Contributions à la robustesse, la sécurité et la confidentialité en apprentissage statistique

Cédric Gouy-Pailler

▶ **To cite this version:**

Cédric Gouy-Pailler. Contributions à la robustesse, la sécurité et la confidentialité en apprentissage statistique. Informatique. Université Paris Saclay, 2025. ⟨tel-05113380⟩

HAL Id: tel-05113380

https://hal.science/tel-05113380v1

Submitted on 15 Jun 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Habilitation à diriger des recherches

université PARIS-SACLAY

Contributions à la robustesse, la sécurité et la confidentialité en apprentissage statistique

Habilitation à diriger des recherches de l'Université Paris-Saclay

présentée et soutenue à Saclay, le 16 mai 2025, par

Cédric GOUY-PAILLER

Composition du jury

Marianne CLAUSEL Professeure des universités, Université de Lorraine	Rapportrice
Aurélien BELLET Directeur de recherche, INRIA (antenne de l'Université de Montpellier)	Rapporteur
Teddy FURON Directeur de recherche, INRIA Rennes	Rapporteur
Sophie ACHARD Directrice de recherche, CNRS, Laboratoire Jean Kuntzmann	Examinatrice
Davy PREUVENEERS Directeur de recherche, KU Leuven	Examinateur
Gaël VAROQUAUX Directeur de recherche, INRIA Saclay	Examinateur
Jamal ATIF Professeur des universités, Université Paris-Dauphine	Examinateur et mentor

Sommaire

Organisation du document

Ce document se décompose en trois parties principales. La première partie, constituée des chapitres 2, *Décomposition, segmentation et débruitage de signaux multivariés*, 3, *Discrimination, compression et approximation*, 4, *Confidentialité et sécurité en contexte centralisé*, et 5, *Sécurité et confidentialité en contexte décentralisé*, retrace les principales pistes de recherche suivies depuis 2010. Une introduction générale de ces travaux, leur positionnement technique et un résumé des principales contributions est accessible au chapitre 1. La deuxième partie, constituée du chapitre 6, présente mon projet de recherche en explicitant de manière détaillée comment il s'inscrit dans le contexte actuel. Une mise en perspective est proposée afin de faire apparaître les impacts du paysage associé à l'apprentissage statistique et aux algorithmes d'intelligence artificielle sur le projet de recherche. La troisième partie, constituée du chapitre 7, présente mon parcours académique. Cette partie détaille en particulier ma formation, mon rôle actuel et dresse un bilan détaillé des activités de recherche, d'encadrement et de valorisation scientifique réalisées depuis ma thèse, soutenue en 2009.

1 Introduction et résumé des contributions

1.1 Contexte

Les 15 années écoulées depuis ma soutenance de thèse, en octobre 2009, ont été riches d'évolutions rapides associées au domaine de l'apprentissage statistique à partir de données. Ces évolutions sont le fruit de plusieurs facteurs. Tout d'abord, les quantités de données générées par les objets, les systèmes et les humains ont explosé [1]. Conjointement, les capacités de transmission et de stockage de ces données ont été décuplées. De plus, des algorithmes et des techniques de traitement plus efficaces (temps de traitement par échantillon) et capables de traiter des données plus volumineuses (capacité de passage à l'échelle, complexité quasi-linéaire vis-à-vis du nombre d'échantillons dans le jeu de données) ont su tirer parti de moyens de calcul de plus en plus puissants (General-purpose Graphical Processor Unit ou Tensor Processor Unit). Enfin de nombreuses initiatives ont visé la mise à disposition sous licences libres (LGPL, Apache, BSD, MIT) des outils et algorithmes permettant à un nombre croissant d'utilisateurs d'avoir accès aux travaux les plus avancés des chercheurs et praticiens expérimentés, par exemple avec scikit-learn [2]. Ces facteurs ont conduit à des vagues techno-médiatiques du « big data » [3] et de l' « Intelligence Artificielle », respectivement entre 2010 et 2017 pour la première et depuis 2015 pour la seconde. Mes travaux de recherche, initialement orientés en sortie de thèse sur les approches statistiques pour le traitement de signaux multivariés, se sont nourris des défis posés par le « big data » puis « l'intelligence artificielle », en essayant de garder les spécificités qui sont propres à mon parcours.

1.2 Cadre général des travaux

Nos travaux s'inscrivent dans le cadre général de l'apprentissage statistique à partir de données, que l'on peut formaliser comme un problème d'optimisation. Considérons un ensemble de n observations indépendantes $z_1, z_2, \dots, z_n \in \mathcal{Z}$. L'objectif de l'apprentissage statistique est, à partir de cet échantillon de données, d'estimer une caractéristique sous-jacente ou un modèle de la distribution inconnue générant les données. Plus formellement, on note $\mathcal{H}$ l'espace des hypothèses (ou espace des modèles) contenant toutes les fonctions candidates ou structures que nous pourrions apprendre. Nous définissons une fonction de perte $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$ qui, pour un modèle donné $h \in \mathcal{H}$ et une observation z, mesure l'écart entre la prédiction et l'observation réelle. Cette fonction de perte dépend de la tâche, mais

le principe d'apprentissage reste identique : nous cherchons un modèle h qui minimise la perte. En pratique, puisque la distribution réelle des données est inconnue, nous minimisons le risque empirique, c'est-à-dire la moyenne de la perte sur l'échantillon d'entraînement. Dans le cadre de la minimisation du risque empirique (MRE), l'apprentissage se réduit ainsi à résoudre le problème d'optimisation

$$\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i) + \lambda\Omega(h) \ .$$

Ici, $\frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i)$ représente le coût empirique (la perte moyenne) associé au modèle h, et $\Omega(h)$ est un terme de régularisation, pondéré par $\lambda \geq 0$, servant à contrôler la complexité du modèle ou à encoder une connaissance a priori. Cette formulation est très générale : en choisissant de manière appropriée l'espace des hypothèses $\mathcal{H}$, la fonction de perte ℓ, et la régularisation $\Omega(.)$, nous pouvons englober une grande variété de paradigmes d'apprentissage. Tout au long de ce document, nous utiliserons le cadre de la MRE pour discuter de méthodes allant de l'apprentissage supervisé à la décomposition non supervisée.

De nombreux problèmes d'apprentissage classiques peuvent être considérés comme des instances particulières de la MRE. Dans chaque cas, le choix de $\mathcal{H}$ (l'espace des fonctions ou des modèles), la fonction de perte ℓ, et la régularisation Ω varient en fonction de la tâche. Nous présentons ci-dessous comment plusieurs paradigmes d'apprentissage s'inscrivent dans cette structure d'optimisation commune :

- **Apprentissage supervisé (régression et classification)**
  Dans l'apprentissage supervisé, chaque observation est de la forme $z_i = (x_i, y_i)$ avec $x_i \in \mathcal{X}$ représentant l'entrée (par exemple, les observations ou caractéristiques) et $y_i \in \mathcal{Y}$ la sortie (étiquette ou réponse). L'espace des hypothèses $\mathcal{H}$ est constitué de fonctions $h: \mathcal{X} \to \mathcal{Y}$ (prédicteurs) qui associent à chaque entrée une prédiction. La fonction de perte $\ell\big(h, (x_i, y_i)\big)$ mesure typiquement l'écart entre la prédiction $h(x_i)$ et la valeur réelle y_i (par exemple, l'erreur quadratique $(h(x_i) - y_i)^2$ pour la régression ou une fonction de perte de classification comme la log-vraisemblance pour les classes). Dans le cas de la classification supervisée, on note en général $\ell\big(h, (x_i, y_i)\big) = \ell(h(x_i), y_i)$ afin de faire apparaître une comparaison entre la prédiction du modèle $h(x_i)$ et la sortie espérée y_i. Le problème de MRE cherche alors à trouver le prédicteur h qui minimise l'erreur moyenne sur l'échantillon d'entraînement

$$\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i) + \lambda\Omega(h) \ .$$

Le terme de régularisation $\Omega(h)$ (tel qu'une norme des paramètres du modèle ou une pénalité de parcimonie) est souvent ajouté pour prévenir le surapprentissage et intégrer des connaissances a priori (par exemple, la régularité ou la simplicité de h). Plusieurs contributions présentées dans ce document s'inscrivent dans l'apprentissage supervisé. Ils sont au cœur des travaux de Rafaël Pinot [4], Arnaud Grivet Sébert [5], Pierre-Emmanuel Clet [6] et Fabiola Espinoza Castellon [7].

- **Apprentissage non-supervisé par décomposition sur dictionnaire**

  Dans les problèmes d'apprentissage de dictionnaires et de décomposition de signaux, les données sont généralement des signaux ou vecteurs non étiquetés $y_i \in \mathbb{R}^M$. Ici, $\mathcal{H}$ représente l'espace des dictionnaires (souvent appelé un ensemble d'atomes). Étant donné un dictionnaire $\Phi \in \mathcal{H}$, chaque échantillon y_i est reconstruit par une combinaison linéaire d'atomes du dictionnaire. La fonction de perte $\ell(\Phi, y_i)$ mesure l'erreur de reconstruction, par exemple,

$$\ell(\Phi, y_i) \ = \ \min_{x_i} \|y_i - \Phi x_i\|^2 + \beta\Psi(x_i)$$

  où x_i représente le code latent associé à y_i et $\Psi(x_i)$ est une pénalité encourageant certaines propriétés souhaitables (comme la parcimonie de la décomposition). Dans cette formulation, le modèle $h \ = \ \Phi$ (et éventuellement l'ensemble des codes x_i pour tous les échantillons) est optimisé pour minimiser l'erreur de reconstruction moyenne

$$\min_{D\in\mathcal{H}} \left[\frac{1}{n}\sum_{i=1}^{n} \left(\min_{x_i} \|y_i - \Phi x_i\|^2 + \beta\Psi(x_i)\right) + \lambda\Omega(\Phi)\right] .$$

  La régularisation $\Omega(\Phi)$ permet d'imposer des contraintes sur le dictionnaire (par exemple, l'énergie des atomes ou leur incohérence). Les méthodes de débruitage et séparation aveugle des sources traitées dans nos travaux se situent dans ce paradigme, où l'on cherche à obtenir une représentation fidèle du signal malgré la présence de bruit. Nos travaux initiaux sur le traitement des signaux électroencéphalographiques, par exemple, ont utilisé des fonctions de coût basées sur les moindres carrés pour extraire des sous-espaces pertinents [Th1] et développer une méthode de suppression des artefacts [J6]. Nous avons également exploré des

méthodes de régression sur dictionnaire appris [C21] afin d'améliorer la qualité des prédictions en utilisant des représentations plus adaptées.

- **Détection de changements (segmentation de séries temporelles)**
  La détection de changements est une tâche d'apprentissage non supervisé appliquée aux données structurées, généralement des séries temporelles. Ici, chaque observation peut correspondre à une série temporelle ou à une séquence, ou l'on considère la série dans son ensemble pour en extraire des segments. Le modèle peut alors être défini comme un ensemble de points de rupture (changements) qui partitionnent la série temporelle en segments où les propriétés statistiques restent constantes. On peut interpréter un modèle $h \in \mathcal{H}$ comme une description par morceaux des données, ou comme un étiquetage des indices temporels en fonction du segment auquel ils appartiennent. La fonction de perte $\ell(h, z_i)$ quantifie ici l'adéquation d'une segmentation h aux données z_i — par exemple, la somme des erreurs au sein de chaque segment (comme la variance intra-segment) ou l'opposé de la log-vraisemblance. Pour éviter une segmentation excessive (trop de changements), un terme de régularisation $\Omega(h)$ est ajouté, pénalisant le nombre de segments ou le nombre de points de rupture. Dans le cadre de la MRE, la détection de changements consiste à trouver la segmentation h minimisant l'erreur totale intra-segment majorée par une pénalité sur le nombre de segments. Nous avons travaillé sur la détection multiple de changements dans des séries temporelles multivariées [J5], développant des méthodes permettant de segmenter automatiquement les séries en sous-parties stationnaires, avec des pénalités pour éviter les ruptures non significatives. Dans ces contributions, la régularisation s'avère essentielle pour obtenir une segmentation pertinente en présence de bruit et de variabilité.

- **Clustering non-supervisé**
  Le clustering constitue une autre instance d'apprentissage non supervisé qui s'inscrit dans le paradigme de la MRE. Ici, chaque z_i est un échantillon dans $\mathbb{R}^d$, et l'objectif est d'assigner ces échantillons à des groupes (clusters) de sorte que les éléments du même groupe soient aussi similaires que possible. L'espace $\mathcal{H}$ peut être défini comme l'ensemble de toutes les partitions possibles des n points. Une fonction de perte naturelle dans ce contexte mesure la distance entre chaque point et son représentant de cluster (centroïde), par exemple

$$\ell(h, z_i) = \min_k \|z_i - \mu_k\|^2$$

où h définit un ensemble de centroïdes $\mu_1, \mu_2, ...$, et la perte est la distance au centroïde le plus proche. Le risque empirique devient alors la somme des distances intra-clusters pour tous les points.

$$\min_{h \in \mathcal{H}} \left[\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \min_k \|z_i - \mu_k\|^2 + \lambda \Omega(\mathrm{h}) \right] .$$

Afin d'éviter des solutions triviales (par exemple, chaque point formant son propre cluster), une régularisation ou une contrainte est imposée, par exemple en fixant le nombre de clusters K, ou en ajoutant une pénalité croissante avec le nombre de clusters par l'intermédiaire de la régularisation $\Omega(h)$. Nos travaux sur le clustering et l'apprentissage non supervisé de grandes masses de données se sont inscrits dans ce cadre [8], avec des méthodes qui intègrent la MRE via des choix particuliers de fonctions de perte et de contraintes.

- **Apprentissage fédéré**

  Dans certaines applications, les données ne peuvent être centralisées pour des raisons de confidentialité, de contraintes réglementaires ou pratiques. L'apprentissage fédéré est un cadre qui répond à ce défi : les données sont partitionnées entre P participants (clients), chacun détenant ses propres données, tandis qu'un serveur central coordonne le processus d'apprentissage. L'objectif global reste la minimisation du risque empirique, c'est-à-dire

$$\min_{h \in \mathcal{H}} \left[\frac{1}{n} \sum_{p=1}^{P} \sum_{i=1}^{n_p} \ell\left(h,\, z_i^{(p)}\right) + \lambda \Omega(h) \right]$$

  mais l'optimisation est réalisée de manière distribuée, chaque participant ne partage que des informations résumées sans exposer les données brutes. Ainsi, l'apprentissage fédéré s'inscrit naturellement dans le cadre de la MRE, avec la particularité d'exiger un protocole de communication et de respecter des contraintes de localité. Nous avons également exploré ce scénario, en étudiant notamment des algorithmes robustes aux attaques et garantissant la confidentialité, comme illustré dans les travaux de Arnaud Grivet Sébert [5] et Fabiola Espinoza Castellon [7].

Un avantage majeur du cadre MRE réside dans la possibilité d'intégrer des connaissances a priori et des contraintes spécifiques au problème directement dans l'objectif d'apprentissage. Dans la pratique, les données sont souvent limitées, bruitées ou collectées dans des conditions contraignantes, de sorte que la simple minimisation du risque empirique sans régularisation peut conduire à du surapprentissage ou à des solutions sous-optimales. En ajoutant un terme de régularisation $\Omega(h)$, nous orientons le processus d'optimisation en fonction de ces connaissances, améliorant ainsi la robustesse et la capacité de généralisation du modèle.

1.3 Résumé des principales contributions

Au fil de nos travaux, nous avons exploré plusieurs instances de ce problème d'optimisation, en considérant des contextes variés et des types de données de plus en plus complexes.

1.3.1 Débruiter les signaux multivariés pour mieux les classer

Mes travaux de thèse ont été dédiés aux interfaces cerveau-machines. Les interfaces cerveau-machines regroupent l'ensemble des systèmes destinés à restaurer un moyen de communication ou de mouvement, afin d'établir un moyen de contrôle direct entre les pensées de l'utilisateur et un système électronique, possédant potentiellement une extension mécanique. Les interfaces cerveau-machines s'appuient sur un système d'enregistrement de l'activité cérébrale, en mesurant des variations électromagnétiques (électroencéphalographie ou magnétoencéphalographie) ou des variations hémodynamiques (oxygène dans le flux sanguin cérébral). En raison d'une réactivité excellente (quelques centaines de millisecondes entre l'activité cérébrale et sa trace sur les capteurs de champs électromagnétiques, contre quelques secondes pour la réponse hémodynamique), son faible coût et son caractère peu invasif[1], la technique privilégiée pendant ma thèse était l'électroencéphalographie au niveau du scalp à l'aide d'un ensemble d'électrodes (entre 8 et 128 capteurs). En dépit de ses nombreux avantages, cette technique possède un inconvénient très lourd : les activités cérébrales sont très largement diffusées et atténuées entre la source de l'activité au niveau des colonnes de neurones pyramidaux et la mesure au niveau du scalp, ce qui provoque un mélange des signaux au niveau des capteurs de champs électriques. Plus précisément, les

[1] Le caractère non invasif est largement accepté pour l'électroencéphalographie de scalp. Notons cependant qu'il existe des variantes beaucoup plus invasives pour capter des signaux électromagnétiques, par exemple en stéréo-encéphalographie ou électrocorticographie.

signaux recueillis sont constitués par un mélange approximativement linéaire[2] entre les signaux utiles (activités cérébrales recherchées), les interférences (signaux cérébraux non spécifiques aux activités recherchées), et différentes sources de bruit (activités musculaires, artefacts oculaires, bruits électromagnétiques). Or le principe d'un tel système est de parvenir à identifier de manière certaine l'activité d'une zone cérébrale particulière à un instant donné (ce principe est globalement le principe des systèmes basés sur l'imagerie motrice, qui utilise le fait qu'on active des zones cérébrales spécifiques au membre du corps pour lequel on imagine un mouvement, ainsi on peut associer par exemple un mouvement d'une souris informatique à gauche ou à droite en imaginant un mouvement de la main gauche ou de la main droite). Les traitements nécessaires consistent donc à séparer le signal utile des interférences et du bruit afin de permettre à un modèle statistique de discriminer entre des activités cérébrales prédéfinies (cette étape passe par un apprentissage machine souvent spécifique à chacun des sujets). Afin de simplifier le processus de traitement et de classification des signaux cérébraux, on utilise des systèmes synchrones, *i.e.* que le sujet a une fenêtre temporelle précise pour réaliser sa tâche cérébrale. A l'opposé, les systèmes asynchrones tentent de supprimer la nécessité d'imposer ces instants au sujet. Les systèmes synchrones sont en conséquence beaucoup plus rigides d'utilisation, imposant des intervalles de commandes. L'un des objectifs de ma thèse était de développer des techniques de traitement et de classification des signaux cérébraux qui puissent être robustes dans le cas des interfaces cerveau-machines asynchrones. Mes contributions de thèse se sont articulées autour de techniques de traitement statistique de signaux multivariés afin de concentrer l'information utile en amont du décodage. En particulier, dans [J9], [J10], nous avons montré que des techniques avancées de traitement du signal, basées sur la diagonalisation conjointe de matrices de statistiques de second ordre, pouvait permettre de débruiter efficacement les signaux EEG relatifs à quatre tâches distinctes, au-delà des résultats de l'état de l'art de l'époque. Ces résultats ont été ensuite complétés par l'approche proposée dans [C27] dans un cas asynchrone. Au-delà des avancées dans le domaine des interfaces cerveau-machines, plusieurs travaux ont découlé des approches proposées dans la thèse. Par exemple dans [J6], nous avons adapté la fonction de coût usuelle en interface cerveau-machines pour proposer une technique permettant le débruitage efficace de signaux EEG contaminés par des artefacts oculaires, comme référencée par exemple dans [9].

[2] L'approximation quasi-statique est valable au regard des fréquences des activités observées. Cette hypothèse forte est justifiée par le fait que les fréquences d'intérêt en EEG (entre 8 et 100 Hertz) sont faibles au regard des vitesses de propagation (cf équations de Maxwell).

1.3.2 La parcimonie pour représenter des signaux multivariés

Malgré des résultats encourageants obtenus pendant la thèse, les avancées proposées n'étaient cependant pas suffisantes pour atteindre le but initial, conférer aux interfaces cerveau-machines une robustesse et une souplesse d'utilisation compatibles avec la vie courante. Nous avons alors proposé, initialement en collaboration avec Quentin Barthélemy, d'explorer des critères de parcimonie pour focaliser le pouvoir de représentation des modèles sur des motifs spatio-temporels prédéfinis ou appris. Cette approche basée sur la décomposition sur un dictionnaire, proposée dans [J7] a été pour la première fois appliquée dans le domaine de l'EEG en utilisant des décompositions invariantes par translation. Bien que cette approche permettait de bénéficier d'une explicabilité idéale, en faisant apparaître les atomes temporels dans les signaux EEG, de nombreuses informations connues a priori sur les signaux EEG de scalp n'étaient pas prises en compte, notamment quant à la régularité spatiale et temporelle des décompositions cherchées. Cette limitation a conduit à proposer le sujet de recherche de Yoann Isaac dans ses travaux de thèse, sous la direction de Jamal Atif. Ces travaux ont conduit à proposer des techniques de décomposition sur dictionnaires incluant des contraintes de parcimonie et des contraintes de structures spatiales, qui sont communes en EEG, dont les plus simples consistent à considérer que des électrodes proches doivent mesurer des informations fortement corrélées [J4]. En parallèle, la problématique de la segmentation de signaux multivariés restait au cœur des préoccupations en EEG, mais également dans de nouveaux domaines que je découvrais alors au CEA, notamment en énergie et en génomique. C'est ainsi que la problématique de la thèse de Flore Harlé est apparue : comment segmenter des séries temporelles multivariées afin d'identifier des changements de régimes dans la dynamique des signaux ? Les travaux de thèse de Flore Harlé, réalisés en collaboration avec Florent Chatelain, Sophie Achard et Olivier Michel, ont permis de mettre au point des techniques innovantes essentiellement par le fait qu'elles créaient un mécanisme d'interaction entre une représentation graphique de données multivariées (représentant les interactions statistiques entre les variables) et une méthode de détection de ruptures prenant en compte la structure a priori des relations entre séries temporelles. Ces travaux ont conduit aux publications [C19], [J5]. Au cours de ces deux lignes de recherche, deux paradigmes fondateurs sont apparus de plus en plus clairs pour mes travaux futurs : d'une part les techniques de traitement du signal avancées devaient être traitées conjointement avec les approches d'apprentissage statistique, afin de considérer les chaines de traitement des données dans leur ensemble ; d'autre part, de nombreux compromis existent entre des critères (efficacité, explicabilité, latence, débit) souvent peu intelligibles au-delà des spécialistes du domaines, et une meilleure maîtrise de ces compromis aidera à l'adoption plus large des techniques développées.

