

Apprentissage Fédéré
entre les EDS du GCS G4
(CHU de Lille, Amiens, Caen, Rouen)

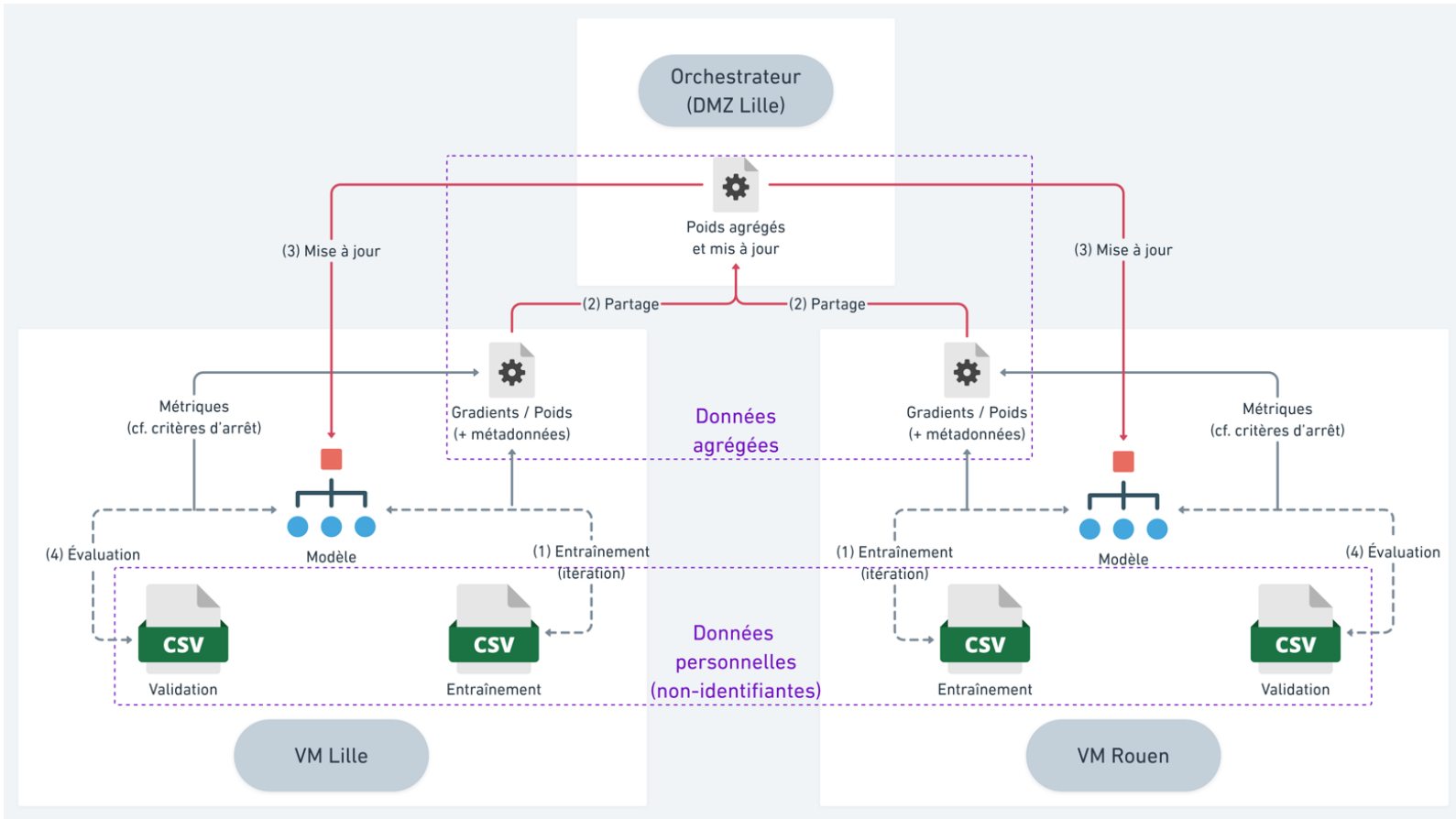
Aperçu du projet

- Projet autour de l'Apprentissage Fédéré et de son application aux données de santé.
- Porté par l'équipe MAGNET (Inria Lille), le CHU de Lille et les CHU du GCS G4.
- Collaboration initiée en 2020 autour de cette thématique.
- Accompagnement par la CNIL via son appel à projets « Bac à Sable » en 2021.
- Dépôt et obtention d'autorisations CNIL associées à deux études en 2022.
- Développement logiciel en cours par MAGNET : package python declearn.
- Travaux plus généraux sur l'apprentissage fédéré :
 - Contribution au développement de Fed-BioMed
 - Inscription dans le DEFI Inria « FedMalin »

Apprentissage Fédéré

- Alternative à la centralisation des données par la distribution des calculs
- Algorithmes pour adapter l'apprentissage statistique au cadre fédéré :
 - Éléments mathématiques sur la convergence du modèle agrégé
 - Recherche nourrie sur les questions d'hétérogénéité des données
- Intérêt pour le traitement de données de santé multicentriques :
 - Juridique : partage d'agrégats statistiques \neq données personnelles
 - Gouvernance : maîtrise par chaque centre de ses données et de leur usage
 - Sécurité / Logistique : flux réseaux ponctuels, identifiés, moins sensibles

Apprentissage Fédéré



Bac à Sable de la CNIL

- Hypothèses formulées :
 - Les autorisations CNIL et solutions techniques locales des EDS permettent (en général) le traitement des données détenues par le centre.
 - Le partage de données agrégées est autorisé tant qu'elles ne permettent pas le retour à une donnée personnelle, *a fortiori* identifiante ou quasi-identifiante.
 - Les agrégats statistiques partagés dans le cadre de l'apprentissage fédéré sont de nature à pouvoir être partagés (selon le point précédent).
