N° 2024HDR011 Année 2024

HABILITATION A DIRIGER DES RECHERCHES

présentée devant

l'Institut National des Sciences Appliqués de Lyon et l'Université Claude Bernard Lyon I

Privacy issues in AI and geolocation : from data protection to user awareness

par

Antoine BOUTET
Docteur en Informatique

Soutenue le 10 décembre 2024 devant la commission d'examen

Composition du jury

Président :	Pr. François Taiani	Université de Rennes 1
Rapporteurs :	Dr. Catuscia Palamidessi	Inria
	Pr. Romain Rouvoy	Université de Lille
	Dr. Aurélien Bellet	Inria
Examinateurs :	Dr. Sonia Ben Mokhtar	CNRS, Insa-Lyon
	Pr. Anne-Marie Kermarrec	EPFL
	Pr. Eddy Caron	Université Lyon 1
	Pr. Sébastien Monnet	Université Savoie Mont-Blanc

Laboratoire CITI

Mis en page avec la classe thesul.

Remerciements

Ce document et les contributions qui y sont rassemblées sont le fruit et le résultat de rencontres avec un grand nombre de personnes. Tout d'abord, je souhaite remercier l'ensemble de l'équipe Inria Privatics qui m'a accueilli à mon arrivée à l'Insa. Cette équipe a été un environnement très riche et propice pour découvrir, échanger, évoluer et monter en compétence et en responsabilité.

Je souhaite également remercier toutes les personnes qui ont contribué aux travaux présentés ici, les collègues, les doctorants, les post doctorants, les ingénieurs, les stagiaires, et les étudiants avec qui j'ai eu un plaisir à travailler. Merci aussi à toutes les personnes qui m'ont suivi dans l'organisation de projets ou d'évènements.

Ces travaux ont principalement été conduits au sein du laboratoire CITI et je remercie l'ensemble de ses membres avec qui j'ai échangé et collaboré. Merci aussi aux collègues du département informatique de l'Insa qui font que cette activité d'enseignement est toujours un réel plaisir.

Mes remerciements vont aussi à mes rapporteurs Catuscia Palamidessi, Aurélien Bellet et Romain Rouvoy qui ont accepté la tâche de rapporter ce travail de synthèse et Anne-Marie Kermarrec, Sonia Ben Mokhtar, François Taiani, Eddy Caron, et Sébastien Monnet de me faire l'honneur de participer à mon jury et d'évaluer mon travail.

Mon parcours a été influencé par différentes personnes que j'aimerais aussi remercier, notamment tous les enseignants passionnés qui ont aiguisé ma curiosité et m'ont orienté vers la recherche. Merci aussi aux personnes qui m'ont apporté beaucoup lors de mes premières années dans la recherche, en particulier Anne-Marie Kermarrec, Rachid Guerraoui, François Taiani, Sonia Ben Mokhtar et tant d'autres.

Finalement, un grand merci à ma famille et mes proches pour leur soutien inconditionnel et tout particulièrement à Maëlle qui éclaire mon quotidien.

Résumé

L'évolution des technologies numériques et leurs adoptions croissantes ont ouvert des opportunités majeures, très bénéfiques pour la société en général et pour les individus en particulier. Cependant, elle pose également des menaces considérables pour la vie privée qui nécessitent des règles juridiques et éthiques appropriées. La vie privée est essentielle pour protéger les individus, par exemple contre d'éventuelles utilisations abusives des données personnelles. La vie privée est également essentielle pour protéger la société, comme l'a montré l'utilisation abusive des données personnelles afin d'influencer les électeurs lors des élections (par exemple, Cambridge Analytica).

Mes travaux de recherche se situent dans ce contexte de développement ultra-rapide des technologies (souvent déployées avant d'être réglementées) et sont axés sur la protection de la vie privée. Plus précisément, je contribue principalement au domaine en proposant des solutions techniques de la vie privée (en quantifiant des risques ou proposant des contremesures par exemple), et également par le biais d'activités interdisciplinaires. En effet, les problèmes de vie privée ne peuvent être résolus par la seule technologie car ils soulèvent également des questions juridiques, éthiques, économiques et sociétales qui nécessitent un dialogue avec des personnes de différentes disciplines.

Mes principales contributions couvrent 1) les problématiques liées à la collecte, à l'exploitation et à la protection des données de localisation, et plus récemment 2) la sécurité et la confidentialité de l'IA. Dans ce deuxième axe, je me suis focalisé sur les « considérations de confidentialité dans le ML », c'est-à-dire l'identification des risques liés aux technologies et contre mesures de ML, et « l'exploitation du ML pour la confidentialité », en utilisant les capacités de ces nouveaux outils pour protéger les individus (avec l'utilisation des modèles de langues pour l'anonymisation des rapports médicaux par exemple).

Face à ces problèmes de vie privée croissants, il est nécessaire de quantifier les nouveaux risques alimentés par l'évolution des technologies et des usages, et d'améliorer la sauvegarde des informations personnelles des utilisateurs en développant des mécanismes de protection. Enfin, il est également nécessaire à la fois de sensibiliser les utilisateurs finaux aux différents risques afin qu'ils adaptent leur utilisation, et de collaborer avec des acteurs du terrain pour l'adoption de meilleures pratiques.

Abstract

The evolution of digital technologies and their increasing adoption have opened major opportunities, highly beneficial for society in general and for individuals in particular. However, it also poses considerable threats to privacy that require appropriate legal and ethical rules. Privacy is essential to protect individuals, for example against possible misuse of personal data. Privacy is also essential to protect society, as shown by the misuse of personal data to influence voters during elections (e.g., Cambridge Analytica).

In this context of ultra-rapid development of technologies (often deployed before being regulated), my research work is focused on privacy protection. More precisely, I mainly contribute to the field by proposing technical solutions to privacy (by quantifying risks or proposing countermeasures for example), and also through interdisciplinary activities. Indeed, privacy issues cannot be solved by technology alone because they also raise legal, ethical, economic and societal questions that require a dialogue with people from different disciplines.

My main contributions cover 1) issues related to the collection, exploitation and protection of location data, and more recently 2) security and confidentiality of AI. In this second axis, I focused on "privacy considerations in ML", i.e., the identification of risks related to ML technologies and countermeasures, and "exploiting ML for confidentiality", using the capabilities of these new tools to protect individuals (with the use of language models for the anonymization of medical reports for example).

To address these growing privacy issues, it is necessary to quantify the new risks fueled by new technologies and new usages, and to improve the safeguarding of users' personal information by developing protection mechanisms. Finally, it is also necessary to both raise awareness among end users about the different risks in order to enable them to adapt their use, and to collaborate with key players in the field to adopt best practices.

Table des matières

Chapitre 1

Curriculum vitae

Antoine BOUTET, né le 4 décembre 1982 à Tours (37).

Affectation et carrière :

Établissement : Insa-Lyon	**Grade :** Maître de Conférences CN
Affectation depuis le : 1^{er} sept. 2017	**Section CNU :** $27^{ème}$ section
Département : 100% Génie Informatique	**Laboratoire :** CITI

Adresse professionelle :

✉ : Insa-Lyon
20 avenue Albert Einstein
69621 Villeurbanne Cedex
☎ : +33 (0)4 72 43 64 82
@ : antoine.boutet@insa-lyon.fr

1.1 Formation

- Doctorat en Informatique, Université de Rennes 1 :
  - Titre : Décentralisation des systèmes de personnalisation de news
  - Encadrant : Anne-Marie Kermarrec, DR Inria, Équipe As Scalable As Possible (ASAP)
  - Durée : du 1er janvier 2010 au 31 mars 2013 ; soutenue le 8 Mars 2013
  - Financement : Projet ERC Gossple
  - Rapporteurs : Amr El Abbadi, Professeur, Université de Californie, US & Peter Triantafillou, Professeur, Université de Glasgow, UK
  - Examinateurs : Rachid Guerraoui, Professeur, EPFL, Suisse & Antonio Carzaniga, Professeur, Université de Lugano, Suisse & Pascale Sébillot, Professeure, Insa-Rennes, France
- Diplôme d'Ingénieur en Informatique, Université de Technologies de Compiègne (UTC)
- Diplôme Universitaire de Technologie (DUT), Systèmes et Réseaux, Université François Rabelais, Blois
- Baccalauréat Sciences et Technologies Industrielles (BAC STI), Lycée Grandmont, Tours
- Brevet d'Etudes Professionnelles (BEP), Lycée Professionnel Henri Becquerel, Tours

1.2 Expériences

- 2015 - 2017 : Chercheur post-doctorant ; Équipe DRIM, LIRIS, Lyon, France
- 2014 - 2015 : Chercheur post-doctorant ; Équipe CI, Laboratoire Hubert Curien, Saint-Étienne, France
- 2013 - 2014 : Chercheur post-doctorant ; Équipe ASAP, Inria Rennes, France
- 2010 - 2013 : Doctorat ; Équipe As Scalable ASAP, Inria Rennes, France
- 2008 - 2009 : Ingénieur de Recherche ; Équipe ASAP, Inria Rennes, France
- 2006 - 2008 : Ingénieur de Recherche ; Équipe DIONYSOS, Inria Rennes, France

1.3 Activités d'enseignement

Voici le détail de mes enseignements passés depuis mon arrivée à l'Insa-Lyon.

2023 – 2024	294 heures (Insa-Lyon)	
2022 – 2023	91 heures (Insa-Lyon)	Délégation Inria 100%
2021 – 2022	212 heures (Insa-Lyon)	
2021 – 2020	280 heures (Insa-Lyon)	
2020 – 2021	160 heures (Insa-Lyon)	1/2 CRCT
2019 – 2020	384 heures d'enseignement (Insa-Lyon), 50 heures Polytech Annecy	
2019 – 2020	364 heures d'enseignement (Insa-Lyon)	
2018 – 2019	60 heures d'enregistrement (Insa-Lyon)	Décharge de 48 heures (nouvel arrivant)

Responsabilités pédagogiques :

- Depuis 2018 : Responsable de l'option thématique CyberSécurité et Vie Privée, 5ème années, Informatique, Insa-Lyon
- 2019 - 2021 : Membre du comité de pilotage pédagogique de l'Insa Lyon

Depuis mon arrivée à l'Insa de Lyon, je suis responsable du parcours Cybersécurité et vie privée en 5ème année du département Informatique de l'Insa de Lyon. Les principaux projets pédagogiques développés dans ce parcours sont décrits dans la section suivante. J'ai également contribué aux enseignements suivants :

- Systèmes d'exploitation, 3ème année département Informatique, encadrement de TP/TD
- Programmation réseau, 4ème année département Informatique, création et encadrement de TP sur la programmation d'un proxy Web
- Méthodologie pour la conception et le déploiement d'architectures Réseaux et de Services, 4ème année département Informatique, encadrement des projets longues durées
- Projet Fil Rouge, 3ème année département Informatique, suivi de projets libres de conception logiciel
- Sécurité et Réseaux, 4ème année département Informatique, mise en place et évolution de la plate-forme de vulnérabilités, création et encadrement de TP
- Enjeux Environnementaux et Sociétaux du Numérique, 4ème année département Informatique, création et encadrement de projets sur les risques liés à l'IA et l'encadrement de la collecte de données personnelles
- Projets P-SAT (Scientifique/Artistique/Technique), 5ème année département Informatique, encadrement de projets scientifiques sur de multiples sujets en lien avec la confi-

dentialité des informations personnelles.

Au cours de mes enseignements, je me suis attaché à faire découvrir le monde de la recherche aux étudiants au travers mes enseignements, notamment dans le cadre des projets P-SAT de 5ème années. Cette découverte s'est effectuée sur l'ensemble des tâches, allant de l'étude de l'état de l'art, à la proposition de solution et à son évaluation ainsi que l'écriture de papier de recherche et la présentation des résultats. Les sujets portaient par exemple sur l'analyse de la personnalisation et du ciblage de contenu sur mobile, l'évaluation du coût énergétique de la publicité en ligne, l'étude de la solution FLOC API de Google [KHPB22], ou l'étude des attaques par canal auxiliaire sur Intel SGX. J'ai également organisé un petit-déjeuner d'échanges pour découvrir le monde de la recherche[1] auprès de tous les étudiants du département en faisant intervenir différents profils : chargé de recherche, ingénieur de recherche, professeur, maître de conférence, doctorant, et post-doctorant. Enfin, j'ai aussi participer à l'organisation de séminaires cybersécurité et de challenges CTF inter Insa[2] pour faire découvrir aux étudiants la recherche dans le domaine de la cybersécurité.

1.4 Projets Pédagogiques

J'adopte principalement une pédagogie par projet pour mes enseignements en mettant en place une pédagogie active qui permet de générer des apprentissages à travers la réalisation d'une production concrète ou un challenge pour allier un axe ludification et apprentissage par la jeu. J'ai également essayé d'apporter un caractère structurant en impliquant différents départements et différents centres Insa sur certains projets. J'ai notamment organisé un exercice de gestion de crise Cyber, un challenge inter Insa d'anonymisation et de réidentification de données, un challenge sur l'apprentissage fédéré, et des projets sur l'IA et l'évaluation de risques.

1.4.1 Exercice de gestion de crise Cyber

L'accroissement des crises et cyber crises touchant les organisations publiques et privées, tant à l'échelle nationale qu'internationale, appelle la création d'une culture du management de crise et de la cybersécurité. Pour former les organisations à ce risque croissant et préparer les professionnels aux cyberattaques, je confronte chaque année depuis 2018 des étudiants de 5ème année du département informatique de l'Insa de Lyon à un exercice de gestion de crise. Cet exercice allie cyberattaques, attaques communicationnelles et atteintes à la réputation, attaques sur le corps métier et kidnapping autour d'un scénario de gestion de crise impliquant une clinique de santé, une ONG ou un industriel en difficulté sous le feu des critiques de toutes parts. Afin de se rapprocher des conditions proches du réel, ces exercices s'effectuent de manière immersive allant de 10 à 24 heures consécutives. Des étudiants de l'Insa Centre Val de Loire ont été invités sur l'édition de 2020 et 2021. Dans l'édition 2021, l'exercice de gestion de crise a été effectué pendant un jour entier et une nuit entière ("24 heures dans la tempête") dans les locaux de la HEG à Genève lors d'un évènement plus important impliquant d'autres équipes d'étudiants ou de professionnels [BD22]. Les retours des participants sont très positifs.

D'un point de vue pédagogique, la simulation de crise est un vecteur qui permet de sensibiliser, de former l'individu au management de crise et de la cyber sécurité et de développer les "soft

[1] https://www.insa-lyon.fr/fr/evenement/petit-dejeuner-d-echanges-pour-decouvrir-monde-recherche

[2] https://www.insa-lyon.fr/fr/evenement/conference-inter-insa-sur-recherche-en-cybersecurite-securite-log

skills". Cet apprentissage de la gestion en terre inconnue est scénarisé, mis en scène et orchestré comme l'est une pièce de théâtre. Mais le participant à la simulation n'est pas spectateur ; il est l'acteur qui se met en jeu, s'expose et apprend in situ, sur lui-même, sur ses interactions avec les autres membres de la cellule de crise, sur sa capacité à affronter l'incertitude avec calme et méthode.

Communications et financements associés :

- Antoine Boutet, Gaëtan Derache, Simulation de crise - 24h dans la tempête. RESSI 2022
- Juin 2018 : " HTTP 403 forbidden " : les 5IF confrontés à une crise globale - actualité INSA-Lyon, article
- Projet IDEXLYON de l'Université de Lyon dans le cadre du Programme Investissements d'Avenir, Innovation pédagogique (ANR-16-IDEX-0005)

1.4.2 Challenge d'anonymisation et de ré-identification de données inter Insa

Afin de sensibiliser et responsabiliser de manière ludique les futurs acteurs du numérique au sujet de la protection des données personnelles, j'ai participé au montage d'un challenge d'anonymisation et de ré-identification de données inter-INSA impliquant des étudiants de $4^{ème}$ et $5^{ème}$ année des départements Informatique de l'INSA de Lyon (IF), Télécommunications de l'INSA de Lyon (TC), et Sécurité et Technologies Informatiques de l'INSA Centre Val de Loire (STI).

Ce challenge consiste à anonymiser un jeu de données provenant d'un site de vente en ligne ou de données de mobilité. Il s'effectue en deux phases. Durant la première phase, les étudiants, organisés en groupes de 4 ou 5 personnes, doivent développer un mécanisme pour anonymiser ces données. La modification d'un jeu de données à des fins de protection induit de manière inhérente une perte d'information. Par conséquent, un jeu de métriques est fourni afin d'évaluer l'information utile conservée (au travers de métriques cherchant à caractériser l'efficacité d'algorithmes d'IA classiques comme le clustering ou la classification), mais également le risque de réidentification en utilisant quelques approches très naïves (par exemple une réidentification basée sur la date et la quantité de produits achetés, sur le produit acheté et le prix, etc.). Ensuite, durant la seconde phase, les étudiants disposent du jeu de données anonymisé des autres groupes et doivent essayer de re-identifier les utilisateurs, c'est à dire les lier avec les utilisateurs du jeu de données d'origine.

Ce projet a plusieurs objectifs. Le principal objectif est de former les étudiants aux enjeux sociétaux et éthiques associés à la révolution numérique et au traitement massif de données. Un second objectif est de stimuler les interactions et mutualiser des moyens entre les différents départements et centres INSA. Et enfin, le troisième objectif est de faire de la transmission de connaissance de manière ludique et d'améliorer la motivation des étudiants.

