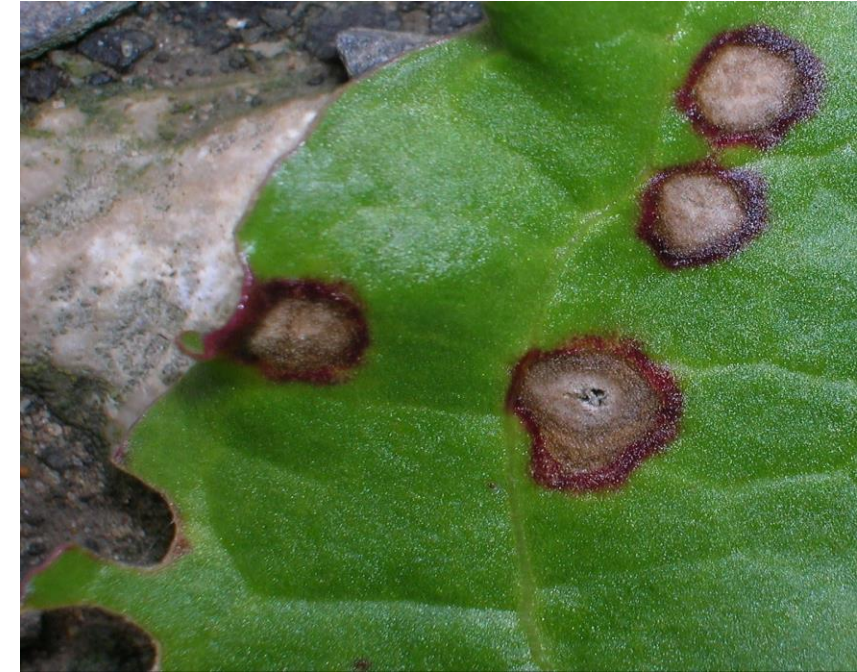


Utilisation de l'apprentissage automatique (machine learning) pour concevoir un nouvel Outil d'aide à la décision prédictif en protection des cultures

François Brun, Lucile Vallet, Olivier Gauriau, Yonghan Liao, François Joudelat, Céline Gouwie, Ghislain Malatesta, Juliette Adrian, Fabienne Maupas

Projet CERCOCAP autour de la cercosporiose de la betterave

- ⇒ créer des outils d'aide à la décision de prévision pour améliorer la gestion des épidémies
Cette présentation : modélisation des épidémies par des approches de *machine learning*
- ⇒ Quantifier les symptômes au champs, en temps réel : voir poster.



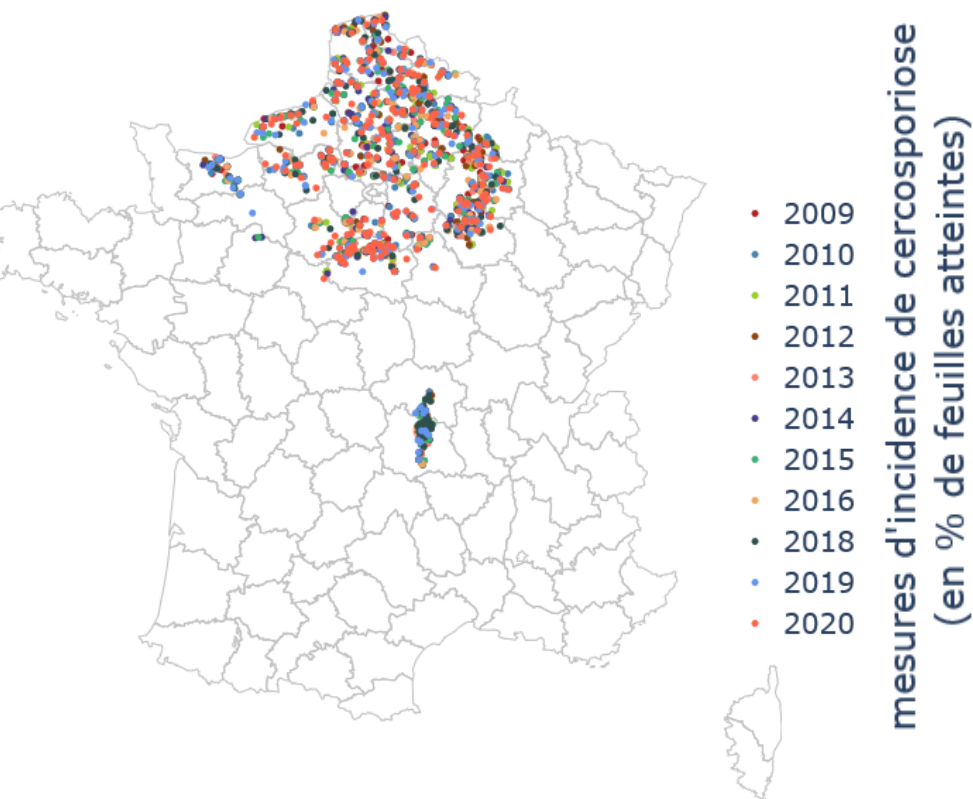
Le principe du machine learning

- **apprentissage machine = machine learning**
- Plus précisément : apprentissage supervisé = supervised learning

Objectif du **machine learning** : « Construire de manière **automatique** une **fonction** qui fait correspondre une entrée à une sortie **de manière performante** en se **basant sur des exemples de paires** {entrée, sortie} »