- Appel à Projets « Bac à Sable » de la CNIL :
 - Constitution d'un dossier formulant et interrogeant ces hypothèses.
 - Accompagnement par la CNIL pour la conception et le cadrage juridique de deux cas-tests relevant de ces questions.

Bac à Sable de la CNIL

- Hypothèse 1 : Les autorisations CNIL et solutions techniques locales des EDS permettent (en général) le traitement des données détenues par le centre.
⇒ pas toujours !
⇒ Retour à la case CNIL / RGPD pour une part conséquente des données et des traitements
- Hypothèse 2 : Le partage de données agrégées est autorisé tant qu'elles ne permettent pas le retour à une donnée personnelle, *a fortiori* identifiante ou quasi-identifiante.
⇒ oui

Bac à Sable de la CNIL

- Hypothèse 3 : Les agrégats statistiques partagés dans le cadre de l'apprentissage fédéré sont de nature à pouvoir être partagés (selon le point précédent).
 - ⇒ pas d'avis tranché de la CNIL
 - ⇒ semble acquis pour des statistiques et modèles « simples »
 - ⇒ question ouverte sur les modèles plus complexes, à moindre interprétabilité
 - ⇒ en lien avec la recherche sur les attaques (pas toutes spécifiques au fédéré)
 - ⇒ à ce stade, pas de critères formalisés pour évaluer / qualifier les risques
 - ⇒ nécessité de présenter une analyse des risques (AIPD) en amont du projet

Etude PARADE

- Prédiction Automatisée de la Ré-hospitalisation par Apprentissage DÉcentralisé
- Tâche d'apprentissage et données :
 - Prédire la ré-hospitalisation à 30 jours de patients, à partir
 - des codes diagnostics (CIM-10) du séjour
 - de l'âge et du sexe du patient
 - de l'existence d'une hospitalisation dans les 6 mois précédents
 - Au moyen de données issues du PMSI local de plusieurs CHU
 - Avec un modèle de régression logistique
 - matrice documents-termes creuse des diagnostics
 - vecteur d'embedding des diagnostics issu d'un travail distinct (CHU de Lille)

