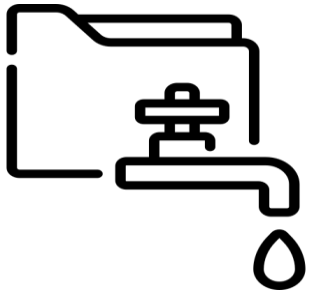
VERS UNE PROTECTION DE LA VIE PRIVÉE ÉQUITABLE POUR LES SERVICES DE SANTÉ

AKRAM BENDOUKHA SAMOVAR, TÉLÉCOM SUD-PARIS/IP-PARIS

DIRECTION : NESRINE KAANICHE SAMOVAR, TÉLÉCOM SUDPARIS/IP-PARIS
ENCADREMENT : AYMEN BOUDGUIGA CEA PARIS-SACLAY
RENAUD SIRDEY CEA PARIS-SACLAY

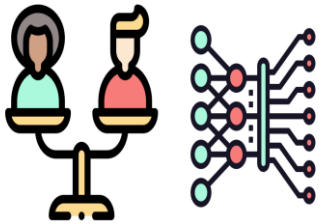
CONTEXTE

LES DEUX AXES DE LA THÈSE



1. Protection de la vie privée :

- Les applications basées sur l'IA sont développées avec un souci limité de la vie privée des utilisateurs.
- L'accès à un modèle (*ML as a service*) permet d'obtenir des informations non-négligeables sur les données dont il est issu : *membership, attribute* ou *property inference attacks*
- Grâce aux protocoles d'apprentissage décentralisé, aucun transfert de données n'est nécessaire, mais les risques de fuite de données à partir des update (Gradients, paramètres …) doivent être considérés



2. Equité des modèles :

- Les biais contenus dans les données sont reproduits et amplifiés par le modèle à travers ses prédictions.
- Etant donnée une tâche d'apprentissage, certains attributs (informations) ne sont pas pertinents dans l'entrainement.
  - ❖ Exemple : « le genre » pour la prédiction du risque lié à un prêt bancaire.
- Plusieurs approches pour améliorer l'équité :
  - ❖ Atténuer les corrélations entre les attributs pertinents des attributs non-pertinents
  - ❖ Définir une fonction de coût qui exprime la non-équité des prédictions et la minimiser
  - ❖ Identifier les composants du modèles responsables du comportement discriminant et les améliorer