1.3.3 Apprentissage statistique pour les flux de données

Cette seconde préoccupation a alors conduit à proposer le sujet de thèse d'Anne Morvan, en collaboration avec Jamal Atif : les méthodes classiques utilisées en traitement du signal et en apprentissage statistique tentent souvent de pousser des critères de précision, en explorant plusieurs hypothèses sur les signaux considérés. Or dans de nombreuses applications, une précision moindre est acceptable, en particulier lorsque des flux de données doivent être exploités avec une latence réduite. Nous avons ainsi proposé de nombreuses approches dans lesquelles un algorithme exact est remplacé par un algorithme approché, en mettant en évidence un gain important en termes de débit ou d'espace nécessaire [C16], [C17], [C15]. Fort de ces contributions à la frontière entre traitement des signaux multivariés et apprentissage statistique, une question est apparue sur la portée des algorithmes de sketching mis en place : alors en plein essor dans le domaine de l'apprentissage statistique, la confidentialité différentielle, qui proposait un cadre théorique pour l'utilisation d'un bruit maitrisé afin de protéger la confidentialité d'éléments individuels d'une base de données, pouvait-elle être obtenue avec des algorithmes de sketching ? Cette réflexion a été le point de départ de la thèse de Rafaël Pinot, en partenariat avec Florian Yger et Jamal Atif, qui a travaillé de manière plus générale sur la confidentialité différentielle pour les algorithmes de clustering de données structurées par des graphes. Les travaux poursuivis alors pendant cette période ont conduit à généraliser les questions de compromis en apprentissage statistique dans le cadre de la robustesse aux attaques adverses. En particulier nous avons proposé des ponts entre le formalisme de la confidentialité différentielle et la robustesse aux attaques adverses, ouvrant des portes à plusieurs approches : des techniques de défense face aux attaques adverses en utilisant des algorithmes incluant des mécanismes d'ajout de bruit ; l'utilisation réciproque d'approches d'un domaine vers le second pour régler et comprendre les compromis. Ces travaux ont donné lieu à plusieurs publications, notamment [C11], [W1], [J2], et nous ont permis de comprendre l'enjeu croissant de la confidentialité dans les critères les plus importants à considérer en apprentissage statistique.

1.3.4 Stratégies de régularisation

Au fil des années, les techniques de d'apprentissage statistique se sont popularisées, leur impact a été de plus en plus important et des échanges continus se sont installés entre les chercheurs développant les techniques d'apprentissage statistique et les utilisateurs de ces techniques. Ainsi des exigences de robustesse vis-à-vis d'attaques, de garanties de confidentialité ou d'interprétabilité ont été mises en exergue. Ces exigences sont aujourd'hui englobées par le terme « intelligence artificielle de confiance ». Elles ont orienté plusieurs

travaux présentés dans ce manuscrit, en particulier avec les thèses de Rafaël Pinot [4], Pierre-Emmanuel Clet [6], Arnaud Grivet Sébert [5] et Fabiola Espinoza Castellon [7]. Au-delà des régularisations classiques, de nouvelles contraintes répondant aux exigences récentes de l'apprentissage statistique — robustesse, confidentialité, interprétabilité — ont été introduites. Par exemple, la robustesse face aux perturbations adverses peut être recherchée en formulant une perte de type min-max, ou en ajoutant des termes pénalisant les prédictions sensibles. Plusieurs de nos contributions récentes traitent de ces enjeux, comme l'introduction de mécanismes de confidentialité différentielle dans les algorithmes de clustering [8] ou le développement de stratégies de défense en classification via l'introduction de bruit [4].

1.3.5 Apprentissage statistique et confidentialité

Une ligne de travaux a alors été ouverte avec l'équipe de Renaud Sirdey au CEA, visant à explorer l'interface entre l'apprentissage statistique et le chiffrement homomorphe, afin de mieux comprendre les compromis existant dans la mise en place conjointe de ces techniques. Plusieurs travaux ont été réalisés dans ce cadre [C2], [C3], [C6]. Ces travaux sont aujourd'hui parmi les plus avancés afin de protéger l'apprentissage statistique face à un spectre de menaces très large. Au cœur des techniques mises en œuvre dans ce cadre, l'apprentissage distribué est un pilier important, dans la mesure où il permet notamment de laisser les jeux de données au niveau du producteur de données, sans avoir à transférer les données au serveur central chargé de l'apprentissage. Les deux autres piliers sont le chiffrement homomorphe et la confidentialité différentielle, qui permettent de protéger le système face à des menaces portées par les participants ou par le serveur d'agrégation. L'apprentissage distribué est ainsi apparu comme un moyen fondamental pour assurer la bonne articulation entre différents compromis, tout en faisant apparaître de nouvelles menaces, qui sont au cœur des travaux de Fabiola Espinoza Castellon [C4], [C5], co-encadrée avec Aurélien Mayoue au CEA.

1.3.6 Des données d'entrée variées

Nos travaux ont initialement été focalisés sur les séries temporelles multivariées, avec les signaux électroencéphalographiques utilisés pour la conception d'interfaces cerveau-machine. Nous avons travaillé dans le cadre de partenariats sur des données énergétiques variées (consommations électriques, production de panneaux solaires, débits et pressions dans les réseaux d'eau). Dans le domaine de la santé, plusieurs actions ont permis de travailler sur des données réelles, par exemple l'électrocardiogramme et les données de santé associées à des transplantations pulmonaires avec l'hôpital Foch. Certains jeux de données

publics d'images ont permis de valider les travaux méthodologiques proposés. Depuis 2017, nous travaillons avec plusieurs partenaires sur des données associées à la cybersécurité, par exemple les données enregistrées par des serveurs DNS (Domain Name Server) fournies par un partenaire interne. Dans le cadre de travaux avec la RATP, nous avons travaillé sur des données permettant d'analyser conjointement l'activité de serveurs en salle de contrôle (métrique d'utilisation et logs systèmes) ainsi que le journal des événements terrains unitaires enregistrés sur deux lignes de métro pendant une durée relativement longue. Nous avons également considéré le cas de données disponible selon un modèle de streaming, notamment dans le cadre des travaux de Anne Morvan [8]. C'est dans ce cadre général, avec une forte sensibilité aux applications pratiques rencontrées dans le cadre des partenariats du CEA[3], que nos travaux ont été conduits depuis la soutenance de ma thèse en 2009.

[3] Au-delà des activités de recherche classiques, mes compétences pratiques ont eu l'occasion d'être testées dans le cadre compétitif de la plateforme kaggle.com (https://www.kaggle.com/cedricgp), entre 2012 et 2016. Parmi les résultats obtenus, en collaboration avec Alexandre Barachant et Rafal Cycon, nous avons remporté la première place à la compétition BCI Challenge @ NER 2015. Cette participation a donné lieu à une publication scientifique [Cn5].

2 Décomposition, segmentation et débruitage de signaux multivariés

Cette partie rend compte des contributions principales réalisées dans le prolongement de mes travaux de thèse. Lors de ma thèse, l'objectif avait été d'utiliser au mieux les informations à disposition sur le processus électrophysiologique afin d'en extraire l'information la plus utile possible. Pour ce faire, les méthodes développées s'appuyaient sur une combinaison de traitements utilisant des informations temporelles, spatiales et fréquentielles. Dans cette partie nous montrons le cheminement des travaux et les contributions s'inscrivant dans la lignée du paradigme issu de ma thèse, ayant conduit aux publications [J6], [J7], [J4], [J5], [Cn3], [Cn6], [C21].

2.1 Traitement de signaux multivariés : cas des signaux électroencéphalographiques

A la suite de la thèse, mes travaux se sont naturellement orientés vers le traitement des signaux électroencéphalographiques. Hormis l'expérience acquise pendant ma thèse sur cette modalité, il est apparu que ce cadre applicatif était relativement exigeant et constituait donc un terrain de jeu idéal pour l'application des méthodes statistiques de décomposition et de débruitage de signaux multivariés. Les signaux électroencéphalographiques (EEG) présentent des caractéristiques uniques qui rendent leur interprétation particulièrement difficile :

- Faible rapport signal/bruit (RSB) : Les signaux EEG associés à des activités cérébrales particulières sont souvent faibles en amplitude et sont facilement masqués par le bruit et les interférences. Ils sont contaminés par des artefacts qui peuvent provenir de sources physiologiques (comme les clignements d'yeux ou les mouvements musculaires) ou d'environnements (comme les interférences électriques). Identifier et retirer ces artefacts sans altérer les données utiles est un défi majeur.
- Non-stationnarité : la distribution des signaux EEG peut varier dans le temps, reflétant des dynamiques à différentes échelles de processus cérébraux. Cette non-stationnarité rend difficile l'application de certaines techniques d'analyse statistique ou de décomposition.

- Variabilité interindividuelle : les signaux EEG peuvent varier considérablement d'une personne à l'autre. Cette diversité rend difficile la création de modèles ou de systèmes qui fonctionnent bien pour tous les utilisateurs.
- Sensibilité aux conditions expérimentales : les signaux EEG peuvent être influencés par de nombreux facteurs, y compris le type de capteurs utilisés, leur emplacement sur le cuir chevelu, et l'état psychologique ou physiologique de la personne.

Ces caractéristiques posent des défis pour le traitement des signaux EEG, nécessitant le développement de méthodes avancées.

2.2 De la séparation aveugle à la séparation informée de sources

Ces travaux ont été réalisés en collaboration avec Reza Sameni. Ils ont fait l'objet d'un article de revue [J6] et font suite à l'article de conférence [C28].

Faisons tout d'abord le lien avec les approches développées dans ma thèse, en partant des travaux réalisés dans [J6], réalisés en collaboration avec Reza Sameni. L'objectif de l'article [J6] est d'étendre une méthode basée sur la décomposition en valeurs propres généralisées pour la détection et l'élimination automatiques des artefacts oculaires des enregistrements d'électroencéphalogramme (EEG) multicanal. Comme expliqué dans [10], plusieurs approches étaient employées mais l'analyse en composantes indépendantes, qui s'appuie sur des statistiques d'ordre supérieur, s'avérait la plus efficace [11], [12]. Afin d'améliorer les résultats obtenus grâce à cette technique, nous avons travaillé sur une approche exploitant les différences d'énergies entre les signaux oculaires et l'EEG. Le schéma de principe de l'approche proposée est représenté en Figure 1.

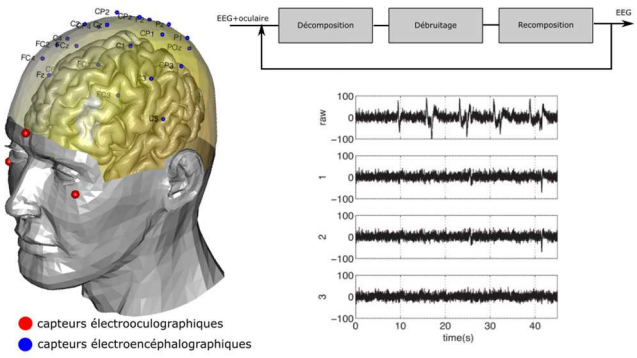


Figure 1 : Illustration de la méthode. A gauche, on voit la position des capteurs dans le cas idéal, c'est-à-dire lorsque des enregistrements EOG sont disponibles (en rouge). En haut à droite, la méthode est illustrée avec la possibilité de réaliser plusieurs itérations en fonction du résultat. En bas à droite, on voit les résultats de la technique après 1, 2 et 3 itérations de la méthode de débruitage, ainsi que les signaux bruts.

L'approche détecte tout d'abord les instants d'activités oculaires afin de construire deux matrices de covariance : une matrice de covariance des signaux EEG lors des activités oculaires, et une matrice de covariance sans activité oculaire. Ces deux matrices sont alors diagonalisées dans une base commune (quotient de Rayleigh) grâce à une décomposition en valeurs propres généralisées. Cette décomposition permet d'obtenir un classement des composantes qui contiennent le plus d'activité oculaire. Celles-ci sont débruitées afin d'enlever les activités oculaires. Nous avons utilisé la notion du nombre effectif de dimensions identifiables pour estimer le nombre de dimensions dominantes de l'espace oculaire, ce qui permet une convergence précise et rapide de l'algorithme. L'information fréquentielle est utilisée dans l'étape de débruitage en utilisant une approche basée sur des ondelettes qui permet de débruiter de manière sélective les premières composantes après décomposition, c'est-à-dire celles qui sont le plus contaminées par les artefacts oculaires. La spécificité de la technique présentée repose sur plusieurs points clés :

- L'approche classique de l'époque pour le débruitage de signaux EEG reposait sur la séparation de sources. Dans notre approche, le classement des composantes les plus ressemblantes aux signaux oculaires contaminants est automatique, s'appuyant sur l'utilisation de la décomposition en valeurs propres généralisées.
- Utilisation de l'information fréquentielle en procédant au débruitage ciblé des bandes de fréquences les plus contaminées par les artefacts oculaires.
- Approche itérative, qui permet de minimiser l'impact d'un mauvais étiquetage initial des instants de contamination.

Même si l'approche est plus rapide lorsque des capteurs oculaires ont été positionnés, celle-ci fonctionne de manière similaire avec uniquement des capteurs EEG. Le principe est identique et les instants d'activité oculaire sont identifiés grâce à l'énergie importante de ces artefacts.

Néanmoins, cette approche repose sur l'existence implicite d'un espace permettant de séparer le signal du bruit. On associe une technique linéaire (décomposition en valeurs propres généralisée) et une approche non-linéaire (débruitage à l'aide d'une transformée en ondelette) afin de séparer au mieux ces deux espaces. Si cette approche s'avère efficace dans le cas de la séparation entre les artefacts oculaires et l'EEG, elle ne peut être efficace dans le cas de la recherche d'activités plus complexes, dont les caractéristiques temporelles, fréquentielles et spatiales ne sont pas assez distinctes. L'utilisation de dictionnaires avait fait ses preuves dans le domaine des images, mais des difficultés importantes demeuraient pour leur utilisation pour des signaux temporels, en particulier dans l'application aux signaux EEG.

2.3 Décomposition sur des dictionnaires

Les travaux sur la décomposition de dictionnaires ont été réalisés dans le cadre de la thèse de Yoann Isaac (encadrement avec Jamal Atif et Michèle Sebag). Les travaux sur l'apprentissage de dictionnaires ont été menés en collaboration avec Quentin Barthélemy, Antoine Souloumiac et Anthony Larue. Ces travaux ont été publiés dans les articles de journaux [J4] et [J7], et les articles de conférence [C21], [Cn3] et [Cn6].

Dans un problème inverse, on cherche à estimer un signal source à partir d'observations bruitées enregistrées par des capteurs. Selon les hypothèses faites sur les dimensions et les propriétés des signaux observés et des signaux sources, de nombreuses méthodes ont été proposées dans le domaine du traitement du signal. Considérons le cas particulier d'une hypothèse de relation linéaire entre les sources et les observations, alors on a

$$\boldsymbol{y} = \Phi\, \boldsymbol{x} + \boldsymbol{e}$$

où $\boldsymbol{y} \in \mathbb{R}^M$ représente les observations, $\boldsymbol{x} \in \mathbb{R}^N$ représente les sources, $\boldsymbol{e} \in \mathbb{R}^M$ représente un bruit additif et $\Phi \in \mathbb{R}^{M \times N}$ est une matrice de mélange.

2.3.1 Décomposition temps-fréquence régularisée spatialement

Soit $Y \in \mathbb{R}^{T \times C}$ un segment de taille T enregistré à l'aide de C capteurs. On suppose que les signaux observés sont issus du mélange linéaire à partir d'un grand nombre d'activités unitaires, regroupées dans un dictionnaire $\Phi \in \mathbb{R}^{T \times N_\Phi}$.

$$Y = \Phi\, X + E$$

où Φ est un dictionnaire temps-fréquence regroupant des atomes de Gabor, $\mathrm{X} \in \mathbb{R}^{N_\Phi \times C}$ contient les coefficients de décomposition et $\mathrm{E} \in \mathbb{R}^{T \times C}$ regroupe les erreurs de mesure au niveau des capteurs.

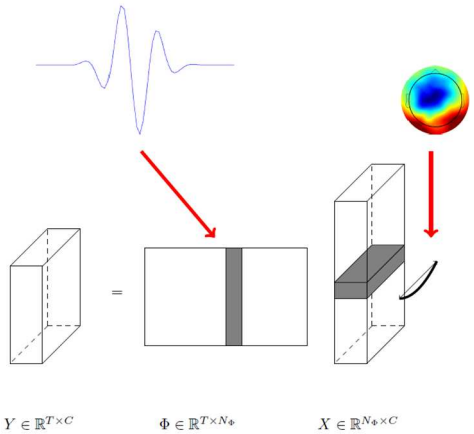


Figure 2 : Modèle de décomposition temporelle d'un signal multivarié, proposé dans [Cn6].

Dans le cadre des signaux EEG, nous nous sommes focalisés sur l'introduction de deux types de régularisations : tout d'abord une régularisation parcimonieuse écrite sous sa forme relâchée $J_{\text{par}}(X) = \|X\|_1$ afin de limiter le nombre d'atomes sollicités et également, en notant $V(c)$ l'ensemble des capteurs proches d'un capteur c, et $x_{i,j}$ étant la composante (i,j) de la matrix X, $J_{\text{spat}}(X) = \sum_{n=1}^{N_\Phi} \sum_{c=1}^{C} \sum_{s \in V(c)} (x_{n,c} - x_{n,s})^2$ qui permet de favoriser des coefficients similaires sur des capteurs proches.

Pour un dictionnaire Φ fixé, le problème de la décomposition des données Y sur ce dictionnaire conduit alors à l'optimisation de la fonction de coût, sous sa forme relaxée :

$$J(X) = \|Y - \Phi X\|_F^2 + \lambda J_{\text{par}}(X) + \mu J_{\text{spat}}(X)$$

L'optimisation de cette forme relaxée de la décomposition parcimonieuse se fait grâce à un algorithme classique appelée FISTA (Fast Iterative Shrinkage Thresholding Algorithm). Nous avons montré que cette approche, basée sur une décomposition temps-fréquence régularisée spatialement, permet dans le cas des signaux EEG de prendre en compte l'effet de diffusion au niveau du crâne et de fournir ainsi des décompositions améliorées par rapport à une approche utilisant uniquement la parcimonie. Nous avons également montré que l'approche, utilisée en débruitage en reconstruisant une approximation $\tilde{Y}$, conduisait à une amélioration du taux de classification lors d'expériences d'interface cerveau-machine appelées P300, qui s'appuient sur un phénomène neurophysiologique modulé par l'attention intervenant environ 300 millisecondes après l'occurrence d'un stimulus visuel généré sur les lignes ou colonnes d'une matrice de lettres.

Bien que cette approche donne des résultats satisfaisants dans ce contexte, elle impose néanmoins la possibilité de segmenter les signaux des capteurs en créant des segments de taille fixe, taille notée T précédemment. Si cette hypothèse est crédible pour des expériences neurophysiologiques du type P300, elle s'avère restrictive pour de nombreux autres cas.

2.3.2 Multi-SSSA : régularisations parcimonieuses structurées

Dans cette seconde approche, l'optique est relativement différente puisque la décomposition se fait à présent sur un dictionnaire de motifs spatiaux et la formulation employée pour la régularisation est plus générique. Reprenons les notations précédentes et considérons le modèle suivant (les dimensions sont inversées) :

$$Y = \Phi X + E$$

où $Y \in \mathbb{R}^{C \times T}$ est la matrice des enregistrements en dimension C ordonnés temporellement, $\Phi \in \mathbb{R}^{C \times N_\Phi}$ représente dans ce cas un dictionnaire surcomplet de N_Φ atomes spatiaux de dimension C, $X \in \mathbb{R}^{N_\Phi \times T}$ est la matrice de décomposition et $E \in \mathbb{R}^{C \times T}$ rend compte des bruits d'enregistrements. La décomposition selon ce modèle a pour objectif de séparer les

signaux d'intérêt du bruit additif, et de disposer d'une matrice de décomposition permettant d'exploiter les contributions unitaires d'intérêt de celles qui doivent être considérées comme des interférences. Le modèle sous-jacent peut-être illustré par le schéma de la Figure 3.

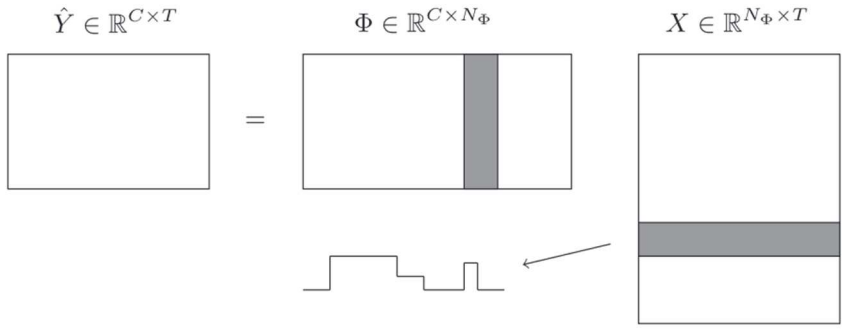


Figure 3 : Illustration d'une décomposition spatiale régularisée par une structure temporelle sur la décomposition cherchée. Ici on illustre avec la prise en compte d'une régularisation visant à créer une fonction par morceaux sur les activations temporelles des atomes. D'après [J4].