Communications et financements associés :

- Antoine Boutet, Mathieu Cunche, Benjamin Nguyen, Sébastien Gambs. DARC : Data Anonymization and Re-identification Challenge. RESSI 2020
- Projet IDEXLYON de l'Université de Lyon dans le cadre du Programme Investissements d'Avenir, Innovation pédagogique (ANR-16-IDEX-0005)

1.4.3 Challenge sur l'apprentissage fédéré

L'apprentissage fédéré est un paradigme d'apprentissage distribué qui entraîne un modèle auprès de plusieurs participants sans échanger directement les données de formation. Bien que le l'apprentissage fédéré suscite beaucoup d'attentes, il souffre de plusieurs limitations en matière de confidentialité et de sécurité. Plus précisément, les mises à jour de modèles envoyées par les participants divulguent des informations sur leurs données d'entraînement et le FL manque de robustesse face aux parties malveillantes qui veulent empoisonner ou abuser du système.

Ce challenge permet aux étudiants de toucher du doigt les limites de l'IA et plus spécifiquement de l'apprentissage fédéré en manipulant des attaques d'inférence, d'appartenance et d'empoisonnement ainsi que l'impact de contre-mesures.

Chapitre 2

Synthèse des activités d'encadrement et de recherche

2.1 Encadrement

2.1.1 Thèses soutenues

- Vincent Primault. Doctorant, équipe DRIM du LIRIS en 3ème et 4ème année. Design d'une solution permettant une configuration dynamique de mécanismes de protection des informations de localisation des utilisateurs. 01/01/2016 – 01/03/2018. **Retombées :** trois publications et un poster dans des conférences internationales, une publication dans une conférence nationale, et un rapport technique. Vincent est actuellement ingénieur R&D à Salesforce, France. Encadrement sans pourcentage officiel.
- Sophie Cerf. Doctorante, équipe DRIM du LIRIS / GiPSA en 2ème et 3ème année. Design d'une solution de contrôle et de configuration des mécanismes de protection de données de localisation des utilisateurs. 01/01/2017 – 31/10/2018. **Retombées :** deux publications et un poster dans une revue et des conférences internationales. Sophie est actuellement chargée de recherche ISFP à l'Inria Grenoble. Encadrement sans pourcentage officiel.
- Albin Petit. Doctorant, équipe DRIM du LIRIS en 3ème année. Design d'attaques et de contre mesures pour assurer la confidentialité des informations personnelles des utilisateurs dans le recherche sur le Web. 01/01/2016 – 15/03/2017. **Retombées :** une publication dans un journal international, une publication dans une conférence internationale et un poster dans une conférence internationale. Albin est actuellement ingénieur de recherche au LIG, Grenoble. Encadrement sans pourcentage officiel.
- Théo Jourdan. Doctorant, équipe Privatics, Inria. Vie privée et transparence dans les systèmes d'apprentissage dans le domaine de la santé. 01/10/2018 – 29/10/2021. **Retombées :** multiples publications dans des journaux, conférences et workshops francophones et internationaux. Doctorant co-encadré avec Carole Frindel (CREATIS). Théo est actuellement post-doctorant dans l'équipe Multi-Scale interaction du laboratoire ISIR (Paris Sorbonne). Encadrement 45%.

2.1.2 Thèses en cours

- Thomas Lebrun. Doctorant, équipe Privatics, Inria. Health Data : Exploring Emerging Privacy Enhancing Mechanisms. 01/11/2021 – 30/11/2024. **Retombées :** plusieurs publications dans des conférences internationales et plusieurs papiers en cours de finalisation. Encadrement 45%.

- Jan Aalmoes. Doctorant, équipe Privatics, Inria. 01/09/2021 – 30/11/2024. **Retombées :** plusieurs publications dans des conférences internationales et plusieurs papiers en cours de finalisation. Encadrement 45%.

2.1.3 Thèses à venir

- Jules Marmier. Doctorant, équipe Privatics, Inria. How to leverage machine unlearning to remove poisoning and backdoors, or sanitize a model. 01/11/2024 – 31/10/2027. Encadrement 50%.

2.1.4 Ingénieurs

- Lucas Magnana. Ingénieur de recherche, équipe Privatics, Inria. Développement d'une librairie de quantification du risque de fuite de confidentialité en lien avec les LLMs dans le milieu médical. Depuis juin 2024.
- Hugo Dabbadie. Ingénieur de recherche, équipe Privatics, Inria. Développement d'outils d'analyse et d'attaques d'empoisonnement en lien avec l'apprentissage fédéré. 01/01/2024 – 30/06/2024.
- Amine Bechorfa. Ingénieur de recherche, équipe Privatics, Inria / Insa-Lyon. Développement d'applications mobiles : suivi de patient post AVC, système fédéré de recommandations. Depuis 2021.
- Nathan Brunet. Ingénieur de recherche, équipe Privatics, Inria. Développement d'un challenge sur l'apprentissage fédéré. 01/02/2022 – 31/05/2024.
- Julien Barnier. Mise à disposition d'un ingénieur de recherche CNRS. Développement d'une librairie de quantification du risque de fuite de confidentialité en lien avec l'anonymisation de données dans le domaine de la santé (collaboration avec le Health Data Hub). 01/10/2022 – 30/06/2023.
- Adrien Baud. Ingénieur de recherche, équipe Privatics, Inria. Développement d'une plateforme web de sensibilisation aux problèmes de vie privée. 01/10/2019 – 30/09/2021.

2.1.4.1 PostDocs

- Mohamed Maouche. Postdoc, équipe Privatics, Inria. Attaque d'appartenance sur apprentissage fédéré avec couches privées et personnalisées. 2021 - 2022. Mohamed est actuellement chargé de recherche ISFP à l'Inria Lyon.

2.1.5 Autres encadrements

- Vasisht Duddu. Bachelor of Technology in Electronics and Communication Engineering from the IIIT Delhi. 01/03/2020 – 31/05/2020. **Retombées :** 2 articles dans un workshop international, 2 articles dans une conférence internationale, et 1 rapport de recherche. Vasisht est actuellement doctorant à l'université de Waterloo au Canada.

2.1.5.1 Stagiaires

- Helain Zimmermann : 1ère année Ensimag. Entraînement de LLms anonymes (c'est à dire sans mémorisation d'information directement et indirectement identifiantes). 24/06/24 -

30/08/24. **Retombées :** un article en préparation.

- Zakaria El Kazdam : 1ère année Ensimag. Quantification des informations indirectement identifiantes dans les comptes-rendus médicaux. 03/06/24 - 09/08/24. **Retombées :** un article en préparation.
- Jules Marmier : 3ème année Ensimag. Exploitation des données synthétiques pour l'audit de systèmes d'IA. 15/04/24 - 30/08/24. **Retombées :** un article en préparation.
- Arno Venaille : 3ème année Insa-Lyon. Impact de la Differential Privacy sur les attaques d'inférence (attaque d'appartenance et inférence d'attribut). **Retombées :** un article en préparation.
- Paul Retourné : 1ème année ISEN. Évaluation de l'impact énergétique de la publicité sur mobile. 01/07/24 - 31/07/24. **Retombées :** un article et un challenge en préparation.
- Murielle Iradukunda : 2ème année Ensimag. Quantification du gain de l'apprentissage fédéré cross silo et développement d'une attaque d'empoisonnement. 22/05/23 - 18/08/23. **Retombées :** un poster dans une école d'été et un article en re-soumission.
- Gaspard Berthelier : 2ème année CentraleSupélec. Développement d'une librairie de quantification du risque en lien avec les LLMs. 16/01/23 - 30/06/23. **Retombées :** publication dans une conférence internationale, poursuite en thèse à Inria (équipe Epione).
- Jan Aalmoes : Master M2 Université Lyon 1. Apprentissage Fédéré. 15/02/2021 – 30/06/2021. **Retombées :** poursuite en thèse et co-direction de thèse.
- Brandon Da Silva Alves : 3ème année Insa-Lyon. Développement d'une plate-forme de gestion de crise. 14/06/2021 - 15/08/2021. **Retombées :** utilisation de la plateforme dans le cadre d'un exercice de gestion de crise inter Insa
- Sohpanna Ngov : 3ème année Insa-Lyon. Attaque par inférence via IoT. 14/06/2021 - 13/08/2021.
- Yann Lafaille : 4ème année Insa-Lyon. Attaque par inférence via IoT. 15/02/2021 – 10/08/2021. **Retombées :** prototype d'une application mobile de sensibilisation utilisateurs aux risques de fuite d'informations personnelles
- Hugo Desgeorges : 3ème année Insa-Lyon. Attaque par inférence via IoT. 15/06/2020 – 28/08/2020
- Jan Aalmoes. Master 1 Mathématiques, Université Lyon 1. Impact de la localisation sur la personnalisation de contenu dans le contexte mobile. 03/06/2019 – 31/08/2019.
- Amine Bahi : Programme pré-doctorale UM6P, Maroc. Privacy-preserving and scalable machine learning using homomorphic encryption. 01/03/2019 – 31/05/2019. **Retombées :** un article dans une revue internationale.
- Hilaire Bouaddi : 3ème année Insa-Lyon. Analyse de trace de mobilité. 17/06/2019 – 30/08/2019.
- Félix Fonteneau : 3ème année Insa-Lyon. Développement d'une plate-forme d'apprentissage fédéré pour de la détection d'activité. 13/06/2019 – 31/08/2019.
- Romain Fournier : 3ème année Insa-Lyon. Développement d'une plate-forme d'attaque Cyber-sécurité. 18/06/2018 – 31/08/2018. **Retombées :** utilisation de la plate-forme d'attaque dans le cadre de cours en 4ème année informatique Insa-Lyon
- Alexandre Ven Beurden : 3ème année INSA-Lyon. Développement d'une plate-forme d'inférence d'information à partir de données de mobilité. 13/06/2018 – 31/08/2018. **Retombées :** une publication dans une conférence internationale.
- Bastien Durand : 3ème année Insa-Lyon. Détection des traits de personnalité à partir de l'historique de mobilité. 11/06/2018 – 31/08/2018
- Jorge Francisco Chong Chang : Master 1 Machine Learning and Data Mining. Transport Mode Detection inside Lyon using mobile phone sensors. 03/04/2017 – 31/08/2017

2.2 Diffusion des travaux et rayonnement scientifique

2.2.1 Media

- Septembre 2024 : IA : La guerre des données aura-t-elle lieu ? - Usbek & Rica - article
- Avril 2024 : L'AI Act, ou comment encadrer les systèmes d'IA en Europe - TheConversation (∼ 6 500 vues) - article
- Février 2024 : Protéger la vie privée des systèmes d'IA : l'ambition du projet IPoP - InCyber News - article
- Janvier 2024 : L'IA générative pourrait aussi servir à exploiter des données personnelles en toute sécurité : la piste des données synthétiques - TheConversation (∼ 5 300 vues) - article
- Juin 2023 : ChatGPT, modèles de langage et données personnelles : quels risques pour nos vies privées ? - The Conversation (∼ 6 700 vues) - article
- Mars 2023 : Données personnelles : rien à cacher, mais beaucoup à perdre - The Conversation (∼ 66 000 vues) - article
- Juin 2020 : L'app StopCovid : guérir le mal par la tech ? - actualité de l'INSA Lyon - article
- Décembre 2019 : L'algorithme : cette formule arbitraire, miroir de l'intention humaine - En Vue, la lettre d'information de l'INSA Lyon – article
- Mai 2019 : Tout compte fait : souriez, vous êtes géolocalisés ! - vidéo
- Juin 2018 : " HTTP 403 forbidden " : les 5IF confrontés à une crise globale - actualité INSA-Lyon, article

2.2.2 Projets de recherche

- *Porteur* – projet **IPoP** (Interdisciplinary Project on Privacy) du PEPR Cybersécurité (2021 - 2027, 5,5M€)
- *Porteur* – projet **INTERFERE** : stressINg sysTems sEcurity thRough on the Fly nEtwork tRaffic gEneration, la Fédération Informatique de Lyon (2021 - 2023, 10k€)
- *Porteur* – projet **Trusty-AI**, Pack Ambition International et FACE Foundation/Thomas Jefferson (2022 - 2023, 60k€)
- *Porteur* – projet **CASCADE** : approChes d'Apprentissage préServant la ConfidentiAlité des Données pErsonnelles, la Fédération Informatique de Lyon (2021 - 2023, 7k€)
- *Participant* – projet **PMR** (Privacy-preserving methods for Medical Research), projet ANR PRC, responsable scientifique INSA-Lyon, (2020 - 2024, 180k€)
- *Porteur* – projet BQR INSA-LYON **TELEMETER** (smarT, and EthicaL rEMotE paTiEnt monitoRing), (2020 - 2021, 16k€)
- *Porteur* – projet IDEX Lyon Innovation pédagogique **DARC** (Data Anonymization and Re-identification Competition), (2020 - 2021, 6k€)
- *Porteur* – projet ADT Inria **PRESERVE** (Plate-foRme wEb de SEnsibilisation aux pRoblèmes de Vie privéE), (2019 - 2021, 105k€)
- *Porteur* – projet d'équipe associée **DATA** entre l'équipe Privatics de l'Inria Grenoble Alpes et le groupe de recherche de Sébastien Gambs de l'UQAM de Montréal, (2018 - 2021, 36k€)
- *Porteur* – projet **ANTIDOT** : ANalyse et proTectIon des Données persOnnelles de mobiliTé, la Fédération Informatique de Lyon. (2018 - 2020, 3,7k€).
- *Participant* – projet **ADAGE** : Anonymous Mobile Traffic Data Generation, FUI (2016 - 2018)

- *Porteur* – projet **SeRUM** : Sensibilisation et Responsabilisation des Utilisateurs face aux risques liés à l'Internet Mobile, projet PEPS Sécurité Informatique et des Systèmes Cyber-physiques (SISC) du CNRS, (2015 – 2017, (12,5k€)
- *Participant* – projet Européen FP7 : **EEXCESS** (2013 - 2016), participant et responsable d'un workpackage
- *Participant* – projet Européan FP7 : **AMADEOS** (2013 – 2016)
- *Participant* – projet **MobiCampus** : Users mobility analysis and protection, Labex IMU, (2016-2019)
- *Participant* – projet **Priva'mov**, Labex IMU, (2013 - 2016)
- *Participant* – projet **C3PO** : Collaborating Creation of Contents and Publishing using Opportunistic Networks, projet ANR CONTINT, (2014 – 2016)

2.2.3 Collaborations internationales

- UQAM, Québec (Sébastien Gambs), **Retombées :** 4 publications, 1 équipe associée Inria, de multiples visites pour travailler sur la transparence algorithmique et la confidentialité des information personnelles, 1 bourse de voyage pour présenter des travaux communs à la conférence DTL en 2017.
- Université de Delft, Pays-bas (Jérémie Decouchant), **Retombées :** 3 publications.
- Université de Neuchâtel, Suisse (Pascal Felber, Valerio Schiavoni), **Retombées :** 2 publications, visite en Janvier 2017 pour travailler sur l'exploitation des enclaves Software Guard Extensions (SGX) d'Intel dans la protection des données personnelles.
- IBM Zurich, Suisse (Lydia Chen), **Retombées :** 2 publications.
- EPFL, Suisse (Rachid Guerraoui), **Retombées :** 7 publications, plusieurs visites.
- Université de Cambridge, UK (Eiko Yoneki), **Retombées :** 4 publications, une mobilité de plusieurs mois en 2011.

2.2.4 Conférencier invité

- Décembre 2024 : conférencier invité au PEPR Cyber Day, Paris – Données de santé : allier confidentialité et partage
- Novembre 2024 : Matinée Cyber & Santé. Programme de Transfert au Campus Cyber, Paris
- Novembre 2024 : Les Échappées inattendues du CNRS, Lyon
- Juin 2024 : Webinaire BERNOULLI LAB – L'anonymisation des données de santé est-elle possible ?
- Février 2024 : Webinaire, Entrepôt de Données de Santé et Confidentialité des données, online
- Janvier 2024 : conférencier invité à l'école d'hiver du PEPR Cybersécurité, Autrans – AI and Privacy, a CNIL perspective
- Décembre 2023 : conférencier invité au PEPR Cyber Day, Paris – Données personnelles, comment se préparer aux évolutions des réglementations ?
- Novembre 2023 : conférencier invité au séminaire EkitIA sur les données synthétiques : encadrement et enjeux des techniques et des usages, online – Techniques de génération de données synthétiques, pertinence et risques de leur utilisation
- Juin 2023 : conférencier invité au 13ème Atelier sur la Protection de la Vie Privée (APVP), Bourgogne Franche-Comté, France – Table ronde : Traitement Automatique des Langues

et Vie Privée
- Avril 2023 : conférencier invité à l'école d'été Deep Learning for Medical Imaging school (DLMI 2023) – Privacy in machine learning : from centralized to federated approaches
- Septembre 2022 : conférencier invité au séminaire du groupe de travail CNRS sur la protection de la vie privée (GT-PVP), online – Inférence d'informations sensibles dans l'apprentissage automatique et contre-mesures
- Avril 2022 : conférencier invité au séminaire de la Fédération Informatique de Lyon (FIL), online – Vers une protection à la source des informations capteurs pour se prémunir des inférences d'informations sensibles
- Mars 2022 : conférencier invité au 6ème workshop franco-japanais de Cybersecurité, online – DYSAN : Dynamically sanitizing motion sensor data against sensitive inferences through adversarial networks
- Février 2021 : conférencier invité au 5ème workshop franco-japanais de Cybersecurité, online – DARC : Data Anonymization and Re-identification Challenge
- Novembre 2020 : conférencier invité au workshop ICT4V Privacy & Anonymization, online – DYSAN : Dynamically sanitizing motion sensor data against sensitive inferences through adversarial networks
- Mai 2018 : conférencier invité au 4ème workshop franco-japanais de Cybersecurité, Annecy, France – ACCIO : How to Make Location Privacy Experimentation Open and Easy
- Mars 2018 : Présentation au Shonan meeting on Anonymization methods and inference attacks : theory and practice, Shonan, Japan – A Privacy-Preserving Mechanism for Requesting Location Data Provider with Wi-Fi Access Points
- Présentation au séminaire du groupe Social Computing Group & Security and Privacy Group, MPI-SWS, Allemagne
- Mai 2015 : conférencier invité au premier 3S workshop du LIRIS, Lyon, France