Etude PARADE

- Objectif : mise en œuvre entre les CHU de Lille et Rouen, par Inria
- Dépôt et obtention d'une autorisation CNIL par le CHU de Lille
- Apports :
 - Déploiement sur une infrastructure et des données réelles
 - Travail de cadrage juridique pouvant être mobilisé à nouveau
 - Bon cas test pour la solution logicielle développée
- Limites :
 - Cas semi-artificiel (existence d'une base nationale du PMSI)
 - Moindre simplification juridique qu'attendu ; pas de nouveauté à ce point de vue

Etude AUDIMAT

- Aide au codage des Diagnostics Médicaux par Analyse Textuelle
- Tâche : prédire les diagnostics (CIM-10) d'un séjour à partir du courrier de sortie
 - classification binaire multi-labels (sur un vocabulaire réduit de ~760 codes)
 - ± extraction d'information (par la nature des courriers)
- Données :
 - courriers médicaux issus de l'EDS du CHU de Lille, 2010-2019 (~ 3.4 M)
 - diagnostics issus du PMSI pour 2019 (~ 83 000 séjours avec courrier)
- Méthodes :
 - BERT pré-entraîné sur le corpus de courriers 2010-2018 ou sur 2019 seulement
 - word2vec + agrégation + régression (mêmes cas de pré-entraînement)
 - classificateurs binaires sur une matrice documents-termes creuse

Etude AUDIMAT

- Objectifs :
 - mise en œuvre locale par le CHU de Lille (étude méthodologique locale)
 - étude des risques de ré-identification par Inria
 - à partir du modèle final (indépendamment de l'apprentissage fédéré)
 - en simulant un entraînement fédéré des modèles de classification
- Dépôt et obtention d'une autorisation CNIL par le CHU de Lille
- Apports :
 - travail sur les questions de risques de ré-identification (nature et évaluation)
 - objectif futur : nourrir l'AIPD d'un projet similaire en fédéré (vie réelle)
- Limites :
 - cas spécifique (manque possible de généralité) et simulé (à ce stade)

DecLearn

- DecLearn : package Python développé par MAGNET, bientôt en open-source
- Double objectif :
 - Utilisation pour la recherche (simulations & expérimentations méthodologiques)
 - Déploiement en vie réelle (e.g. pour l'étude PARADE)
- Objectifs de design :
 - Modularité : sur les frameworks, l'algorithmie, le protocole réseau...
 - Extensibilité : outils pour étendre le support du backend à de nouveaux objets
 - Couches d'abstraction pour limiter et isoler le code framework-dépendant

DecLearn

- Principales fonctionnalités :
 - Support pour TensorFlow, PyTorch et les modèles linéaires de Scikit-Learn
 - Système de plug-ins d'optimisation (agnostiques au framework du modèle)
 - Algorithmes implémentés : FedAvg, FedNova, FedOpt, FedProx, Scaffold
- En cours de développement :
 - Differential Privacy : algorithme DP-SGD, Local DP, Central DP
 - Éléments d'utilisabilité : configuration d'expérience par fichier
- Objectif à court/moyen-terme :
 - Apprentissage décentralisé (sans orchestrateur central)

DecLearn & Fed-BioMed

- Contributions à Fed-BioMed, pour :
 - Proposer des éléments d'algorithmie de DecLearn dans Fed-BioMed
 - Continuer à faire de Fed-BioMed la référence pour les applications en recherche bio-médicale dans un contexte « cross-silo » (typiquement, entre EDS)
 - Repenser l'orientation de DecLearn à court/moyen-terme
 - pour la recherche
 - pour des applications « cross-device » ?
 - pour des applications décentralisée

Merci pour votre attention

Pour accéder ou contribuer à declearn : paul.andrey@inria.fr