La fonction de coût permettant de réaliser la décomposition parcimonieuse associée à ce modèle s'écrit alors :

$$\text{argmin}_{X \in \mathbb{R}^{N_\Phi \times T}} \|Y - \Phi X\|_2^2 + \lambda_1 \|X\|_1 + \lambda_2 \|X\,P\|_1$$

dans laquelle on introduit deux coefficients de régularisation λ_1 et λ_2, et $P \in \mathbb{R}^{T \times N_P}$ est une matrice encodant la connaissance a priori concernant la structure de la décomposition recherchée (N_P est une dimension quelconque de la matrice P, permettant de coder la structure de la régularisation efficacement). La matrice P peut être interprétée comme un ensemble de filtres linéaires sur lesquels la projection de la décomposition doit être parcimonieuse. Plusieurs points sont fondamentaux dans cette nouvelle formulation. Tout d'abord la fonction de coût ne permet par l'utilisation d'algorithmes d'optimisation classiques en raison des deux termes non-différentiables, utilisant la norme ℓ_1. Bien que des approches existaient pour des dimensions faibles et des dictionnaires non-redondants, aucune technique n'avait été mise en place pour la résolution efficace de ce type de problèmes. De plus, on peut constater la généricité de la formulation précédente en observant qu'il est possible de retrouver le classique problème « Fused-LASSO » dans un contexte multidimensionnel en choisissant les matrices ci-dessous, approximations discrètes d'une régularisation appelée « variation totale » :

$$P^{FL_1} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & \ddots & 0 \\ 0 & 0 & \ddots & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ ou } P^{FL_2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & -2 & \ddots & 0 \\ 0 & 1 & \ddots & 1 \\ 0 & 0 & \ddots & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

La technique d'optimisation mis en place s'appuie sur un schéma d'itérations appelé « split Bregman » et utilisé dans le cas de l'optimisation de fonctions de coût faisant intervenir des normes ℓ_1 par [13]. Des résultats sur signaux simulés ainsi que sur des données réelles ont été présentés dans l'article [J4].

Application à la recherche de micro-états en EEG

L'article [Cn3] nous a permis de présenter une application du modèle précédent pour décomposer l'EEG enregistré pendant une expérience de P300. Dans cette expérience, un dictionnaire est appris à partir des signaux enregistrés pendant les premières sessions de l'expérience, résultant en 576 atomes (cf. article pour les détails de construction du dictionnaire). Cette expérience a mis en évidence la capacité du modèle proposé à capturer des activités jusqu'alors difficiles à extraire, comme en témoigne la Figure 4, en particulier grâce à l'atome 2 qui capture une activité à la fréquence de la stimulation visuelle.

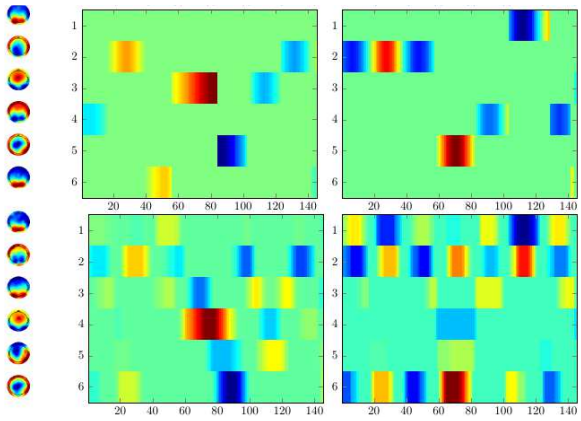


Figure 4 : Représentation de la réponse illustrant l'amplitude de décomposition sur les atomes de dictionnaires en présence du potentiel évoqué P300 (gauche) et l'amplitude des coefficients en l'absence du potentiel évoqué P300 (droite). Pour chaque colonne, on représente la décomposition et les atomes topographiques associés, pour la décomposition utilisée dans la littérature (haut) et pour le modèle proposé (bas). Le temps est exprimé en millisecondes.

Généricité du modèle

Notons enfin que le modèle proposé avec Multi-SSSA permet de combiner des approches de manière inédite. Reprenons le modèle le plus général en transposant Y afin de réaliser une décomposition temporelle ($Z = Y^\top \in \mathbb{R}^{T \times C}$, $\Phi \in \mathbb{R}^{T \times N_\Phi}$), avec sa fonction de coût associée :

$$\operatorname{argmin}_{X \in \mathbb{R}^{N_\Phi \times C}} \|Z - \Phi X\|_2^2 + \lambda_1 \|X\|_1 + \lambda_2 \|X\, P\|_1$$

Nous avons montré qu'en prenant :

- Φ un dictionnaire d'atomes de Gabor ;

- $P \in \mathbb{R}^{C \times N_P}$ est construit de la manière suivante : on génère environ 5000 topologies réalistes à partir de la résolution du problème EEG direct grâce à OpenMEEG [14]. Puis on sélectionne 350 topologies grâce à une approche gloutonne en maintenant une cohérence de dictionnaire inférieure à 0.9. Enfin P est calculé en prenant l'inverse de Moore-Penrose de ce sous-ensemble de topologies, comme noté dans [15],

alors on obtient à la fois une régularisation parcimonieuse temporelle et spatiale, qui généralise les modèles considérés précédemment.

Apprentissage de dictionnaires et de régularisations

En parallèle des travaux réalisés avec Yoann Isaac dans le cadre de sa thèse, j'ai eu l'occasion de collaborer avec un autre doctorant du CEA, Quentin Barthélemy. Dans le cadre de sa thèse, il a travaillé sur des décompositions parcimonieuses invariantes par des transformations telles que des rotations (par exemple pour l'écriture manuscrite) ou des translations. Dans le cadre de [J7], un algorithme d'apprentissage de dictionnaires pour des signaux multivariés a été proposé. Cet algorithme alterne une étape de projection basée sur l'algorithme « Orthogonal Matching Pursuit » incluant une étape de recherche du décalage temporel optimal, puis une descente de gradient pour la mise à jour du dictionnaire. Nous avons montré dans [J7] que cette approche permettait d'expliquer la variance des signaux EEG avec une quantité réduite d'atomes (par rapport à un dictionnaire temps-fréquence utilisé de manière classique). L'approche a également permis de mettre en évidence des potentiels évoqués P300, qui constituent un signal physiologique simple et robuste pour la mise en place d'une interface cerveau machine.

Par ailleurs des expériences ont été menées au cours de la thèse de Yoann Isaac [16, p. 71] pour apprendre des structures particulières de régularisations. La formulation proposée conduit à la résolution d'une équation de Sylvester pour un problème d'apprentissage de métrique. Bien que les expériences aient été conduites pour montrer les bonnes performances de la technique d'apprentissage de régularisation face à une régularisation laplacienne, ces travaux n'ont pas donné lieu à des investigations plus poussées.

2.4 Segmentation et extraction de connaissances à partir de signaux multivariés

Les travaux ci-dessous ont été réalisés dans le cadre de la thèse de Flore Harlé (encadrement avec Florent Chatelain, Sophie Achard et Olivier Michel). Ils ont été publiés dans un article de journal [J5], et les articles de conférence [C19] et [Cn4].

En parallèle des travaux entrepris sur l'utilisation de dictionnaires pour la décomposition de signaux multivariés, un autre axe de recherche a été lancé en collaboration avec Sophie Achard, Florent Chatelain et Olivier Michel : la segmentation robuste de séries temporelles multivariées. Complémentaire de l'utilisation de dictionnaires, les travaux entrepris ont été principalement conduits par Flore Harlé dans le cadre de sa thèse [17]. Plusieurs contributions ont résulté de ces travaux :

- Le Bernoulli Detector : il s'agit d'un modèle pour la détection de multiples ruptures dans un signal univarié. Il utilise les p-valeurs d'un test statistique non paramétrique basé sur les rangs des données considérées (le test de Wilcoxon-Mann-Whitney). Il est défini dans un cadre bayésien, qui rend la méthode applicable à plusieurs distributions sans nécessiter de modification du modèle. Il introduit un paramètre pour contrôler le risque de fausses détections.
- Une approche de segmentation multivariée sous contrainte de liaisons connues entre les signaux : cette deuxième contribution est une adaptation du Bernoulli Detector pour traiter simultanément plusieurs signaux. Elle introduit dans le modèle de détection la possibilité d'introduire des *a priori* sur les relations de dépendances connues entre les séries temporelles. La méthode proposée conduit ainsi à une segmentation jointe de l'ensemble des séries temporelles. Ces ruptures peuvent être communes à plusieurs signaux ou apparaître de manière isolée sur un unique signal.

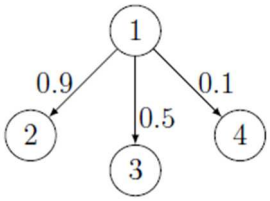


Figure 5 : D'après [17], graphe de dépendance imposé entre des signaux synthétiques. Les nœuds représentent l'indice du signal où se produit la rupture et les arêtes indiquent l'existence d'un lien entre les ruptures. Ainsi, dans le cas représenté, une rupture dans le signal 1 entraîne l'apparition d'une rupture sur le signal 2 avec une probabilité de 0.9, sur le signal 3 avec une probabilité de 0.5 et sur le signal 4 avec une probabilité de 0.1.

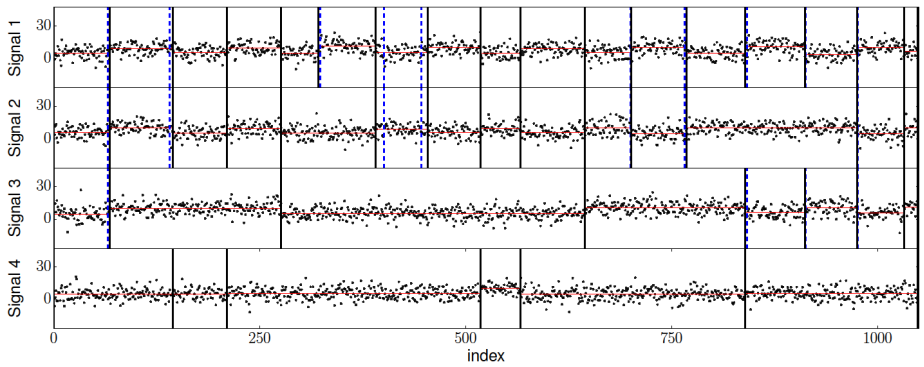


Figure 6 : Détection de rupture obtenue avec un a priori non informatif, sans restriction du nombre de configurations possibles.

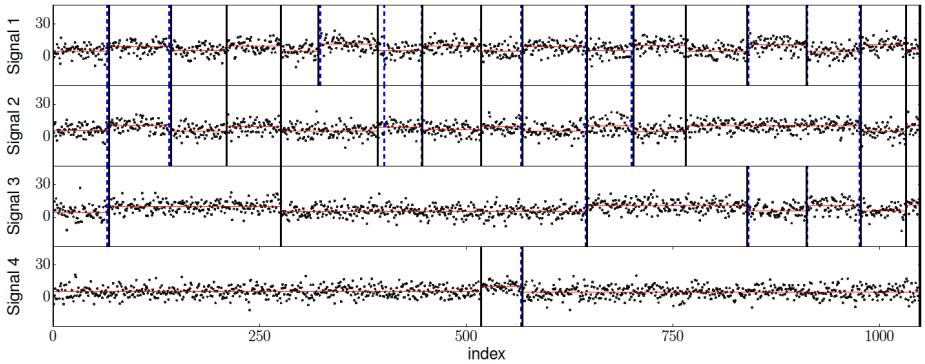


Figure 7 : Segmentation obtenue avec un a priori informatif, utilisant la connaissance du graphe de la Figure 5, et restreignant les configurations possibles de rupture jointe.

- En contexte multivarié, les travaux ont conduit à proposer une méthodologie pour exploiter les ruptures trouvées pour inférer la structure de dépendance entre les séries temporelles : cette approche tente de retrouver le modèle d'indépendance conditionnelle entre les variables. Elle s'appuie sur le Bernoulli Detector.

2.5 Conclusion et impact sur les axes de recherche

Les travaux entrepris dans les domaines de la décomposition, de la segmentation et du débruitage des séries temporelles multivariées se sont inscrits en continuité assez forte avec mes travaux de thèse sur les interfaces cerveau-machines asynchrones et mon ancrage sur les techniques de traitement du signal. Ces différents travaux ont conduit à des techniques utiles et efficaces, qui ont été testées dans plusieurs projets plus appliqués, notamment dans le cadre du projet européen eCo-FEV (détection d'anomalies) ou du projet interne Subénergie visant la détection d'anomalies de production dans des champs de panneaux photovoltaïques. Trois constats ont néanmoins été faits en 2015-2016 : tout d'abord des

efforts importants étaient encore nécessaires dans ces voies pour conduire à des techniques passant à l'échelle (vis-à-vis du nombre de séries temporelles et de la taille des séries) ; de plus, les efforts orientés sur la compréhension des signaux n'étaient pas alignés avec les nouveaux paradigmes observés pour exploiter des données produites en continu ; enfin les techniques développées donnaient souvent lieu à une accumulation d'étapes pour conduire à la résolution de la tâche finale (classification, régression, prévision), rendant parfois difficile la possibilité d'évaluer simplement les résultats produits.

Ces réflexions ont conduit à proposer de nouvelles pistes de recherche visant à simplifier la chaîne globale de traitement de séries temporelles et à prendre en compte l'utilisation *in fine* des algorithmes et des modèles développés en intégrant le nouveau paradigme associé au traitement de données produites en continu.

3 Discrimination, compression et approximation

Cette partie se focalise sur certains compromis inhérents à l'apprentissage statistique. Pour illustrer la transition vis-à-vis des travaux précédents, considérons tout d'abord le compromis entre l'interprétabilité et le pouvoir de discrimination. Dans l'article [C23], nous avons commencé les premières investigations afin d'utiliser la décomposition sur dictionnaires dans des tâches de classification. Une décomposition parcimonieuse donne lieu à des représentations appelées « spikegrams », qui indiquent l'utilisation d'atomes du dictionnaire avec un coefficient calculé selon des algorithmes tels que ceux présentés dans la partie précédente. Un exemple d'une telle décomposition est illustré dans la Figure 8.

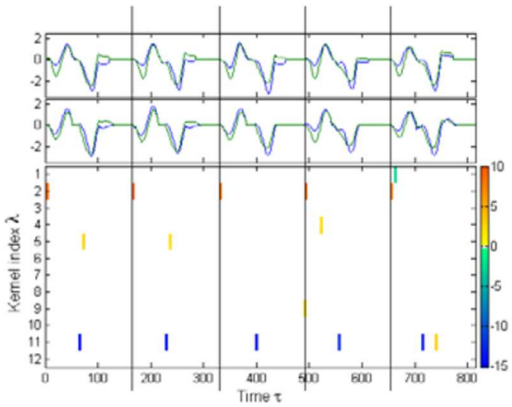


Figure 8 : Décomposition parcimonieuse d'un signal de dimension 2 sur un dictionnaire de taille 12. Extrait de [C23].

Une telle représentation est extrêmement riche d'un point de vue de l'interprétation du signal initial sur des dictionnaires de motif, mais lorsqu'elle doit être utilisée dans des tâches de classification, les classifieurs classiques ne vont pas être robustes face à des utilisations instables d'atomes proches. Les dictionnaires offrent une grande souplesse, mais n'offrent pas les propriétés mathématiques associées au théorème de projection, et s'avèrent donc plus sensibles que les bases orthonormées face à des bruits dans les signaux sources. Dans les travaux mentionnés ici, nous avons proposé l'utilisation d'une métrique plus adaptée aux données résultant de la décomposition sur des dictionnaires. Néanmoins, l'analyse a posteriori de l'approche fait apparaître une gestion perfectible du compromis entre interprétabilité et pouvoir de discrimination : dans une première étape, la puissance du modèle est consommée pour projeter au mieux les signaux sources sur un dictionnaire,

tandis que dans l'étape de classification, la méthodologie mise en place s'efforce de discriminer entre des classes distinctes.

Nous collectons, enregistrons et tentons d'analyser de plus en plus de données années après années. Aujourd'hui, nous créons chaque jour environ 2500 pétaoctets de données, qui proviennent d'applications extrêmement variées telles que le suivi des variables climatiques, les activités sur les médias sociaux, les images et les vidéos, ou encore les signaux GPS émis par les téléphones portables. Si les réseaux de télécommunication permettent la collecte d'une telle quantité de données et les infrastructures informatiques rendent possible leur stockage, l'analyse en ligne de gros volumes de données reste limitée par les complexités des algorithmes nécessaires au traitement. Or de nombreux contextes applicatifs (industriels, services ou sécurité) ne peuvent se contenter d'analyses de données rétrospectives (sur des gigantesques bases de données distribuées par exemple dans des clusters Hadoop) mais nécessitent une intégration continue des nouvelles données. Deux raisons principales peuvent être avancées pour expliquer ce besoin :

- Le contexte est la plupart du temps non-stationnaire, ainsi garder un modèle à jour (que ce soit un modèle de prédiction ou un modèle de représentation) nécessite d'intégrer au plus vite les données les plus récentes ;
- Les complexités en espace des algorithmes d'apprentissage rendent souvent prohibitif le traitement de grosses quantités de données de manière conjointe (in-memory), tandis que les approches "en ligne" permettent de s'affranchir de ces limitations en considérant les observations une par une (ou par petits blocs).

Les travaux de thèse de Anne Morvan [8] se sont inscrits dans le paradigme de traitement de données en flux : les données acquises doivent être traitées le plus efficacement possible en temps et en espace, l'information issue des traitements réalisés doit permettre de reconstruire avec une bonne approximation la donnée initiale, et la reconstruction doit être possible avec une complexité faible. Ce paradigme repose sur l'hypothèse de parcimonie des observations, c'est-à-dire qu'à un instant t, seules certaines dimensions du vecteur d'observation sont non nulles. Cette hypothèse n'est néanmoins pas très contraignante puisqu'il suffit de trouver une base adaptée à la représentation parcimonieuse des observations.

D'une part, en se basant sur une hypothèse de parcimonie, Candès, Romberg et Tao (2006) ont donné les conditions sur la parcimonie d'un vecteur pour qu'il soit possible de le reconstruire à partir de l'observation d'un petit nombre de points. Ces travaux ont permis l'essor d'approches générales dénommées « compressive sensing » consistant à observer un nombre restreint de points, construits à partir de la multiplication des observations par une matrice aléatoire, en conservant une erreur limitée sur la qualité de la reconstruction.

Plusieurs approches ont depuis été proposées afin d'aller au-delà des limites mentionnées dans cet article, en utilisant deux pistes principales : rendre le processus d'échantillonnage adaptatif, ou encore rajouter des hypothèses supplémentaires sur la parcimonie du vecteur d'entrée. D'autre part des travaux importants ont été réalisés depuis les années 1995 sur l'utilisation de fonctions de hashage spécifiques pour limiter la complexité en espace liée au traitement de flux de données. Le « linear sketching » permet par exemple de réaliser des opérations spécifiques (heavy hitters, frequency moment estimation) en utilisant un espace mémoire extrêmement limité et des algorithmes de reconstruction permettant d'avoir une approximation contrôlée de la quantité à estimer. En pratique, les grands volumes de données ainsi que la vitesse d'acquisition des données soulèvent des défis importants pour l'application de techniques d'apprentissage statistique :

1. Les grands volumes de données posent en réalité deux problèmes. Dans certains cas les jeux de données d'entraînement ne tiennent pas dans la mémoire centrale d'un unique ordinateur. On considère qu'un traitement classique peut être réalisé si la mémoire vive d'un ordinateur peut stocker environ 3 fois le jeu de données d'entraînement. Au-delà il faut faire appel à des techniques spécialisées pour appliquer l'apprentissage statistique. Dans une seconde famille de situations, le jeu de données peut rentrer dans la mémoire vive, mais la complexité des algorithmes à appliquer sur le jeu de données a une complexité temporelle non-linéaire. Par exemple dans le cas d'une complexité temporelle ou spatiale de $\mathcal{O}(N^2)$, avec N le nombre d'exemples dans le jeu de données, il est impossible d'appliquer cet algorithme sur des jeux de données qui rentrent très largement sur la mémoire vive d'un ordinateur (par exemple dans le cas de la multiplication de matrices).
2. Les exemples peuvent nécessiter de grandes dimensions, regrouper de nombreux attributs. La conséquence peut être une difficulté des algorithmes à correctement évaluer les densités spatiales dans l'espace, et ainsi à appliquer des fonctions dans des endroits dépeuplés de l'espace (malédiction de la dimension).
3. Les données peuvent être observées de manière itérative (streaming) dans un flux potentiellement infini.
4. Si les données sont de plus en plus nombreuses, les étiquettes associées à ces données, qui permettent de mettre en place des techniques d'apprentissage supervisé, sont en général rares.

Dans le cadre de la thèse de Anne Morvan, nous nous sommes intéressés à deux problèmes importants de l'apprentissage statistique : la recherche de plus proches voisins, et l'apprentissage non-supervisé. Dans ces deux cas, on cherche à segmenter ou à rapprocher des données en groupes en utilisant des proximités dans leurs propriétés.

Réaliser ces tâches de manière efficace dans le contexte décrit précédemment implique de s'appuyer sur des représentations compactes des données, préservant approximativement les distances et/ou les structures, et réduisant les complexités algorithmiques associées au traitement sur ces structures. Nous avons focalisé nos travaux sur les méthodes de sketching (réduction de dimension et/ou échantillonnage) qui sont adaptées aux données observées en flux. Les représentations compactes visent à réduire l'impact spatial (mémoire) des algorithmes de traitement et la complexité temporelle par exemple (augmentant ainsi le débit possible). L'utilisation de ces techniques a un coût : une précision moindre du résultat du traitement. C'est ainsi que la notion de compromis apparaît de manière cruciale. Les besoins et les contraintes liés à l'application doivent être adaptés afin de proposer les compromis les plus efficaces entre le **coût spatial de la structure de données**, le **débit associé au traitement des données** et la **précision du résultat du traitement**.