2.2.5 Responsabilités scientifiques

2.2.5.1 Organisation

- Co-organisateur d'APVP 2024, Juin 2024, le domaine Lou Capitelle
- Co-organisateur de l'atelier IPoP : Audits de l'IA, Mai 2024, Paris
- Co-organisateur de l'école d'hiver Réseaux et Systèmes Distribués, 2024 - 2017, Le Pleynet
- Co-organisateur de l'atelier IPoP : Données de santé - Entre Partage et Protection, Mai 2023, Paris
- Organisateur du workshop Health and Privacy-Preserving Machine, Octobre 2021, Lyon
- Organisateur du workshop francophone "Atelier "algorithmes en boîte-noire", Octobre 2019, Lyon
- Organisateur du workshop francophone sur la transparence et l'opacité des systèmes d'information, Avril 2018, Lyon

2.2.5.2 Comité de programme, relecteur

- Relecteur, TheWebConf (WWW), 2025
- Membre du comité de programme, La Cybersécurité en santé : allier sécurité et partage, 2024
- Relecteur, The Lancet, 2024
- Relecteur, Springer Data Mining and Knowledge Discovery, 2024
- Membre du comité de programme, SRDS 2023

- Membre du comité de programme, APVP 2024, 2023, 2022, 2021, 2020, 2019, 2018
- Membre du comité de programme, Compas 2023, 2019, 2018, 2017, 2016
- Relecteur, ACM Digital Threats, 2021
- Relecteur, IEEE Internet of Things Journal, 2021
- Relecteur, MobiQuitous, 2020
- Relecteur, PLOS ONE, 2021 – 2020
- Membre du comité de programme, IWPE 2021, 2020
- Membre du comité de programme, LPW 2021, 2019
- Membre du comité de programme, DIAS 2020
- Relecteur, IEEE Transactions on Big Data, 2020
- Relecteur, IEEE Transactions on Dependable and Secure Computing, 2020 - 2019
- Relecteur, IEEE Transactions on Emerging Topics in Computing, 2020
- Relecteur, IEEE Transactions on Network and Service Management, 2020
- Relecteur, IEEE/ACM Transactions on Networking, 2018
- Relecteur, MDPI, 2020, 2019, 2018
- Relecteur, Social Network Analysis and Mining, 2019
- Relecteur, Nature, 2019
- Relecteur, IEEE Transactions on Services Computing, 2019
- Membre du comité local d'organisation, SRDS 2019
- Publicity Co-Chair, SRDS 2018
- Membre du comité de programme, Middleware, 2017
- Membre du comité de programme, PANS 2018, 2017
- Membre du comité de programme, DSN Student Forum, 2018
- Membre du comité de programme, BiDAS, 2018
- Membre du comité de programme, DeSN 2016
- Membre du comité local d'organisation, SSS 2016
- Relecteur, Computer Networks, 2017
- Relecteur, Adhoc Networks, 2016

2.2.6 Autres activités et responsabilités :

- Depuis 2018 : Membre élu de l'ACM Sigops France (ASF) – Trésorier
- Depuis 2021 : Membre du conseil du laboratoire CITI
- Depuis 2018 : Responsable communication et bibliométrie du laboratoire CITI

Dossier Recherche

Chapitre 3

Introduction

3.1 Context

In a very short time, we have gone from a world where “ambient privacy” was the rule, to one dominated by massive, ubiquitous and precise data collections, where trying to protect our privacy requires constant efforts. If 50 years ago the perceived threat was that of state surveillance (e.g., the SAFARI project led to the creation of the French privacy regulation and the DPA, CNIL in 1978), today, “capitalist surveillance”, a term popularized by Shoshana Zubboff, is a concern of equal, if not greater importance. It was made possible by the ultra-rapid development of the web in the 1990s, smartphones a decade later, and now IoT devices and AI systems of all kinds, and all these technological advances have led to the creation of highly profitable giant companies, most of which exploit user data for profiling and targeting.

Undoubtedly, this digital world has opened up major opportunities, very beneficial for society in general and for individuals in particular. But it also presents considerable risks for privacy that can potentially turn these new technologies into a nightmare if they are not accompanied by appropriate legal and ethical rules. As the French "Loi Informatique et Liberté" (1978) states in its first chapter : "Information technology must be at the service of every citizen. [...] It must not infringe on human identity, human rights, privacy, or personal or public freedom."

Privacy is therefore essential to protect individuals, for example against possible misuse of personal data. Privacy is also essential to protect society, as shown by the misuse of personal data to influence voters in elections (e.g., Cambridge Analytica). But privacy is too important to be left solely in the hands of individuals or companies : the role of regulators and data protection authorities is fundamental from this point of view, leading to regulations (e.g., GDPR) that protect all citizens by default.

In this landscape, public research has a key role to play. By working in a complementary manner, from highly theoretical subjects up to the reverse engineering of deployed systems, or the design of privacy enhancing technologies, public research also contributes to making the world – a little bit – better.

3.2 My contributions

In this context of super-fast development of technologies (often deployed before being regulated), my research work is focused on the protection of privacy. Specifically, I mainly contribute to the domain through technical aspects of privacy, and also through interisciplinary activities. Indeed, privacy issues cannot be resolved through technology alone because they also raise legal, ethical, economic and societal issues that require dialogue with people from different disciplines and the end users at some point.

3.2.1 Contributions presented in this manuscript

Since my recruitment at Insa-Lyon as an assistant professor in fall 2017, my main contributions cover 1) the data leaks associated with Web, smartphone, and IoT, and 2) the security and privacy of AI. For the first part on the data leaks associated with Web, smartphone and IoT, I have mainly focused my research work on **issues related to the collection, the exploitation, and the protection of location data**. For the second part on the security and privacy of AI, developed more recently, my works includes both the "**privacy considerations in ML**" (i.e., identification of risks related to ML technologies and countermeasures), and the "**exploitation of ML for privacy**" (e.g., using NLP for medical report anonymisation).

Contributions presented in this thesis are organized in two parts. The first part focuses on my contributions on the security and privacy of IA (Chapter 4), while the second part presents my contributions on location privacy (Chapter 5). My perspectives for my future work and contributions are mentioned in the last part of the manuscript (Chapter 6).

3.2.2 Other contributions not included in this manuscript

This section presents a brief overview of contributions that are not presented in detail in this manuscript but are related to the general problematic of this thesis.

3.2.2.1 Private and Secure GWAS

Genome-Wide Association Studies (GWAS) identify the genomic variations that are statistically associated with a particular phenotype (e.g., a disease). The confidence in GWAS results increases with the number of genomes analyzed, which encourages federated computations where biocenters would periodically share the genomes they have sequenced. However, for economical and legal reasons, this collaboration could only happen if biocenters cannot learn each others' data. In addition, GWAS releases should not jeopardize the privacy of the individuals whose genomes were used. In this contribution, we introduce DyPS, a novel framework to conduct dynamic privacy-preserving federated GWAS. DyPS leverages a Trusted Execution Environment (TEE) to secure GWAS computations. Moreover, DyPS enables dynamic releases of privacy-preserving GWAS results according to the evolving number of genomes used in the study, even if individuals retract their participation consent. Lastly, DyPS is also tolerant to colluding biocenters without privacy leaks.

We also highlight that privacy-preserving releases of multiple GWASes remain vulnerable if they utilize overlapping sets of individuals and genomic variations. In such conditions, we show that even when relying on state-of-the-art techniques for protecting releases, an adversary could reconstruct the genomic variations and that the released statistics of genomic variations would enable membership inference attacks. We introduce I-GWAS, a novel framework that securely computes and releases the results of multiple possibly interdependent GWASes. I-GWAS continuously releases privacy-preserving and noise-free GWAS results as new genomes become available.

Publications :

- I-GWAS : Privacy-Preserving Interdependent Genome-Wide Association Studies. T Pascoal, J Decouchant, A Boutet, M Völp. Proceedings on Privacy Enhancing Technologies 1, 437-454, 2023.
- Dyps : Dynamic, private and secure gwas. T Pascoal, J Decouchant, A Boutet, P Esteves-Verissimo. Proceedings on Privacy Enhancing Technologies, 2021.

3.2.2.2 Contact Tracing

Contact tracing in case of pandemic is becoming an essential mitigation tool for national health services to break infection chains and prevent the virus from spreading further. To support manual tracing, several countries have been developing contact tracing apps that detect nearby mobile phones using Bluetooth. Such data collection raised privacy concerns and calls for privacy-preserving protocols. During the covid epidemic, I worked on several privacy-preserving exposure notification systems and was lucky to be involved in the development of the French solution, TousAntiCovid.

Publications :

- Vincent Roca, Antoine Boutet, Claude Castelluccia. "The Cluster Exposure Verification (Cléa) Protocol : Specifications of the Lightweight Version." (2021).
- Antoine Boutet, Claude Castelluccia, Mathieu Cunche, Alexandra Dmitrienko, Vincenzo Iovino, Markus Miettinen, Thien Duc Nguyen, Vincent Roca, Ahmad-Reza Sadeghi, Serge Vaudenay, Ivan Visconti, Martin Vuagnoux. Contact Tracing by Giant Data Collectors : Opening Pandora's Box of Threats to Privacy, Sovereignty and National Security. Diss. EPFL, Switzerland ; Inria, France ; JMU Würzburg, Germany ; University of Salerno, Italy ; base23, Geneva, Switzerland ; Technical University of Darmstadt, Germany, 2020.
- Claude Castelluccia, Nataliia Bielova, Antoine Boutet, Mathieu Cunche, Cédric Lauradoux, Daniel Le Métayer, Vincent Roca "DESIRE : A Third Way for a European Exposure Notification System Leveraging the best of centralized and decentralized systems." arXiv preprint arXiv :2008.01621 (2020).
- Antoine Boutet, Claude Castelluccia, Mathieu Cunche, Cédric Lauradou, Vincent Roca, Adrien Baud, Pierre-Guillaume Raverdy "DESIRE : Leveraging the best of centralized and decentralized contact tracing systems." Digital Threats : Research and Practice (DTRAP) 3.3 (2022) : 1-20.
- Guillaume Kessibi, Mathieu Cunche, Antoine Boutet, Claude Castelluccia, Cédric Lauradoux, Vincent Roca. "Analysis of Diagnosis Key distribution mechanism in contact tracing applications based on Google-Apple Exposure Notification (GAEN) framework." (2020).
- Castelluccia, Claude and Bielova, Nataliia and Boutet, Antoine and Cunche, Mathieu and Lauradoux, Cédric and Le Métayer, Daniel and Roca, Vincent. ROBERT : ROBust and privacy-presERving proximity Tracing. 2020.

— M Cunche, A Boutet, C Castelluccia, C Lauradoux, V Roca. On using Bluetooth-Low-Energy for contact tracing. Diss. Inria Grenoble Rhône-Alpes ; INSA de Lyon, 2020.

— Antoine Boutet, Nataliia Bielova, Claude Castelluccia, Mathieu Cunche, Cédric Lauradoux, Daniel Le Métayer, Vincent Roca. Proximity tracing approaches-comparative impact analysis. Diss. INRIA Grenoble-Rhone-Alpes, 2020.

3.2.2.3 Privacy-preserving live streaming

Video consumption is one of the most popular Internet activities worldwide. The emergence of sharing videos directly recorded with smartphones raises important privacy concerns. In this contribution we propose P3LS, the first practical privacy-preserving peer-to-peer live streaming system. To protect the privacy of its users, P3LS relies on k-anonymity when users subscribe to streams, and on plausible deniability for the dissemination of video streams. Specifically, plausible deniability during the dissemination phase ensures that an adversary is never able to distinguish a user's stream of interest from the fake streams from a statistical analysis (i.e., using an analysis of variance).

Publications :

— Jérémie Decouchant, Antoine Boutet, Jiangshan Yu, Paulo Esteves-Verissimo. P3LS : Plausible deniability for practical privacy-preserving live streaming. Symposium on Reliable Distributed Systems (SRDS), 1-109, 2019.

3.2.2.4 Private Web search

By regularly querying Web search engines, users (unconsciously) disclose large amounts of their personal data as part of their search queries, among which some might reveal sensitive information (e.g., health issues, sexual, political or religious preferences). Several solutions exist to allow users querying search engines while improving privacy protection. However, these solutions suffer from a number of limitations : some are subject to user re-identification attacks, while others lack scalability or are unable to provide accurate results. In this contribution, we propose X-Search, a novel private Web search mechanism, and CYCLOSA, a secure, scalable and accurate private Web search solution, both relying on the disruptive Software Guard Extensions (SGX) proposed by Intel.

Publications :

— Rafael Pires, David Goltzsche, Sonia Ben Mokhtar, Sara Bouchenak, Antoine Boutet, Pascal Felber, Rüdiger Kapitza, Marcelo Pasin, Valerio Schiavoni. CYCLOSA : Decentralizing private web search through SGX-based browser extensions. International Conference on Distributed Computing Systems (ICDCS), 2018.

— Sonia Ben Mokhtar, Antoine Boutet, Pascal Felber, Marcelo Pasin, Rafael Pires, Valerio Schiavoni. X-search : revisiting private web search using intel sgx. International Middleware Conference, 2017.

— Albin Petit, Thomas Cerqueus, Antoine Boutet, Sonia Ben Mokhtar, David Coquil, Lionel Brunie, Harald Kosch. SimAttack : private web search under fire. Journal of Internet Services and Applications 7, 1-17, 2016.

— Antoine Boutet, Albin Petit, Sonia Ben Mokhtar, Léa Laporte. Leveraging Query Sensitivity for Practical Private Web Search. International Middleware Conference, 2016.

3.3 Impact of this research

Scientific impact : First of all, my contributions have a scientific impact with the production of a large number of papers published in national and international conferences and journals, or technical reports. When possible, I try to involve students and trainees as much as possible in this scientific production. Moreover, to ensure the reproducibility of results, a fundamental aspect of the scientific approach, I make available and well documented all key resources such as data, code as well as detailed descriptions of the experimental setup.

Collaborating with regulators : I believe that the most efficient way to combat the "surveillance capitalism" doctrine big tech companies, such as GAFA, created is to work with regulators. I am fortunate to work closely with the French Data Protection Agency, CNIL (Commission nationale de l'informatique et des libertés) thanks to the links built with the Privatics team, and more recently in the coordination of the IPoP project where the CNIL is directly involved.

Actions towards the general public : Scientific outreach towards the general public is one of our missions as a researcher. I take it to heart to "vulgarize" and help our fellow citizens understand this highly complex domain, with so many societal implications. I regularly write articles for the non-scientific public and I am interviewed sometimes by journalists (Section 2.2.1). I expect to do more in the next years by proposing mini-conferences and working sessions with the general and young public to promote science and privacy (e.g., upcoming participation to Les Échappées inattendues of the CNRS in November, Section 2.2.4). I would also like to develop participatory research by enlisting the general public in the collection and analysis of (unethical) practices that take place on the Web and mobile applications (Section 6.2.9).

Actions in support of public authorities : Helping public authorities is also part of the missions of an academic researcher. During the COVID19 crisis, I worked on contact and presence tracing protocols, in the context of the public/private StopCovid project-team, contributing to a successful "crisis application". I also contributed, with all the member of this StopCovid project-team, to strengthen the technological and digital sovereignty of the Nation, with a solution focused on the health authority, respectful of our values and choices.

Bring my expertise to health committees and working groups : I am a member of the CIA (Artificial Intelligence committee) of the HCLs (Hospices Civils de Lyon) and bring my expertise in security and privacy of IA. I also regularly bring my expertise on anonymization and AI to a working group gathering several French health data warehouses.

Contributing to international standards : Being able to contribute to standards in order to promote our views and research outputs, is a highly efficient manner to have concrete impacts and balances the influence of lobbying actions from certain companies. I have recently been involved in the Joint standardization on Cybersecurity for AI systems of the CEN/CLC/JTC 21/WG 5.

Chapitre 4

AI Security & Privacy

4.1 Introduction

Advances in Machine Learning (ML) have massively democratized these technologies which currently power a majority of innovative AI services and contribute to numerous scientific discoveries in data-driven fields by performing complex tasks. To ensure trustworthy ML models, several challenges still need to be addressed such as :

- Utility : the model must correctly perform the task for which it was designed ;
- Privacy : in many cases, learning models are trained with datasets that contain personal data and potentially sensitive information, which raises privacy concerns for participants ;
- Security : malicious actors (who are very real in practice) must not be able to abuse a model by biasing or influencing its decisions or its learning process ;
- Fairness : the model must ensure equitable decisions across different demographic subgroups and individuals, and avoid discriminatory behaviour ;
- Explainability : although the model is very complex, its decisions should be interpretable by a human.

Obviously, the main goal is to address all these challenges holistically. However, these challenges may not be aligned and may contradict each other. For example, handling data leakage or bias in decisions may negatively impact model performance. In my work, I have mainly studied the tradeoff between privacy and utility, and I have studied the tension of this tradeoff with other challenges individually (i.e., fairness, explainability, and security).

To safeguard personal information, my contributions in this AI research axis are mainly divided into two parts : 1) **exploitation of ML to improve confidentiality**, and 2) **improvement of the confidentiality of ML**.