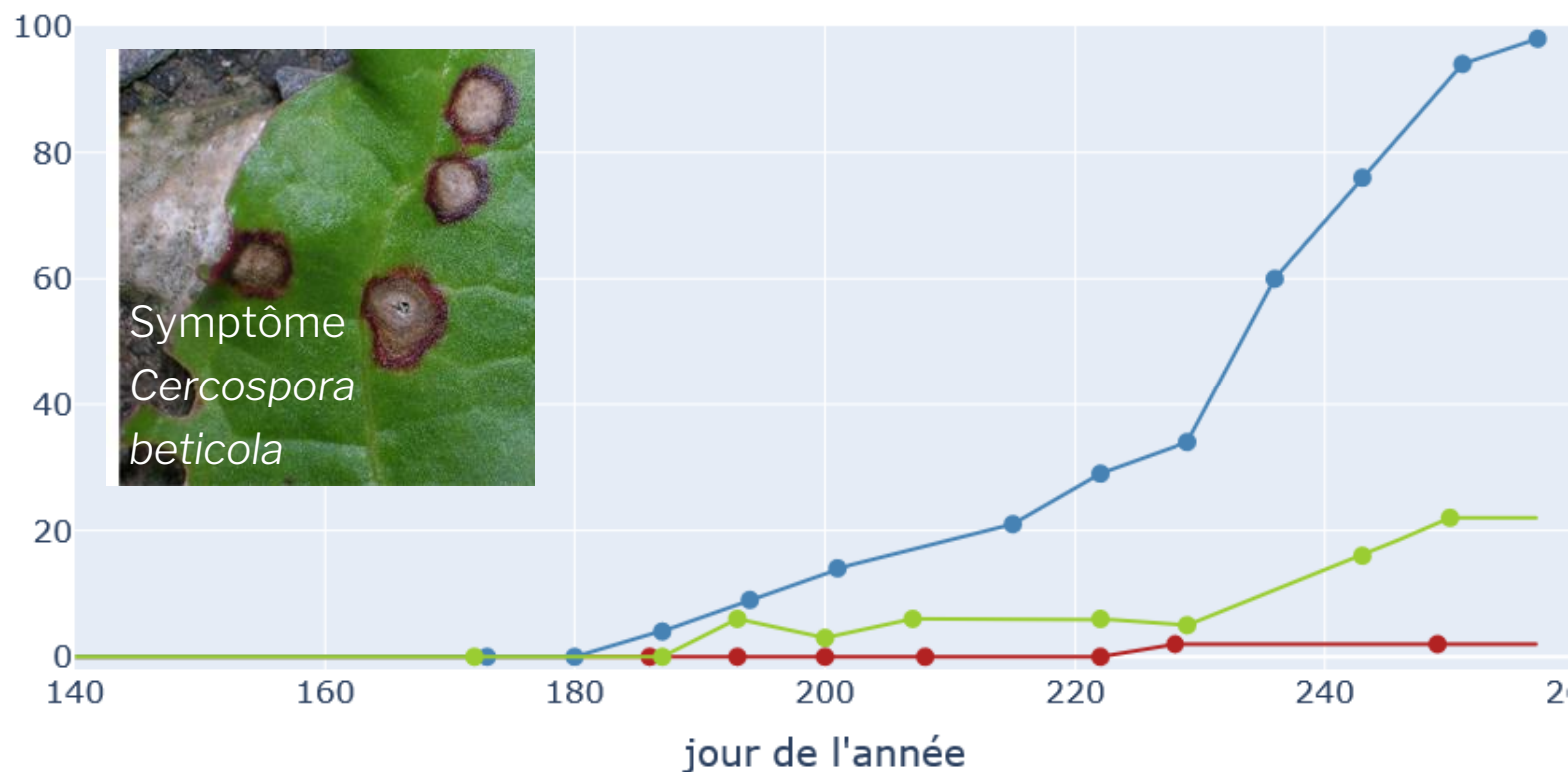
- **sortie** (output) : une variable bien définie que l'on cherche à prédire => Y
- **entrée** (input) : un ensemble des variables explicatives (nombre de colonne) => X
- **exemples de paires** : le jeu de données pour apprendre. Le nombre d'exemple = le nombre d'individus = le nombre de ligne
- **performante** => évaluation





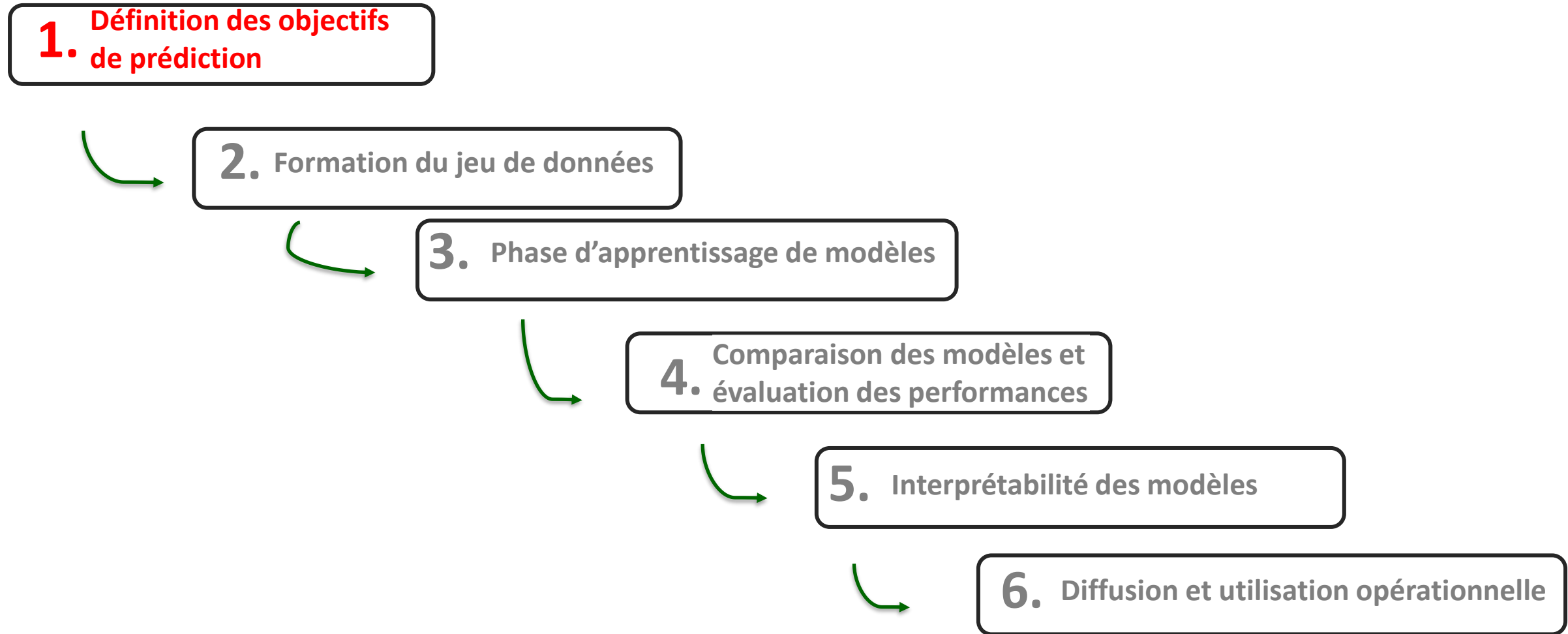
Collectées sur 2009-2020
via Vigicultures pour le BSV