3.1 Projections linéaires aléatoires pour l'apprentissage statistique en grande dimension

Ces travaux ont fait l'objet d'un article de conférence à AISTATS 2017 dans le cadre de la thèse de Anne Morvan [C17] (co-encadrée avec Jamal Atif).

3.1.1 Projection aléatoire avec TripleSpin

Soit $\boldsymbol{x} \in \mathbb{R}^d$ un vecteur dans un espace de dimension $\boldsymbol{d}$. On souhaite projeter ce vecteur sur un espace défini par une famille de $\boldsymbol{p}$ vecteurs $\{u_1, \dots, u_p\}$ où chaque vecteur est de dimension $\boldsymbol{d}$. Dans le cas qui nous intéresse, on considère les approches capables de réduire la dimension du problème considéré, ainsi $p \ll d$. Créons alors $U = [u_1, \dots, u_p] \in \mathbb{R}^{d \times p}$ la matrice agrégeant selon les colonnes les vecteurs u_i. Le sous-espace engendré par la famille de vecteurs $\{u_1, \dots, u_p\}$ est aussi notée $\mathrm{span}\{u_1, \dots, u_p\}$. La reconstruction de la projection de $\boldsymbol{x}$ sur $\mathrm{span}\{u_1, \dots, u_p\}$ dans $\mathbb{R}^d$ est alors définie comme :

$$\mathbb{R}^d \ni \hat{\boldsymbol{x}} = U\,\beta$$

Où les coefficients $\beta \in \mathbb{R}^p$ sont inconnus. Si $x \in \mathrm{span}\{u_1, \dots, u_p\}$ alors ce système linéaire a une solution exacte et $\boldsymbol{x} = \hat{\boldsymbol{x}} = U\,\beta$. Dans le cas général, $\boldsymbol{x}$ et $\hat{\boldsymbol{x}}$ sont différents et on peut uniquement dire que le résidu $\boldsymbol{r} = \boldsymbol{x} - \hat{\boldsymbol{x}}$ est le plus petit possible quand il est orthogonal à $\mathrm{span}\{u_1, \dots, u_p\}$. Ainsi

$$\boldsymbol{x} - U\,\beta \;\perp\; U \implies U^T(\boldsymbol{x} - U\,\beta) = 0$$
$$\implies \beta = (U^T U)^{-1} U^T \boldsymbol{x}$$

Et par conséquent, on a

$$\hat{\boldsymbol{x}} = U\,(U^T U)^{-1}\,U^T \boldsymbol{x}$$

qui permet de définir le projecteur $\Pi = U\,(U^T U)^{-1}\,U^T \in \mathbb{R}^{d\times d}$.

Une projection linéaire aléatoire est une projection linéaire pour laquelle les éléments de la matrice de projection sont des variables aléatoires dont les éléments sont indépendants, identiquement distribués (i.i.d). Ils sont échantillonnés selon une distribution de moyenne nulle et dont la variance est 1. Il est important de noter que le choix des éléments de la matrice aléatoire U est indépendant des données, contrairement à de nombreuses méthodes comme l'analyse en composantes principales ou l'analyse en composantes indépendantes. De manière surprenante, la projection de données sur des matrices aléatoires fonctionne de manière très satisfaisante pour de nombreuses applications, en dépit du fait que les éléments de la matrice soient aléatoires. Cette performance s'appuie en réalité sur le lemme de Johnson-Lindenstrauss [18]. Celui-ci lie la distance entre deux points de l'espace initial (en grande dimension) avec la distance entre ces deux mêmes points après projection aléatoire dans un espace de dimension plus faible. Dans la version de 1984, ce théorème donne des bornes aux probabilités de succès de cette approximation pour une matrice aléatoire dont les éléments sont i.i.d de moyenne nulle et de variance unitaire, par exemple une distribution Gaussienne $\mathcal{N}\left(0, \frac{1}{p}\right) = \frac{1}{\sqrt{p}}\,\mathcal{N}(0,1)$. D'autres analyses s'appuient sur le formalisme de la concentration de la mesure [19] pour expliquer pourquoi la projection aléatoire permet en effet de donner des approximations satisfaisantes en dimensions largement inférieures à la dimension de l'espace d'origine. De nombreux résultats ont étendu le lemme de Johnston-Lindenstrauss à des normes différentes de ℓ_2 et au cas où on choisit des vecteurs de projection parcimonieux notamment.

S'il est possible de réaliser des projections de données en grande dimension en perdant une quantité limitée d'information, il faut néanmoins observer que le coût calculatoire de la projection aléatoire n'est pas négligeable. Ainsi la projection aléatoire nécessite $\mathcal{O}(dp)$ opérations pour projeter un vecteur de dimension d dans un espace réduit de dimension p. Pour améliorer le coût calculatoire associé à cette opération, plusieurs auteurs ont proposé l'utilisation de matrices aléatoires structurées, qui permettent d'utiliser des simplifications algorithmiques s'appuyant sur la transformée de Fourier (cas de matrices Gaussiennes circulantes) ou des transformées de Hadamard (cas des matrices de Hadamard). Néanmoins lorsqu'on introduit ce type de structure dans la matrice de projection aléatoire, on sort des conditions classiques garantissant la qualité du résultat. Dans [C17], les familles de vecteurs de projection basées sur la famille TripleSpin ont été considérées. Celles-ci sont construites comme le produit de trois blocs de matrices structurées (Toeplitz, Hankel, Hadamard, ...).

Au-delà de l'accélération pratique sur le coût conputationnel et l'occupation mémoire (la technique avait été proposée auparavant notamment par [20]), nous avons donné les conditions sur les trois blocs de matrices afin de fournir des garanties similaires à la version basée sur l'homologue non-structurée. Les preuves s'appuient notamment sur des outils de la concentration de la mesure, le Théorème Central Limite de type Berry-Esseen [21].

Plusieurs applications ont été utilisées pour démontrer la capacité des projections sur matrices structurées du type $HD_3HD_2HD_1$ (matrices de Hadamard et matrices diagonales avec des variables de Rademacher) à assurer des performances similaires aux matrices non-structurées : cross-polytope Locallity-Sensitive Hashing (LSH), approximation de noyaux, optimisation convexe avec des Newton sketches, ou encore l'optimisation des calculs dans les réseaux de neurones.

3.1.2 Expériences numériques : Locality-Sensitive Hashing

Les techniques de Locality-Sensitive Hashing (LSH) [20], [22], [23], [24], [25] s'avèrent être une approche efficace pour aborder les problèmes de recherche de plus proches voisins dans des espaces de grande dimension. Le principe consiste à approximer la recherche de plus proche voisin classique (impliquant un calcul exhaustif des distances dans l'espace original) par une recherche de plus proche voisin approchée, consistant à réaliser la recherche dans un espace plus propice au calcul des distances. Dans [26], les données sont projetées sur un espace binaire en utilisant une projection aléatoire structurée de type $HD_3HD_2HD_1$, avec H une matrice de Hadamard, et D_i des matrices de Rademacher. Cette approche permet d'obtenir une accélération du calcul et une diminution de l'espace mémoire nécessaire. Dans notre validation expérimentale [C17], on a montré la qualité des fonctions de hashage dans la méthode de Crosspolytope LSH [24]. La Figure 9 reporte les résultats qui permettent de conclure à l'équivalence entre la matrice aléatoire gaussienne G de taille 256×64 et cinq autres types de matrices de la famille TripleSpin structurés :

- $G_{\text{circ}}\, K_2\, K_1$,
- $G_{\text{Toeplitz}}\, D_2\, H\, D_1$,
- $G_{\text{skew-circ}}\, D_2\, H\, D_1$,
- $H\, D_{g_1,\ldots,g_n}\, H\, D_2\, H\, D_1$,
- $H\, D_3\, H\, D_2\, H\, D_1$,

où K_i, G_{Toeplitz}, $G_{\text{skew-circ}}$ et G_{circ} sont respectivement une matrice de Kronecker, une matrice de Toeplitz gaussienne, une matrice anti-circulante gaussienne et une matrice gaussienne circulante.

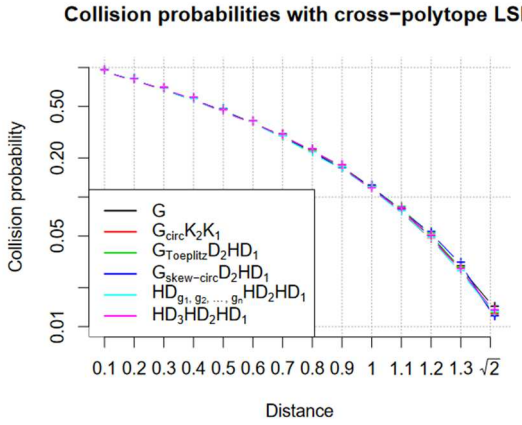


Figure 9 : Cross-polytope LSH. Probabilités de collision pour des distances comprises entre 0 et $\sqrt{2}$.

3.2 Apprentissage de codes binaires compacts pour la recherche des plus proches voisins

Ces travaux, réalisés dans le cadre de la thèse de Anne Morvan, ont donné lieu à la publication [C15]. Une version étendue est disponible [W2].

Alors que la section précédente s'est intéressée à une approche indépendante des données d'entrée, nous abordons dans ce travail une approche permettant également d'aborder la recherche de plus proches voisins, mais en utilisant les caractéristiques des données. On souhaite à nouveau être en mesure de faire une recherche par similarité efficace.

3.2.1 Formalisation du problème

Considérons un flux de données de moyenne nulle $\{x_t \in \mathbb{R}^d\}_{1 \leq t \leq n}$. On considère une méthode de projection linéaire sur un code binaire, i.e. $b_t = \text{sign}(\widetilde{W}\, x_t) \in \{0,1\}^c$ pour tout $t \in [n]$, où c est la taille du code final recherché avec $c \ll d$, et $\widetilde{W} \in \mathbb{R}^{c \times d}$ est l'opérateur de projection linéaire. Nous nous plaçons dans le cadre des fonctions de hashage hypercubique. Ainsi $\widetilde{W} = R\, W$, où W est un opérateur linéaire de réduction de dimension et R une matrice orthogonale carrée de dimension c. Alors que W est une matrice classique de réduction de dimension, le rôle de R est de corriger l'orientation des données afin que l'application de la fonction signe subséquente ne crée pas des distorsions artificielles de codes binaires pour des points en réalité proches dans l'espace de projection. Anne Morvan a proposé une approche fonctionnelle dans les cas hors-ligne et en ligne, mais nous focalisons

ici la présentation sur le cas en ligne, qui porte le résultat le plus marquant par rapport à la littérature lorsque la thèse a été réalisée.

3.2.2 UnifDiag Hashing

Pour un batch de nouvelles données ou un unique échantillon x_t, l'algorithme proposé dans [C15] repose sur :

- L'estimation du sous-espace de projection permettant de concentrer la variance sur un nombre réduit de composantes (projection sur les c premières composantes estimées de la matrice de covariance), $v_t = W\, x_t$ où $v_t \in \mathbb{R}^c$,
- L'application d'une matrice de rotation $R \in \mathbb{R}^{c\times c}$ afin d'équilibrer les densités des points selon des critères que nous allons préciser par la suite, $y_t = R\, v_t$, où $y_t \in \mathbb{R}^c$
- Construction du code binaire selon le signe de chacune des composantes de y_t, $b_t = \text{sign}\,(y_t)$ avec $b_t \in \{0,1\}^c$.

L'approche proposée par Anne Morvan s'appuie sur OPAST [27] pour le calcul des c premières composantes principales de manière itérative. Au-delà de la possibilité de calcul en ligne des composantes principales, l'originalité du travail de Anne Morvan réside dans l'approche consistant à appliquer une matrice de rotation de manière efficace et en amenant des justifications théoriques soutenant le choix du critère proposé. L'algorithme proposé s'appuie sur l'idée qu'il est nécessaire d'équilibrer les variances entre les c composantes sélectionnées. Notons $\Sigma_{V,t}$ la matrice de covariance estimée de v_t à l'instant t. La rotation R a pour rôle d'équilibrer la variance sur les composantes, ainsi chaque coefficient diagonal de $\Sigma_{Y,t}$ doit valoir $\tau = \frac{\text{Tr}(\Sigma_{V,t})}{c}$, où $\text{Tr}(.)$ est la trace de la matrice. Cette opération d'équilibrage des contributions des composantes s'appuie sur le produit de $c-1$ matrices de Givens $G_r(i_r, j_r, \theta_r)$ pour $r \in [c-1]$ et i_r, j_r, θ_r étant les paramètres déterminant une matrice de Givens.

Définition 1 : Matrice de rotation de Givens.

En dimension m, une matrice de rotation de Givens $G(i,j,\theta) \in \mathbb{R}^{m\times m}$ est une matrice de la forme

$$G(i,j,\theta) = \begin{bmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \dots & c & \dots & -s & \dots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \dots & s & \dots & c & \dots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{bmatrix}$$

où $c = \cos\theta$ et $s = \sin\theta$, c'est-à-dire qu'elle est égale à la matrice identité de dimension $m \times m$, sauf pour les entrées $[G(i,j,\theta)]_{i,i} = c$, $[G(i,j,\theta)]_{j,j} = c$, $[G(i,j,\theta)]_{i,j} = -s$ et $[G(i,j,\theta)]_{j,i} = s$, en prenant $i < j$.

L'algorithme UnifDiag est ainsi donné dans la Figure 10. Notons que la justification de l'approche a été donnée dans l'article [W2], contrairement aux approches analogues de l'état de l'art de l'époque, qui manquaient d'arguments théoriques.

Inputs : Σ_V ($c \times c$, symmetric), tolerance: tol
$R \leftarrow I_c$ // $c \times c$ Identity matrix; $\tau \leftarrow \mathrm{Tr}(\Sigma_V)/c$; $it = 0$
$iInf = \{ l \in \{1, ..., c\} \mid \Sigma_{V,l,l} < \tau - tol \}$
$iSup = \{ l \in \{1, ..., c\} \mid \Sigma_{V,l,l} > \tau + tol \}$
while $it < c - 1$ & not isEmpty($iInf$) & not isEmpty($iSup$) **do**
  // Givens rotation parameters computation:
  $j \leftarrow pop(iInf)$; $i \leftarrow \mathrm{pop}(iSup)$; $a \leftarrow \Sigma_V[j,j]$; $b \leftarrow \Sigma_V[i,j]$; $d \leftarrow \Sigma_V[i,i]$; c, s (Th. 3.1); $it \leftarrow it + 1$
  // Σ_V *update*:
  $row_j \leftarrow \Sigma_V[j,:]$; $row_i \leftarrow \Sigma_V[i,:]$
  $\Sigma_V[j,:] = c \times row_j - s \times row_i$;
  $\Sigma_V[i,:] = s \times row_j + c \times row_i$
  $\Sigma_V[:,j] = \Sigma_V[j,:]$; $\Sigma_V[:,i] = \Sigma_V[i,:]$
  $\Sigma_V[j,j] = a'$; $\Sigma_V[i,i] = d'$; $\Sigma_V[j,i] = b'$ (Th. 3.1)
  // Rotation update:
  $col_j \leftarrow R[:,j]$; $col_i \leftarrow R[:,i]$
  $R[:,j] = c \times col_j - s \times col_i$
  $R[:,i] = s \times col_j + c \times col_i$
  // Indices list update:
  if $\frac{a+d}{2} < \tau - tol$ **then**
    $\mathrm{add}(iInf, i)$
  if $\frac{a+d}{2} > \tau + tol$ **then**
    $\mathrm{add}(iSup, i)$
return R

Figure 10 : Algorithme permettant l'uniformisation des contributions sur les c composantes. Algorithme extrait de [C15].

3.2.3 Expériences

Nous avons évalué la méthode proposée sur un problème de recherche de plus proches voisins. Le code binaire de taille c est utilisé pour rechercher les plus proches voisins. La vérité terrain est donnée par le calcul de la distance euclidienne dans l'espace initial. On compare le résultat obtenu avec le résultat théorique en calculant l'indice MAP (Mean Average Precision). Les tests sont réalisés sur CIFAR-10[4]. CIFAR-10 (CIFAR) contient 60000 images colorées de taille 32 × 32, distribuées de manière égale en 10 classes. On utilise les descripteurs GIST de dimension 960. La Figure 11 présentent les résultats pour le dataset CIFAR pour différentes longueurs de code. La méthode proposée fonctionne au moins aussi

[4] http://www.cs.toronto.edu/~kriz/cifar.html

bien que l'état de l'art (et mieux dans plusieurs conditions) pour des méthodes de hashing binaire en streaming, tout en étant plus simple, justifiée théoriquement et avec une complexité temporelle plus faible.

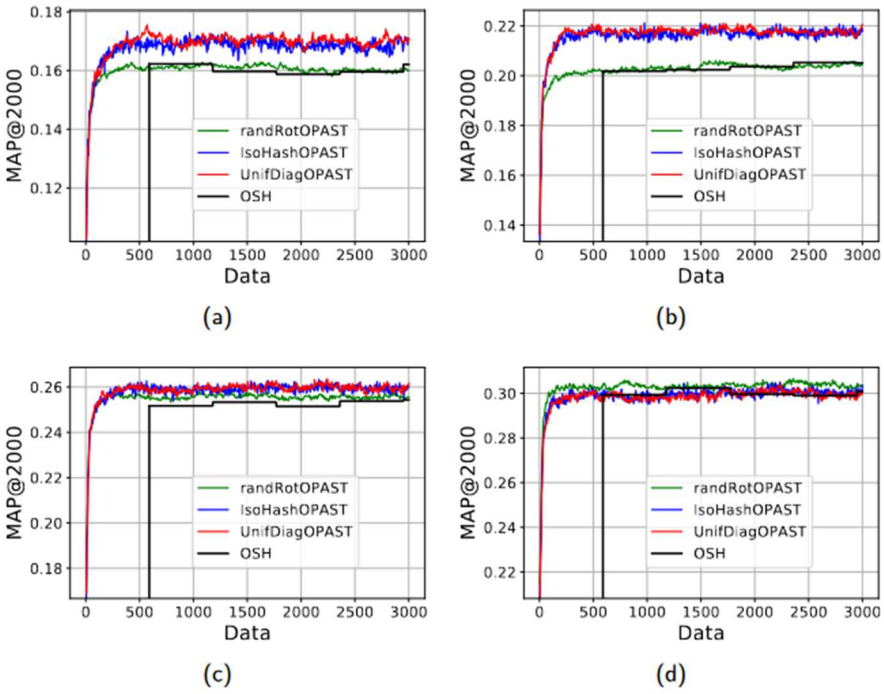


Figure 11 : MAP@2000 dans le cadre streaming pour différentes tailles de code et le dataset CIFAR : (a) c = 8, (b) c = 16, (c) c = 32, (d) c = 64.

3.3 Arbre recouvrant minimum approché pour le clustering

Ces travaux, réalisés dans le cadre de la thèse de Anne Morvan, ont été publiés dans [C16]. Des garanties théoriques de l'algorithme ont été obtenues dans [C14] par Anne Morvan et Rafaël Pinot (co-encadré avec Florian Yger et Jamal Atif).

3.3.1 Positionnement du problème

On suppose qu'on observe un graphe pondéré non-dirigé $\mathcal{G} = (V, E)$, où l'ensemble des nœuds est noté V , l'ensemble des arêtes est noté E et il existe une fonction $w : V \times V \to]0\,,1]$ donnant le poids de chacune des arêtes. On note $|V| = N$ et $|E| = M$. On suppose qu'il est possible de connaître le poids de chaque arête entre deux nœuds, caractérisant la distance qui les sépare. A partir d'un jeu de données représenté sous la forme d'un tel graphe, on souhaite former des groupes de nœuds (clusters). Le clustering à partir d'une

représentation graphique a fait l'objet de plusieurs travaux historiques, notamment DenGraph [28] ou encore des méthodes basés sur le modèle de « Stochastic Block Model » [29], [30], [31]. La Figure 12 présente le principe de clustering à partir de l'observation d'un graphe.

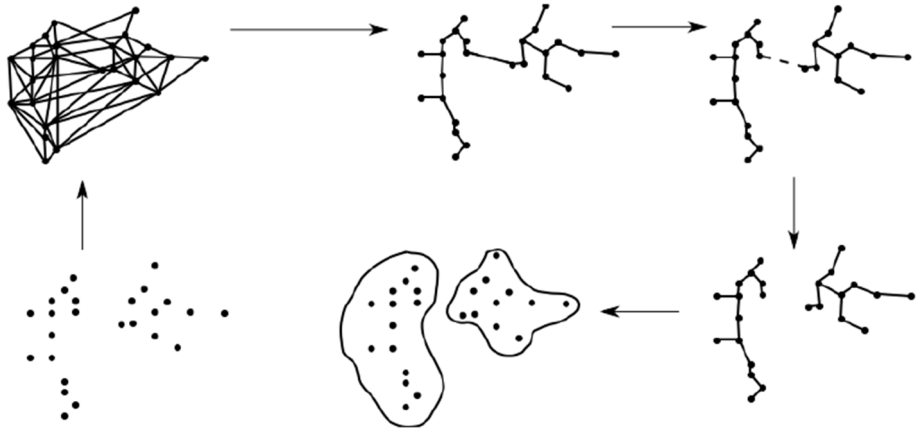


Figure 12 : Principe de clustering à partir d'un graphe. On observe les arêtes pondérées pour chaque paire de nœuds. Puis on calcule l'arbre recouvrant minimum. L'algorithme proposé dans cette partie permet à partir de l'arbre recouvrant minimum de regrouper les nœuds dans des clusters homogènes. La figure est extraite de la présentation de soutenance de Anne Morvan en novembre 2018.