In the first part, I leverage the capabilities of ML to propose anonymization or sanitation schemes, or synthetic generation mechanisms for different types of data such as motion data or MRI [RMD+21, BFM23, DJM+20, JBF18]. For instance, to **address the privacy issues related to the widespread adoption of the quantified self movement**, we proposed DySan [BFG+21], a privacy-preserving framework to dynamically sanitize motion sensor data against unwanted sensitive inferences (i.e., improving privacy) while limiting the loss of accuracy on the physical activity monitoring (i.e., maintaining data utility). To ensure a good trade-off between utility and privacy, DySan leverages the framework of Generative Adversarial Network (GAN) to sanitize the sensor data. Compared to static inputs (e.g., images), here the sanitizing scheme dynamically adapts protection based on data from motion sensors which may change depending on user activity. I also started the development of mobile applications relying on the Federated Learning (FL) scheme to process locally the user data without sharing them [BBB+23]. Recently, I also proposed a new synthetic data generation based on conditional sampling [LBA+24] as well as started collaborations with the APHP (a group of hospitals at Paris) and the Léon-Bérard Cancer Center (CLB) in order to propose a new methodology to train NLP models for the anonymization of medical reports (Section 6.2.4). Contributions in this part are further developed Section 4.2.

In the second part, I studied the ML attack surface in order to identify the different ways of inferring personal information and proposed countermeasures. I also analyzed more broadly the limitations of the different ML techniques in terms of explainability, robustness and fairness. Here we considered centralized approaches, as well as the federated approaches that avoid sharing and centralizing data but where inference attacks on the exchanged models remain possible. For instance, I **deeply investigated the Federated Learning (FL)** approach and proposed a protection strategy against inference attacks relying on a proxy that mixes neural network layers [LBAB22], and conducted a privacy assessment of FL schemes using private personalized layers [JBF21]. I also studied the **privacy risks** (e.g., through attribute and membership inference attacks) **introduced by model explanations** [DB22], **graph neural networks** [DBS21], **deep learning for embedded systems** [DBS22], **synthetic data generation** [LBA+24], and mechanisms to ensure **fairness** in the decisions [ADB24]. More recently, I also started **auditing privacy leaks in NLP models** and proposed guidelines to account for and reduce these leaks during model training [BBR23]. This last work is conducted jointly with the HCL (a group of hospitals around Lyon) in order to evaluate the leaks and improve the anonymization pipeline developed internally. Contributions in this part are developed Section 4.3.

4.2 AI for Privacy

In this section, I detail my contributions which exploit the capabilities of ML to propose mechanisms or schemes to improve the privacy for different types of data and use cases. Specifically, I described a new sanitizing scheme to dynamically protect motion sensor data against sensitive inferences in Section 4.2.1, and a new synthetic data generation based on conditional sampling in latent space in Section 4.2.2.

4.2.1 Dynamically sanitizing motion sensor data against sensitive inferences

The integration of motion sensors in smartphones and wearables has been accompanied by the growth of quantified-self movement [YYZC16]. For instance nowadays, users increasingly exploit these devices to monitor their physical activities. Usually, the motion sensor data is not analyzed directly on the device but are rather transmitted to analytics applications hosted in the cloud. These analytics applications leverage machine learning models to compute statistical indicators related to the status of users that are sent back to them. While these analyses can bring many benefits from the health perspective [FS11, PCN17, QYFD15], they can also lead to privacy breaches by exposing personal information regarding the individual concerned. Indeed, a large range of inferences can be done from motion sensor data including sensitive ones such as demographic and health-related attributes [LM02, HON+12, KRB19].

Consider for instance the scenario in which Alice, a woman, uses a fitness application on her smartphone to monitor her physical activity. The application performs the activity recognition as well as the activity monitoring in the cloud. However even if the service provider declares it explicitly in the terms of conditions of the service, Alice has no formal guarantees that her data will not be processed to infer other information about her (e.g., for targeting or marketing purposes). Another possible scenario is related to the new trend of insurance companies that propose discounts to clients if they accept to use a connected device to follow their daily activity [TBO17]. This data can be used to provide personalized coaching for better health management but also for early detection of a pathology, which can negatively impact the insurance cost or lead to other types of discrimination. To address the issues raised by these two scenarios, in this work we propose a solution sanitizing the motion sensor data in such a way that it hides sensitive attributes while preserving the activity information contained in the data.

To achieve this objective, we design DySan, inspired from the framework of Generative Adversarial Networks (GANs) [PYY+19] to sanitize the sensor data. More precisely, by learning in a competitive manner several networks, DySan is able to build models for sanitizing motion data to prevent inferences on a specified sensitive attribute while maintaining a high level of activity recognition. One of the objectives of DySan is also to limit the distortion between the raw and sanitized data, thus also maintaining a high level of utility with respect to other analysis tasks related to activity monitoring (*e.g.*, steps counting). DySan has thus to learn sanitizing models in order to find the best trade-off to cope with these conflicting optimizing goals.

Furthermore, our approach aims at addressing the heterogeneous aspect of data collected by motion sensors. Indeed, these sensor data are user dependent and inherently reflect the way each user moves, to the characteristics of the sensors used for data collection and to the evolution of activity during the day. Thus, one unique sanitizing model cannot cope with the heterogeneity of data and provide the best utility/privacy trade-off for all users over time. To solve this issue, DySan builds a set of diverse sanitizing models by exploring different combinations of hyperparameters leading to different balance in privacy protection with respect to sensitive inference, utility preservation in terms of the loss induced for activity recognition, and the data distortion. By doing so, DySan is able to assess the trained sanitized models and to dynamically select the model providing the best trade-off over time according to the incoming sensor data.

The evaluation of DySan on real datasets, in which the *gender* is considered as the sensitive information to hide due to the possible risk of discrimination, demonstrates that DySan can drastically limit the gender inference up to 41% while only inducing a drop of 3% on the accuracy of activity recognition. In addition to preserving activity recognition, by limiting data distortion, DySan also preserves the sensor data utility for other analytical tasks such as estimating the number of steps. Moreover, we show that the dynamic model selection of DySan successfully provides an adaptation of the sanitization according to the incoming user data. This dynamic model selection is especially useful to generalize the sanitization capacity learnt from the dataset used to build the sanitizer models to another dataset with new users with potentially different behaviours. Our dynamic sanitizing method overcomes several shortcomings of the state-of-the-art approaches, namely the use of the same sanitizing model for all users over time, which may lead to a poor privacy-utility trade-off for atypical users. Lastly, we evaluate the cost of operating DySan on a smartphone and show that the introduced overhead is compatible with real-time processing and that the energy consumption remains reasonable.

See paper at : https://inria.hal.science/hal-02512640v3/

Impacts : This work was done as part of Théo Jourdan's thesis and was published in the AsiaCCS conference in 2021. Several other works on risk quantification as well as anonymization and protection of motion data have complimented this contribution, here is the exhaustive list :

- Loise Bart, El Amine Bechorfa, Antoine Boutet, Jan Ramon, Carole Frindel. A Smartphone-based Architecture for Prolonged Monitoring of Gait. AIMHC 2024.
- Pierre Rougé, Ali Moukadem, Alain Dieterlen, Antoine Boutet, Carole Frindel. Generalizable Features for Anonymizing Motion Signals Based on the Zeros of the Short-Time Fourier Transform. Journal of Signal Processing Systems, 2022.
- Pierre Rougé, Ali Moukadem , Alain Dieterlen, Carole Frindel, Antoine Boutet. Anonymizing motion sensor data through time-frequency domain. MLSP 2021.
- Theo Jourdan, Antoine Boutet, Carole Frinder, Sébastien Gambs, Claude Rosin Ngueveu. DYSAN : Dynamically sanitizing motion sensor data against sensitive inferences through adversarial networks. ASIACCS, 2021.
- Noëlie Debs, Theo Jourdan, Antoine Boutet, and Carole Frindel. Motion sensor data anonymization by time-frequency filtering. EUSIPCO 2020.
- Théo Jourdan, Antoine Boutet, Carole Frindel. Vers la protection de la vie privée dans les objets connectés pour la reconnaissance d'activité en santé. Technique et Science Informatiques (TSI), 2019.
- Théo Jourdan, Antoine Boutet, Carole Frindel. Toward privacy in IoT mobile devices for activity recognition. MOBIQUITOUS 2018.
- These works also led to discussions and collaboration with the ARRPAC center from HCL around the monitoring of post-stroke patients using accelerometric data from phones.

4.2.2 Synthetic Data : Generate Avatar Data on Demand

The collection of personal data has grown to a tremendous proportion and is done through diverse sources such as credit cards, medical records, digital photographs, emails, websites, social media, Internet of Things (IoT), smartphones, wearable technologies, to name a few. All of this data has enormous value for improving the understanding of human behavior and creating useful societal applications, but it also raises serious privacy concerns. For instance, healthcare generates massive amounts of data whose sharing and re-using is essential for accelerating research and to develop robust machine learning algorithms methods that can be deployed in clinical settings. Specifically, this health data can be used to improve the quality of care and knowledge of the health system, identify disease risk factors, assist in diagnosis, monitor the effectiveness of treatments, deliver personalized healthcare value, etc [Len23]. However, this data is very sensitive and must be anonymized before it can be used beyond the purpose of its initial collection.

Anonymization is a complex task that requires calibrating a trade-off between the privacy guarantees (e.g., robustness to privacy attacks) and the remaining usefulness of anonymized data, which is difficult to control and depends on the data and the analysis considered. Thus in practice, a high privacy protection often results in a limited utility. To overcome this limitation, the use of synthetic data that resemble the real data (i.e., which preserves global statistical properties and task-specific performance) is increasingly recognized as a promising way to enable such reuse while addressing personal data privacy concerns [CLC+21]. For example, some projections predict that synthetic data will completely eclipse real data in AI models by 2030 [3]. However, there is still no consensus on a standard approach to systematically and quantitatively assess the privacy gain and residual utility of synthetic data, which slows their adoption. Nonetheless, to shed some light on the real guarantees of synthetic data and help hospitals assess the prospects of this new technology, some papers have started to assess the privacy [ALP+22] and utility [VBM+24] of synthetic data for medical data analyses.

Recently, new approaches based on avatar data have attracted attention for generating synthetic patient-data [GRP+23]. For each individual observation, this approach identifies the k nearest neighbors in a latent space and leverages this neighborhood to generate an avatar through a local stochastic modeling. While appealing these avatar-based approaches lack a proper privacy assessment [KDM+23]. To overcome this limitation, in this contribution we present an extensive utility and privacy assessment of avatar data based on a wide variety of metrics and datasets. More precisely, we quantify the privacy of synthetic data through criteria used to evaluate anonymization schemes according to the GDPR, namely singling-out, linkability and inference. In addition, we have also implemented a re-identification attack (i.e., mapping a synthetic data record to a close original data record) and a membership inference attack (i.e., inferring data records leveraged to generate a synthetic dataset), and focus on the most vulnerable data. We evaluate the utility and compared the avatar approach to different synthetic data generation methods as well as anonymization schemes. Our main objective is to provide a comprehensive assessment of the utility and privacy of avatar data to subsequently facilitate their use in the medical field under the best conditions. We also propose an improvement based on conditional sampling in the latent space, which allows synthetic data to be generated on demand (i.e., of arbitrary size), and which depicts utility and privacy trade-off aligned with the state-of-the-art.

See paper at : https://inria.hal.science/hal-04715055

Impacts :

- Thomas Lebrun, Louis Beziaud, Tristan Allard, Antoine Boutet, Sébastien Gambs, and Mohamed Maouche. Synthetic Data : Generate Avatar Data on Demand. WISE 2024.
- Collaboration with the Octopize startup to have access to their API and to conduct a privacy challenge around their anonymization solution.

4.3 Privacy for AI

In this section, I detail my main contributions which study the ML attack surface in order to identify the different ways of inferring personal information and proposed countermeasures. More specifically, I start to present MixNN, a privacy-preserving service for FL against inference attacks from a curious aggregation server (Section 4.3.1), the interaction between the utility and privacy trade-off and the fairness (Section 4.3.2), the explinability (Section 4.3.3), and the efficiency in embedded deep learning (Section 4.3.4). Finally, I also present a privacy assessment in Graph Embedding (Section 4.3.5) and a training method for NLP models to take into account privacy leakages (Section 4.3.6).

[3] https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai

4.3.1 Protection of Federated Learning Against Inference Attacks

Recently, Federated Learning (FL) [BEG+19] has emerged as promising privacy-by-design alternatives to decentralized learning schemes. In such a collaborative scheme, personal data never leaves the user device. Instead, devices (computing and refining a learning model with their own data) and a central server (aggregating models) work together to build a global learning model. This new ML scheme has attracted much attention these last few years, not only from the research community but also from major Internet companies, suggesting future deployments. For instance, Google already exploits FL for next-word prediction in a virtual keyboard for smartphones [HRM+18]. While the FL scheme is a clear step forward towards enforcing users' privacy, it still suffers from a large ML-based attack surface including membership, property and attribute inference from participants or from the server. Different protection mechanisms to limit inference capabilities of an adversary have been proposed [NHC21]. For instance, some solutions [vdMH20, YBS20] are based on perturbation in order to reveal only noisy information to the server, such as differential privacy. However, these solutions significantly damage the accuracy of the model and its capacity to converge [TB20]. Secure aggregation relying on a cryptographic scheme has been also proposed [BIK+16, BIK+17, BBG+20]. Similar to MixNN, this solution ensures that the server is only aware of the aggregate of all models, keeping the model of each participant (and the associated inference) private. Although the overhead of this solution remains low, the underlying cryptographic scheme requires the participation of the server in the protection. We argue that such solutions are not deployed in practice. Indeed, few companies accept to afford the additional cost of the protection. For instance, Private Information Retrieval (PIR) protocols which follow similar cryptographic schemes to protect the profile of users are not widely adopted in practice. Moreover, a curious or malicious server trying to infer information from participants will certainly not adopt such a protection.

In this context, we propose MixNN, a new privacy-preserving service for FL against inference attacks from a curious aggregation server. To achieve that, MixNN relies on a proxy mixing the layers of the model updates (also named parameter updates) among participants before sending them to the aggregate server. Like Mixnets to ensure anonymity in information routing [Cha81], mixing the layers of the participants' updates of neural network prevents inference attacks (i.e., both membership and attribute inference attacks) without decreasing the accuracy of the aggregated model. This solution, albeit simple, leads to drastically improving privacy without any trade-off with utility. In addition, MixNN is transparent to the FL service, participants only need to configure a web proxy for the associated traffic. To make the deployment of MixNN easier by anyone (e.g., operated by an individual or non profit organizations willing to protect privacy) and possibly on an untrusted infrastructure, the proxy mixing the neural network layers is running inside an SGX enclave ensuring confidentiality and attestation on its behavior. Interestingly enough, the behavior of the proxy can be adapted according to the expected security and privacy guarantees. For instance, the proxy can aggregate itself the model updates or can adopt another aggregation scheme to improve the robustness against model poisoning or backdoor attacks [BVH+20a] by replacing averaging with robust estimators such as geometric median. In this case, the utility-privacy-performance trade-off can change.

To illustrate the capability of MixNN to protect privacy while maintaining the same level of utility, we implemented MixNN and experimentally evaluated it with several datasets and neural network architectures. Moreover, we show that reconstructing the update of a participant (by identifying the layers of an individual among the mixed updates) is costly and a difficult task which gives poor accuracy. Finally, we show that the MixNN proxy is scalable and introduces only a small latency on the model updates.

See paper at : https://inria.hal.science/hal-03795818v1

In another contribution, we quantify the utility and privacy of a FL scheme using private personalized layers [JBF21]. In such a scheme, only the lower layers of the model (capturing coarse grain information) are exchanged with the server while the upper layers of the model (capturing fine grain information) are personalized and kept private on each user. This scheme is known to improve the accuracy of the model in presence of heterogeneous data across users. However, the privacy impact of sharing only a sub part of the model has never been measured. To assess privacy leakage of this scheme, we consider both an attribute and a membership inference attack. Evaluations have been conducted using two datasets of motion sensor data collecting in real-life conditions. Results show that FL with personalized layers speeds up the convergence compared to vanilla FL and slightly increases the activity accuracy while decreasing the gender and the overweight inference for membership. This utility and privacy trade-off is better than a defense scheme using local differential privacy which decreases the inference of the gender and the overweight but at the cost of the activity accuracy. These results tend to show that minimizing the information exchanged with the server is an interesting avenue for both personalizing the model (i.e., improving accuracy) while limiting potential inferences (i.e., improving privacy).

See paper at : https://inria.hal.science/hal-03354722

Impacts :

- Thomas Lebrun, Antoine Boutet, Jan Aalmoes, Adrien Baud. MixNN : protection of federated learning against inference attacks by mixing neural network layers. International Middleware Conference, 2022.
- Théo Jourdan, Antoine Boutet, Carole Frindel. Privacy assessment of federated learning using private personalized layers. International Workshop on Machine Learning for Signal Processing, 2021.

4.3.2 On the Alignment of Group Fairness with Attribute Privacy

Machine learning (ML) models have been adopted for several high-stakes decision-making applications, such as criminal justice and healthcare. To govern this massive deployment of ML models, new AI regulations have highlighted the design of trustworthy models. Trustworthy models rely on several pillars such as privacy, safety, fairness [KDK21, Hou20, Law18, TBH+19, Hig19]. The design of models ensuring all these (potentially conflicting) properties remains an open challenge and requires an understanding of the relationship among them [DSA23].

For instance, to avoid models susceptible to discriminatory behaviour [MMS+21], group fairness algorithms train an ML model optimized for a fairness metric (e.g., equalized odds, demographic parity) to ensure equitable behaviour across different demographic subgroups [ZLM18, ABD$^+$18]. This optimization ensures the conditional independence between the model's predictions and sensitive attributes [ZVGRG19, HPS16]. However, training with group fairness may conflict with different notions of privacy [DSA23]. For instance, group fairness increases the susceptibility to membership inference attacks [CS21] and conflicts with differential privacy [CGKM19, FTVHZ22]. However, there is limited literature on the interaction of group fairness with privacy of sensitive attributes, as measured using Attribute Inference Attacks (AIAs) where an adversary infers sensitive attributes (e.g., race and sex) from model predictions [FLJ$^+$14, DM20, YGFJ18a, SS20, MBG21, MDK$^+$22]. Ferry et al. [FAG$^+$23] indicate a conflict in a restricted setting : fair models are susceptible to AIAs when the adversary knows the fairness metric that the model was optimized on. This assumption is unlikely in practice since companies do not reveal their proprietary training procedures. Hence, the interaction in a blackbox setting where the adversary has no knowledge about the target model is more realistic, but missing in the literature. On the contrary, Zhang et al. [ZOC22] only speculate the alignment of group fairness with attribute privacy without any evaluation. Hence, it is still unclear how different fairness metrics influence the interaction [ZOC22]. Indeed, despite the GDPR's emphasis on safeguarding individuals against attribute inference, this specific privacy risk has not been thoroughly evaluated concerning its trade-off with fairness in ML models.