Observations issues du réseau pour le Bulletin de Santé du Végétal



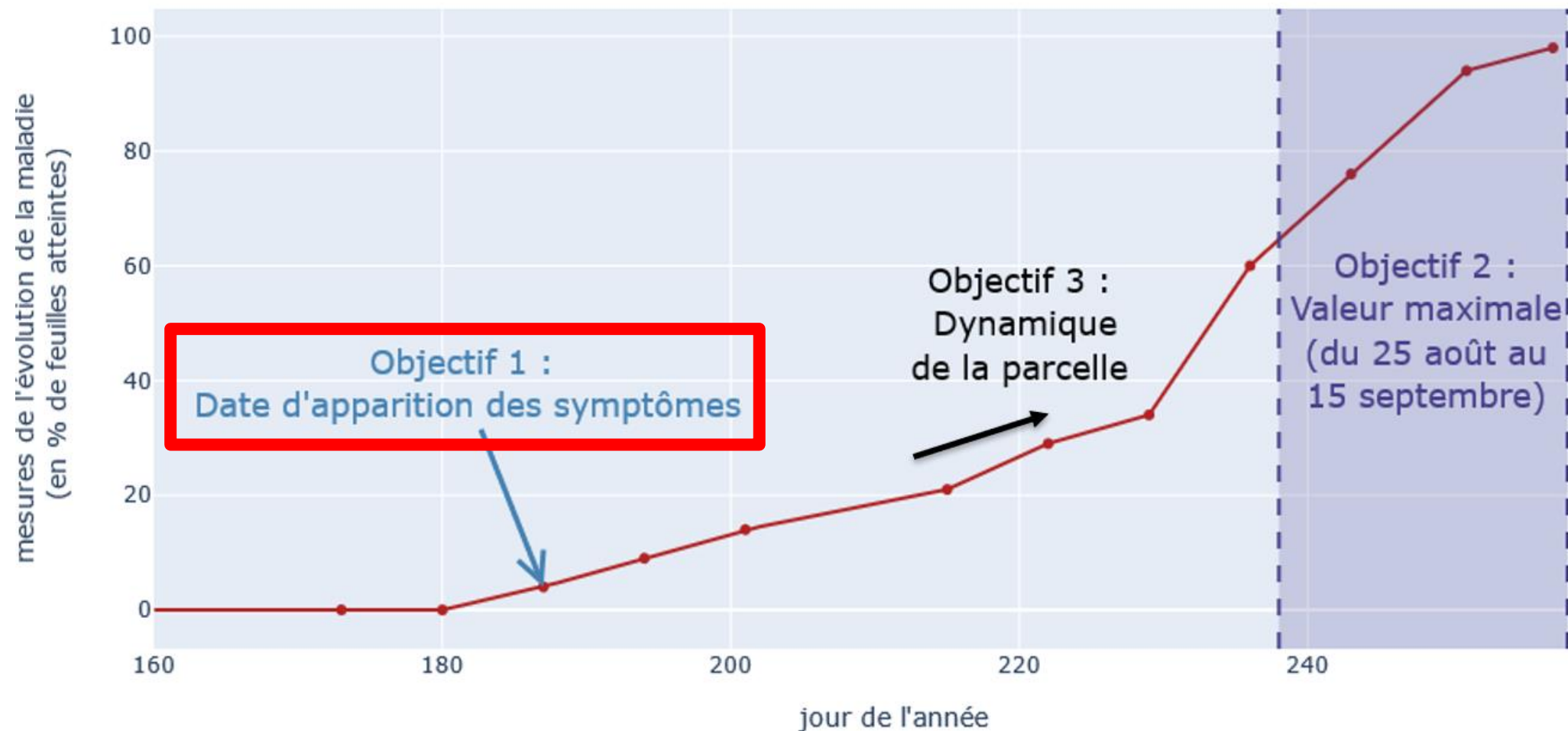
Exemples de 3 séries temporelles d'observation
de l'incidence (% de feuilles atteintes)

Démarche Machine Learning



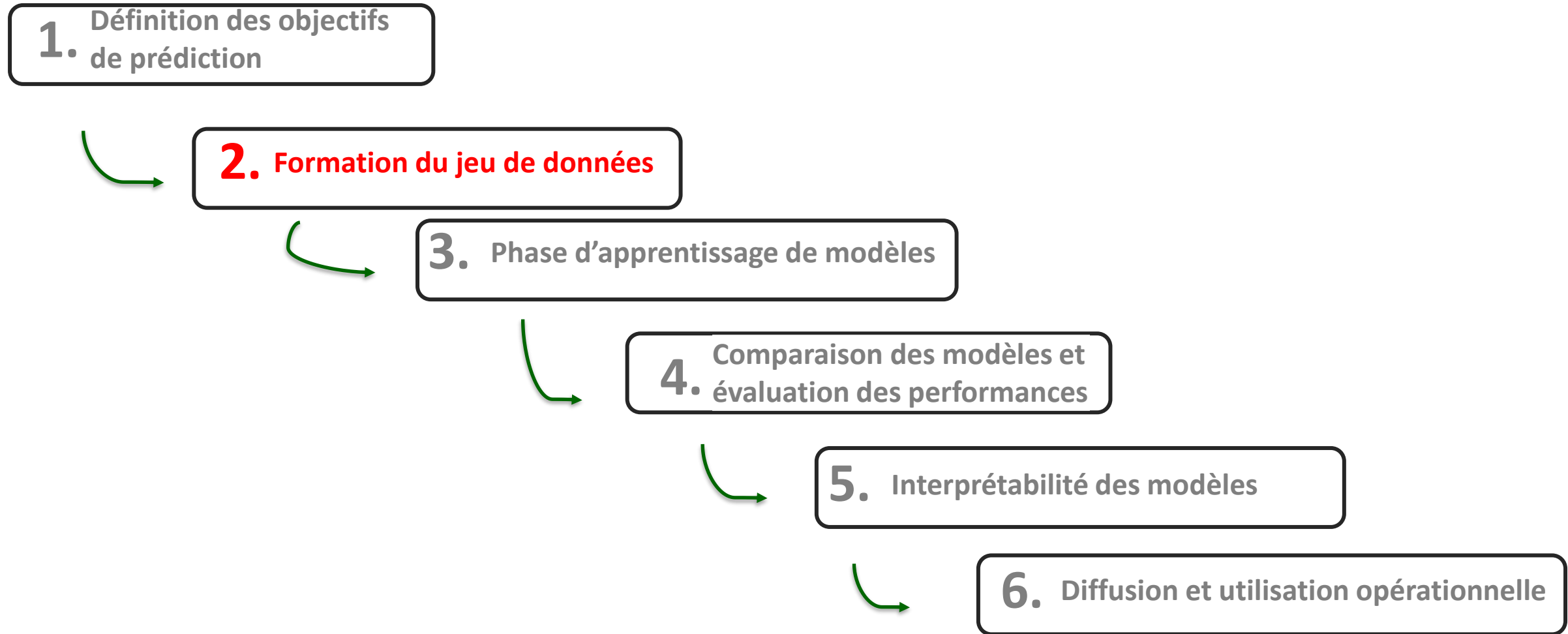
Définition de 3 objectifs

Echanges avec les experts importants pour bien cerner les attentes opérationnelles



=> Focus sur la date d'apparition des symptômes

Démarche Machine Learning



Formation du jeu de données : croisement des données

Données du réseau d'observations d'épidémiosurveillance

- 2349 parcelles suivies.
- 11 années : de 2009 à 2016 et de 2018 à 2020.
- 23 départements.

Données météorologiques

Données horaires SAFRAN

- Humidité relative
- Température
- Vitesse du vent
- Précipitations.

Calcul de la date d'apparition
(jour de l'année)
Calcul de la valeur d'incidence
maximum

X

Données
agronomique
(variété, date de
semis,...)