3.3.2 Construction en streaming de l'arbre recouvrant minimum

Dans une première étape, l'arbre recouvrant minimum est construit selon l'approche proposée par [32]. Cette approche nécessite $\mathcal{O}(N\log^3(N))$ en complexité spatiale et une complexité temporelle de $\mathrm{polylog}(N)$ pour chaque mise à jour du graphe. Cette approche s'appuie sur une observation itérative des poids du graphe pour construire le l'arbre recouvrant minimum approché.

3.3.3 Clustering à partir de l'arbre recouvrant minimum

Dans l'article [C16], nous proposons une approche sans paramètre pour procéder à des coupes successives dans le graphe recouvrant minimum. L'algorithme démarre la procédure avec un seul cluster, contenant l'ensemble des nœuds. Une métrique est alors calculée à partir des notions de dispersion et de séparation afin de choisir l'arête à couper. La notion de dispersion d'un cluster C_i est

$$\forall i \in [K] \qquad DISP(C_i) = \begin{cases} \max(w_j)_{j\ ,\ e_j \in C_i} & si\ |E(C_i)| \neq 0 \\ 0 & sinon \end{cases}$$

Tandis que la notion de séparation d'un cluster C_i est définie comme la distance minimum entre les nœuds de C_i et ceux des autres clusters C_j. Notons que si le nombre de clusters est

de 1, alors la séparation est de 1. La séparation d'un cluster est notée $SEP(C_i)$. On définit alors l'indice de validité d'un cluster en fonction de la dispersion et de la séparation comme :

$$V_C(C_i) = \frac{SEP(C_i) - DISP(C_i)}{\max(SEP(C_i), DISP(C_i))}.$$

L'indice de validité d'un cluster est compris entre -1 et 1. Enfin l'indice de validité d'une partition en clusters Π est notée

$$DBCVI(\Pi) = \sum_{i=1}^{K} \frac{|C_i|}{N} V_C(C_i)$$

L'Algorithme 1 donne alors la procédure itérative permettant, à partir de l'arbre recouvrant minimum, de calculer une partition des nœuds en clusters.

Algorithme 1 : Algorithme de clustering DBMSTClu

```
1: Input: 𝒯, the MST
2: dbcvi ← −1.0; clusters = [ ]; cut_list ← [E(𝒯)]
3: while dbcvi < 1.0 do
4:    cut_tp ← None; dbcvi_tp ← dbcvi
5:    for each cut in cut_list do
6:       newDbcvi ← evaluateCut(𝒯, cut)
7:       if newDbcvi ≥ dbcvi_tp then
8:          cut_tp ← cut; dbcvi_tp ← newDbcvi
9:    if cut_tp ≠ None then
10:      clusters ← cut(clusters, cut_tp)
11:      dbcvi ← dbcvi_tp; remove(cut_list, cut_tp)
12:   else
13:      break
14: return clusters, dbcvi
```

Au-delà de la procédure de clustering donnée ci-dessus, plusieurs preuves théoriques ont été fournies pour cet algorithme. Nous invitons le lecteur intéressé à consulter [C14] pour plus de détails sur ces résultats théoriques.

3.3.4 Expérience

Plusieurs expériences ont été réalisées afin de montrer les bonnes performances pratiques de l'approche, les capacités de l'algorithme à fonctionner avec en entrée un arbre recouvrant minimum approché, et enfin sa capacité de passage à l'échelle. L'approche a été comparée à :

- SEMST [33], [34] : méthode qui effectue $K - 1$ coupes parmi les arêtes de poids le plus élevé dans l'arbre recouvrant minimum,
- DBSCAN [35], une méthode classique de clustering.

Les expériences ont montré une bonne qualité du regroupement des points et une capacité à détecter des anomalies ponctuelles, voir Figure 13 et Figure 14.

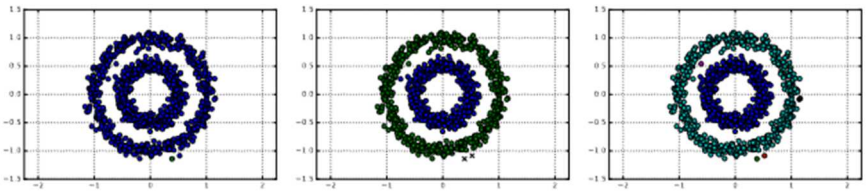


Figure 13 : Cercles bruités : SEMST, DBSCAN ($\epsilon = 0.15$, nombre de points minimum de 5), DBMSTClu avec un arbre couvrant minimum approché.

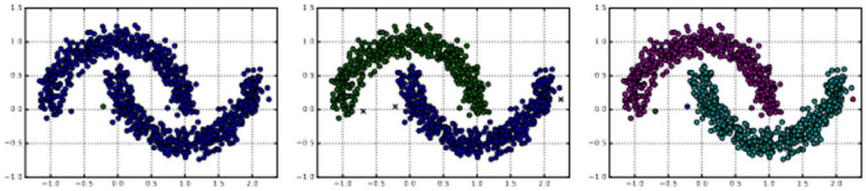


Figure 14 : Bananes bruitées : SEMST, DBSCAN ($\epsilon = 0.15$, nombre de points minimum de 5), DBMSTClu avec un arbre couvrant minimum approché.

La capacité de passage à l'échelle est démontrée dans la figure ci-dessous. Des expériences ont été conduites à partir d'un modèle stochastique par blocs. K clusters ont été créés pour un nombre de nœuds donné N. Les temps de calcul associés pour retrouver exactement les K clusters sont donnés dans la Figure 15.

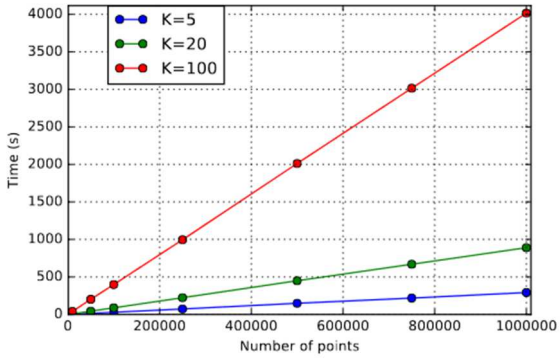


Figure 15 : Temps d'exécution de DBMSTClu avec $N \in \{1K, 10K, 50K, 100K, 250K, 500K, 750K, 1M\}$.

3.4 Applications dans le domaine de la cybersécurité

Nous décrivons à présent comment les approches d'apprentissage statistique présentées dans les parties précédentes ont été utilisées dans un contexte de cybersécurité, initialement dans un projet interne CEA en cours depuis 2017. Certains aspects de ces travaux ont été publiés dans [C10]. Ces travaux ont été réalisés en collaboration avec les collègues de CEA

Tech en région Toulousaine, notamment Radhouene Azzabi, Hubert Dubois et Philippe Limousin.

3.4.1 Contexte

Les menaces cyber sont aujourd'hui au cœur de nombreuses préoccupations. Devant une sophistication sans précédent des techniques des attaquants, l'apprentissage statistique apparaît comme une technique de défense prometteuse. Néanmoins, la diversité des attaques cyber réelles doit conduire à la plus grande humilité devant l'ampleur des défis. Depuis 2017, nous avons travaillé en interne CEA pour comprendre le mode d'opération des experts cyber et voir comment les assister au mieux. Dans un premier temps nous avons identifié un besoin d'appréhender des données extrêmement volumineuses en un temps minimum. Aujourd'hui de nombreuses règles données par des opérateurs régaliens permettent d'alimenter des règles de détection d'alertes, qui conduisent à un nombre d'alertes important face aux capacités d'un expert cyber à investiguer en détails le déroulé des événements. Fort heureusement la plupart des alertes sont de fausses alertes, mais un certain nombre nécessite une investigation poussée afin de comprendre l'origine de l'alerte, isoler l'intervalle temporel des événements suspicieux et identifier le périmètre des machines potentiellement concernées. Ces analyses se font grâce à des données très volumineuses provenant de bases de données hétérogènes.

Afin d'assister l'expert cyber dans ses tâches, nous travaillons sur un système combinant la visualisation de données avec l'apprentissage statistique afin de permettre le filtrage, le regroupement et l'identification. Le type de graphe manipulé par l'expert est représenté sur la Figure 16. Selon l'objectif visé, les significations des nœuds et des arêtes peuvent être adaptées (adresses IP, noms de domaine).

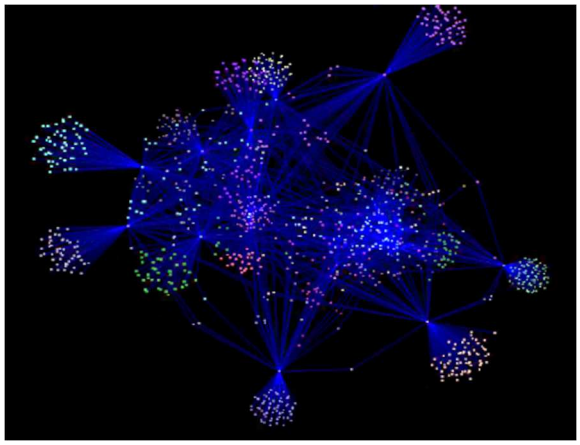


Figure 16 : Graphe représentant des données cyber dans un espace 3D, permettant à l'analyste d'explorer le contexte d'une alerte.

3.4.2 Architecture logicielle

La mise en place d'outils dans le domaine de la cybersécurité nécessite de développer des produits logiciels en adéquation avec les usages actuels des experts cyber. Afin d'intégrer des possibilités de traitement à l'interface entre visualisation et apprentissage statistique, une plateforme a été développée par Radhouene Azzabi à Toulouse (ADViz). Nos algorithmes d'apprentissage statistique interagissent grâce à des API REST avec les bases de données du système, afin de produire des données à visualiser ou des informations en superposition des données. Ces résultats de traitement sont envoyés au module de visualisation en gardant l'objectif d'une interaction fluide avec l'expert cyber.

3.4.3 Interactions entre apprentissage statistique et visualisation

Différents algorithmes d'apprentissage automatique sont disponibles dans les outils pour permettre une exploration interactive efficace. Fonctionnellement, l'objectif de ces algorithmes est de trois types :

- Modélisation et résumé du comportement : il s'agit de développer des approches pour décomposer l'historique agrégé des activités d'un ordinateur en différentes classes, en utilisant des graphiques pour modéliser les événements et ainsi résumer et caractériser les comportements.
- Réduction de dimensionnalité : cette technique est utilisée pour gérer les espaces de haute dimension générés par l'analyse de diverses activités, comme les demandes de noms de domaine par les ordinateurs d'un réseau interne. Selon les contraintes de latence de calcul et les dimensions manipulées, nous avons utilisé le hachage de caractéristiques et les projections aléatoires structurées pour permettre le calcul efficace de proximité entre éléments.
- Recherche de similarité, requêtes top-k et clustering : une fois que les utilisateurs et les événements ont été correctement décrits et résumés, il est crucial de permettre à l'opérateur d'effectuer diverses actions de comparaison, telles que la recherche de similarité pour trouver des entités ou des événements similaires à un objet déjà identifié, ou le clustering pour regrouper utilisateurs ou événements en plusieurs groupes. Pour cela, des techniques de clustering (k-means et clustering hiérarchique), utilisant des distances de Jaccard et Euclidiennes, ont été implémentées.

Ces approches se sont appuyées sur les travaux de la partie précédente pour permettre des approches efficaces sur des données en grande dimension. Les modes d'interaction sont résumés dans la Figure 17.

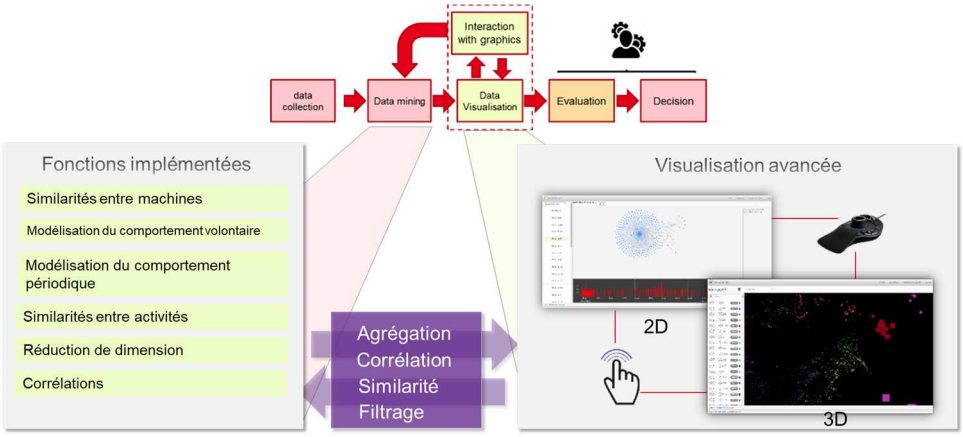


Figure 17 : Interactions entre visualisation et apprentissage statistique

3.5 Conclusion et impact sur les axes de recherche

Ces travaux ont permis d'explorer en détails les compromis entre le coût spatial d'une structure de données, le débit de traitement des données et la précision du résultat de l'algorithme. Différents algorithmes ont été proposés pour les tâches de recherche de plus proches voisins et de clustering, avec des applications dans de nombreux domaines. Nous avons en particulier illustré les premières exploitations des algorithmes conçus dans le domaine de la cybersécurité. Ces travaux ont permis d'initier plusieurs pistes de recherche qui sont actuellement encore en cours d'exploration, menées par différents collègues du laboratoire. Sandra Garcia Rodriguez a travaillé sur l'exploitation de Streamer[5] pour rejouer et traiter des données cyber historiques avec des algorithmes de streaming. Gabriel Rilling et Rémi Tschupp, dans le cadre du projet européen FED AINCEPTION, ont proposé une technique de classification d'anomalies cyber à partir d'une représentation graphique des activités. Cette approche a été prise en main et adaptée dans le cadre des projets européen STARLIGHT et KINAITICS pour des données provenant des forces de maintien de l'ordre et l'extension à des réseaux hétérogènes (IoT et serveurs). Dans le cadre du projet PEPR Cyber/SUPERVIZ, Gabriel Rilling s'intéresse, avec Yu-Fei Han, à la robustesse des

[5] https://streamer-framework.github.io. Streamer est une plateforme, financée en partie dans le cadre du projet StreamOPS (DataIA), pour la mise à disposition d'algorithmes de traitement de données en streaming à des experts métiers. Dans le cadre de StreamOPS, nous avons travaillé avec l'Université Versailles Saint-Quentin (équipe DAVID) et l'hôpital Foch pour l'exploitation dans le cadre de données de santé (dispositifs de santé connectés). Dans le cadre d'un projet financé en interne, nous avons abordé l'utilisation de streamer pour le traitement de données de cybersécurité.

algorithmes de détection basés sur l'apprentissage statistique face à des attaques cyber adverses. Enfin le laboratoire est impliqué dans la standardisation des messages décrivant des événements de détection en cybersécurité, dans le cadre du projet européen SAFE4SOC.

D'un point de vue des approches d'apprentissage statistique, les applications mentionnées ont fait apparaître de nouvelles propriétés désirées, en particulier pour le respect de la confidentialité des données et la sécurité des algorithmes. Les compromis en résultant font l'objet des parties suivantes.

4 Confidentialité et sécurité en contexte centralisé

Dans cette partie, nous synthétisons les travaux réalisés sur des propriétés particulières d'algorithmes d'apprentissage statistique relatives au respect de la vie privée et la sécurité face à des participants malveillants.

La protection des données personnelles contre de potentielles fuites prend ses sources dans les années 80 [36], [37], [38]. Cependant c'est en 2008 que ce sujet a attiré à nouveau l'attention lorsque Narayanan et Shmatikov [39] ont présenté une approche pour retrouver des informations personnelles à partir de données anonymes publiées pour la compétition « Netflix Prize » qui se voulait une compétition dédiée aux algorithmes de recommandation. Mentionnons par ailleurs que l'Union Européenne s'est emparée du sujet en adoptant en 2016 le Règlement Général sur la Protection des Données (RGPD) [40]. Ce règlement définit les obligations légales associées au stockage et au traitement de données personnelles. Dans la mesure où cette réglementation s'applique sur l'ensemble du cycle de vie de la donnée, elle impacte l'entraînement et l'utilisation des modèles entrainés par apprentissage statistique. Dans ce second cas, il est important de noter qu'un modèle peut avoir mémorisé des données personnelles, qui sont alors exposées au risque de fuite pour un attaquant disposant uniquement du modèle statistique destiné à la phase d'inférence. Ainsi plusieurs définitions se sont attachées à caractériser la notion de protection des données et de confidentialité dans le contexte de l'apprentissage statistique [41]. Parmi ces définitions, la confidentialité différentielle [42] est aujourd'hui le standard le plus répandu. Dans cette approche, on considère qu'un algorithme offre des garanties de confidentialité vis-à-vis d'un traitement statistique lorsque le résultat de ce traitement ne peut être distingué statistiquement pour deux bases de données similaires $\mathcal{D}$ et $\mathcal{D}'$, pour lesquelles elles ne diffèrent que par la présence ou l'absence d'un individu (si on souhaite protéger la confidentialité de chacun des individus). Nous reviendrons sur la définition mathématique de cette notion dans le cas de graphes. Le lecteur intéressé pourra se référer au texte de référence [43] pour des définitions formelles dans le cas général.

La sécurité d'un modèle appris par apprentissage statistique englobe beaucoup d'autres propriétés, et est en conséquence plus complexe à définir. D'après [44], sur lequel nous reviendrons plus en détail dans la dernière partie de ce manuscrit, la sécurité de l'apprentissage statistique doit englober l'ensemble des menaces auxquelles peut être confronté le cycle de vie d'un modèle, depuis l'acquisition des données jusqu'au monitoring de son fonctionnement suite à sa mise en production. Dans ce chapitre nous nous focaliserons sur la protection de la confidentialité pour le cas particulier des poids des arêtes

d'un arbres dans une tâche de clustering puis nous aborderons les attaques par évasion, qui ont lieu lors de l'exploitation d'un modèle, lorsqu'un attaquant introduit volontairement de subtiles perturbations sur les entrées du système, afin de provoquer des modifications substantielles de la sortie du modèle.

4.1 Clustering et confidentialité différentielle

Ces travaux résultent de la collaboration entre Anne Morvan et Rafaël Pinot, doctorants encadrés avec Jamal Atif et Florian Yger. Ils ont été publiés dans [C14].

Nous avons vu dans la partie précédente un algorithme de clustering basé sur l'arbre recouvrant minimum. Les travaux dans le cadre de la thèse de Anne Morvan ont permis de travailler sur le compromis entre l'empreinte mémoire d'un algorithme et sa capacité à fournir un clustering à partir de l'arbre minimum recouvrant. Notre première contribution sur la préservation de la confidentialité en apprentissage statistique concerne l'introduction d'un mécanisme permettant de préserver la confidentialité différentielle dans cet algorithme de clustering. Ce travail a été publié dans [C14].

Arbre minimum recouvrant approché sous contrainte de confidentialité différentielle

Soit $\mathcal{G} = (V, E, w)$ un graphe pondéré non-dirigé composé d'un ensemble de nœuds V, un ensemble d'arêtes E, et une fonction de poids qui associe à chaque arête un poids $w : E \longrightarrow \mathbb{R}$. Afin de mettre en place la confidentialité différentielle, nous avons besoin de définir le type d'information à protéger. Ces travaux se focalisent sur la protection de l'information portée par les poids des arêtes. Ainsi on définit mathématiquement que pour tout ensemble d'arêtes E, les fonctions de poids w et w' sont voisines si $\|w - w'\|_\infty \coloneqq \max_{e \in E} |w(e) - w'(e)| \leq \mu$. On note alors $w \sim w'$. Dans cette définition, le paramètre μ représente un paramètre de sensibilité de la fonction de poids, elle permet d'adapter la notion de voisinage selon les applications visées. A partir de cette définition, on considère la notion de voisinage entre deux graphes :

Définition 2 : Graphes voisins

Deux graphes pondérés $\mathcal{G} = (V, E, w)$ et $\mathcal{G}' = (V', E', w')$ sont dits voisins si $V = V'$, $E = E'$ et $w \sim w'$.

Pendant son stage de Master 2, Rafaël Pinot a proposé un algorithme permettant, à partir d'un graphe pondéré non-dirigé, de fournir un arbre recouvrant quasi-minimal

respectant les contraintes de confidentialité différentielle. Cet algorithme, appelé PAMST, est donné ci-dessous dans Algorithme 2, d'après [45].

Algorithme 2 : Arbre recouvrant minimal approché respectant les contraintes de confidentialité PAMST($\mathcal{G}, u_{\mathcal{G}}, w, \epsilon$).

Require: A graph topology $\mathcal{G} = (V, E)$, a weight function w, a degree of privacy ϵ, utility function $u_{\mathcal{G}}$.
Ensure: The topology of an approximated minimum spanning tree represented by S_E a set of edges.

Pick $v \in V$ arbitrarily
$S_V \leftarrow \{v\}$
$S_E \leftarrow \varnothing$
while $S_V \neq V$ **do**
  $r = \mathcal{M}_{Exp}(\mathcal{G}, w, u_{\mathcal{G}}, \mathcal{R}_{S_V}, \frac{\epsilon}{|V|-1})$
  $S_V \leftarrow S_V \cup \{r_{\rightarrow}\}$
  $S_E \leftarrow S_E \cup \{r\}$
end while
return S_E

Clustering sous contrainte de confidentialité différentielle

En combinant PAMST et l'algorithme DBMSTClu de la partie précédente, un algorithme privé de clustering est alors proposé dans Algorithme 3.