To address this lack of understanding, we study the following research question : *How does group fairness interact with attribute privacy ?* We formally define attribute privacy as the indistinguishability in the model's predictions for different sensitive attribute values [ZOC22]. Empirically, we evaluate this by checking whether the AIAs are close to random guess. Intuitively, the conditional independence of sensitive attributes on using group fairness should be equivalent to indistinguishability in model predictions. Since this meets the requirement for attribute privacy, we conjecture that there is alignment.

Our goal is to validate this conjecture by examining whether the success of AIAs is close to random guessing, which would imply attribute privacy and indicate alignment with group fairness. However, this is challenging. First, none of the current AIAs account for real-world datasets with significant class imbalance in sensitive attributes making them ineffective in practice. Second, group fairness algorithms can either output soft labels (probability that an input belongs to different classes) as seen for adversarial debiasing [ZLM18, LKC17] or hard labels (most likely class from soft labels) as seen in exponentiated gradient descent [ABD$^+$18]. We have to design AIAs for both settings. We address these by proposing a state-of-the-art AIA, AdaptAIA, to measure attribute privacy. We theoretically analyze the bounds of different fairness metrics to protect against AIAs, and validate these bounds through an extensive experimental analysis using several state-of-the-art datasets.

See paper at : https://hal.science/hal-04739862

Impacts :

— Jan Aalmoes, Vasisht Dudu and Antoine Boutet. On the Alignment of Group Fairness with Attribute Privacy. WISE 2024.

4.3.3 Inferring Sensitive Attributes from Model Explanations

ML models are used for high-stakes decision making for several real-world applications. For instance, these models assist decision makers such as doctors and judges in healthcare and criminal justice [Rud19]. However, the model's high complexity makes it difficult for human interpretation into the decision making process. This creates the need for *transparency* into the model behaviour. Model explanations release additional information to explain the behaviour of complex ML models. Specifically, attribute based model explanations explain the model's prediction on an input by releasing the influence of different input attributes responsible for the prediction [SGK17, ACOG18, LL17, STY17, STK+17].

Some of the input attributes can be sensitive (e.g., race and sex). This raises the data privacy concerns when an adversary can leverage model explanations as an attack surface. For instance, Shokri et al. [SSZ21] show that explanations can be exploited for membership inference (i.e., inferring whether input record was part of training data) and data reconstruction. Additionally, releasing model explanations could leak the values of sensitive attributes which is a privacy risk, not considered in literature. For instance, consider the setting where an ML model is trained to predict the likelihood that a criminal will re-offend as an aid to judges in a court. In addition to output predictions, the model reveals explanations on why it made the prediction on that input. Attribute inference attacks could reveal race and sex from model explanations which individual prefers to keep their private to avoid biased decisions.

However, this quantification of privacy risk of model explanations to *attribute inference attacks* is lacking in current literature. An analysis of this trade-off between privacy and transparency is necessary so that a model builder can make appropriate choices to train ML models for high-stakes applications. In this work, we ask the following research question : *can an adversary exploit model explanations to infer sensitive attributes of individual data records ?* We design an *attribute inference attack* to infer sensitive attributes from model explanations in two threat models : 1) Sensitive attributes are included in the training dataset and the input and the adversary only sees the output predictions but not their inputs. The adversary has no control over passing the inputs but has to infer sensitive attributes from only the observed predictions. And 2) sensitive attributes are not included in training data or input (censored by the model builder for privacy). This corresponds to real-world applications such as ML as a Service (MLaaS).

In this work, we design the first attribute inference attack, to infer sensitive attributes, e.g., race and sex, of the data records from corresponding model explanations. The adversary trains an ML attack model to map model explanations to sensitive attributes. We additionally calibrate the threshold over the attack model's predictions to increase the adversary's power. We then show that our attack successfully infers the sensitive attributes from model explanations. In addition, despite censoring the sensitive attributes, we show that our attack can successfully infer them using model explanations of other non-sensitive attributes. Finally, we show that exploiting model explanations has a higher success than prior state-of-the-art attribute inference attacks which exploit model predictions. This indicates that releasing model explanations increases the attack surface enabling the adversary to mount strong attribute inference attacks.

Impacts :

— - Vasisht Duddu, Antoine Boutet. Inferring Sensitive Attributes from Model Explanations. CIKM 2022.

4.3.4 Reconciling Privacy, Accuracy and Efficiency in Embedded Deep Learning

Embedded systems demand on-device processing of data using Neural Networks (NNs) while conforming to the memory, power and computation constraints, leading to an efficiency and accuracy tradeoff. To bring NNs to edge devices, several optimizations such as model compression through pruning, quantization, and off-the-shelf architectures with efficient design have been extensively adopted. These algorithms, when deployed to real world sensitive applications, are required to resist inference attacks to protect the privacy of user's training data. However, resistance against inference attacks is not accounted for designing NN models for embedded systems. In this work, we analyse the three-dimensional privacy-accuracy-efficiency tradeoff in NNs for embedded systems and propose Gecko training methodology where we explicitly add resistance to private inferences as a design objective. We optimize the inference-time memory, computation, and power constraints of embedded devices as a criterion for designing NN architecture while also preserving privacy. We choose quantization as design choice for highly efficient and private models. This choice is driven by the observation that compressed models leak more information compared to baseline models while off-the-shelf efficient architectures indicate poor efficiency and privacy tradeoff. We show that models trained using Gecko methodology are comparable to prior defences against blackbox membership attacks in terms of accuracy and privacy while providing efficiency.

Impacts :

— Vasisht Duddu, Antoine Boutet and Virat Shejwalkar. Towards Privacy Aware Deep Learning for Embedded Systems. ACM/SIGAPP Symposium On Applied Computing - System Software and Security - Embedded Systems Track, April, 2022.

4.3.5 Quantifying Privacy Leakage in Graph Embedding

Graph embeddings have been proposed to map graph data to low dimensional space for downstream processing (e.g., node classification or link prediction). With the increasing collection of personal data, graph embeddings can be trained on private and sensitive data. For the first time, we quantify the privacy leakage in graph embeddings through three inference attacks targeting Graph Neural Networks. Our *membership inference attack* aims to infer whether a graph node corresponding to an individual user's data was a member of the model's private training data or not. We consider a *blackbox* setting where the adversary exploits the output prediction scores and a *whitebox* setting where the adversary also has access to the released node embeddings. Our attack provides accuracy up to 28% (blackbox) and 36% (whitebox) beyond the random guess by exploiting the distinguishable footprint between train and test data records left by the graph embedding. In our *graph reconstruction* attack, the adversary aims to reconstruct the target graph given the corresponding graph embeddings. Here, the adversary can reconstruct the graph with more than 80% of accuracy and infer the link between two nodes with $\sim$30% more accuracy than the random guess. Finally, we propose an *attribute inference attack* where the adversary aims to infer the sensitive node attributes corresponding to an individual user. We show that the strong correlation between the graph embeddings and node attributes allows the adversary to infer sensitive information (e.g., gender or location).

Impacts :

— Vasisht Duddu, Antoine Boutet, Virat Shejwalkar. Quantifying Privacy Leakage in Graph Embedding. Mobiquitous 2020.

4.3.6 Toward training NLP models to take into account privacy leakages

Healthcare is an important generator of massive data collected from different sources. The use of this valuable data has many advantages and promises : improve quality of care, acquisition of a better knowledge of the health system, identification of disease risk factors, assistance in diagnosis, choice and monitoring of the effectiveness of treatments, deliver personalized healthcare value, epidemiology, etc. A large part of this data corresponds to text documents (e.g., medical reports). With the rise of machine learning and the advent of Natural Language Processing (NLP) models are increasingly used to automate the processing of medical documents and reports [CLA22, YKS+22, WZA+22].

In recent years, a need to share medical data between various healthcare centers has emerged. This need was all the more felt during the SARS-Cov-2 pandemic for example, where the objective was to propose epidemiological models taking into account data from all over the world. However, patient medical records are extremely sensitive and private data. Their use and distribution is therefore subject to numerous regulations such as HIPAA, for the USA, or GDPR for Europe. In these regulations, one of the main prerequisites for the dissemination of medical data is to remove any elements that can be used to trace a patient directly (i.e., de-identification) or indirectly (i.e., anonymization).

The de-identification of medical documents is a complex task, costly in time and sometimes requiring several doctors which can slow down research. However, recent advances in NLP [SM21] based on neural networks have shown encouraging results. Indeed, NLP has grown in popularity since the advent of ChatGPT, yet NLP-models are not limited to text generation, and can include multiple tasks including classification, named entity recognition, and thus the de-identification of free texts. Johnson et al. for example proposed to use a neural network based on a BERT architecture [ea20] to detect a certain number of identifying elements in medical documents. More recently, different hospitals have also explored the feasibility of using NLP-models to pseudonymize (i.e., hiding specific direct identifiers) text documents from their clinical data warehouse [TWC+23, RTG23].

Although the use of language models to automate the processing of medical documents and to remove personal information (or pseudonymize them by replacing direct identifier to pseudonym) in order to facilitate their sharing is appealing [DJT+22], the attack surface of these models trained on personal and highly sensitive data is still poorly understood [WUR+22, LJP+21]. Thus, additionally to the evaluation of the quality of the de-identification itself, using NLP-models to process medical reports still poses a number of threats related to leakage of sensitive information used during the training of models. Specifically, there are a few known privacy vulnerabilities involving the training data (i.e., a large corpus of medical documents) associated with machine learning and NLP-models [LXW+21] mostly considered individually such as counterfactual memorization [ZIL+21] (i.e., memorisation of rare data), data extraction or reconstruction [GKR+23, PB21], and membership inference attacks [JRY21, MGU+22, WXH+22] (i.e., identifying elements of the training data). To reduce these risks, mitigation techniques have been proposed such as Differential Drivacy [ACG+16] (DP) or pruning strategy [WXH+22]. However, these mitigation techniques drastically degrade the accuracy of the model making them unusable in practice.

Compared to the state-of-the-art which addresses privacy risks individually, in this contribution we audit and quantify all these risks as well as the impact of a DP-based mitigation technique through a complete study using a NLP-model on medical reports. Moreover, we propose a training methodology to limit privacy leaks of sensitive information directly during the training phase. Specifically, instead of using cross validation to define the value of hyper-parameters only according to the performance of the model, in this methodology hyper-parameters are defined according to both the model accuracy and the information leakage. We argue that a model calibrated to also take into account the potential information leakage provides a better utility and privacy trade-off than using a mitigation strategy based on DP which degrades the utility of the model to much. In addition, sensitive information subject to counterfactual memorisation can be also discarded during the training phase.

See paper at : `https://hal.science/hal-04299405`

Impacts :

- Gaspard Berthelier, Antoine Boutet, and Antoine Richard. Toward training NLP models to take into account privacy leakages. BigData 2023.
- Gaspard Berthelier, Antoine Boutet, and Antoine Richard. Privacy leakages on NLP models and mitigations through a use case on medical data. ComPAS, 2023.
- Collaboration with the HCL to improve their pipeline of anonymisation of medical reports.

Chapitre 5

Location Privacy : from data protection to user awareness

5.1 Introduction

As the profiling has become the norm on the Internet, the personal data of users is massively collected without the consent of the individuals concerned [EN16]. Due to the wide adoption of mobile devices, the location data of users is obviously part of the tracking and the most extensively collected data [ASS+15]. This tracking is usually performed through the usage of mobile applications exploiting the location of users. In these Location-Based Services (LBS for short), the position of the user is usually sent to a distant server, which process it to provide contextual and personalized answers or simply to store this information for profiling purpose (e.g., sometimes the location is collected even if it is not necessary for the application). Although these LBS can provide useful information for users, once the data has been collected by a third party nothing prevents it from analyzing and possibly sharing the collected information for commercial purpose, which opens the door to many privacy threats. Examples of such leaks of personal information are regularly covered by news media and can include sensitive information such as the HIV status of the users (e.g., Grindr [Moy18]).

The providers of Internet services and mobile applications have a strong incentive to profile users based on the personal information collected and to monetize these profiles for targeting purposes. Indeed, the monetization of user profiles is the main source of funding for most of these providers. Despite the fact that web tracking has been a very well investigated field of study for almost 20 years, in the mobile context determining which information is collected and how it is processed and used remains a challenging issue [DTD15]. As a result, this lack of transparency coupled with the emergence of controversial practices such as discrimination [HSL+14] raises serious concerns. A recent study shows that the most sensitive and valued category of personal information is location [SOL+14]. However, end users are usually not aware about this profiling as well as the type and the accuracy of the information that can be inferred from their location histories as well as the associated privacy and discrimination issues.

The privacy issues raised by location data have received quite a lot of attention in the last few years. In particular, recent works have demonstrated that mobility is a very rich contextual information in the sense that it has a strong inferential potential in terms of information that can be predicted about the individuals whose movements are recorded. For instance, researchers have shown that analyzing mobility traces can reveal personal data about individuals such as their points of interests (*e.g.*, home and place of work) [GKNdPC11], their race and gender [ZYZ+15], their social network [SD14] as well as to predict their mobility [SK12], to link accounts of the same user across different datasets [RKC+16] and to uniquely identify users from anonymous datasets or to conduct a de-anonymization attack [GKdPC13]. Moreover, it is possible to analyze the semantics of these mobility traces to infer even more sensitive information such as their religion [FB15]. In addition, other studies have also demonstrated that the location is used for price discrimination [MGEL12].

To mitigate these privacy problems, many location privacy protection mechanisms (LPPMs for short, or just protection mechanisms) have been proposed in the literature [PBMB18]. Their goal is to protect location privacy of users while still allowing them to enjoy geolocated services. There is a rich literature about existing LPPMs. Some of them are rather generic and can adapt to a lot of situations while others are very specific to a single use case. LPPMs rely on a wide array of techniques, ranging from data perturbation (e.g., [ABCP13, GKS07, JSB+13]) to data encryption (e.g., [ZGH07, MFB+11, PBBL11]), and including fake data generation (e.g., [QLM+11, PGDV+11, KYS05]).

In this context, I have mainly contributed by proposing **LLPMs enabling dynamic configuration** to automatically adjust a set of privacy and utility objectives [PBMB16, CPB+17, CBR+18], the production of a **public dataset** of location data [MBB+17], a **framework to ease the design and evaluation** of protection mechanisms [PMB+18], a **quantification of the uniqueness of human mobility** captured from multiple sensors [BBM21], a solution to preserve **users' privacy when requesting users' location from Wi-Fi** [BC21], and more recently tools and studies to **raise user awareness about the privacy threats** associated with the disclosure of their location data [BG19].

In this section, I present a sample of my contributions divided into two parts : 1) **improving location data protection by taking into account the dynamic aspect of location data** (Section 5.2), and 2) **raising users' awareness of the privacy threats associated with the disclosure of their location data** (Section 5.3).

5.2 Adaptive Location Privacy

With the increasing amount of mobility data being collected on a daily basis by location-based services (LBSs) comes a new range of threats for users, related to the over-sharing of their location information. To deal with this issue, several location privacy protection mechanisms (LPPMs) have been proposed in the past years. However, each of these mechanisms comes with different configuration parameters that have a direct impact both on the privacy guarantees offered to the users and on the resulting utility of the protected data. In this context, it can be difficult for non-expert system designers to choose the appropriate configuration to use. Moreover, these mechanisms are generally configured once for all, which results in the same configuration for every protected piece of information. However, not all users have the same behaviour, and even the behaviour of a single user is likely to change over time. To address this issue, we propose different frameworks enabling the dynamic selection and configuration of LPPMs based on utility and privacy modelling [CPB+17, CBR+18].

Impacts :

- Antoine Boutet, Mathieu Cunche. Privacy Protection for Wi-Fi Location Positioning Systems. Journal of Information Security and Applications, 2020.
- Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, Brunie Lionel. The Long Road to Computational Location Privacy. IEEE Communications Surveys and Tutorials, 2018
- Vincent Primault, Mohamed Maouche, Antoine Boutet, Sonia Ben Mokhtar, Sara Bouchenak, Lionel Brunie. ACCIO : How to Make Location Privacy Experimentation Open and Easy. ICDCS 2018, Vienna, Austria.
- Sophie Cerf, Sara Bouchenak, Bogdan Robu, Nicolas Marchand, Vincent Primault, Sonia Ben Mokhtar, Antoine Boutet, Lydia Y. Chen. Automatic privacy and utility preservation for mobility data : A nonlinear model-based approach. IEEE Transactions on Dependable and Secure Computing. 2018.
- Sophie Cerf, Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, Robert Birke, Lydia Y. Chen, Sara Bouchenak, Nicolas Marchand and Bogdan Robu. Achieving privacy and utility trade-off in mobility database with PULP. SRDS 2017, Hong kong, China.
- Sonia Ben Mokhtar, Antoine Boutet, Louafi Bouzouina, Patrick Bonnel, Olivier Brette, Lionel Brunie, Mathieu Cunche, Stephane D'Alu, Vincent Primault, Patrice Raveneau, Herve Rivano, Razvan Stanica. PRIVA'MOV : Analysing Human Mobility Through Multi-Sensor Datasets. NetMob Conference, 2017, Milan, Italy.
- Sophie Cerf, Bogdan Robu, Nicolas Marchand, Antoine Boutet, Vincent Primault, Sonia Ben Mokhtar, Sara Bouchenak. Poster : Toward an Easy Configuration of Location Privacy Protection Mechanisms. Middleware, 2016.
- Sophie Cerf, Vincent Primault, Bogdan Robu, Nicolas Marchand, Antoine Boutet, Sonia Ben Mokhtar, Sara Bouchenak. Données de mobilité : protection de la vie privée vs. utilité des données. ComPAS, 2017.
- Vincent Primault, Sonia Ben Mokhtar, Antoine Boutet and Lionel Brunie. Adaptive Location Privacy with ALP. SRDS 2016, September 2016, Budapest, Hungary.