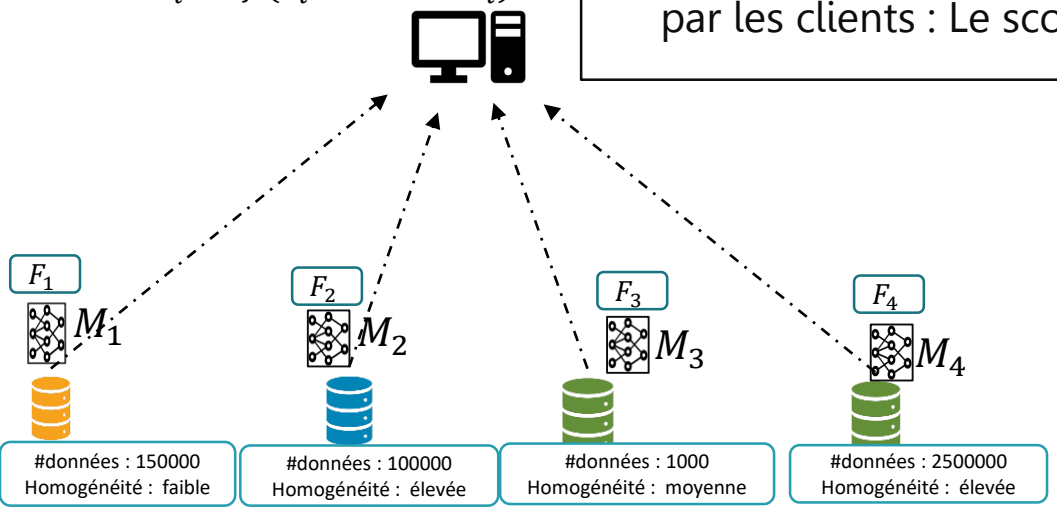
DIRECTION DE RECHERCHE

PRIVACY-PRESERVING & FAIRNESS-AWARE FEDERATED LEARNING

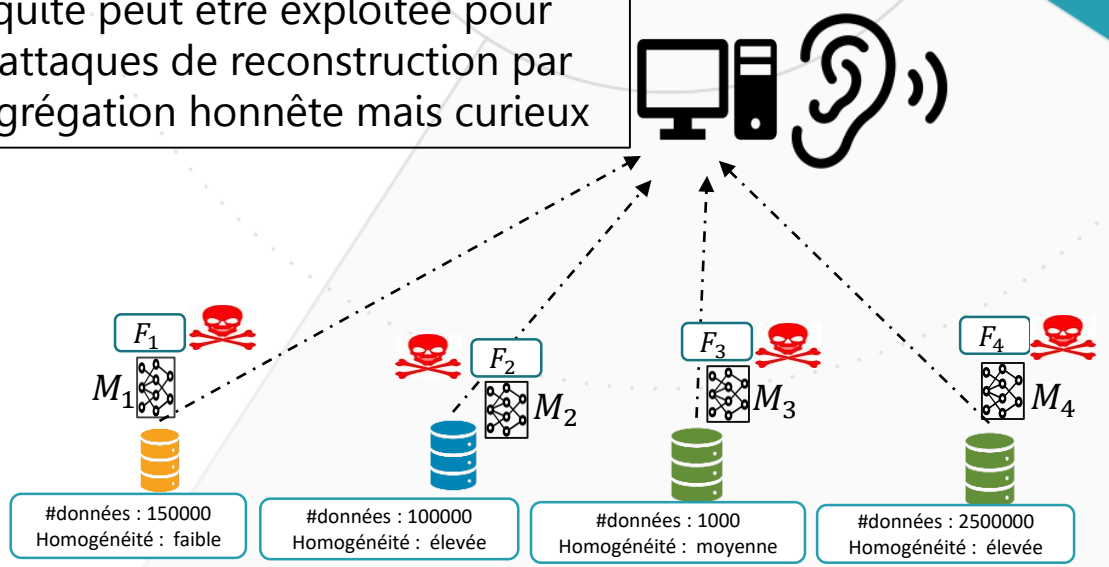
$$M_{global} = \sum_{i=1}^{4} \omega_i \cdot M_i$$

$$\omega_i = f(F_i, \#données_i)$$

- Définition d'une fonction d'agrégation prenant en compte l'équité des modèles produits par les clients.
- Nécessite la partage d'une information supplémentaire par les clients : Le score d'équité de leur modèle



- La mesure d'équité peut être exploitée pour améliorer des attaques de reconstruction par un serveur d'agrégation honnête mais curieux



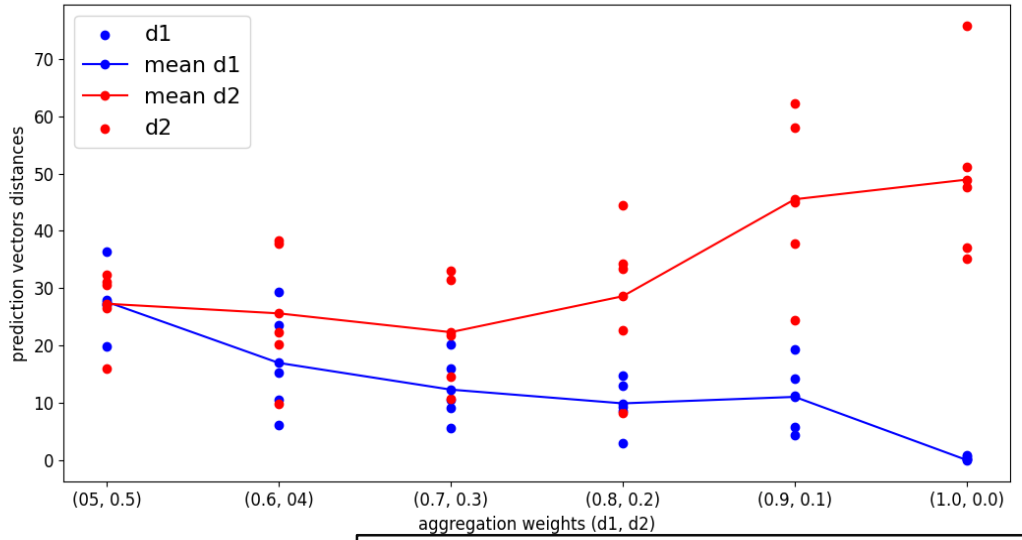
- Défi : Demander davantage d'informations aux détenteurs de données tout en leur garantissant une protection robuste de leur données