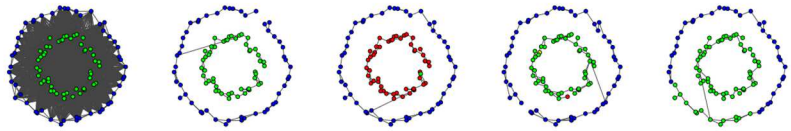
Algorithme 3 : PTClust($\mathcal{G}, w, u_{\mathcal{G}}, \epsilon, \tau, p$)

1: **Input:** $\mathcal{G} = (V, E, w)$ a weighted graph (separately the topology G and the weight function w), ϵ a degree of privacy and u_G utility function.
2: $T = \text{PAMST}(G, w, u_G, \epsilon/2)$
3: $\mathcal{T}' = \mathcal{M}_{w.r}(T, w_{|E(T)}, \frac{2\mu}{\epsilon}, \tau, p)$
4: **return** DBMSTCLU($\mathcal{T}'$)

Nous avons alors montré dans [C14] que l'algorithme PTClust respecte la confidentialité différentielle de paramètre ϵ sur le graphe G. En compléments, des théorèmes importants ont été fournis pour évaluer les compromis entre la confidentialité et la précision pour cet algorithme [C14, section 3.4].

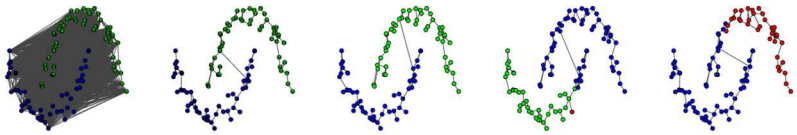
Expériences

Plusieurs expériences ont été réalisées afin de montrer la capacité de l'algorithme à fournir un clustering satisfaisant sous contraintes de confidentialité différentielle.



(a) Homogeneous graph (b) DBMSTCLU (c) PTCLUST, $\epsilon = 1.0$ (d) PTCLUST, $\epsilon = 0.7$ (e) PTCLUST, $\epsilon = 0.5$

Figure 18 : Expériences de clustering pour des données organisées sur des cercles avec $n = 100$. Les paramètres de PTClust sont : $w_{\min} = 0.1$, $w_{\max} = 0.3$, $\mu = 0.1$.



(a) Homogeneous graph (b) DBMSTCLU (c) PTCLUST, $\epsilon = 1.0$ (d) PTCLUST, $\epsilon = 0.7$ (e) PTCLUST, $\epsilon = 0.5$

Figure 19 : Expériences de clustering pour des données organisées sur des arcs avec $n = 100$. Les paramètres de PTClust sont : $w_{\min} = 0.1$, $w_{\max} = 0.3$, $\mu = 0.1$.

4.2 Confidentialité différentielle et robustesse aux attaques adverses

Ces travaux, relatifs au lien entre la confidentialité différentielle et la robustesse aux attaques adverses, réalisés dans le cadre de la thèse de Rafaël Pinot, ont été publiés dans [W1]. La présentation des attaques adverses est plus détaillée dans [J2] et [C11].

4.2.1 Confidentialité différentielle et espaces métriques

La confidentialité différentielle, dans sa forme la plus simple, s'interprète comme la capacité à rendre statistiquement indistinguable des résultats d'un traitement algorithmique pour des données d'entrée voisine. Par exemple dans le cas de bases de données il s'agit de s'assurer que le résultat d'un traitement statistique ne sera pas statistiquement distinguable selon qu'on utilise la base de données initiale ou alors la base de données dans laquelle une seule ligne a été supprimée. Cette intuition, formalisée dans un premier temps par [42], a par la suite fait l'objet de reformulations permettant de préciser les notions sous-jacentes de proximité entre les entrées, les sorties, et le mécanisme aléatoire chargé de rendre certains éléments indistinguables. Dans l'article [W1], nous avons adopté une formulation générique capable d'englober les définitions plus classiques. On considère ainsi une tâche quelconque, impliquant deux espaces métriques arbitraires $(\mathcal{X}, d_{\mathcal{X}})$ et $(\mathcal{Y}, d_{\mathcal{Y}})$. Soit $\sigma(\mathcal{Y})$ une σ-algèbre sur $\mathcal{Y}$. On note $\mathcal{P}(\mathcal{Y})$ l'ensemble des mesures de probabilités sur $(\mathcal{Y}, \sigma(\mathcal{Y}))$. On considère la divergence de Rényi (aussi appelée I-divergence de Rényi dans [46]) entre deux mesures de

probabilités $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{Y})$, toutes deux dominées par une mesure de référence[6] ν, définie par :

$$D_\lambda(\mu_1, \mu_2) = \frac{-1}{1-\lambda} \log \int_y g_2(y) \left(\frac{g_1(y)}{g_2(y)}\right)^\lambda d\nu(y)$$

où g_1 et g_2 sont deux densités de probabilités de μ_1 et μ_2, par rapport à ν. La divergence de Rényi est définie sur $(1, +\infty)$. Elle a été choisie dans les travaux de Rafaël Pinot car elle permet d'englober plusieurs autres divergences classiques (nous y reviendrons dans la partie sur les attaques adverses), notamment la divergence de Kullback-Leibler D_{KL} quand λ tend vers 1 et la maximum divergence D_∞ quand λ tend vers l'infini. La confidentialité différentielle consiste à introduire suffisamment de bruit dans une opération de traitement pour que l'information individuelle dans les données d'entraînement puisse ne pas être retrouvée. Ainsi nous allons définir cette opération dans la Définition 3, qui va nous conduire à la définition de la confidentialité différentielle au sens de Rényi dans la Définition 4.

Définition 3 : Mécanisme randomisé

Un mécanisme randomisé $\mathcal{M}$ de $\mathcal{X}$ vers $\mathcal{Y}$ est une application $\mathcal{M} : \mathcal{X} \longrightarrow \mathcal{P}(\mathcal{Y})$. Etant donnée x dans l'espace d'entrées $\mathcal{X}$, $\mathcal{M}$ fournit une mesure de probabilité $\mathcal{M}(x)$. Afin d'obtenir une quantité numérique y dans l'espace des sorties $\mathcal{Y}$, il est nécessaire d'échantillonner selon $y \sim \mathcal{M}(x)$.

Définition 4 : Confidentialité différentielle au sens de Rényi [48]

Soit $\epsilon > 0$, $(\mathcal{X}, d_\mathcal{X})$ un espace arbitraire d'entrées et $\mathcal{Y}$ l'espace de sortie. Un mécanisme randomisé $\mathcal{M}$ de $\mathcal{X}$ vers $\mathcal{Y}$ garantit la $(\lambda, \epsilon, \alpha)$-$d_\mathcal{X}$ confidentialité différentielle au sens de Rényi si pour tout x, x' tels que $d_\mathcal{X}(x, x') < \alpha$, alors on a $D_\lambda\big(\mathcal{M}(x), \mathcal{M}(x')\big) < \epsilon$.

Nous avons détaillé dans [W1] en quoi cette définition permet de retrouver la définition classique de [43] et la définition s'appuyant sur des espaces métriques donnée dans [49].

4.2.2 Classification en présence d'un adversaire

Dans le cadre standard de la classification supervisée, les espaces $\mathcal{X}$ et $\mathcal{Y}$ sont supposés être liés par une distribution de probabilités $\mathcal{D}$. Comme introduit dans le premier chapitre, le problème de la classification supervisée consiste alors à construire une hypothèse h (modèle de classification) qui permet de décrire le lien entre ces deux espaces. On peut par

[6] Cette mesure de référence est nécessaire en théorie de la mesure, voir [47].

exemple considérer la tâche de classification, présentée en Figure 20, qui associe un ensemble d'images $\mathcal{X}$ (encodées dans $\mathbb{R}^d$) à des étiquettes $\mathcal{Y}$ correspondants à ces images (encodés dans $\{-1, 1\}$).

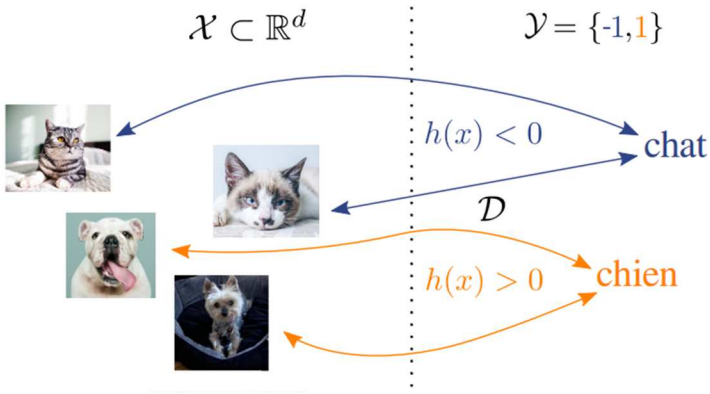


Figure 20 : Illustration d'un problème de classification d'images de chiens et de chats. Les chiens sont encodés par le label 1 et les chats par le label −1. Droits : Rafaël Pinot.

L'objectif d'un algorithme d'apprentissage supervisé est donc de construire une hypothèse $h : \mathcal{X} \longrightarrow \mathbb{R}$ qui associe à toute image x une valeur positive si le label de x est 1 et négative sinon. Pour ce faire, il cherche à sélectionner h dans un espace fonctionnel $\mathcal{H}$ (aussi appelé classe d'hypothèses) de manière à minimiser le risque de h, c'est-à-dire la probabilité que h ne réussisse pas à faire correspondre une image avec son label. Le risque d'une hypothèse h est défini comme ceci[7] :

$$\mathcal{R}(h) = \mathbb{E}_{x,y\sim\mathcal{D}}\big(\ell(h(x), y)\big)$$

où $\ell(.)$ est une fonction de coût qui compte le nombre de fois où h échoue à faire correspondre x et y. En pratique l'algorithme d'apprentissage n'a pas accès à la distribution des données $\mathcal{D}$ mais seulement à un jeu de données de taille n. L'algorithme d'entraînement va ainsi minimiser le risque empirique[8] :

$$\hat{\mathcal{R}}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)$$

Notons que pour évaluer la distance entre l'hypothèse sélectionnée par minimisation du risque empirique et l'hypothèse optimale, on cherche à majorer la différence entre le risque et le risque empirique pour n'importe quelle hypothèse de notre classe $h \in \mathcal{H}$. Cette différence est appelée l'écart de généralisation. Au cours des 20 dernières années, un

[7] Tandis que dans l'introduction, nous avons limité la présentation au risque empirique, nous présentons dans cette partie la formulation théorique dans un premier temps. On suppose donc l'existence d'une distribution jointe $(x, y) \sim \mathcal{D}$ qui gouverne les liens entre l'espace des entrées $\mathcal{X}$ et la classe de l'image.

[8] L'introduction présente la version régularisée du risque empirique. On retrouve bien l'expression de l'introduction en prenant $\forall h \in \mathcal{H},\ \Omega(\mathrm{h}) = 0$.

ensemble de classes d'hypothèses particulières a montré une grande flexibilité et une grande efficacité dans des tâches telles que la classification d'images : les réseaux de neurones. Par exemple on peut représenter une classe de réseaux de neurones à N couches selon la Figure 21.

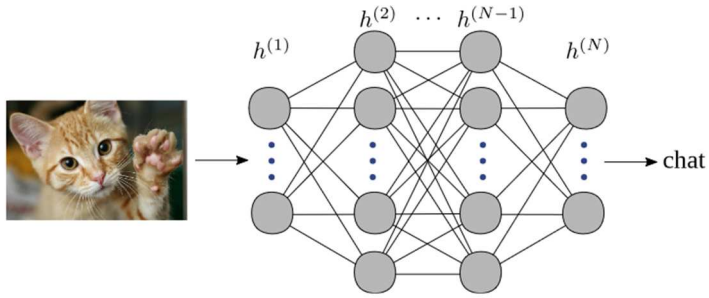


Figure 21 : Illustration d'un réseau de neurones à N couches ($h(i)$). Chaque noeud est une composition d'une application linéaire simple et d'une fonction non linéaire. Image construite par R. Pinot.

Étant donnée une hypothèse $h \in H$ et une image accompagnée de sa classe (x, y), le but d'un adversaire est d'introduire une perturbation τ telle que (ces deux points de définition sont repris d'après [4]) :

- La perturbation doit être imperceptible. Cela signifie qu'un humain ne peut pas distinguer visuellement l'image standard x de l'image adverse $x + \tau$.
- La perturbation modifie suffisamment x pour que l'hypothèse se trompe. Plus formellement, l'adversaire recherche une perturbation τ telle que $h(x + \tau)$ et y soient de signe différent.

Comme évoqué dans [4], la notion d'imperceptibilité humaine est difficile à formaliser et mesurer. Elle dépend également des individus. En s'appuyant sur les outils mathématiques accessibles et facilement manipulables, on peut néanmoins dire qu'une condition suffisante pour qu'une attaque ne soit pas détectée est de la contraindre en norme. On formalise cette notion par l'introduction d'un seuil $\alpha > 0$ pour lequel une perturbation τ est considérée comme imperceptible dès lors que $\|\tau\| \leq \alpha$. En pratique pour obtenir une perturbation de petite norme, il suffit de multiplier la perturbation choisie par une très petite constante, comme illustrée en Figure 22.

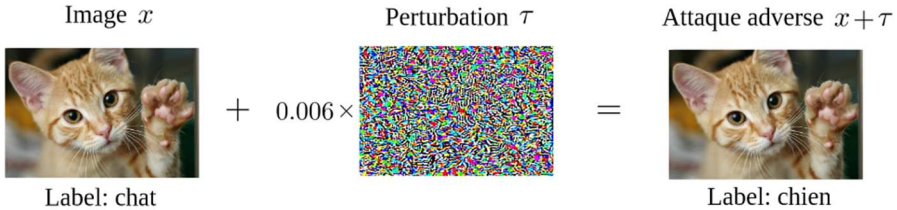


Figure 22 : Illustration d'un exemple adverse. À gauche, une image de chat classique qu'un réseau de neurones état-de-l'art associe facilement à la classe « chat ». Au milieu, une perturbation multipliée par une très

petite constante pour avoir une norme faible. À droite l'image adverse qui est identique à l'œil nu mais que le réseau n'est plus capable de classifier correctement. Image fournie par Rafaël Pinot.

Depuis les travaux initiaux de [50], [51], de nombreuses techniques d'attaque ont été développées pour perturber des réseaux de neurones. En général, un système dont la conception n'a pas prévu l'éventualité d'une telle attaque (système non défendu) aura 99% de chance de se tromper lorsqu'il tente de classifier une image adverse. Le problème de génération d'un exemple adverse revient à résoudre, en notant $\tau \in \mathcal{X}$ la perturbation recherchée :

$$\min d_{\mathcal{X}}(x, x+\tau) \text{ avec } h(x+\tau) \neq h(x)$$

On appelle risque adverse la quantité :

$$\mathcal{R}_{\text{adv}}(h\,;\alpha) = \mathbb{E}_{x,y\sim\mathcal{D}}\big(\max_{\tau\in B(\alpha)} \ell(h(x+\tau), y)\big)$$

où $B(\alpha) := \{\tau \in \mathcal{X} \quad \text{tel que} \quad \|\tau\| \leq \alpha\}$.

4.2.3 Lien entre confidentialité différentielle et robustesse aux attaques adverses

Une contribution importante de [W1] a été de donner une définition claire de la robustesse aux attaques adverses, et surtout de faire le lien avec le formalisme de la confidentialité différentielle. En effet en repartant de l'expression précédente, on peut définir le risque de changement de prédiction comme $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[\exists x' \in B(x,\alpha) \text{ tel que } h(x') \neq h(x)\,]$, où $B(x,\alpha) = \{x' \in \mathcal{X} \mid d_{\mathcal{X}}(x,x') < \alpha\}$ est la boule de rayon α centrée en x, et $\mathcal{D}_{\mathcal{X}}$ est la distribution marginale de $\mathcal{D}$ par rapport à $\mathcal{X}$. Ainsi une définition naturelle de la robustesse aux attaques adverses est

Définition 5 : Robustesse d'un classifieur.

Un classifieur h est (α,γ)-robuste si $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[\exists x' \in B(x,\alpha) \text{ tel que } h(x') \neq h(x)\,] < \gamma$.

Si bien qu'en réutilisant la notion de mécanisme randomisé de la Définition 3, on peut donner une définition plus générale de la robustesse adverse.

Définition 6 : Robustesse adverse généralisée

Soit $D_{\mathcal{P}(\mathcal{Y})}$ une divergence sur $\mathcal{P}(\mathcal{Y})$. Un classifieur randomisé $\mathcal{M}$ est $D_{\mathcal{P}(\mathcal{Y})}$-$(\alpha, \epsilon, \gamma)$-robuste si $\mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}\left[\exists x' \in B(x, \alpha) \text{ tel que } D_{\mathcal{P}(\mathcal{Y})}\big(\mathcal{M}(x') \neq \mathcal{M}(x)\big) > \epsilon\right] < \gamma$.

Ainsi d'après les définitions de la confidentialité différentielle au sens de Rényi et de la robustesse adverse généralisée, il y a une équivalence stricte entre ces deux notions (en choisissant la divergence de Rényi pour la robustesse adverse, choix qui est justifié dans [C11], [J2]).

4.3 Classifieurs aléatoires contre les attaques adverses

Ces travaux, réalisés dans le cadre de la thèse de Rafaël Pinot, ont été publiés dans [J2] et [C11].

Ce formalisme a donné les bases à la proposition de l'utilisation d'algorithmes randomisés pour obtenir des propriétés théoriques sur les hypothèses randomisées, et proposer une nouvelle approche simple de randomisation pour défendre les réseaux de neurones. On peut interpréter la Définition 6 comme la volonté d'imposer à h d'être localement Lipschitz par rapport à une norme $\|.\|$ sur l'espace des entrées, et par rapport à une métrique de probabilité sur l'espace des sorties (e.g. la divergence de Rényi). On note $\mathcal{H}_{D_{\mathcal{P}(\mathcal{Y})}}(\alpha, \epsilon)$ la classe des hypothèses qui respectent la condition :

$$\forall x, x' \in \mathcal{X} \quad \|x - x'\| < \alpha \implies D_{\mathcal{P}(\mathcal{Y})}\big(h(x), h(x')\big) < \epsilon \ .$$

Deux résultats principaux, exposés ci-dessous, ont été fournis dans [C11], [J2].

4.3.1 Résultat théorique : risque classique et risque adverse

Lorsque $\mathcal{D}$ est soit la distance de variation totale, soit la divergence de Renyi, nous montrons que pour toute hypothèse $h \in \mathcal{H}_{\mathcal{D}}(\alpha, \epsilon)$, il est possible de contrôler l'écart entre le risque classique et le risque adverse de h. Notamment, si $\mathcal{D}$ est la distance de variation totale, pour toute $h \in \mathcal{H}_{\mathcal{D}}(\alpha, \epsilon)$ nous avons

$$\mathcal{R}_{\text{adv}}(h\,; \alpha) - \mathcal{R}(h) \leq \epsilon$$

Par conséquent, ϵ contrôle le compromis maximal entre les erreurs robuste et standard pour une hypothèse aléatoire donnée. Cela signifie que, pour la classe d'hypothèses aléatoires $\mathcal{H}_{\mathcal{D}}(\alpha, \epsilon)$, la solution du problème de minimisation du risque, donne une solution approchée pour le problème de minimisation du risque adverse (à un facteur ϵ près). Nous démontrons

également qu'il est aussi possible, sous certaines hypothèses, de contrôler l'écart de généralisation de toute hypothèse $h \in \mathcal{H}_{\mathcal{D}}(\alpha, \epsilon)$.

4.3.2 Méthode simple et efficace basée sur l'injection de bruit

En s'appuyant sur les fondements théoriques partagés entre la confidentialité différentielle et la robustesse aux attaques adverses, nous avons construit une méthode simple de défense contre les attaques adverses, illustrée en Figure 23.

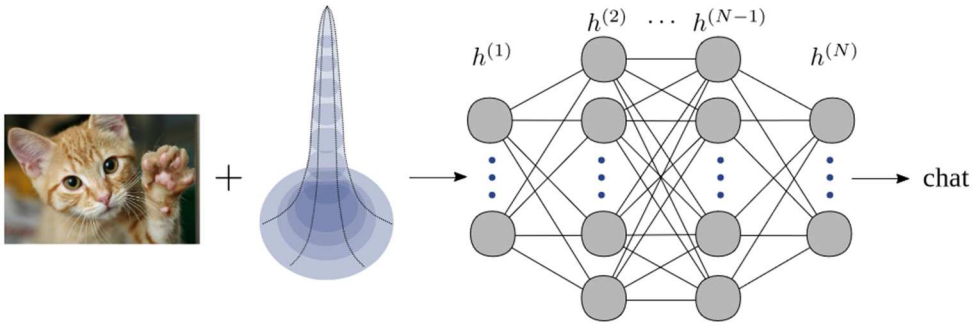


Figure 23 : Illustration d'une méthode simple de défense contre les attaques adverses. Avant de demander au modèle de classifier l'image, on ajoute un bruit sur celle-ci. Image fournie par R. Pinot.

L'idée est simplement d'injecter un bruit aléatoire (tiré d'une distribution Gaussienne) sur chaque image avant de demander au modèle de lui associer un label. Ainsi, les perturbations adverses sont noyées dans la masse des changements aléatoires, les rendant moins efficaces en moyenne. Nous démontrons que tous les modèles construits avec ce type d'injection de bruit font partie des classes d'hypothèses $\mathcal{H}_{\mathcal{D}}(\alpha, \epsilon)$ que nous avons précédemment étudiées. Cela nous permet de démontrer que cette méthode simple garantit une protection minimale pour un grand nombre de modèles d'IA, quelle que soit l'attaque contre laquelle ils doivent se défendre. Nous illustrons certains des résultats obtenus par cette défense en Figure 24, pour le jeu de donnée CIFAR10. Cette figure illustre la garantie théorique de la précision sous attaque de notre méthode pour différents niveaux de bruit. Cette fois-ci, le compromis entre la précision et la robustesse apparaît clairement en fonction de l'intensité du bruit (σ). Avec de petits bruits (courbes jaunes à oranges), la précision du modèle est élevée, mais la précision garantie chute rapidement en fonction de la taille de la perturbation adverse (α). Inversement, avec des bruits plus importants (courbes violette à bleu), la précision standard est plus faible, mais diminue lentement en fonction de l'ampleur de la perturbation adverse. Dans l'ensemble, nous obtenons de solides garanties de précision contre les petites perturbations adverses, mais lorsque la perturbation est trop grande les garanties ne sont toujours pas suffisantes. Notons toutefois que le modèle non défendu (en

vert) n'offre aucune garantie théorique. Cela montre donc que notre méthode d'injection de bruit augmente considérablement la robustesse du modèle, même pour des petites valeurs de bruit. La Figure 24 montre également que pour un choix adapté de bruit (e.g., $\sigma = 0.32$), il est possible de construire un modèle qui obtienne simultanément une prédiction relativement précise et une robustesse raisonnable.

