**COCLICO** : COmparaison de CLassIfication par rapport à une référence COnnue

- Comparer les résultats de 2 processus
- Une **note relative** pour savoir si c'est les résultats du processus A ou ceux du processus B qui sont les plus proches de la donnée de référence.



# Evaluer les évolutions

- Un jeu de données d'évaluation servant de référence
  - 71 critères de représentativité du paysage
  - Une liste de 67 dalles avec leur descriptif
  - Des corrections interactives pour avoir une donnée de référence
- COCLICO
  - Des métriques (comparaison point à point, proximité planimétrique, altimétrique, présence d'objets, ...)
  - Des pondérations par classe

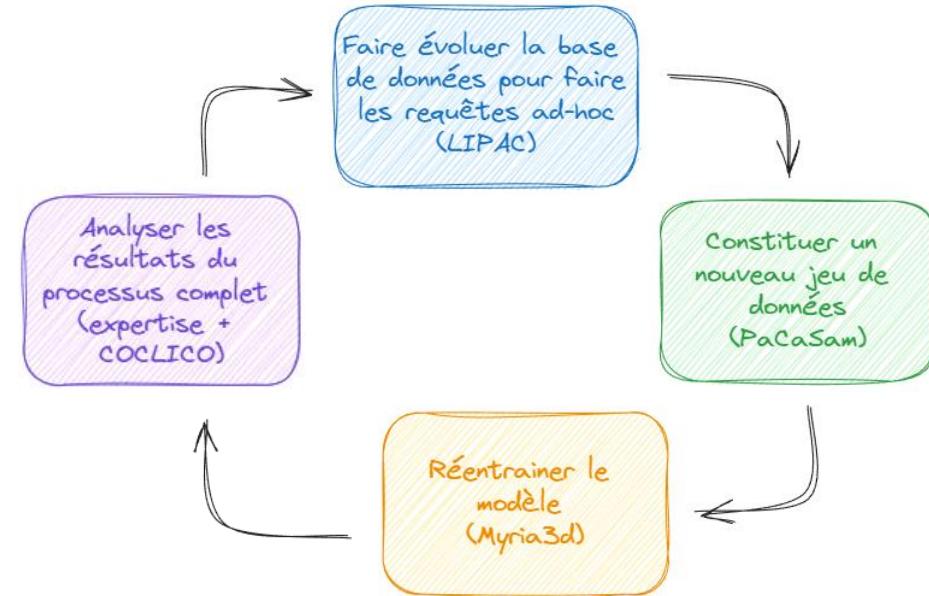


## 4. Au cœur du processus

# L'amélioration continue côté processus

## Intérêt complémentaire:

- Encapaciter l'équipe chargée de faire évoluer le processus pour lui donner les moyens de faire également évoluer le processus IA.
- Cycle itératif d'amélioration du processus



## 5. Bilan & Perspectives

# Bilan & Perspectives

## Bilan

- L'amélioration des modèles passe aussi par l'amélioration des jeux de données.
- Des meilleurs résultats (données issues de PaCaSam) à jeu de données de taille constante
- Des outils pour agir sur les données d'entraînement et améliorer les modèles par ce biais
- Identifier les défauts des modèles (et donc améliorer les jeux de données) nécessite une connaissance du produit et du territoire.
- Une méthode et un outil qui « sécurise » les évolutions du processus de classification.

## Perspectives

- Davantage de données en base (DROM?)
- Des évolutions des données d'entraînement – et donc des modèles – au plus proche de la définition du processus

## A noter

- Importance de la qualité des données
- Importance d'évaluer le processus final, et pas uniquement le modèle.
- Utilisation de la connaissance du territoire via la BD Topo et le Lidar. D'autres descriptions du territoire possible (OCS, BD Forêt, ...).
- D'autres usages pour Lipac & PaCaSam ? D'autres produits? Une sélection pour annoter? (attention à la difficulté d'annoter des confettis).

## 6. Et encore... quelques liens utiles

# POC forêt : Les résultats

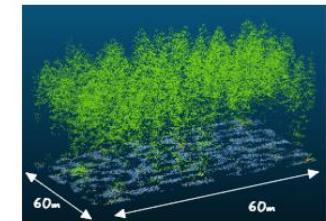
**Problématique:** L'IA appliquée au Lidar HD apporte-t-il quelque chose dans la distinction des essences ?

## Cadrage

- Détection de l'espèce sur une **vignette 50 x 50m de forêt fermée**, supposée composée d'une seule **essence pure**.
- Expérimentation en 2 temps: hiver 2022-2023 pour la constitution du jeu de données, hiver 2023-2024 pour la reprise des données, adaptation de l'architecture du modèle et apprentissage.

## Livrables

- jeu de données [IGNF/PureForest · Datasets at Hugging Face](#)
- Data paper [\[2404.12064\] PureForest: A Large-Scale Aerial Lidar and Aerial Imagery Dataset for Tree Species Classification in Monospecific Forests](#)



## Résultats

- Un jeu de données utilisé par l'équipe Forêt pour initier les travaux sur BD forêt niveau 3
- Reprise et adaptation de Myria3D.
- Intérêt de l'altitude (présente en Lidar), meilleurs résultats sans la couleur issue de l'ortho : réelle complémentarité Lidar - Image
- Des résultats meilleurs que le modèle orthoimages & Sentinel 2 multi-temporel à date en décembre 2023

## Perspectives

- Reprise des résultats par l'équipe «forêt»
- Possibilité de transfert de connaissance ? (à l'arbre, forêt ouvertes, mixtes)

# Quelques liens utiles

- PaCaSam, pour échantillonner le jeu de données (code disponible sur [Github](#))
- FRACTAL, FRench ALS Clouds from TArgetted Landscapes : le jeu de données issu des outils et publié au printemps 2024
  - Le dataset : <https://huggingface.co/datasets/IGNF/FRACTAL>
  - Le data paper: <https://arxiv.org/abs/2405.04634>
  - Le modèle entraîné: [IGNF/FRACTAL-LidarHD\\_7cl\\_randlanet/](#)

*En interne IGN, transmissible si demandé*

- Documentation de LiPac et mode opératoire pour une utilisation en interne IGN
- La page sur le site du SIMV recensant les documents relatifs aux besoins, choix techniques, ateliers utilisateurs, bilans, présentations ...
- La vidéo de présentation à l'agora de l'IA (2eme partie, à partir de 31'05'')





RÉPUBLIQUE  
FRANÇAISE

*Liberté  
Égalité  
Fraternité*

**IGN**  
INSTITUT NATIONAL  
DE L'INFORMATION  
GÉOGRAPHIQUE  
ET FORESTIÈRE

CHANGER  
D'ÉCHELLE

# DES QUESTIONS ?