5.3 Raising user awareness on privacy threats associated with disclosing his location data

Effectively protecting personal data is challenging. A large number of protection mechanisms have been proposed in the academic literature but very few have been adopted by application providers and companies in the field and even fewer by users. Furthermore, protecting data inherently also comes with a loss of service and utility (i.e., there is no free food). Data anonymization is rarely perfect, we are talking more about the probability of risk depending on the means used by an adversary to carry out a re-identification attack (e.g., collection of auxiliary information, etc.). For instance, the risk is necessarily greater for atypical people than for people more similar to the rest of the population.

However, before asking the question of how to effectively protect our location data and what risk remains, perhaps it is better to ask whether it is relevant to generate this data and let third parties collect it. Indeed, without collected personal data, the risk of disclosing information in an uncontrolled manner does not exist. Raising user awareness by better explaining the risks would reduce the amount of data generated and collected. In this context, I mainly developed a demonstration of awareness tools and conducted a study of user perceptions of the risks of information leaks [BG19]. These contributions are described below.

The main objective of the demonstration is to raise the user's awareness about the profiling capabilities related to the disclosure of their personal location data and the associated privacy and discrimination threats. More precisely, users are invited to analyze their own location history collected and provided by Google for ensuring data portability (i.e., one of the new right that has appear with the General Data Protection Regulation) and to inspect the information that can be inferred from the collected data. Then, we build and present to users a contextual profile combining location data with semantic information deduced from the mobility traces such as their points of interests, their home and work places, and the associted demographic information. Moreover, the rationale behind each information appearing in the contextual profile presented to users is also detailed to users so that can better understand how such inference was possible.

In addition to the inference attacks presented previously, an important literature has been devoted to developing protection mechanisms for location data this last decade. However, none of the proposed LPPMs have been adopted by mobile applications and the location data of users are still collected without any protection in real-life. In this demonstration, we will propose the users to apply an LPPM on their mobility traces implementing the privacy notion of Geo-indistinguishability [ABCP13]. Thus, users can both visualize the mobility traces as outputted by the LPPM and inspect the associated inferred information and contrast it with inferences performed on the raw mobility traces. Finally, users are also invited to provide feedback on the accuracy of the information shown as well to quantify their level of (un)comfort with the disclosure of this personal information.

See paper at : `https://inria.hal.science/hal-02421828/`

With this increasing exposure of privacy risks, understanding how young users share their data, and the extent to which they are aware of security and privacy risks, are also important properties to assess these risks and to develop effective Privacy and Transparency Enhancing Technologies matching current users' expectations to increase adoption. Although user perceptions of technology [Bec03] and the privacy paradox [4] [BdJ+19, KJ21] have received a lot of attention, users' self-reported behaviors in mobility contexts associated with smartphones, and the impact of an awareness demonstration platform have been less studied in the academic literature. To fill this gap, we address two research questions : 1) what are the perceptions and the understanding of young users' privacy and its protection in a mobility context ?, and 2) what is the impact of a demonstrator for the visualization of location traces and associated privacy risks on these perceptions and understanding ?

In order to answer our research questions, we designed and deployed survey questionnaires answered by $n = 99$ young participants from Insa students (i.e., digital natives, persons who grew up in the information age, aged between 20 and 26 with an average of 21 in our case). Specifically, we explored the participants' behavior, self-reported behavior, and awareness regarding their own data-sharing practices. We devised a first questionnaire to study their perception of privacy and to their permission management with respect to location, in which we emphasized the eventual discrepancies between their remembered practices and their actual behaviors. We also surveyed participants' understanding of privacy risks before and after exposing them to location traces demonstrating what information can be inferred from this data (through the de-

[4] The privacy paradox refers to self-reported concerns about privacy appear to be in contradiction with often careless online behaviors.

monstration tool presented above), as well as their awareness of protection tools, in conjunction with a second questionnaire.

Our results show that participants have **risky practices in terms of privacy** where more than half of participants underestimate the number of mobile applications to which they have granted access to their data. In addition, most of the participants tend to forget about disabling location access permission for apps that are not actively used. Our results also show that participants are **poorly aware of the privacy risks** and are unable to list cases of personal data leaks or scandals linked to their uses despite the media coverage. Moreover, by using a demonstrator to perform inferences from location data, more than half of the participants (57%) are surprised by the extent of potentially inferred information and 47% intend to reduce access to their data via permissions. Finally, most of the participants are inclined to **better use protection tools in the future**, even if they are still little aware of the available tools to improve privacy today.

Impacts :

— Antoine Boutet, and Victor Morel. "I'm not for sale" – Digital natives are still unaware of privacy risks associated with sharing mobility data. Under submission (ICWSM 2015).

— Antoine Boutet, Sébastien Gambs. Demo : Inspect what your location history reveals about you ; Raising user awareness on privacy threats associated with disclosing his location data. CIKM 2019, Beijing, China.

Chapitre 6

Conclusion & Perspectives

6.1 A short conclusion

The collection of personal data is a subject firmly grounded in public debates. The growing awareness of the population on privacy issues led to stronger regulations on data protection (e.g., GDPR, HIPAA) and contributed to the appearance of new services making privacy an incentive vector such as privacy-based search engine (e.g., Duckduckgo, Qwant), web browsing (e.g., Web Proxy, Tor, Brave), or mailing (e.g., Protonmail).

However, the ever-increasing digitalization of our society exposes individuals to omnipresent data collection. This massive collection of information exposes individuals to new risks ranging from the disclosure of sensitive information to increased manipulation through personal cognitive biases facilitated by the exposure of our personality. The emergence and development of AI have only amplified these risks despite the regulatory effort at different levels (e.g., AI Act) which highlights fundamental freedoms. Between opportunities and threats, the economic and societal issues are enormous, monopolizing many influential actors with different objectives. In this context, academic research on privacy is at the heart of these issues and fuels multiple work perspectives in this field of research. Furthermore, academic research, which aims to be removed from commercial considerations, helps to objectively shed light on debates on the real risks to privacy and to counterbalance the biased positions taken by industrialists, who are often influential with decision-making bodies.

6.2 Perspectives

In this section, I list some future work that I would like to pursue in the coming years.

6.2.1 Towards an evolution in the characterization of the risk of re-identification of medical images

With the rapid advances in technology and the development of AI in recent years, increasingly powerful tools are being made available to the general public (and potentially used by malicious actors) for a multitude of tasks such as face recognition systems. Face recognition systems are increasingly deployed for the authentication process as well as mass surveillance programs [Hil23]. These systems are typically built by scraping publicly available images from social media. Smart cameras equipped with facial recognition are becoming a new threat to privacy. However, this risk does not only concern images and facial recognition tools can be used also on medical imaging.

In recent years, medical professionals have increasingly relied on various imaging technologies for diagnosing patients. Brain imaging, in particular, has seen remarkable advancements, with Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Computed Tomography (CT) being among the key imaging modalities. Moreover, MRI, which continues to evolve in terms of both modalities and contrast techniques, provide diverse capabilities for visualizing brain tissues, ranging from highlighting fat (T1-weighted), fluids (T2-weighted), and lesions (FLAIR, DWI) to mapping functional activity (fMRI) and vascular structures (SWI, MRA) and assessing macromolecular content (MTI).

However, this rapid enhancement in brain imaging quality, along with the widespread sharing of tagged personal images on social media, raises concerns regarding the potential for re-identification using facial recognition software. Instances of re-identification have been reported for anatomical MRI images [MRZ+21, SKT+19, SPJJ20], PET images [SKL+22, SKL+22], CT images [PM17], as well as functional MRI (fMRI, dMRI, and ASF) images [SKA+23]. To mitigate this risk, de-facing software has been developed to obscure or replace part of the human faces [GNjl+22, SKW+21]. Nevertheless, these mitigation measures often compromise image utility and do not provide complete protection against re-identification [AE19, SKT+20].

In the pursuit of automating face recognition from medical images, methods for face reconstruction typically involve creating an isosurface and applying skin rendering [MRZ+21, SKT+19]. Recognition is then achieved through geometric structural properties [LSZ+10] or dedicated neural networks [HYY+15]. Commercial face recognition software, such as Amazon Rekognition [5], Deep Vision AI [6], Face++ [7] or Microsoft Azure Cognitive Services Face API [8] is increasingly integrated into face databases.

The rapid evolution of imaging technologies and the accessibility of certain tools gives rise to significant privacy questions. Legal frameworks like the European General Data Protection Regulation (GDPR), the Canadian Personal Information Protection and Electronic Documents Act (PIPEDA), and the California's Consumer Privacy Act (CCPA) mandate the quantification of privacy risks and the effective protection and anonymization of personal data, particularly sensitive health-related information. These laws take into account evolving practices, available tools, and adversaries's capacities for identification. For example, under GDPR Article 29 [2914], there is a strong emphasis on considering contextual factors and all potential identification methods, especially in light of recent technological advancements and increased computing power. Therefore, the evolving landscape of face reconstruction and recognition tools calls a reassessment of re-identification risk and the implementation of appropriate protection measures.

While re-identification is a central concern, privacy violations encompass more than just this risk. GDPR and similar regulations emphasize safeguarding against three primary risks : singling out, linkability, and inference. These risks extend beyond mere identification, allowing data to be linked, and potentially disclosing highly sensitive information.

[5] https://aws.amazon.com/fr/rekognition/
[6] http://deepvisionai.in/
[7] https://www.faceplusplus.com/
[8] https://azure.microsoft.com/en-gb/products/cognitive-services/face

Imaging data, although not publicly available, resides in health data warehouses. Yet, this data is increasingly shared for research purposes or with private laboratories. Hospitals, in particular, face cybersecurity attacks, leading to data breaches in recent years [9]. Therefore, safegaurding this medical data at its source is crucial to prevent re-identification risks.

It is therefore important to reassess privacy violations related to brain imaging, considering the evolving landscape of available tools and data. Indeed, to establish the privacy risk of both assessments, we take into account the severity of the impacts (which depends on the significance of the consequences and how difficult it would be for subjects to overcome them) and likelihood (which depends on the feasibility of the threat and its motivation). With the evolution of facial recognition and reconstruction tools, as well as their easier accessibility, we need to consider a higher likelihood.

6.2.2 Identify limits of Federated Learning

FL is becoming very popular and fuels a lot of expectations from the industry (the global FL market size was USD 112.7 million in 2022 and is expected to register a mean annual growth rate of 10.5% until 2032 [10]). It is intended to be used either to enable collaborative training of a learning model between different centers (i.e., cross-silo such as bio centers) or user devices (i.e., cross-device such as mobile phones). Although the FL drains a lot of expectations, it still faces many challenges [KMA+21, HHL22] including several privacy and security limitations. Specifically, model updates sent by participants leak information on their training data [YGFJ18b, MPP+21, GBDM20, CML+23]. In addition, FL lacks robustness against malicious parties poisoning or abusing the system [BCMC19, BVH+20b, FVL21, SCZ+23, GGP23]. These limitations are serious barriers to the broad adoption of FL. Indeed, the confidentiality of the information used and the truthfulness in the system are essential not only for privacy purposes but also to protect the competitive insights of the participating companies (i.e., specifically in the case of cross silo FL).

Beyond privacy or proprietary data constraints, the incentive for a participant to adopt a cross silo FL relies on the possibility of learning a model that produces a better accuracy and generalisation than if it follows an isolated training using only their own local data. Without this learning gain, a participant has no interest in using FL. The privacy constraint does not occur with isolated and local learning because no information (i.e., data or model) is shared. Although this gain has been shown in medical use cases bringing together several actors (e.g., hospitals) where the number of patients in each of the participants was very small [HMM+0, PBE+22], it is not clear if a gain is achieved in other use cases. To address this limitation, [CJL+24] has proposed an optimization scheme for improving the accuracy gain of the global model from FL for most participants compared to the local model [CJL+24]. However, this paper does not consider a cross silo FL, the proposed scheme required to share more information with the aggregation server which increases the information leakage, and the considered datasets are artificial and do not depict an heterogeneity. Further research on learning gains is needed to better understand in which situations a cross silo FL is beneficial and in which not, and which participants would benefit most.

[9] https://www.upguard.com/blog/biggest-data-breaches-in-healthcare
[10] https://www.emergenresearch.com/industry-report/federated-learning-market

6.2.3 Leveraging unlearning to remove poisoning

The concept of machine unlearning has been introduced [NHN+22, SAKS21, TCMK23] to implement the right to be forgotten covered by new privacy regulations. While removing data from back-end databases satisfies the regulations, doing so is not sufficient in the AI context as machine learning models often 'remember' the old data.

Investigating how machine unlearning mechanisms could be leveraged in other use-cases than removing personal data used in the training could be an interesting perspective. Specifically, unlearning mechanisms could be used to remove poisoning, backdoors or bias introduced by a malicious party in a learning model. Unlearning could also be used to remove the influence of data coming from outliers which have a negative impact on the models. Similarly, unlearning could be used locally by a participant in a FL system to personalize the model received by the server by removing information in the model which does not correspond to him. Lastly, by inspecting the impact of unlearning mechanisms on a model, we could also be able to infer if some data points have been used in the training. This observation could be used to implement a Membership Inference Attack or to audit the model and evaluate if some data (protected ones for instance) have been actually used in the training or not.

6.2.4 Going beyond pseudonymization with language models

A strong need for sharing medical data between different health centers has emerged. Sensitive data by nature, medical data are subject to strict supervision and must be protected before being shared via pseudonymization or anonymization for example. NLP technologies are increasingly used to create specialized language models through training from medical reports to perform classification tasks or automatic processing of reports such as pseudonymization [TWC+23, RTG23]. Before sharing these models, it is necessary to ensure that these models cannot leak personal information from training. However, various studies highlight the possibility of extracting data from language models refined on a corpus.

In order to ensure that models do not regurgitate personal information (and allow medical structures to publish and share their models), we propose a masking method (i.e., training with the aim of specializing a foundation model) to avoid the memorization of directly and indirectly identifying information. This ongoing work is conducted in collaboration with AP-HP, the Léon Bérard center (Unicancer), and the CNIL.

6.2.5 Training learning models with synthetic data : assessing risk propagation

The deployment of ML and AI is becoming widespread in all areas ranging from cars, medicine, IoT, multimedia, cybersecurity to name a few. This race is fueled by a strong hope for innovation and new services such as autonomous cars, personalized medicine, better understanding of life, advanced and personalized service for instance. This generalization of AI requires being fed with a large quantity of data in order to train the underlying learning models. At the heart of a large amount of innovative service is the human, therefore, these learning models are fed by a lot of personal information which is continuously and massively collected and resold.

This omnipresence of personal information in training data opens up a new attack surface that is still poorly understood. A large number of attacks have appeared in recent years which showed a risk of reconstruction of training data, inference of sensitive attributes on individuals, inference of membership on training data for instance. These new privacy risks come up against regulations governing the use of personal data and are framed in new regulations on AI (e.g., IA Act).

In order to reduce these risks on privacy and better comply with regulations, the generation of synthetic data has been largely adopted. This technique relies on a generative model which is learned to artificially generate data which has the same statistical properties to the training data. Hence, instead of sharing raw personal data (or anonymized), synthetic data points are not linked to any individual and so can be shared with less restrictions. This new El Dorado of synthetic data for the sale or sharing of data outside the GDPR is attracting a lot of interest and many startups or services have emerged to meet the demand. This new economy also fuels the need for data to feed learning models through training with synthetic data.

Although training from synthetic data reduces the risk, synthetic data also carries a risk of leaking information compared to the raw data used such as membership inference and attribute inference. However, no study has focused on the propagation of the risks of using synthetic data instead of real data for model training.

Investigating this risk propagation before a large adoption of such practice could be interesting. Specifically, studying how membership and attribute inference are impacted by synthetic data would shed light on the limitations and risks and provide recommendations for the generation of synthetic data for this use.

6.2.6 Helping hospitals meet their challenges

The healthcare sector generates huge amounts of data. The exploitation of this valuable data would allow the development of a large number of clinical research projects. However, this data is sensitive and subject to numerous regulations, which limits its use. Several obstacles prevent the exploitation of this data from being facilitated : 1) the lack of tools to objectify and quantify the risks to help researchers to better establish their requests for access to certain data, and to assist in the decision-making of the ethics committees that must rule on these requests ; 2) a lack of expertise to evaluate and improve the protection measures implemented when accessing data ; 3) a lack of expertise and tools to evaluate the risks associated with sharing learning models trained on sensitive data. This lack of expertise and audit tools does not allow health centers to position themselves on a potential adoption of emerging AI-based solutions such as natural language processing models, federated learning, or synthetic data generation. Given the strong impact that greater exploitation of health data would have, it seems important to help these key players to meet their challenges.