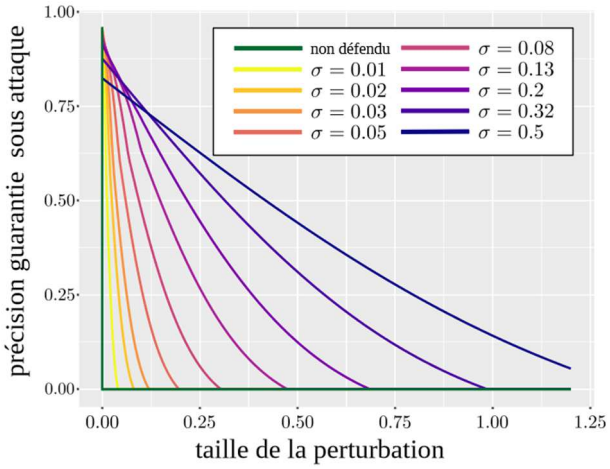


Figure 24 : Garanties obtenues par une défense basée sur l'injection de bruit, pour le jeu de donnée CIFAR10, pour différents niveaux de bruit σ. Plus σ est grand, plus la variance du bruit est importante.

4.4 Conclusion et impact sur les axes de recherche

La fin des travaux de thèse de Anne Morvan ainsi que les travaux de thèse de Rafaël Pinot ont permis d'aborder les questions de la confidentialité différentielle et de la robustesse aux attaques adverses. De nouvelles techniques ont été proposées et des résultats théoriques ont été démontrés. Nos contributions sur l'utilisation de méthodes randomisées sont un premier pas vers des processus de défense efficaces contre les attaques adverses. Grâce à ces travaux, de nombreuses pistes ont été ouvertes, notamment dans le cadre de différentes thèses à l'Université Paris Dauphine-PSL. Côté CEA, différentes thèses et projets ont également été lancés à la suite de ces premiers travaux, suivis actuellement par des collègues du laboratoire. Certains domaines d'application du laboratoire s'appuient sur des décisions critiques fournis par des algorithmes (par exemple le contrôle non-destructif pour le domaine nucléaire). Bien que l'apprentissage statistique offre des perspectives intéressantes en termes de performances par rapport à certaines approches traditionnelles, les propriétés non-fonctionnelles des algorithmes (confidentialité, robustesse, sécurité) ont une importance considérable pour permettre l'adoption de ces techniques.

5 Sécurité et confidentialité en contexte décentralisé

Cette dernière partie s'appuie sur les travaux de Arnaud Grivet Sébert, Fabiola Espinoza Castellon et plus marginalement Pierre-Emmanuel Clet. Alors que les travaux présentés jusqu'ici se sont focalisés sur des contextes centralisés, cette partie résume nos contributions pour la sécurisation et le respect de la vie privée dans des contextes décentralisés.

L'apprentissage décentralisé regroupe des approches dans lesquelles le traitement des données et l'optimisation du modèle sont réalisés de manière distribuée, sans que l'intégralité des données ne soit centralisée sur un serveur unique. Ce paradigme est particulièrement utile lorsque les données sont générées dans des environnements dispersés – par exemple sur des dispositifs mobiles ou dans des réseaux de capteurs – et que leur centralisation pose des problèmes de confidentialité, de bande passante ou de coût de communication. En outre, le traitement décentralisé permet de respecter les contraintes réglementaires et de garantir que les données sensibles restent localement sur chaque appareil. Ce cadre général ouvre la voie à des méthodes collaboratives où plusieurs participants contribuent à l'apprentissage d'un modèle commun tout en conservant leurs données privées. Rappelons que, dans le cadre général de la minimisation du risque empirique (MRE) présenté en Section 1.2, on considère une famille d'hypothèses $\mathcal{H}$. On introduit ici un vecteur de paramètres w. Autrement dit, chaque hypothèse est représentée par h_w, et la fonction de perte associée à une observation z est notée $\ell(h_w, z) = \ell(w, z)$, ce qui correspond à la perte encourue par le modèle h_w sur l'observation z.

Supposons que l'on dispose de P participants, chacun détenant un jeu de données local D_p composé de n_p observations, pour $p = 1, \dots, P$. Pour chaque participant, la fonction de perte locale s'exprime alors par

$$\mathcal{R}_p(w) = \frac{1}{n_p} \sum_{z \in D_p} \ell(w, z)$$

où $\ell(w, z)$ mesure l'écart entre la prédiction réalisée par h_w et l'observation z. L'objectif global de l'apprentissage décentralisé est de minimiser le risque empirique pondéré, c'est-à-dire de résoudre le problème d'optimisation suivant :

$$\min_w \sum_{p=1}^{P} \alpha_p \mathcal{R}_p(w) + \lambda\, \Omega(w)$$

où les coefficients α_p reflètent la proportion d'observations détenues par chaque participant, $\Omega(w)$ est un terme de régularisation permettant d'intégrer des connaissances a priori ou de contrôler la complexité du modèle, et $\lambda \geq 0$ est le paramètre d'équilibrage entre le risque et la régularisation. Dans un contexte purement décentralisé, chaque participant procède localement à l'optimisation de son risque $R_p(w)$ et communique périodiquement avec les autres participants (ou avec un serveur d'agrégation dans certains cas) pour mettre à jour le modèle global w. Cette communication permet de converger vers un modèle commun qui minimise le risque global, tout en évitant le transfert de données brutes.

Apprentissage fédéré

Dans ce contexte décentralisé, l'apprentissage fédéré constitue une approche particulièrement attractive. Comme présenté en Section 1.2, l'apprentissage fédéré (FL) permet à un ensemble de participants (clients) de collaborer pour entraîner un modèle global sans transférer leurs données brutes vers un serveur central. Chaque client calcule des mises à jour sur ses propres données, et un serveur agrège périodiquement ces mises à jour pour constituer le modèle global.

Mathématiquement, ce processus peut être formulé dans le cadre de la minimisation du risque empirique (MRE). Autrement dit, l'objectif est de trouver les paramètres du modèle w qui minimisent la somme pondérée des pertes locales calculées sur chacun des jeux de données des participants. Si chaque participant p dispose d'un jeu de données local D_p et d'une fonction de perte $\ell_p(w, D_p)$, l'objectif global peut être formulé comme :

$$\min_w \frac{1}{P} \sum_{p=1}^{P} \frac{1}{|D_p|} \sum_{(x,y)\in D_p} \ell_p\big(w, (x,y)\big)$$

ce qui correspond à la minimisation du risque empirique global. Ce processus distribué offre une première barrière de confidentialité puisque seules des mises à jour (gradients ou paramètres) sont communiquées, et non les données brutes elles-mêmes. Il s'appuie sur l'orchestration d'un nœud central qui rythme les étapes d'apprentissage et sélectionne les participants tour après tour. Après une initialisation arbitraire commune du modèle global côté serveur, le processus d'apprentissage fédéré consiste en des rounds successifs de communication entre le serveur et les clients. À chaque itération, une fraction des clients reçoit les paramètres actuels du modèle et les met à jour en minimisant une fonction de perte locale. Ensuite, ils renvoient les mises à jour au serveur qui les agrège pour actualiser les paramètres du modèle global. Cette méthode est illustrée dans la Figure 25.

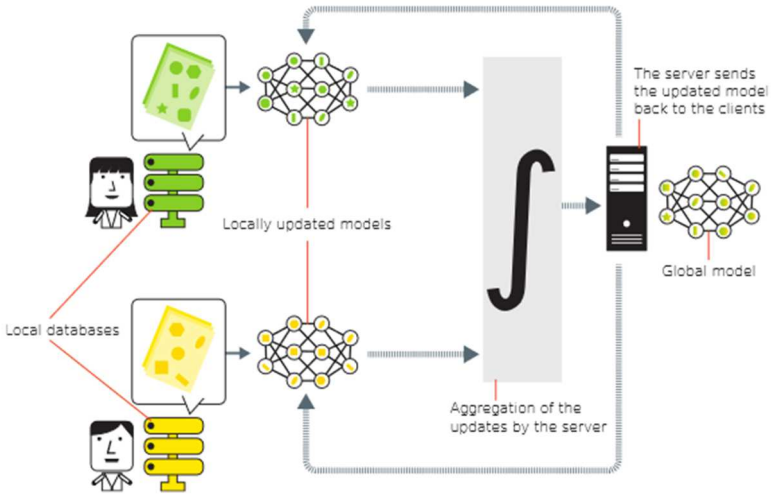


Figure 25 : Principe de l'apprentissage fédéré, d'après [5], [W3], reproduit avec la permission de Florent Robert, Industrie et Technologie.

L'algorithme suivant, proposé en 2016 par McMahan et ses collaborateurs [52], permet d'entraîner un modèle dans un contexte fédéré.

Algorithme 4 : Federated Averaging (FedAvg) d'après [52].

Server executes:
Input : M: total number of clients
K: number of participants per round
n_k: number of data points of participant k
w_t: model parameters at round t
Output: model parameters w_{t+1} at final round
initialise w_0;
for *each round* t **do**
  $K_t \leftarrow$ random set of $K \leq M$ clients;
  for *each client* $k \in K_t$ *in parallel* **do**
    $u_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$;
  $w_{t+1} \leftarrow w_t + \sum_{k=1}^{K} \frac{n_k}{n} u_{t+1}^k$ where $n = \sum_{k=1}^{K} n_k$

ClientUpdate(k, w):
Input : k: id number of the participant
D_k: training set of participant k
B: local mini-batch size
E: number of local epochs
η: learning rate
w: global model parameters
L: local loss function
Output: updates $w_k - w$ after last epoch
initialise $w_k = w$;
$\mathcal{B} \leftarrow$ split the n_k samples of D_k into batches of size B;
for *each epoch from* 1 *to* E **do**
  for *each batch* $b \in \mathcal{B}$ **do**
    $w_k \leftarrow w_k - \eta \nabla L(w_k; b)$;

Menaces sur la confidentialité et la sécurité dans l'apprentissage collaboratif

L'apprentissage fédéré ne garantit pas à lui seul la sécurité ni la confidentialité du processus d'apprentissage. La nature décentralisée du système crée de nouvelles surfaces d'attaque et des vulnérabilités potentielles. En particulier, deux grandes catégories de menaces se dégagent :

- Attaques sur la Confidentialité : même si les données brutes restent locales, les mises à jour du modèle peuvent révéler des informations sensibles. Des attaques telles que les attaques à l''inférence permettent à un attaquant, qu'il soit le serveur ou un client malveillant, d'exploiter ces mises à jour pour déterminer si un certain échantillon a été utilisé durant l'entraînement (attaques d'inférence d'appartenance) ou même pour reconstruire des parties des données d'entraînement (attaques par inversion de modèle).
- Attaques d'Intégrité (Backdoor) : le processus d'agrégation implique que le serveur combine les mises à jour de nombreux participants. Dans ce scénario, des participants malveillants peuvent injecter des mises à jour trompeuses (poisoning) dans le but d'insérer une porte dérobée dans le modèle global. Une attaque par backdoor consiste à altérer le comportement du modèle sur des entrées spécifiques (contenant un trigger), sans affecter sa performance sur les données normales.

Le modèle de menace en apprentissage fédéré doit donc prendre en compte des acteurs malveillants du côté du serveur (serveur curieux ou compromis qui tente d'inférer des informations privées) ainsi que des participants potentiellement corrompus ou collusifs qui cherchent à manipuler l'agrégation pour implanter des portes dérobées.

Défis de l'optimisation sécurisée et privée en apprentissage fédéré

Garantir la sécurité et la confidentialité dans ce cadre d'optimisation décentralisée pose plusieurs défis :

- Protection de la confidentialité : bien que l'échange des mises à jour permette de ne pas transférer les données brutes, des techniques comme l'agrégation sécurisée ou le recours à des méthodes de chiffrement homomorphe doivent être mises en œuvre pour empêcher le serveur ou d'autres participants de déduire des informations sensibles. Par ailleurs, l'intégration de la confidentialité différentielle (ajout de bruit aux mises à jour) est nécessaire pour limiter l'extraction d'information à partir des gradients communiqués.
- Robustesse contre les attaques par portes dérobées : afin de contrer les mises à jour malveillantes, il est indispensable de mettre en place des règles d'agrégation robustes ou des mécanismes de détection d'anomalies qui puissent identifier et

neutraliser les contributions suspectes. Cependant, distinguer des mises à jour légitimes des mises à jour malveillantes est particulièrement difficile dans un environnement caractérisé par l'hétérogénéité des données et l'absence d'accès aux données locales.

- Trade-offs entre sécurité et performance : l'intégration de techniques de protection, telles que le chiffrement homomorphe ou l'ajout de bruit pour la confidentialité différentielle, peut induire des surcoûts en termes de complexité computationnelle et d'efficacité de la convergence du modèle. Ainsi, l'un des grands défis est de parvenir à un compromis équilibré entre la robustesse et la protection de la confidentialité, tout en maintenant une performance optimale du modèle global.

Ces défis nous ont conduit à considérer différentes solutions d'apprentissage fédéré qui pourraient améliorer la protection de ces approches. Ces travaux sont présentés dans les sections suivantes.

5.1 Protection de l'apprentissage fédéré face aux attaques par portes dérobées

Ces travaux, réalisés dans le cadre de la thèse de Fabiola Espinoza Castellon (co-encadrée avec Aurélien Mayoue), sont extraits de [Cn1]. La version anglaise est disponible dans [C4].

Cette partie reprend certains passages de l'article [Cn1]. L'apprentissage fédéré (AF) est un paradigme d'apprentissage automatique qui permet à plusieurs entités, que nous appelons clients, d'entraîner un modèle de façon collaborative mais sans jamais partager leurs données. Ce processus est coordonné par un serveur central et nécessite plusieurs tours d'apprentissage pour réaliser l'entraînement du modèle fédéré. Ainsi, à chaque tour, le serveur transmet le modèle courant aux clients qui vont alors mettre à jour ses paramètres en l'entraînant indépendamment à partir de leurs propres données pour ensuite les retourner au serveur qui va les agréger. Dans sa version initiale, l'AF se base sur FedAvg [52], qui utilise la moyenne pondérée comme règle d'agrégation. Au tour T , les paramètres calculés par le serveur sont $w_T = \sum_{k \in C_T} \frac{n_k}{N} w_T^k$, où C_T est l'ensemble des clients participant au tour T , w_T^k sont les paramètres envoyés par le client k, n_k est le nombre de donnés du client k et $N = \sum_{q \in C_T} n_q$. Le serveur d'agrégation n'ayant pas accès aux données, celui-ci n'a pas de moyens directs de vérifier les informations envoyées par les participants. Le processus d'apprentissage est donc exposé à de nombreuses attaques par des participants qui

pourraient avoir intérêt à perturber la convergence ou à influencer le modèle dans une mauvaise direction. Dans [C4], nous nous sommes intéressés à une attaque appelée « attaque par porte dérobée » [53]. Le principe est schématisé dans la Figure 26. Pendant la phase d'entraînement, un groupe de clients malveillants ajoute un motif d'attaque, également nommé déclencheur, à leurs données et leur associe une classe cible t. Lors de l'inférence, le modèle va ainsi incorrectement prédire la classe cible t pour les données empoisonnées tandis que les données bénignes seront correctement classées ce qui rend les attaques par porte dérobée difficilement détectables.

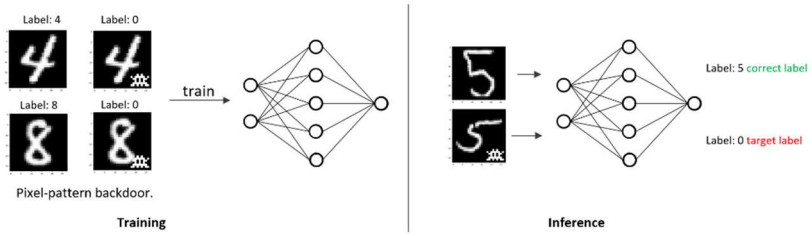


Figure 26 : Principe d'une attaque par porte dérobée.

La méthode proposée dans [C4] repose sur une estimation du motif déclencheur à partir des paramètres du modèle fédéré, puis utilise cette estimation pour introduire une étape de d'atténuation des risques. La méthode proposée s'appuie sur une technique d'inversion de modèle [54], [55]. En prenant en considération plusieurs observations réalisées sur le comportement des fonctions de coût en présence d'attaques par porte dérobée, une estimation fiable du motif de l'attaque, supposé constant, ainsi que de l'étiquette cible, est fournie. A partir de cette estimation, le serveur d'agrégation met alors en place une procédure d'atténuation des risques en modifiant localement la zone potentiellement impactée par le motif d'attaque. Une évaluation quantitative de la méthode proposée a été réalisée. Les résultats sont détaillés dans le Tableau 1. La « Clean accuracy » rend compte de la précision du réseau attaqué sur les échantillons non modifiés. Le score de « Backdoor success » représente la proportion d'échantillons du jeu de données empoisonné prédits avec la classe cible. Enfin le score « poisoned accuracy » représente la proportion d'échantillons empoisonnés correctement prédits comme leur classe d'origine.

		Before defense		With defense		
Dataset	Pattern	Clean accuracy	Backdoor success	Clean accuracy	Backdoor success	poisoned accuracy
MNIST	Square	0.98 ± 0.8	0.99 ± 0.3	0.98 ± 1.5	0.10 ± 1.2	0.98 ± 1.4
	Cross	0.98 ± 0.9	0.99 ± 0.8	0.98 ± 2.4	0.10 ± 2.7	0.98 ± 2.7
	Copyright	0.98 ± 1.0	0.99 ± 1.0	0.98 ± 1.3	0.15 ± 67	0.88 ± 70
FashionMNIST	Square	0.86 ± 2.5	0.99 ± 1.1	0.85 ± 4.9	0.10 ± 6.0	0.85 ± 4.0
	Cross	0.86 ± 3.6	0.99 ± 2.1	0.85 ± 4.0	0.13 ± 56	0.83 ± 41
	Copyright	0.86 ± 2.1	0.93 ± 11	0.85 ± 3.5	0.22 ± 83	0.74 ± 65
GTSRB	Yellow post-it	0.86 ± 4.4	0.98 ± 2.4	0.83 ± 5.7	0.02 ± 19	0.83 ± 5.5

Tableau 1 : résultats de bonnes classification moyens. Les écartypes ont été multipliés par 1000.

Si de nombreuses recherches sont encore nécessaires pour lutter de manière plus générique face à ce type d'attaques, ces résultats constituent un premier pas dans la défense face aux attaques par portes dérobées en contexte d'apprentissage fédéré.

5.2 SPEED : apprentissage collaboratif et vie privée

Ces travaux, réalisés dans le cadre de la thèse de Arnaud Grivet Sébert (co-encadré avec Renaud Sirdey), ont été publiés dans [J3].

Dans SPEED (Secure PrivatE and Efficient Deep learning), nous avons proposé dans [J3] un cadre d'apprentissage profond capable de satisfaire de fortes exigences de confidentialité. S'appuyant sur l'apprentissage distribué, la confidentialité différentielle et le chiffrement homomorphe, l'approche proposée permet une protection face à une large gamme de menaces, en particulier l'hypothèse d'un serveur honnête mais curieux. Basée sur la confidentialité différentielle distribuée et un opérateur argmax homomorphe, notre méthode est conçue pour assurer des coûts de communication faibles. Les parties suivantes expliquent le principe de fonctionnement de SPEED et nous rappelons les résultats théoriques et expérimentaux obtenus dans l'article [J3].

5.2.1 Le protocole PATE (Private Agregation of Teachers Ensemble)

Le protocole PATE (Private Aggregation of Teacher Ensembles) est un type d'apprentissage collaboratif. Dans cette solution destinée à une tâche de classification, les propriétaires des données, appelés professeurs, sont supposés avoir entraîné un modèle local à l'aide de leurs données. On fait aussi l'hypothèse que l'on dispose d'une base de données publique mais non étiquetée (on ne connaît pas les classes des échantillons). L'objectif est alors d'étiqueter la base de données publique qui sera ensuite utilisée pour entraîner le modèle global, nommé modèle étudiant. Pour ce faire, chaque échantillon à étiqueter est envoyé aux professeurs dont les modèles locaux infèrent une classe pour l'échantillon considéré. Les classes inférées par les modèles locaux, que l'on peut assimiler aux votes des professeurs, sont envoyées à un serveur qui compte le nombre de votes pour chaque classe et envoie la classe majoritaire au modèle étudiant. Ce dernier peut alors étiqueter l'échantillon et l'utiliser pour son propre entraînement. Ce processus est illustré dans la Figure 27.