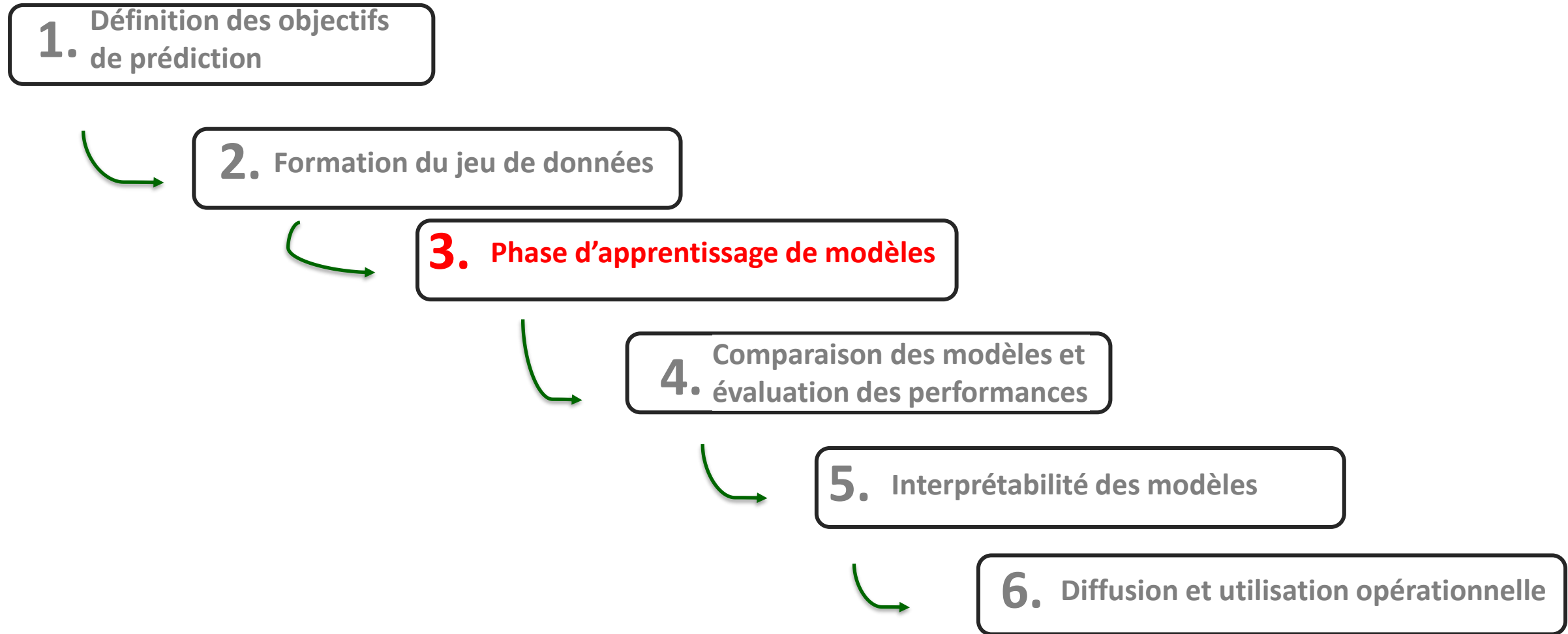
X

Indicateurs
météos agrégés
bimensuel

Exemple d'indicateur météo : MO5_2_STmoy10

La somme des températures en base 10 sur du 16 au 30 mai

Démarche Machine Learning

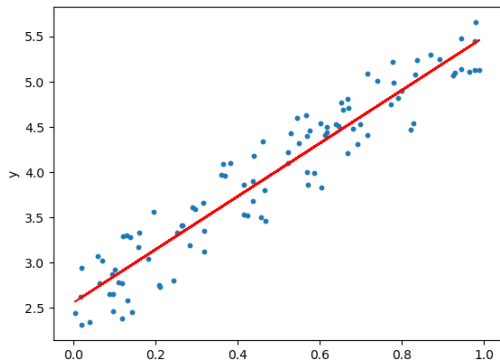


- **Notion d'ajustement**

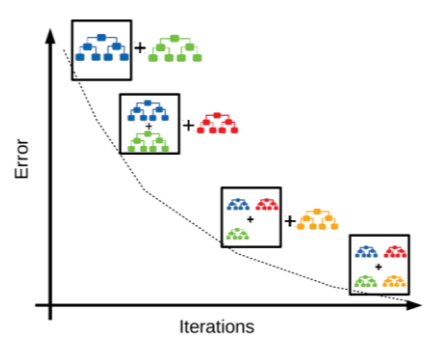
les méthodes de machine learning consiste à ajuster les paramètres, la forme, d'un modèle à un jeu de données d'observations (aussi appelé jeu de données d'apprentissage) pour prédire de nouveaux individus.

- **Diversité de modèles existants**

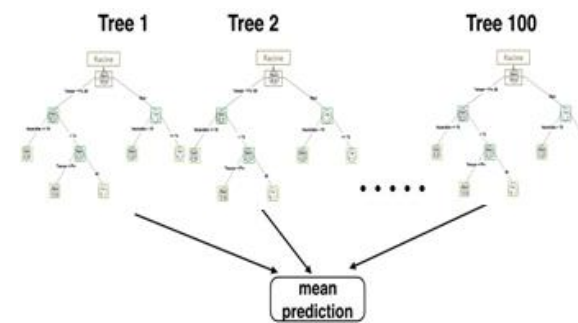
- **Méthodes de regression (avec ou sans pénalisation)**



- **Méthodes d'ensemble**

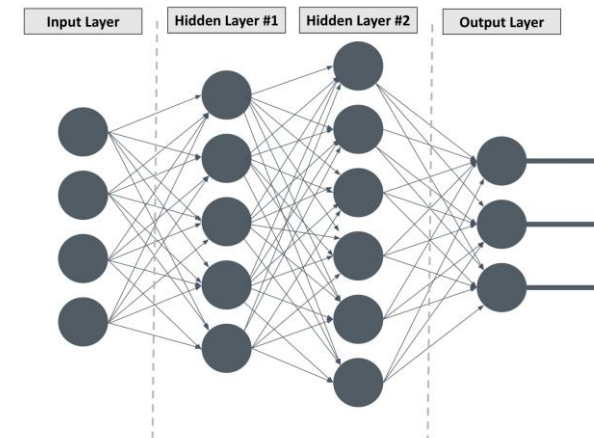


Gradient Boosting



Random Forest

- **Deep learning**



Ne pas se limiter à un modèle, tester plusieurs méthodes puis comparer pour sélectionner le/les meilleurs modèles

Méthodes d'ensemble (forêts aléatoires)

- **Le principe de forêt aléatoire (random forest)**

Principe : construction d'une multitude d'arbres indépendants (chacun ayant une vision partie des données)

Random forest = bagging + feature sampling

Random forest = bagging + feature sampling.