6.2.7 Accounting the carbon footprint in the design

The list of challenges associated with the massive deployment of AI mentioned in Section 4.1 is not exhaustive. An important point is the adequacy with finite planetary resources. On the Web, tracking largely fuels the advertising targeting ecosystem. Online advertising not only has an impact on privacy, it also plays a role in the planet's carbon footprint. Indeed, on the user side, the tracking generates a significant additional cost (i.e., an overhead) in terms of communication, latency in displaying useful content and higher consumption of our phones' batteries. Moreover, on the server side, the cost of the infrastructures set up for tracking as well as the cost related to the expensive learning of AI models are also very important. Highlighting this significant carbon footprint associated with the collection of personal information through tracking could be a lever for changing practices. Since planetary resources are not inexhaustible, the carbon footprint of the collection and choice of protection mechanisms should also be taken into account in the design of the solution.

6.2.8 Investigate new form of tracking

New forms of tracking are evolving as fast as technologies. On the web, we are seeing the emergence of tracking based on browser fingerprints or tracking carried out on the server side without leaving a trace on the client side. On mobile (for physical tracking), we are seeing the emergence of tracking practices based on the microphone for geolocation (with the democratization of voice control, a request for microphone access seems less intrusive and raises less concern from users than a direct request for location). All these new forms of tracking are still poorly understood and poorly supervised by regulations. Indeed, regulation is not evolving as quickly as the big tech sector. In this context, I was interested in the tracking capacity of the phone's motion data (i.e., accelerometer and gyroscope data). This data gives a vibration footprint of transport lines (e.g., metro, tram) and our preliminary results show that they could be used for geolocation purposes (based on a library of already known fingerprints, a bit like the shazam application for music). Access to this data is by default accessible by applications, which raises questions. Having a better understanding of the privacy risks associated with this type of data (particularly in relation to physical tracking) would allow us to better regulate practices.

6.2.9 Participatory Research

In order to analyze the personalization mechanisms on the web and mobile applications and to better understand the targeting of content of all kinds, we initially developed bots to control and simulate human interactions. This approach showed its limits quite quickly. Indeed, we spent a lot of time instrumenting smartphones (especially geolocation), creating bots bypassing bot detection mechanisms and capturing content targeting on different platforms. But a cat and mouse game set in, where our workarounds proved obsolete after a while and we had to find another solution to bypass the restrictions again with each new bot detection mechanism.

Finally, we changed our approach. With the increasing awareness of privacy issues in society, we decided to put users and citizens in the loop by inviting them to participate in our studies. After all, involving users in issues of understanding customization mechanisms on the web, a question that crystallizes many concerns where users are at the center of the issues, made perfect sense. For instance, who has not heard a relative or a friend asking questions about the increase in the price of trains or planes after a few requests on websites (i.e., price personalization).

We have therefore developed a first version of a mobile application involving users. This application presents users with a list of open analyses where each analysis is described with the purpose of the analysis (e.g., is there price personalization on a certain platform), the methodology and associated data collection, as well as what is expected from the user. When a user wishes to participate in a study, she is redirected to the web page or mobile application (e.g., redirected to the page associated with an item on sale), and the user presses a button to take a screenshot. This screenshot is then sent to a server that analyzes all user feedback. A dashboard on the server allows the collections to be administered (e.g., creation, manual analysis, results, etc.).

Involving users in privacy research is a very important element if we want to engage and cooperate with citizens on important issues, raise awareness of evolving risks in society, and design protection tools that will ultimately be adopted. Participatory research is a great tool that I would like to develop in the future.

Bibliographie

[2914] Article 29 data protection working party, 2014. URL https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

[ABCP13] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. Geo-indistinguishability : differential privacy for location-based systems. In *CCS*, pages 901–914, 2013.

[ABD$^+$18] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, volume 80, pages 60–69, 2018.

[ACG$^+$16] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. oct 2016. doi:10.1145/2976749.2978318.

[ACOG18] M. Ancona, E. Ceolini, C. Oztireli, and M. Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sy21R9JAW.

[ADB24] J. Aalmoes, V. Duddu, and A. Boutet. On the Alignment of Group Fairness with Attribute Privacy. In *The International Web Information Systems Engineering conference (WISE)*, Doha, Qatar, Dec. 2024. URL https://hal.science/hal-04739862.

[AE19] D. Abramian and A. Eklund. Refacing : Reconstructing anonymized facial features using GANs. In *ISBI*, pages 1104–1108, 2019.

[ALP$^+$22] A. Appenzeller, M. Leitner, P. Philipp, E. Krempel, and J. Beyerer. Privacy and utility of private synthetic data for medical data analyses. *Applied Sciences*, 12(23), 2022.

[ASS$^+$15] H. Almuhimedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. F. Cranor, and Y. Agarwal. Your location has been shared 5,398 times ! : A field study on mobile app privacy nudging. In *CHI*, pages 787–796, 2015.

[BBB$^+$23] L. Bart, E. A. Bechorfa, A. Boutet, J. Ramon, and C. Frindel. A Smartphone-based Architecture for Prolonged Monitoring of Gait. Technical report, Insa Lyon ; Inria Lyon, Dec. 2023. URL https://hal.science/hal-04355370.

[BBG+20] J. H. Bell, K. A. Bonawitz, A. Gascón, T. Lepoint, and M. Raykova. Secure single-server aggregation with (poly)logarithmic overhead. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 1253–1269, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3372297.3417885.

[BBM21] A. Boutet and S. Ben Mokhtar. Uniqueness assessment of human mobility on multi-sensor datasets. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, pages 1–10, 2021.

[BBR23] G. Berthelier, A. Boutet, and A. Richard. Toward training NLP models to take into account privacy leakages. In *BigData 2023 - IEEE International Conference on Big Data*, pages 1–9, Sorrento, Italy, Dec. 2023. IEEE. URL `https://hal.science/hal-04299405`.

[BC21] A. Boutet and M. Cunche. Privacy protection for wi-fi location positioning systems. *Journal of information security and applications*, 58 :102635, 2021.

[BCMC19] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo. Analyzing federated learning through an adversarial lens. *arXiv preprint arXiv :1811.12470*, 2019.

[BD22] A. Boutet and G. Derache. Simulation de crise-24h dans la tempête. In *RESSI 2022-Rendez-vous de la Recherche et de l'Enseignement de la Sécurité des Systèmes d'Information*, pages 1–4, 2022.

[BdJ+19] S. Barth, M. D. de Jong, M. Junger, P. H. Hartel, and J. C. Roppelt. Putting the privacy paradox to the test : Online privacy and security behaviors among users with technical knowledge, privacy awareness, and financial resources. *Telematics and Informatics*, 41 :55–69, 2019. doi:https://doi.org/10.1016/j.tele.2019.03.003.

[Bec03] R. Beckwith. Designing for ubiquity : the perception of privacy. *IEEE Pervasive Computing*, 2(2) :40–46, 2003. doi:10.1109/MPRV.2003.1203752.

[BEG+19] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, B. McMahan, T. Van Overveldt, D. Petrou, D. Ramage, and J. Roselander. Towards federated learning at scale : System design. In A. Talwalkar, V. Smith, and M. Zaharia, editors, *Proceedings of Machine Learning and Systems*, volume 1, pages 374–388, 2019. URL `https://proceedings.mlsys.org/paper/2019/file/bd686fd640be98efaae0091fa301e613-Paper.pdf`.

[BFG+21] A. Boutet, C. Frindel, S. Gambs, T. Jourdan, and R. C. Ngueveu. DYSAN : Dynamically sanitizing motion sensor data against sensitive inferences through adversarial networks. In *ACM ASIACCS 2021 - 16th ACM ASIA Conference on Computer and Communications Security*, Hong Kong (Virtuel), China, June 2021. doi:10.1145/3433210.3453095. `https://arxiv.org/abs/2003.10325`.

[BFM23] A. Boutet, C. Frindel, and M. Maouche. Towards an evolution in the characterization of the risk of re-identification of medical images. In *BigData 2023 - IEEE International Conference on Big Data*, pages 1–6, Sorrento, Italy, Dec. 2023. IEEE. URL `https://hal.science/hal-04299422`.

[BG19] A. Boutet and S. Gambs. Inspect what your location history reveals about you : Raising user awareness on privacy threats associated with disclosing his location data. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2861–2864, 2019.

[BIK+16] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv :1611.04482*, 2016.

[BIK+17] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 1175–1191, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3133956.3133982.

[BVH+20a] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, pages 2938–2948. PMLR, 2020.

[BVH+20b] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov. How to backdoor federated learning. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2938–2948. PMLR, 26–28 Aug 2020. URL `https://proceedings.mlr.press/v108/bagdasaryan20a.html`.

[CBR+18] S. Cerf, S. Bouchenak, B. Robu, N. Marchand, V. Primault, S. B. Mokhtar, A. Boutet, and L. Y. Chen. Automatic privacy and utility preservation for mobility data : A nonlinear model-based approach. *IEEE Transactions on Dependable and Secure Computing*, 18(1) :269–282, 2018.

[CGKM19] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern. On the compatibility of privacy and fairness. In *Conference on User Modeling, Adaptation and Personalization*, page 309–315, 2019.

[Cha81] D. L. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Commun.*, 24(2) :84–90, Feb. 1981.

[CJL+24] Y. J. Cho, D. Jhunjhunwala, T. Li, V. Smith, and G. Joshi. Maximizing global model appeal in federated learning. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=8GI1SXqJBk`.

[CLA22] X. Chen, J. Lin, and Y. An. Dl-bert : a time-aware double-level bert-style model with pre-training for disease prediction. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1801–1808, 2022. doi:10.1109/BigData55660.2022.10020513.

[CLC+21] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. Williamson, and F. Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6) :493–497, 2021.

[CML⁺23] Y. Cui, S. I. A. Meerza, Z. Li, L. Liu, J. Zhang, and J. Liu. Recup-fl : Reconciling utility and privacy in federated learning via user-configurable privacy defense. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, ASIA CCS '23, page 80–94, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3579856.3582819.

[CPB⁺17] S. Cerf, V. Primault, A. Boutet, S. B. Mokhtar, R. Birke, S. Bouchenak, L. Y. Chen, N. Marchand, and B. Robu. Pulp : Achieving privacy and utility trade-off in user mobility data. In *2017 IEEE 36th symposium on reliable distributed systems (SRDS)*, pages 164–173. IEEE, 2017.

[CS21] H. Chang and R. Shokri. On the privacy risks of algorithmic fairness. *European Security & Privacy*, pages 292–303, 2021.

[DB22] V. Duddu and A. Boutet. Inferring sensitive attributes from model explanations. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 416–425, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3511808.3557362.

[DBS21] V. Duddu, A. Boutet, and V. Shejwalkar. Quantifying privacy leakage in graph embedding. In *MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems : Computing, Networking and Services*, MobiQuitous '20, page 76–85, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3448891.3448939.

[DBS22] V. Duddu, A. Boutet, and V. Shejwalkar. Towards privacy aware deep learning for embedded systems. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, page 520–529, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3477314.3507128.

[DJM⁺20] N. Debs, T. Jourdan, A. Moukadem, A. Boutet, and C. Frindel. Motion sensor data anonymization by time-frequency filtering. In *28th European Signal Processing Conference (EUSIPCO 2020)*, Amsterdam, Netherlands, Aug. 2020. URL `https://inria.hal.science/hal-02888083`.

[DJT⁺22] B. Dura, C. Jean, X. Tannier, A. Calliger, R. Bey, A. Neuraz, and R. Flicoteaux. Learning structures of the french clinical language :development and validation of word embedding models using 21 million clinical reports from electronic health records, 2022, 2207.12940.

[DM20] A. S. Divyat Mahajan, Shruti Tople. Does learning stable features provide privacy benefits for machine learning models ? In *NeurIPS PPML Workshop*, 2020.

[DSA23] V. Duddu, S. Szyller, and N. Asokan. Sok : Unintended interactions among machine learning defenses and risks. *arXiv preprint arXiv :2312.04542*, 2023.

[DTD15] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings. *PoPETs*, 2015(1) :92–112, 2015.

[ea20] J. et al. Deidentification of free-text medical records using pre-trained bidirectional transformers. 2020. URL `https://pubmed.ncbi.nlm.nih.gov/34350426/`.

[EN16] S. Englehardt and A. Narayanan. Online tracking : A 1-million-site measurement and analysis. In *CCS*, pages 1388–1401, 2016.

[FAG+23] J. Ferry, U. Aivodji, S. Gambs, M. Huguet, and M. Siala. Exploiting fairness to enhance sensitive attributes reconstruction. In *Conference on Secure and Trustworthy Machine Learning*, pages 18–41, feb 2023.

[FB15] L. Franceschi-Bicchierai. Redditor cracks anonymous data trove to pinpoint muslim cab drivers. `http://mashable.com/2015/01/28/redditor-muslim-cab-drivers/`, Jan. 2015.

[FLJ+14] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. Privacy in pharmacogenetics : An end-to-end case study of personalized warfarin dosing. In *USENIX Conference on Security Symposium*, page 17–32, 2014.

[FS11] G. D. Fulk and E. Sazonov. Using sensors to measure activity in people with stroke. *Topics in Stroke Rehabilitation*, 18(6) :746–757, 2011.

[FTVHZ22] F. Fioretto, C. Tran, P. Van Hentenryck, and K. Zhu. Differential privacy and fairness in decisions and learning tasks : A survey. In *International Joint Conference on Artificial Intelligence*, pages 5470–5477, 7 2022.

[FVL21] Y. Fraboni, R. Vidal, and M. Lorenzi. Free-rider attacks on model aggregation in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1846–1854. PMLR, 2021.

[GBDM20] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller. Inverting gradients-how easy is it to break privacy in federated learning ? *Advances in Neural Information Processing Systems*, 33 :16937–16947, 2020.

[GGP23] R. Guerraoui, N. Gupta, and R. Pinot. Byzantine machine learning : A primer. *ACM Comput. Surv.*, aug 2023. doi:10.1145/3616537. Just Accepted.

[GKdPC13] S. Gambs, M. O. Killijian, and M. N. d. P. Cortez. De-anonymization attack on geolocated data. In *TrustCom*, pages 789–797, July 2013.

[GKNdPC11] S. Gambs, M.-O. Killijian, and M. Núòez del Prado Cortez. Show me how you move and i will tell you who you are. *Trans. Data Privacy*, 4(2) :103–126, Aug. 2011.

[GKR+23] K. Gu, E. Kabir, N. Ramsurrun, S. Vosoughi, and S. Mehnaz. Towards sentence level inference attack against pre-trained language models. *PoPETS*, 2023 :62–78, 2023. doi:https://doi.org/10.56553/popets-2023-0070.

[GKS07] G. Ghinita, P. Kalnis, and S. Skiadopoulos. Prive : anonymous location-based queries in distributed mobile systems. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 371–380, New York, NY, USA, 2007. Association for Computing Machinery. doi:10.1145/1242572.1242623.

[GNjl+22] O. F. Gulban, D. Nielson, john lee, R. Poldrack, C. Gorgolewski, Vanessasaurus, and C. Markiewicz. poldracklab/pydeface : Pydeface v2.0.2, July 2022. doi:10.5281/zenodo.6856482.

[GRP+23] M. Guillaudeux, O. Rousseau, J. Petot, Z. Bennis, C.-A. Dein, T. Goronflot, N. Vince, S. Limou, M. Karakachoff, M. Wargny, and P.-A. Gourraud. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *npj Digital Medicine*, 6(1) :37, Mar. 2023.

[HHL22] C. Huang, J. Huang, and X. Liu. Cross-silo federated learning : Challenges and opportunities. *arXiv preprint arXiv :2206.12949*, 2022.

[Hig19] High-Level Expert Group on AI. Ethics guidelines for trustworthy ai. Report, European Commission, Brussels, Apr. 2019.

[Hil23] K. Hill. Meet clearview ai, the secretive company that might end privacy as we know it, 2023. URL https://www.chicagotribune.com/nation-world/ct-nw-nyt-clearview-facial-recognition-20200119-dkdqz7ypaveb3id42tpz7ymase-story.html.

[HMM+0] W. Heyndrickx, L. Mervin, T. Morawietz, N. Sturm, L. Friedrich, A. Zalewski, A. Pentina, L. Humbeck, M. Oldenhof, R. Niwayama, P. Schmidtke, N. Fechner, J. Simm, A. Arany, N. Drizard, R. Jabal, A. Afanasyeva, R. Loeb, S. Verma, S. Harnqvist, M. Holmes, B. Pejo, M. Telenczuk, N. Holway, A. Dieckmann, N. Rieke, F. Zumsande, D.-A. Clevert, M. Krug, C. Luscombe, D. Green, P. Ertl, P. Antal, D. Marcus, N. Do Huu, H. Fuji, S. Pickett, G. Acs, E. Boniface, B. Beck, Y. Sun, A. Gohier, F. Rippmann, O. Engkvist, A. H. Göller, Y. Moreau, M. N. Galtier, A. Schuffenhauer, and H. Ceulemans. Melloddy : Cross-pharma federated learning at unprecedented scale unlocks benefits in qsar without compromising proprietary information. *Journal of Chemical Information and Modeling*, 0(0) :null, 0. doi:10.1021/acs.jcim.3c00799. PMID : 37642660.

[HON+12] J. Han, E. Owusu, L. T. Nguyen, A. Perrig, and J. Zhang. Accomplice : Location inference using accelerometers on smartphones. In *COMSNETS*, pages 1–9, 2012.

[Hou20] W. House. Guidance for regulation of artificial intelligence applications. In *Memorandum For The Heads Of Executive Departments And Agencies*, 2020.

[HPS16] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, page 3323–3331, 2016.

[HRM+18] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv :1811.03604*, 2018.

[HSL+14] A. Hannak, G. Soeller, D. Lazer, A. Mislove, and C. Wilson. Measuring price discrimination and steering on e-commerce web sites. In *IMC*, pages 305–318, 2014.

[HYY+15] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, and T. Hospedales. When face recognition meets with deep learning : an evaluation of convolutional neural networks for face recognition. In *international conference on computer vision workshops*, pages 142–150, 2015.

[JBF18] T. Jourdan, A. Boutet, and C. Frindel. Toward privacy in iot mobile devices for activity recognition. In *MobiQuitous*, pages 155–165, 2018.