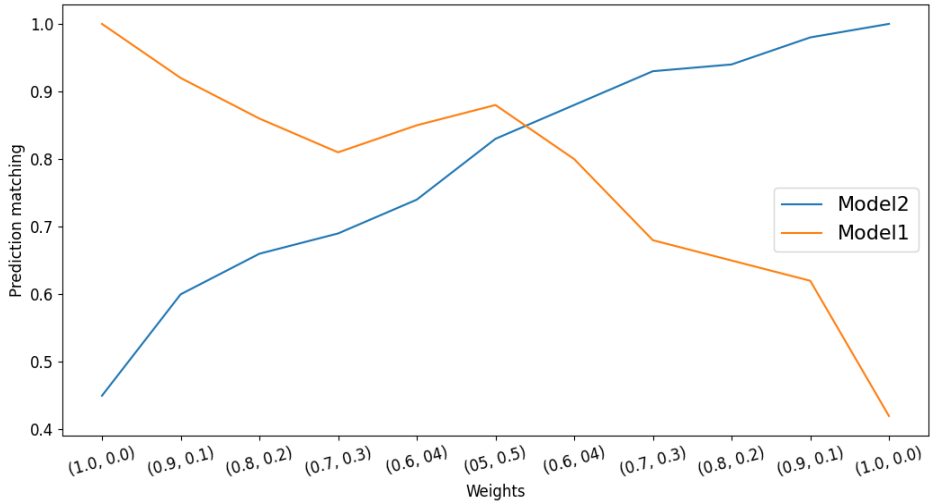
DIRECTION DE RECHERCHE

PRIVACY PRESERVING & FAIRNESS-AWARE FEDERATED LEARNING

- <u>Second risque :</u> l'efficacité des attaques sur le modèle agrégé et une base de données D_i est proportionnelle au poids ω_i attribué au modèle M_i lors de la phase d'agrégation.

➢ Améliorer l'équité d'un modèle aura pour conséquence l'augmentation des risques des attaques d'inférence (membership/attribute inference) sur la base de données associée.





Distance entre les prédictions du modèle agrégé et les prédictions de deux modèles provenant de deux data-sets en fonction de leurs poids respectifs.

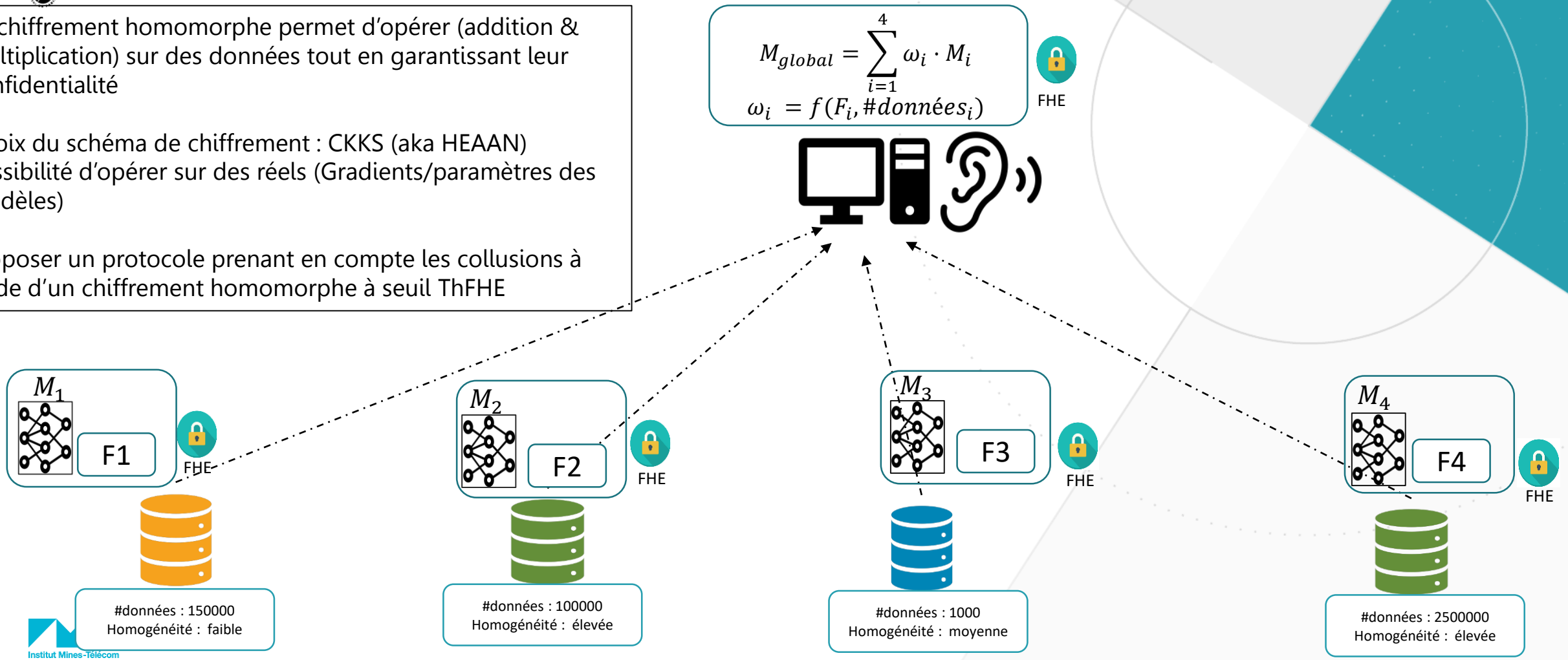
DIRECTION DE RECHERCHE

PRIVACY PRESERVING & FAIRNESS-AWARE FEDERATED LEARNING

Le chiffrement homomorphe permet d'opérer (addition & multiplication) sur des données tout en garantissant leur confidentialité

Choix du schéma de chiffrement : CKKS (aka HEAAN)
Possibilité d'opérer sur des réels (Gradients/paramètres des modèles)

Proposer un protocole prenant en compte les collusions à l'aide d'un chiffrement homomorphe à seuil ThFHE



MERCI DE VOTRE ATTENTION