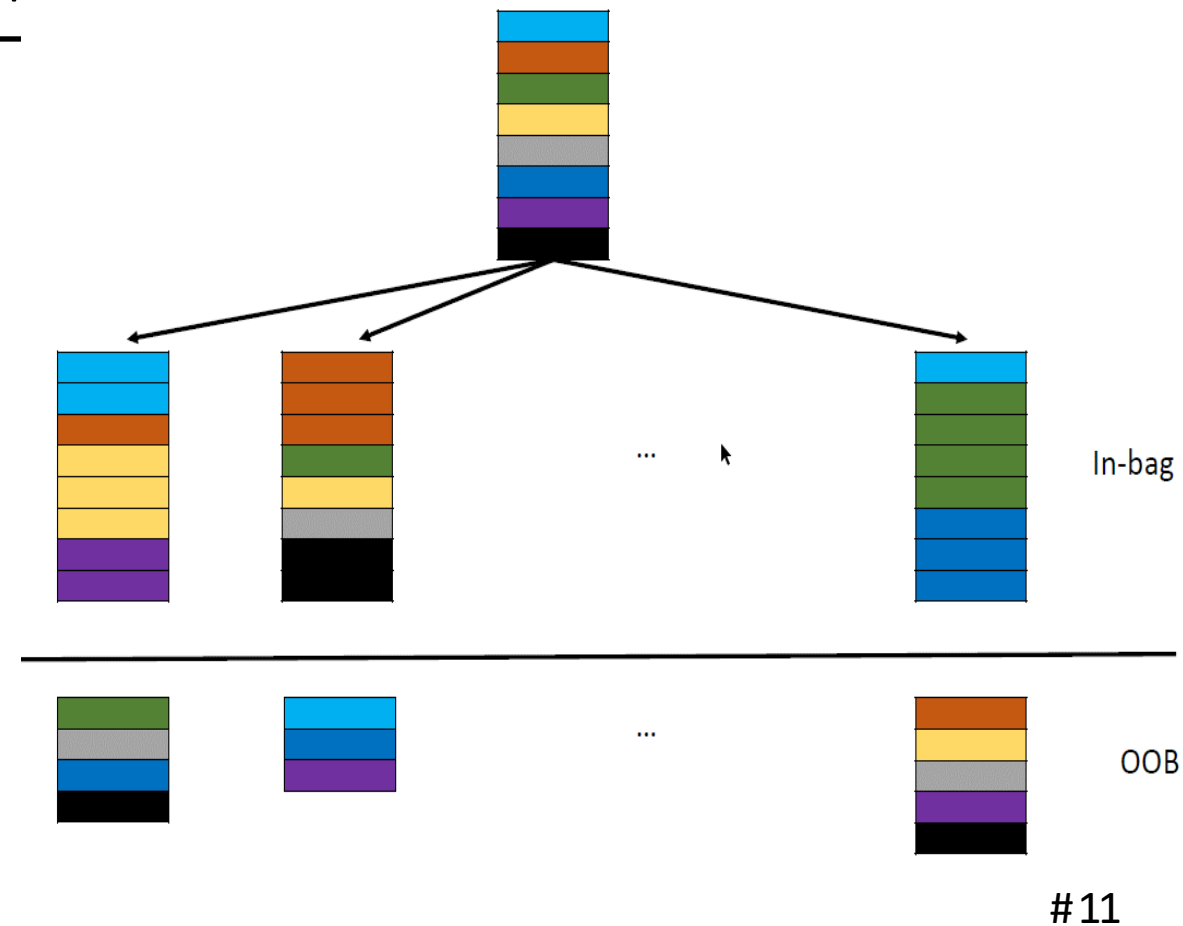
bagging = bootstrap + agregation.

- **Bootstrap :**

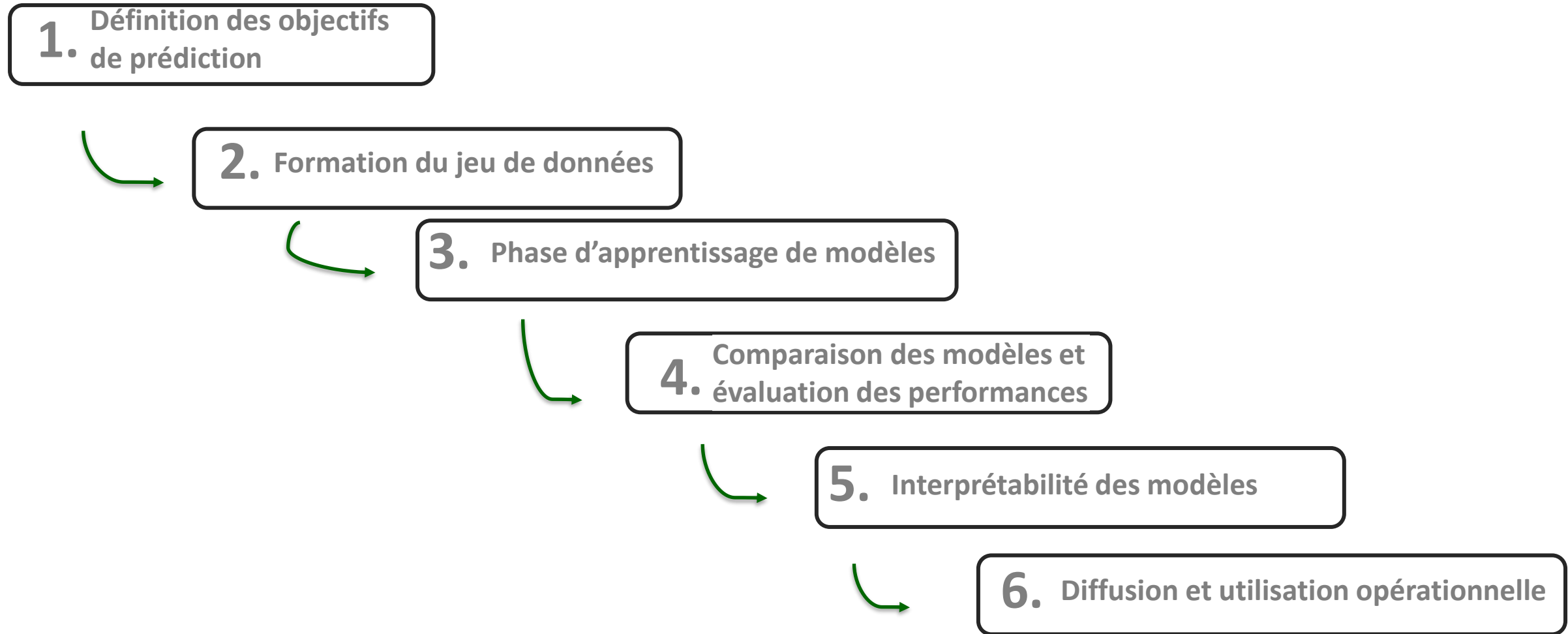
- On tire avec remise

Individus «In-Bag»: individus présents dans l'échantillon bootstrapé (~63%) = échantillon d'apprentissage

Individus «Out-Of-Bag» (OOB): individus non présents dans l'échantillon bootstrapé (~37%)
= échantillon test