[JBF21] T. Jourdan, A. Boutet, and C. Frindel. Privacy assessment of federated learning using private personalized layers. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2021. doi:10.1109/MLSP52302.2021.9596237.

[JRY21] A. Jagannatha, B. P. S. Rawat, and H. Yu. Membership inference attack susceptibility of clinical language models, 2021, 2104.08305.

[JSB+13] K. Jiang, D. Shao, S. Bressan, T. Kister, and K.-L. Tan. Publishing trajectories with differential privacy guarantees. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, SSDBM '13, New York, NY, USA, 2013. Association for Computing Machinery. doi:10.1145/2484838.2484846.

[KDK21] E. Kazim, D. M. T. Denny, and A. Koshiyama. AI auditing and impact assessment : according to the uk information commissioner's office. *AI and Ethics*, Feb 2021.

[KDM+23] B. Kaabachi, J. Despraz, T. Meurers, K. Otte, M. Halilovic, F. Prasser, and J. L. Raisaro. Can we trust synthetic data in medicine ? a scoping review of privacy and utility metrics, 11 2023.

[KHPB22] G. Kessibi, A. O. Hamouda, C. Poirier, and A. Boutet. A complementary utility and privacy trade-off evaluation of Google's FloC API. working paper or preprint, May 2022. URL `https://inria.hal.science/hal-03953308`.

[KJ21] H. Kang and E. H. Jung. The smart wearables-privacy paradox : A cluster analysis of smartwatch users. *Behaviour & Information Technology*, 40(16) :1755–1768, 2021.

[KMA+21] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. Advances and open problems in federated learning. *arXiv preprint arXiv :1912.04977*, 2021.

[KRB19] J. L. Kröger, P. Raschke, and T. R. Bhuiyan. Privacy implications of accelerometer data : a review of possible inferences. In *ICCSP*, pages 81–87, 2019.

[KYS05] H. Kido, Y. Yanagisawa, and T. Satoh. Protection of location privacy using dummies for location-based services. In *21st International conference on data engineering workshops (ICDEW'05)*, pages 1248–1248. IEEE, 2005.

[Law18] E. U. Law. Art. 35 GDPR data protection impact assessment. In *General Data Protection Regulation (GDPR)*, 2018. URL `https://gdpr-info.eu/art-35-gdpr/`.

[LBA$^+$24] T. Lebrun, L. Béziaud, T. Allard, T. Allard, A. Boutet, S. Gambs, and M. Maouche. Synthetic Data : Generate Avatar Data on Demand. working paper or preprint, 2024. URL `https://hal.science/hal-04715055`.

[LBAB22] T. Lebrun, A. Boutet, J. Aalmoes, and A. Baud. MixNN : Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers. In *MIDDLEWARE 2022 - 23rd ACM/IFIP International Middleware Conference*, pages 1–11, Quebec, Canada, Nov. 2022. doi:10.1145/3528535.3565240.

[Len23] M. Lenharo. An ai revolution is brewing in medicine. what will it look like ? *Nature*, 622(7984) :686–688, 2023.

[LJP$^+$21] E. Lehman, S. Jain, K. Pichotta, Y. Goldberg, and B. Wallace. Does BERT pretrained on clinical notes reveal sensitive data ? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 946–959, Online, June 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-main.73.

[LKC17] G. Louppe, M. Kagan, and K. Cranmer. Learning to pivot with adversarial networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[LL17] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, 2017.

[LM02] S.-W. Lee and K. Mase. Activity and location recognition using wearable sensors. *IEEE pervasive computing*, 1(3) :24–32, 2002.

[LSZ$^+$10] Y.-A. Li, Y.-J. Shen, G.-D. Zhang, T. Yuan, X.-J. Xiao, and H.-L. Xu. An efficient 3D face recognition method using geometric features. In *International Workshop on Intelligent Systems and Applications*, pages 1–4, 2010.

[LXW$^+$21] X. Liu, L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, and A. V. Vasilakos. Privacy and security issues in deep learning : A survey. *IEEE Access*, 9 :4566–4593, 2021. doi:10.1109/ACCESS.2020.3045078.

[MBB$^+$17] S. B. Mokhtar, A. Boutet, L. Bouzouina, P. Bonnel, O. Brette, L. Brunie, M. Cunche, S. D'Alu, V. Primault, P. Raveneau, et al. Priva'mov : Analysing human mobility through multi-sensor datasets. In *NetMob 2017*, 2017.

[MBG21] M. Malekzadeh, A. Borovykh, and D. Gündüz. Honest-but-curious nets : Sensitive attributes of private inputs can be secretly coded into the classifiers' outputs. In *Conference on Computer and Communications Security*, page 825–844, 2021.

[MDK$^+$22] S. Mehnaz, S. V. Dibbo, E. Kabir, N. Li, and E. Bertino. Are your sensitive attributes private ? novel model inversion attribute inference attacks on classification models. In *USENIX Security Symposium*, pages 4579–4596, 2022.

[MFB$^+$11] S. Mascetti, D. Freni, C. Bettini, X. S. Wang, and S. Jajodia. Privacy in geo-social networks : proximity notification with untrusted service providers and curious buddies. *The VLDB journal*, 20 :541–566, 2011.

[MGEL12] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Detecting price and search discrimination on the internet. In *HotNets*, pages 79–84, 2012.

[MGU+22] F. Mireshghallah, K. Goyal, A. Uniyal, T. Berg-Kirkpatrick, and R. Shokri. Quantifying privacy risks of masked language models using membership inference attacks, 2022, 2203.03929.

[MMS+21] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *Comput. Surv.*, 54(6), jul 2021.

[Moy18] B. Moylan. Grindr was a safe space for gay men. its hiv status leak betrayed us. `https://www.theguardian.com/commentisfree/2018/apr/04/grindr-gay-men-hiv-status-leak-app`, 2018.

[MPP+21] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115 :619–640, 2021.

[MRZ+21] E. Mikulan, S. Russo, F. M. Zauli, P. d'Orio, S. Parmigiani, J. Favaro, W. Knight, S. Squarza, P. Perri, F. Cardinale, et al. A comparative study between state-of-the-art mri deidentification and anonyMI, a new method combining re-identification risk reduction and geometrical preservation. Technical report, Wiley Online Library, 2021.

[NHC21] M. Naseri, J. Hayes, and E. D. Cristofaro. Toward robustness and privacy in federated learning : Experimenting with local and central differential privacy. *arXiv preprint arXiv :2009.03561*, 2021.

[NHN+22] T. T. Nguyen, T. T. Huynh, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen. A survey of machine unlearning. *arXiv preprint arXiv :2209.02299*, 2022.

[PB21] R. Panchendrarajan and S. Bhoi. Dataset reconstruction attack against language models. 2021.

[PBBL11] R. A. Popa, A. J. Blumberg, H. Balakrishnan, and F. H. Li. Privacy and accountability for location-based aggregate statistics. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 653–666, 2011.

[PBE+22] S. Pati, U. Baid, B. Edwards, M. Sheller, S.-H. Wang, G. A. Reina, P. Foley, A. Gruzdev, D. Karkada, C. Davatzikos, et al. Federated learning enables big data for rare cancer boundary detection. *Nature communications*, 13(1) :7346, 2022.

[PBMB16] V. Primault, A. Boutet, S. B. Mokhtar, and L. Brunie. Adaptive location privacy with alp. In *2016 IEEE 35th Symposium on Reliable Distributed Systems (SRDS)*, pages 269–278. IEEE, 2016.

[PBMB18] V. Primault, A. Boutet, S. B. Mokhtar, and L. Brunie. The long road to computational location privacy : A survey. *IEEE Communications Surveys & Tutorials*, 21(3) :2772–2793, 2018.

[PCN17] E. Park, H.-J. Chang, and H. S. Nam. Use of machine learning classifiers and sensor data to detect neurological deficit in stroke patients. *J Med Internet Res*, 19(4) :e120, 2017.

[PGDV+11] N. Pelekis, A. Gkoulalas-Divanis, M. Vodas, D. Kopanaki, and Y. Theodoridis. Privacy-aware querying over sensitive trajectory data. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 895–904, 2011.

[PM17] C. L. Parks and K. L. Monson. Automated facial recognition of computed tomography-derived facial images : patient privacy implications. *Journal of digital imaging*, 30(2) :204–214, 2017.

[PMB+18] V. Primault, M. Maouche, A. Boutet, S. B. Mokhtar, S. Bouchenak, and L. Brunie. Accio : How to make location privacy experimentation open and easy. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pages 896–906. IEEE, 2018.

[PYY+19] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng. Recent progress on generative adversarial networks (gans) : A survey. *IEEE Access*, 7 :36322–36333, 2019.

[QLM+11] D. Quercia, I. Leontiadis, L. McNamara, C. Mascolo, and J. Crowcroft. Spotme if you can : Randomized responses for location obfuscation on mobile phones. In *2011 31st International Conference on Distributed Computing Systems*, pages 363–372. IEEE, 2011.

[QYFD15] J. Qi, P. Yang, D. Fan, and Z. Deng. A survey of physical activity monitoring and assessment using internet of things technology. In *CIT/IUCC/DASC/PICOM*, pages 2353–2358, 2015.

[RKC+16] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi. Linking users across domains with location data : Theory and validation. In *WWW*, pages 707–719, 2016.

[RMD+21] P. Rougé, A. Moukadem, A. Dieterlen, A. Boutet, and C. Frindel. Anonymizing motion sensor data through time-frequency domain. In *MLSP 2021 - Machine Learning for Signal Processing*, pages 1–6, Queensland, Australia, Oct. 2021. doi:10.1109/MLSP52302.2021.9596442.

[RTG23] A. Richard, F. Talbot, and D. Gimbert. Anonymisation de documents médicaux en texte libre et en français via réseaux de neurones. In *Plate-forme Intelligence Artificielle 2023 (PFIA2023) - Journée Santé & IA*, Starsbourg, France, July 2023. Association française pour l'Intelligence Artificielle (AfIA) and Université de Strasbourg and Association française d'Informatique Médicale (AIM). URL https://hal.science/hal-04139391.

[Rud19] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5) :206–215, 2019.

[SAKS21] A. Sekhari, J. Acharya, G. Kamath, and A. T. Suresh. Remember what you want to forget : Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34 :18075–18086, 2021.

[SCZ+23] A. Sharma, W. Chen, J. Zhao, Q. Qiu, S. Bagchi, and S. Chaterji. Flair : Defense against model poisoning attack in federated learning. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, ASIA CCS '23, page 553–566, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3579856.3582836.

[SD14] K. Sharad and G. Danezis. An automated social graph de-anonymization technique. In *WPES*, pages 47–58, 2014.

[SGK17] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153, 2017.

[SK12] A. Sadilek and J. Krumm. Far out : Predicting long-term human mobility. In *AAAI*, pages 814–820, 2012.

[SKA+23] C. G. Schwarz, W. K. Kremers, A. Arani, M. Savvides, R. I. Reid, J. L. Gunter, M. L. Senjem, P. M. Cogswell, P. Vemuri, K. Kantarci, D. S. Knopman, R. C. Petersen, and C. R. Jack. A face-off of MRI research sequences by their need for de-facing. *NeuroImage*, 276 :120199, 2023.

[SKL+22] C. G. Schwarz, W. K. Kremers, V. J. Lowe, M. Savvides, J. L. Gunter, M. L. Senjem, P. Vemuri, K. Kantarci, D. S. Knopman, R. C. Petersen, et al. Potential for re-identifying brain PET research participants using face recognition. *Alzheimer's & Dementia*, 18 :e063651, 2022.

[SKT+19] C. G. Schwarz, W. K. Kremers, T. M. Therneau, R. R. Sharp, J. L. Gunter, P. Vemuri, A. Arani, A. J. Spychalla, K. Kantarci, D. S. Knopman, et al. Identification of anonymous MRI research participants with face-recognition software. *New England Journal of Medicine*, 381(17) :1684–1686, 2019.

[SKT+20] C. G. Schwarz, W. K. Kremers, T. M. Therneau, R. R. Sharp, J. L. Gunter, P. Vemuri, A. Arani, A. J. Spychalla, K. Kantarci, D. S. Knopman, et al. Popular MRI de-facing software does not sufficiently protect participants from re-identification via face recognition : Neuroimaging/optimal neuroimaging measures for tracking disease progression. *Alzheimer's & Dementia*, 16 :e045157, 2020.

[SKW+21] C. G. Schwarz, W. K. Kremers, H. J. Wiste, J. L. Gunter, P. Vemuri, A. J. Spychalla, K. Kantarci, A. P. Schultz, R. A. Sperling, D. S. Knopman, R. C. Petersen, and C. R. J. Jr. Changing the face of neuroimaging research : Comparing a new MRI de-facing technique with popular alternatives. *NeuroImage*, 231 :117845, 2021.

[SM21] S. Singh and A. Mahmood. The NLP cookbook : Modern recipes for transformer based deep learning architectures. *CoRR*, abs/2104.10640, 2021, 2104.10640. URL `https://arxiv.org/abs/2104.10640`.

[SOL+14] J. Staiano, N. Oliver, B. Lepri, R. de Oliveira, M. Caraviello, and N. Sebe. Money walks : A human-centric study on the economics of personal mobile data. In *UbiComp*, pages 583–594, 2014.

[SPJJ20] C. G. Schwarz, R. C. Petersen, and C. R. Jack Jr. Identification from MRI with face-recognition software. Reply. *The New England journal of medicine*, 382(5) :490–490, 2020.

[SS20] C. Song and V. Shmatikov. Overlearning reveals sensitive attributes. In *International Conference on Learning Representations*, 2020.

[SSZ21] R. Shokri, M. Strobel, and Y. Zick. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 231–241, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3461702.3462533.

[STK+17] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg. Smoothgrad : removing noise by adding noise. *ArXiv*, abs/1706.03825, 2017.

[STY17] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328, 2017.

[TB20] F. Tramèr and D. Boneh. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv :2011.11660*, 2020.

[TBH+19] E. Tabassi, K. J. Burns, M. Hadjimichael, A. Molina-Markham, and J. Sexton. A taxonomy and terminology of adversarial machine learning. In *NIST Interagency/Internal Report*, 2019.

[TBO17] S. Tedesco, J. Barton, and B. O'Flynn. A review of activity trackers for senior citizens : Research perspectives, commercial landscape and the role of the insurance industry. *Sensors*, 17(6) :1277, 2017.

[TCMK23] A. K. Tarun, V. S. Chundawat, M. Mandal, and M. Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[TWC+23] X. Tannier, P. Wajsbürt, A. Calliger, B. Dura, A. Mouchet, M. Hilka, and R. Bey. Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse, 2023, 2303.13451.

[VBM+24] V. B. Vallevik, A. Babic, S. E. Marshall, S. Elvatun, H. M. Brøgger, S. Alagaratnam, B. Edwin, N. R. Veeraragavan, A. K. Befring, and J. F. Nygård. Can i trust my fake data – a comprehensive quality assessment framework for synthetic tabular data in healthcare. *International Journal of Medical Informatics*, 185 :105413, May 2024.

[vdMH20] L. van der Maaten and A. Hannun. The trade-offs of private prediction. *arXiv preprint arXiv :2007.05089*, 2020.

[WUR$^+$22] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, and I. Gabriel. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3531146.3533088.

[WXH$^+$22] Y. Wang, N. Xu, S. Huang, K. Mahmood, D. Guo, C. Ding, W. Wen, and S. Rajasekaran. Analyzing and defending against membership inference attacks in natural language processing classification. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5823–5832, 2022. doi:10.1109/BigData55660.2022.10020711.

[WZA$^+$22] Q. Wei, X. Zuo, O. Anjum, Y. Hu, R. Denlinger, E. V. Bernstam, M. J. Citardi, and H. Xu. Clinicallayoutlm : A pre-trained multi-modal model for understanding scanned document in electronic health records. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2821–2827, 2022. doi:10.1109/BigData55660.2022.10020569.

[YBS20] T. Yu, E. Bagdasaryan, and V. Shmatikov. Salvaging federated learning by local adaptation, 2020.

[YGFJ18a] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning : Analyzing the connection to overfitting. In *Computer Security Foundations Symposium*, pages 268–282, 2018.

[YGFJ18b] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning : Analyzing the connection to overfitting, 2018.

[YKS$^+$22] H. Yeo, E. Khorasani, V. Sheinin, I. Manotas, N. P. An Vo, O. Popescu, and P. Zerfos. Natural language interface for process mining queries in healthcare. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4443–4452, 2022. doi:10.1109/BigData55660.2022.10020685.

[YYZC16] H. Yang, J. Yu, H. Zo, and M. Choi. User acceptance of wearable devices : An extended perspective of perceived value. *Telematics and Informatics*, 33(2) :256–269, 2016.

[ZGH07] G. Zhong, I. Goldberg, and U. Hengartner. Louis, lester and pierre : Three protocols for location privacy. In N. Borisov and P. Golle, editors, *Privacy Enhancing Technologies*, pages 62–76, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[ZIL$^+$21] C. Zhang, D. Ippolito, K. Lee, M. Jagielski, F. Tramèr, and N. Carlini. Counterfactual memorization in neural language models. *CoRR*, abs/2112.12938, 2021, 2112.12938. URL `https://arxiv.org/abs/2112.12938`.

[ZLM18] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Conference on AI, Ethics, and Society*, page 335–340, 2018.

[ZOC22] W. Zhang, O. Ohrimenko, and R. Cummings. Attribute privacy : Framework and mechanisms. In *Fairness, Accountability, and Transparency*, page 757–766, 2022.

[ZVGRG19] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints : A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75) :1–42, 2019.

[ZYZ$^+$15] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie. You are where you go : Inferring demographic attributes from location check-ins. In *WSDM*, pages 295–304, 2015.