Démarche Machine Learning



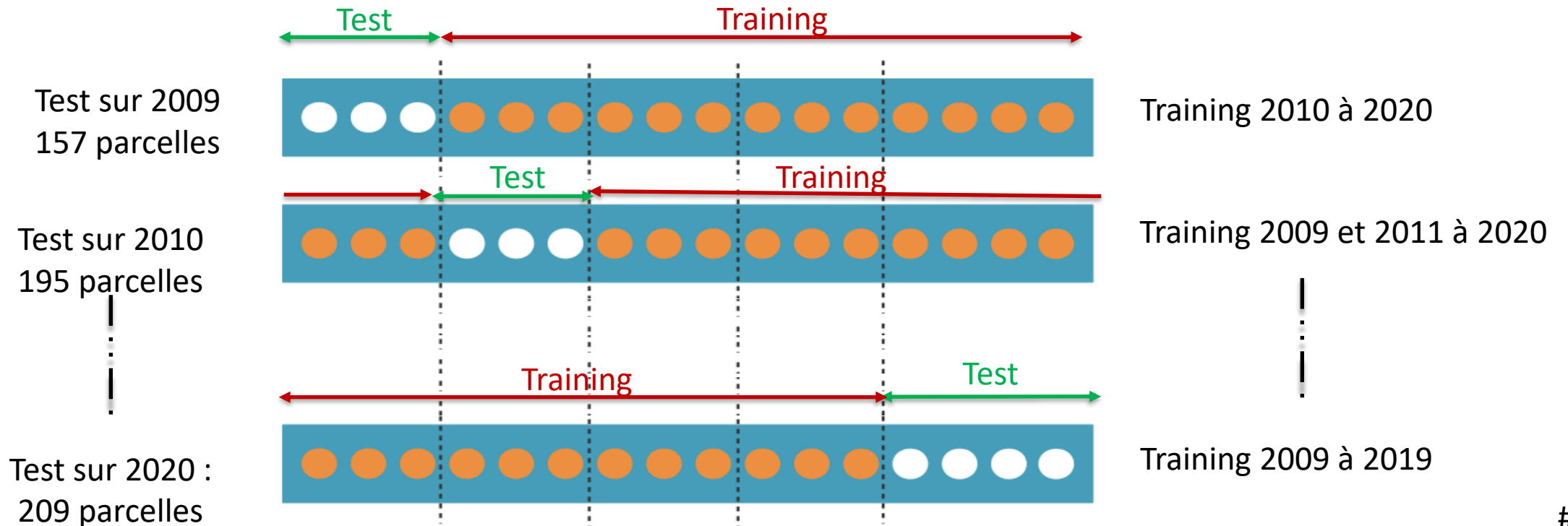
Comparaison des modèles et évaluation des performances

- **2 mesures de performance**

- RMSE : erreur moyenne des prédictions.
- R2 : part de variabilité expliquée des données.

- **Validation croisée par année**

- 11 apprentissages-tests pour les 11 années.



Etape 4 : Comparaison des modèles et évaluation des performances

ETUDE DE CAS

Meilleur RMSE = 15,73 jours

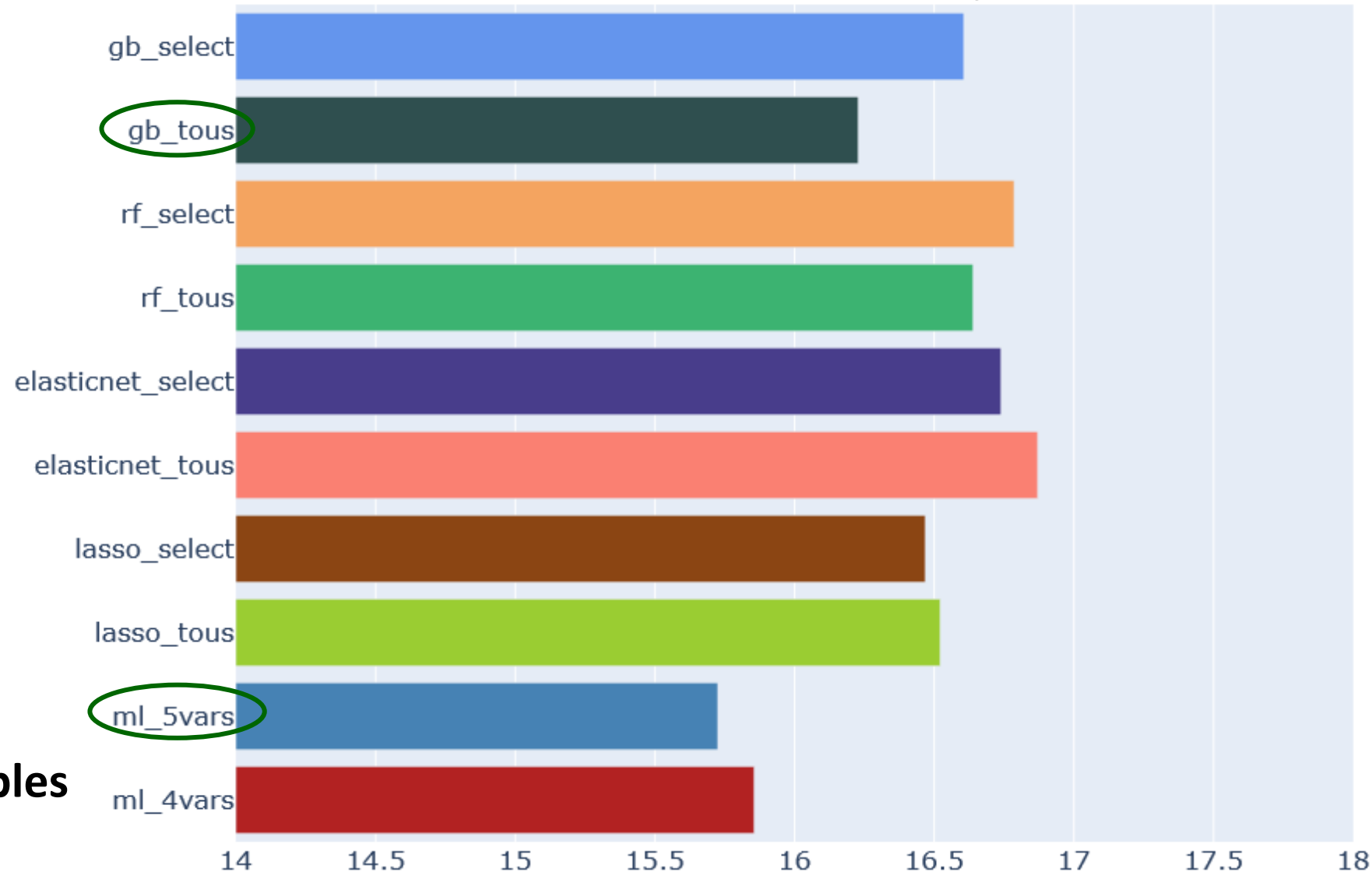
Meilleur R2 = 0,33

Modèle informatif, mais il reste pas mal d'erreur

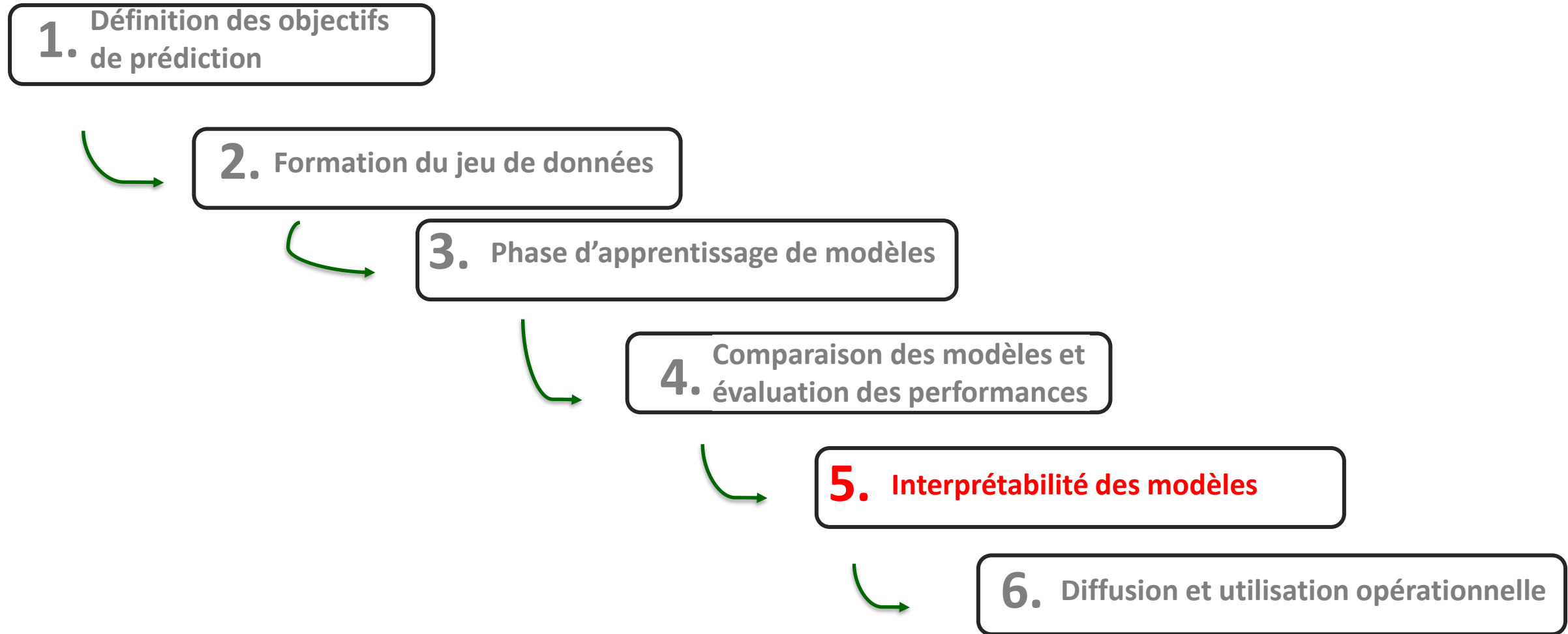
Modèles retenus

→ **Modèle linéaire à 5 variables**

→ **Gradient boosting sans sélection de variables**

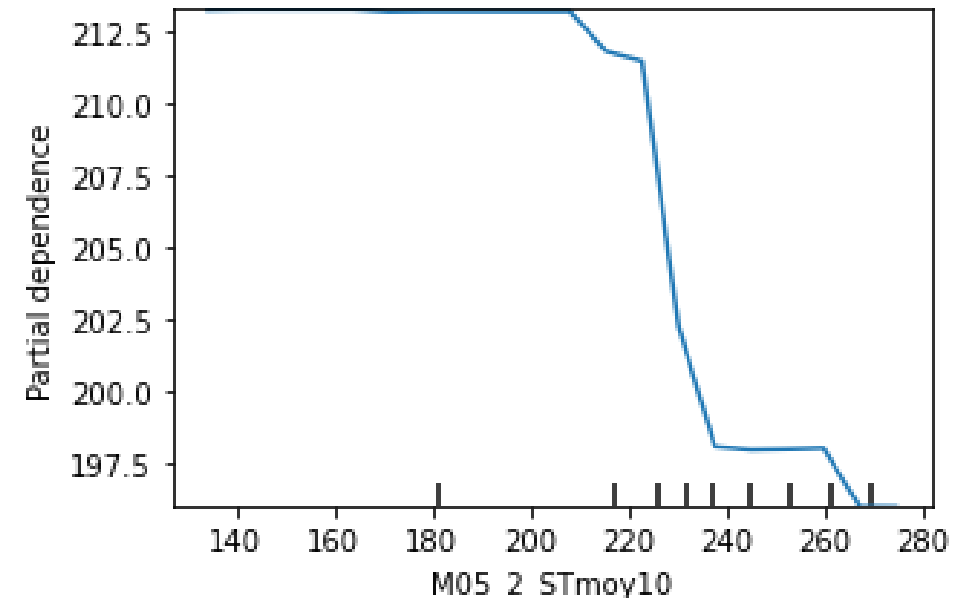


Démarche Machine Learning

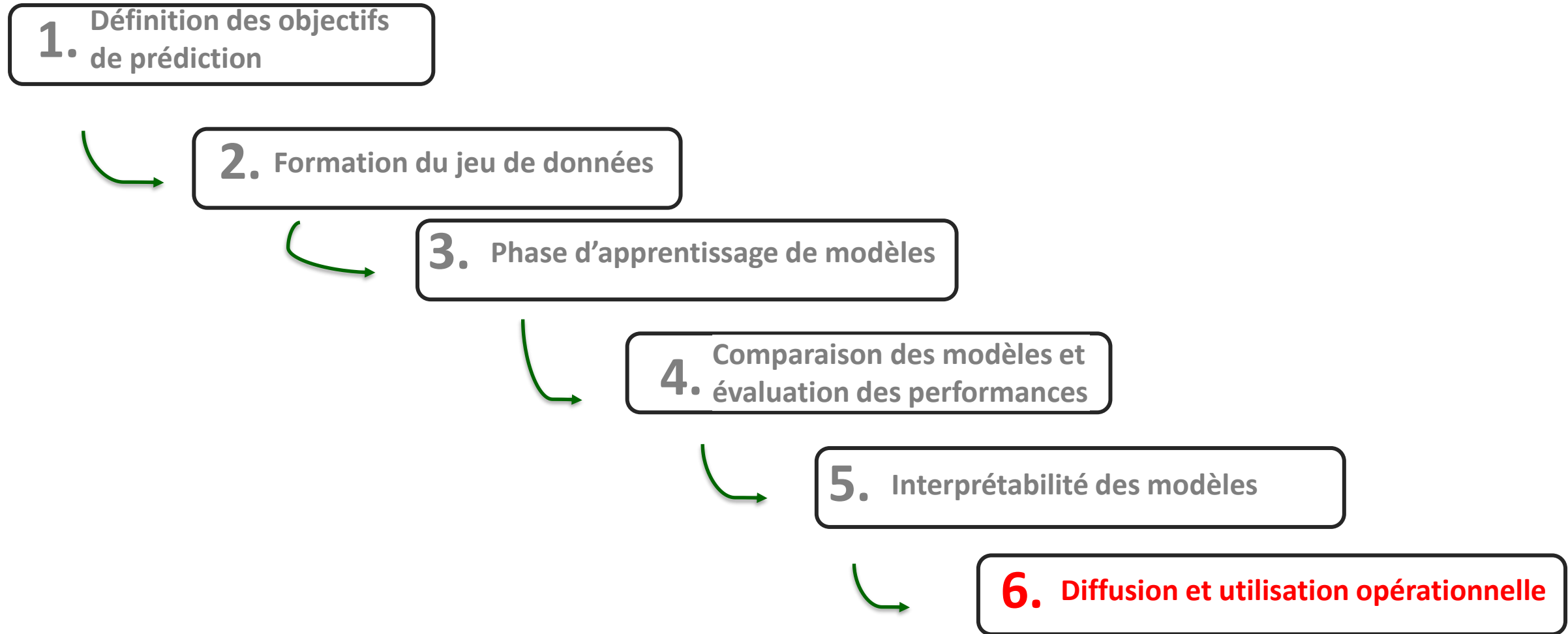


Interprétabilité des modèles

- **Le machine learning ne cherche pas à trouver les causes d'un phénomène mais à le prédire.**
- **l'interprétation n'est pas au cœur, mais c'est un aspect intéressant de l'interaction avec les experts**
- **Etude des effets des variables explicatives**
 - Régression classique: étude des signes des coefficients
 - Forêts aléatoires : graphe de dépendances partielles
- **Etude de l'importance des variables explicatives**
 - Forêts aléatoires : mesure basée sur le nombre de fois où la variable a été utilisée comme critère de séparation au niveau d'un nœud.

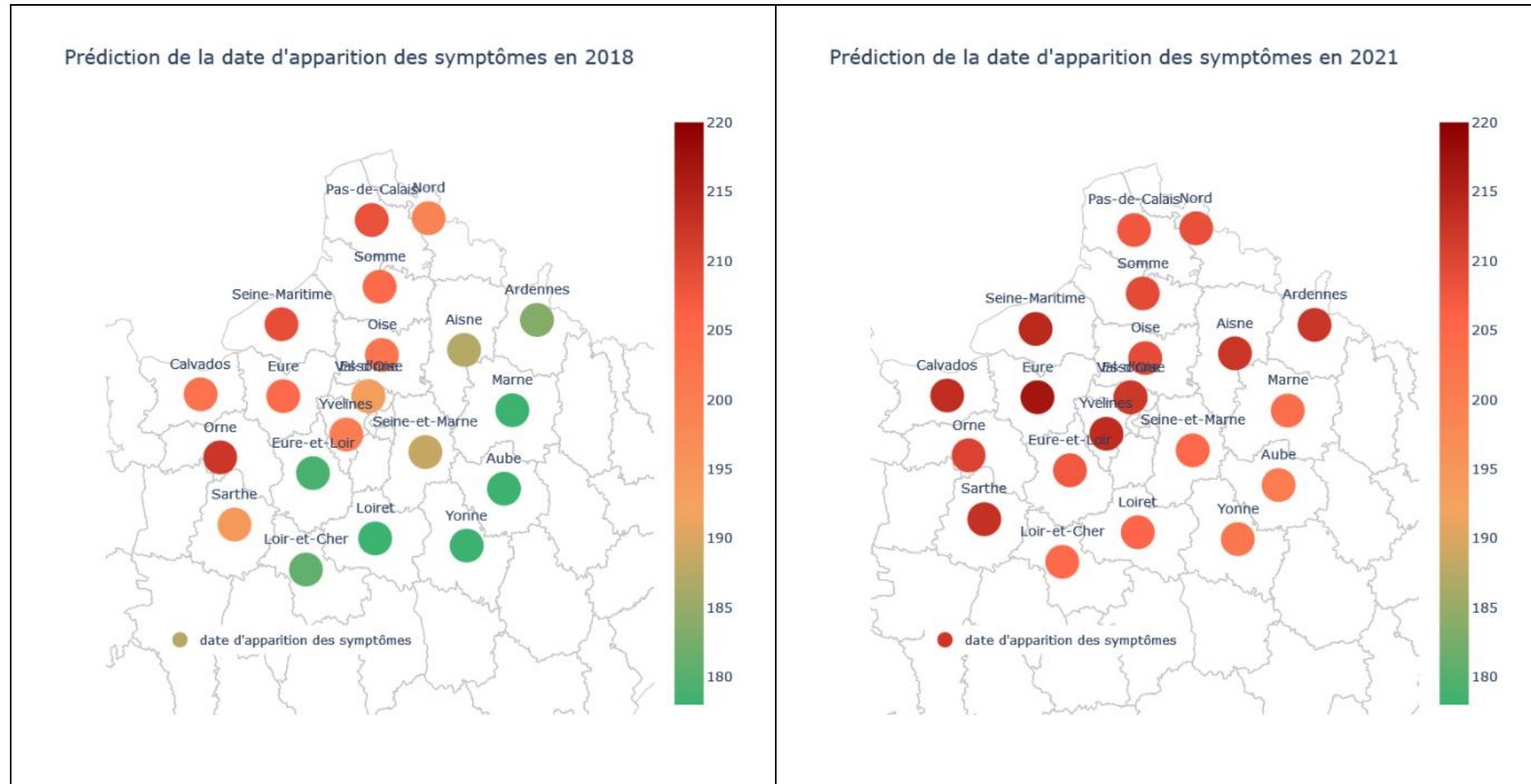


Démarche Machine Learning

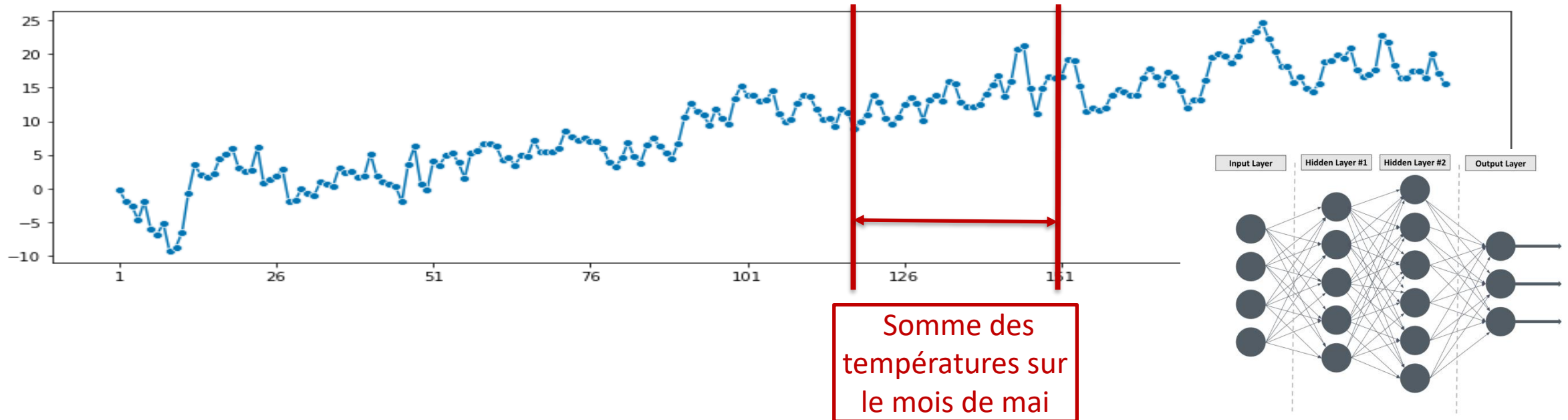


Sur 2021 et 2022 : test opérationnel

Comparaison des prédictions de l'apparition des premiers symptômes en 2021 (à droite) par rapport à 2018 (à gauche)



- Modèles en cours d'intégration dans un outil informatique par l'ITB.
 - Lien avec les données météo d'entrée + Interface utilisateur
- Mieux profiter de la complexité du signal météorologique ?



=> de premiers résultats encourageant avec des réseaux de neurones

- Les questions d'interprétabilité : thèse REGEDI d'Olivier Gauriau (Acta-INRIA-INRAE) – DigiTag et Ecophyto.