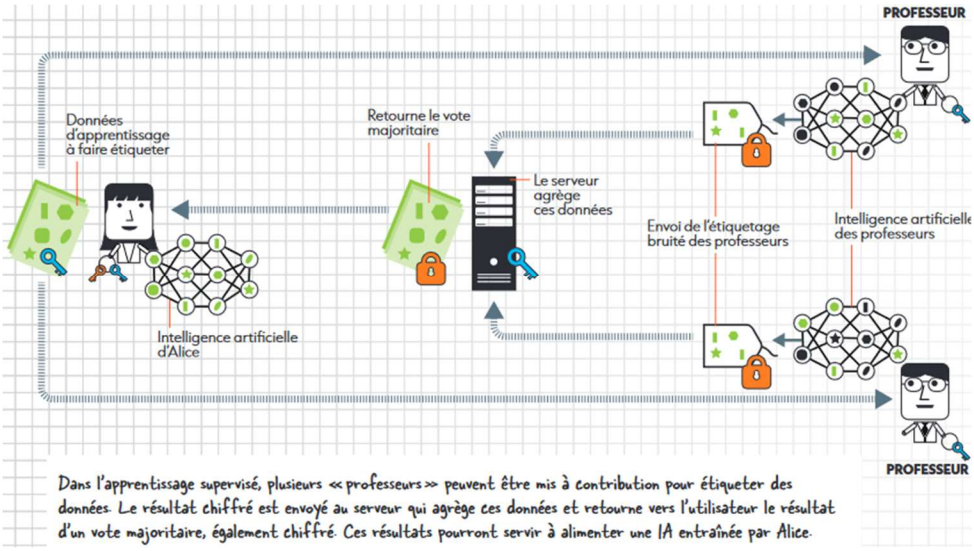


Figure 27 : Le protocole d'apprentissage distribué PATE, d'après [W3], reproduit avec la permission de Florent Robert, Industrie et Technologie.

5.2.2 SPEED : protection face à une surface d'attaque augmentée

Dans SPEED, nous considérons deux éléments complémentaires par rapport à PATE. Premièrement nous prenons en compte que le serveur peut être malveillant ou corrompu, ouvrant la possibilité au fait que les votes des professeurs peuvent être utilisés à mauvais escient, ou mal protégés vis-à-vis des étudiants. Deuxièmement, nous prenons en compte la possibilité de professeurs qui pourraient s'entendre lors de la phase d'introduction du bruit pour la confidentialité différentielle. Les contributions proposées dans SPEED proposent des solutions techniques pour ce modèle étendu de menaces. Le protocole est présenté dans la Figure 28.

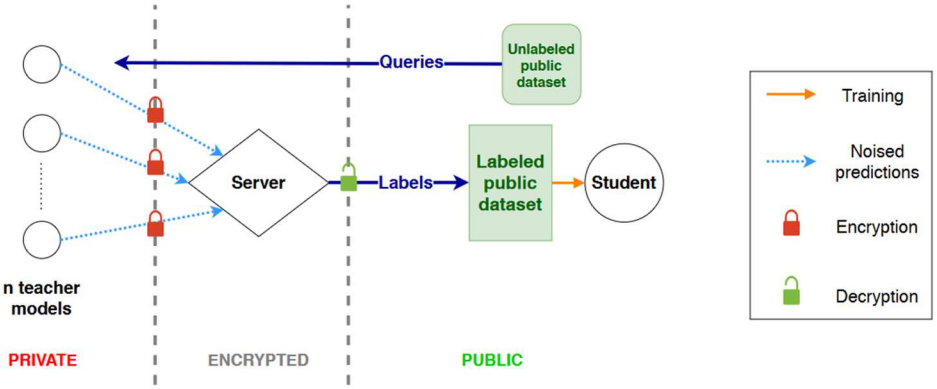


Figure 28 : SPEED - Les enseignants envoient au serveur d'agrégation leurs réponses bruitées et chiffrées aux requêtes de l'étudiant. Le serveur effectue de manière homomorphe l'agrégation dans le domaine chiffré et envoie le résultat au modèle étudiant qui le déchiffre et l'utilise pour l'entraînement.

On considère un ensemble de n possesseurs de données, les professeurs. On suppose disponible également une base de données publique D. On souhaite construire un modèle collaboratif (le modèle étudiant) de $\mathcal{X} \longrightarrow [K] = \{1, \dots, K\}$, exploitant les connaissances des différents professeurs. Le processus est le suivant :

1. Pour chaque élément $x \in D$, celui-ci est envoyé aux professeurs pour leur demander d'étiqueter cet exemple, i.e. donner $k \in [K]$. Chaque professeur i utilise alors son propre modèle privé h_i pour estimer k_i, l'étiquette prédite par le professeur i. Pour des raisons liées au calcul de l'agrégation et pour introduire le bruit nécessaire à la protection de la confidentialité (confidentialité différentielle), le professeur envoie alors un vecteur $z^{(i)}$ de taille K qui contient

$$\left[z^{(i)}\right]_{j\in[1,K]} = \begin{cases} G_{j,1}^{(i)} - G_{j,2}^{(i)} & \text{si} \quad j \neq k_i \\ 1 + G_{j,1}^{(i)} - G_{j,2}^{(i)} & \text{si} \quad j = k_i \end{cases}$$

   où, comme montré dans [J3], $G_1^{(i)}$ et $G_2^{(i)}$ sont des variables aléatoires de dimension K iid dont chaque élément suit une distribution Gamma. Nous avons exploité ici un théorème de probabilités liant la différence de deux variables aléatoires indépendantes distribuées selon une loi Gamma, avec une loi de Laplace, voir [56] pour les détails et conditions. Les vecteurs $z^{(i)}$ sont envoyés à l'agrégateur après chiffrement (chiffrement homomorphe) des composantes.
2. L'agrégateur reçoit l'ensemble des vecteurs envoyés dans le domaine chiffré par les professeurs. Il réalise alors une somme entre tous ces éléments. La propriété principale du chiffrement homomorphe est qu'il permet de réaliser des opérations dans le domaine chiffré, sans jamais nécessiter aucun déchiffrement. Cette somme est donc réalisée par l'agrégateur dans le domaine chiffré, sans qu'il puisse obtenir de l'information sur les quantités manipulées, ni le résultat obtenu. Après la somme, l'agrégateur doit alors appliquer un opérateur consistant à fournir l'indice de la composante la plus grande du vecteur, l'argmax. Grâce à l'opérateur argmax développé dans [57], nous disposons d'une approche relativement efficace (comparativement aux approches disponibles dans l'état de l'art) pour réaliser cette opération. L'étiquette associée à la donnée x peut alors être envoyée, dans le domaine chiffré, à l'étudiant à l'origine de la requête initiale.
3. L'étudiant, qui dispose de la clé de déchiffrement, peut alors décoder le résultat et exploiter l'information reçue pour se constituer un nouvel exemple dans son jeu de données d'entraînement.

La répétition de ces opérations permet ainsi à l'étudiant de se constituer une base de données exploitant la base de données publique ainsi que les connaissances des professeurs.

5.2.3 Résultats expérimentaux

Plusieurs expériences ont été menées afin de valider ce protocole d'apprentissage collaboratif, et d'évaluer les coûts computationnels engendrés par l'utilisation du chiffrement homomorphe. Le Tableau 2 rend compte des expériences menées. La ligne « Non-private » correspond à une technique d'apprentissage similaire à celle employée dans SPEED, mais sans confidentialité différentielle ni chiffrement homomorphe. « Trusted » correspond à PATE. Enfin les lignes suivantes indiquent les résultats obtenus lorsqu'une partie des professeurs n'incluent pas le bruit permettant d'assurer la confidentialité différentielle ($\tau = 1$ signifie que tous les professeurs incluent bien le bruit Gamma).

Tableau 2 : résultats des expériences menées sur les bases de données MNIST (gauche) et SVHN (droite).

Framework	ϵ	Acc. ($\pm$ std) [%]	HE overhead
Non-private	-	96.22 ($\pm$2.27)	-
Trusted	1.41	95.95 ($\pm$2.97)	-
$\tau = 1$	1.41	95.91 ($\pm$2.57)	6.5 min
$\tau = 0.9$	1.66	96.02 ($\pm$2.92)	
$\tau = 0.7$	2.37	96.06 ($\pm$2.61)	

Framework	ϵ	Acc. [%]	HE overhead
Non-private	-	84.7	-
Trusted	4.73	83.7	-
$\tau = 1$	4.73	83.5	32.5 min
$\tau = 0.9$	5.59	83.8	
$\tau = 0.7$	8.16	84.6	

5.3 SHIELD : garantie de confidentialité différentielle par construction pour un opérateur probabiliste homomorphe

Ces travaux, réalisés dans le cadre de la thèse de Arnaud Grivet Sébert (co-encadré avec Renaud Sirdey), ont été publiés dans [C3].

Dans [C3], nous avons proposé l'opérateur SHIELD (Secure and Homomorphic Imperfect Election via Lightweight Design) dans l'optique suivante : le design d'un opérateur non-linéaire homomorphe efficace étant intrinsèquement approximatif, peut-on choisir une conception dont le bruit introduit par l'approximation pourrait garantir des propriétés de confidentialité différentielle ? Il s'agit par exemple de concevoir un protocole d'apprentissage distribué, tel que celui de la section précédente (SPEED) bénéficiant de protection accrue pour la protection de la vie privée, dans lequel les opérations homomorphes (approximant dans SPEED un opérateur argmax) seraient suffisantes pour assurer des garanties de confidentialité différentielle. Tel qu'illustré dans la Figure 29, il s'agit dans SPEED de remplacer l'argmax approché et le bruit introduit par les professeurs par un opérateur unique assurant ces deux propriétés.

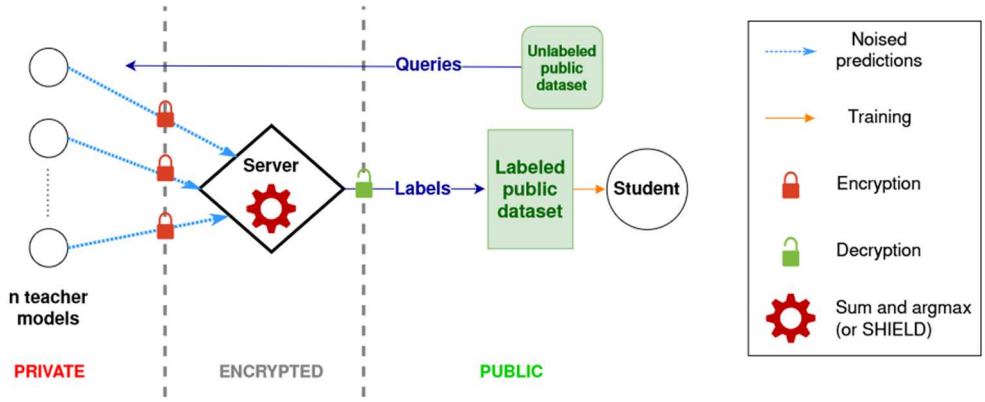


Figure 29 : Protocole SPEED dans lequel l'opérateur d'agrégation est remplacé par SHIELD. Comparativement à la figure précédente, le serveur utilise ici l'opérateur SHIELD.

Afin de satisfaire ces deux objectifs, nous avons proposé un nouvel opérateur construit sur le principe d'un vote majoritaire approché (argmax approché). Il s'agit d'un algorithme itératif qui prend en entrée un ensemble de votes, voir Algorithme 5, et fournit en sortie un vecteur indicateur de la classe majoritaire. Un résultat important de [C4] est que cet opérateur compatible avec une implémentation homomorphe efficace fournit des garanties de confidentialité différentielle, sans ajout supplémentaire de bruit. Bien qu'il s'agisse d'une première approche relativement directe de ce principe, nous pensons qu'elle permet de mieux contrôler le compromis entre le surcoût computationnel impliqué par l'utilisation du chiffrement homomorphe et les nécessaires protections de la confidentialité des données.

Algorithme 5 : SHIELD

Input : number of vectors n, number of classes K, list of encrypted votes Z, polynomial $(a_p)_{p\in[D]}$, offset ω
Output : $res = z^{(i_0)}$ where $i_0 \in [n]$

$Z \leftarrow Z$ augmented by ω encrypted one-hot encodings for each class;
$res \leftarrow (0, \dots, 0) \in (\mathbb{Z}_2)^K$;
$found\_not\_null \leftarrow 0$;
for p *from* D *to* 1 **do**
  for j *in* $[a_p]$ **do**
    $\pi \leftarrow (1, \dots, 1) \in (\mathbb{Z}_2)^K$;
    for l *in* $[p]$ **do**
      Draw a vector z of Z uniformly at random;
      $\pi \leftarrow \pi \otimes z$;
    end
    $res \leftarrow res \oplus (1 \oplus found\_not\_null) \otimes \pi$;
    $is\_not\_null \leftarrow \bigoplus_{k=1}^{K} \pi_k$;
    $found\_not\_null \leftarrow found\_not\_null \veebar is\_not\_null$;
  end
end

Différentes implémentations ont été réalisées dans l'article [C4], montrant, en fonction du polynôme utilisé pour le calcul de l'opérateur, un bon comportement entre précision et efficacité computationnelle. De nombreuses questions restent à investiguer pour comprendre l'exacte influence de chaque paramètre sur les résultats, mais cette première étape est très encourageante.

5.4 Combiner confidentialité différentielle et chiffrement homomorphe en apprentissage fédéré

Ces travaux, réalisés dans le cadre de la thèse de Arnaud Grivet Sébert (co-encadré avec Renaud Sirdey), ont été publiés dans [C2].

Combiner confidentialité différentielle et chiffrement homomorphe dans l'apprentissage fédéré se heurte à des difficultés pratiques importantes. Considérons le contexte de l'apprentissage fédéré pour lequel il est nécessaire de se protéger face au serveur d'agrégation et à des participants malintentionnés qui chercheraient à extraire de l'information à partir des gradients ou des modèles renvoyés par le serveur.

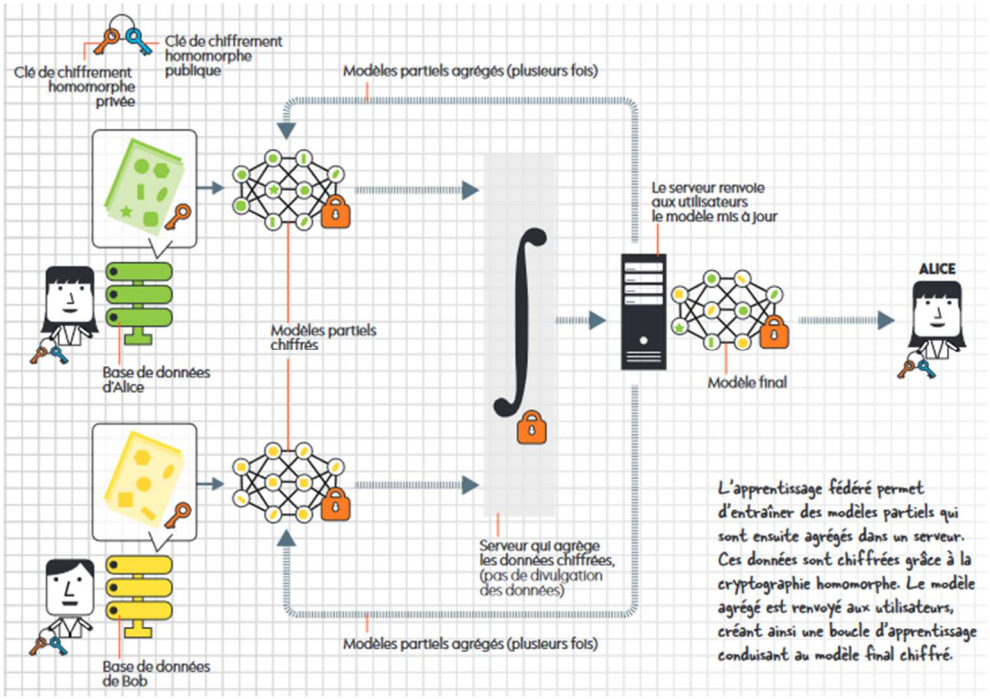


Figure 30 : Apprentissage fédéré sécurisé par chiffrement homomorphe. D'après [W3], reproduit avec la permission de Florent Robert, Industrie et Technologie.

Ce cas de figure, de manière similaire aux deux parties précédentes, laisse penser qu'une combinaison de confidentialité différentielle et de chiffrement homomorphe doit être adaptée à ce contexte. Malheureusement cette intuition se heurte à une difficulté de taille : une telle combinaison nécessite d'envoyer une version chiffrée du gradient de chacun des participants. Or le chiffrement homomorphe nécessite, pour être efficace, une information codée de manière discrète, finie, et sur un nombre de bits le plus réduit possible. A l'opposé, la confidentialité différentielle offre des garanties théoriques lorsque la variance du bruit ajouté est la plus grande possible. Ces deux aspects constituent donc un obstacle important à la mise en œuvre d'une approche combinant ces deux techniques en apprentissage fédéré. Plusieurs approches existent dans la littérature [58], [59], [60], [61], [62] mais nous avons montré dans [C2] qu'aucune n'était complètement satisfaisante. Dans cet article, nous proposons un nouvel opérateur de quantification stochastique pour combiner ces deux techniques. Celui-ci nous permet d'établir des garanties de confidentialité différentielle lorsque le bruit est à la fois quantifié et à support borné. Ces deux conditions (support borné et quantification) sont nécessaires pour l'utilisation dans un contexte de chiffrement homomorphe.

L'approche proposée s'appuie sur la distribution de Poisson, qui commute avec l'agrégation. Nous avons montré dans [C2] que cet opérateur est équivalent à un post-traitement. Ainsi la discrétisation n'a aucun impact sur les garanties de confidentialité qui s'avèrent donc être les mêmes que celles du très classique mécanisme gaussien.

5.5 Conclusion

Cette partie a présenté nos travaux les plus récents, qui s'intéressent de manière générale aux compromis nécessaires en apprentissage distribué, notamment entre l'efficacité, le respect de la confidentialité ou encore la robustesse à des comportements malicieux coordonnés des participants. Plusieurs contributions ont été présentées, fournissant une base à de futurs travaux sur la notion générale de compromis entre différentes propriétés nécessaires à l'adoption plus large de l'apprentissage distribué pour de nombreux cas d'usage.

6 Mise en perspective et projet de recherche

Nous avons retracé les grandes pistes de recherche explorées depuis une dizaine d'années. Si le paysage scientifique et technique autour des travaux présentés a connu des évolutions majeures, il m'apparaît aujourd'hui difficile de concevoir des travaux de recherche en apprentissage statistique sans prendre en compte des facteurs stratégiques, économiques, réglementaires et éthiques. Avant de présenter notre projet de recherche dans la section 6.2, nous revenons dans un premier temps sur les facteurs qui nous apparaissent cruciaux à prendre en compte pour positionner des recherches futures.

6.1 Mise en perspective des travaux de recherche

6.1.1 Enjeux stratégiques

Dans [63], l'auteure rappelle que « la notion de souveraineté est définie traditionnellement comme le pouvoir suprême exercé sur un territoire, à l'égard d'une population, par un Etat indépendant, libre de s'autodéterminer ». Il est donc naturel que le pouvoir actuel des multinationales du monde du numérique soulève de nombreuses questions sur les articulations entre le pouvoir des Etats et le pouvoir de ces géants du numérique. En particulier, Pauline Türk note que les réflexions sur la notion de souveraineté numérique nait de la volonté de refuser « de voir les peuples, les communautés d'utilisateurs, les États, les individus perdre le contrôle de leur destin au profit d'entités mal identifiées, non légitimes, et dont l'objectif n'est pas la promotion de l'intérêt général » [63, p. 20]. Au-delà des impacts sur le cyberespace, le numérique, a fortiori l'apprentissage statistique, a aujourd'hui une place centrale dans l'ensemble des chaines d'approvisionnement [64].

Au niveau français, le sujet de la souveraineté numérique a focalisé de nombreuses initiatives ces dernières années. La commission d'enquête sur la souveraineté numérique[9], et le rapport qui en a découlé ont par exemple fait un ensemble de recommandations de mesures à prendre rapidement dans l'objectif d'affirmer que « la souveraineté numérique est un devoir national et, à ce titre, engage nos compatriotes, toutes responsabilités confondues ; aussi serait mis en place un Forum national du numérique [...] pour sortir de la situation peu satisfaisante dans laquelle les attributs traditionnels de la souveraineté nationale et nos valeurs démocratiques sont malmenés ». Le même rapport propose

[9] https://www.senat.fr/travaux-parlementaires/structures-temporaires/commissions-denquete/commissions-denquete/commission-denquete-sur-la-souverainete-numerique.html

l'adoption d'une loi triennale d'orientation et de suivi de la souveraineté numérique (LOSSN) afin que « le parlement puisse exercer pleinement son rôle de gardien de la souveraineté numérique nationale ». De nombreuses actions ont ensuite suivi cette analyse, nous y reviendrons dans les paragraphes suivants.

Au niveau européen, de nombreuses actions ont été lancées et les conséquences politiques, économiques et réglementaires de la prise de conscience de la nécessité de défendre une souveraineté numérique européenne [65] sont encore en développement (voir paragraphe 6.1.3). L'adoption par les Etats-Unis du CLOUD Act (Clarifying Lawful Overseas Use of Data) en 2018 a notamment conforté cette nécessité. Cette loi permet en effet aux agents fédéraux américains, dans le cadre d'enquêtes, de saisir les données hébergées par un acteur américain du monde du numérique, indistinctement de la position géographique du serveur. Cette loi est symptomatique d'un phénomène qui a pris de l'ampleur dans les années 2010 : l'extraterritorialité des lois, voir par exemple [66], qui consiste à créer et utiliser des lois qui s'appliquent au-delà du territoire national de l'Etat l'ayant promulguée. L'extraterritorialité est utilisée aujourd'hui comme un outil pour des enjeux économiques dans de nombreux sujets, mais nous voulons pointer ici que le monde du numérique, de par ses aspects transfrontaliers, est un terrain de jeu où le nombre de ces lois va probablement augmenter dans les années à venir, créant ainsi une complexité et une instabilité pour les acteurs de ce domaine.

Au cœur du numérique, l'intelligence artificielle[10] est aujourd'hui identifiée comme un élément central, et la mise en production par OpenAI de son agent conversationnel générique, ChatGPT, a conforté cette position. Sans attendre ChatGPT, toutes les grandes puissances mondiales avaient identifié les opportunités et risques stratégiques liés à l'intelligence artificielle, et elles avaient travaillé sur des plans de développement ambitieux. En France, la rapport [67] a élaboré une stratégie nationale pour l'intelligence artificielle. L'Europe a publié sa vision dans un « white paper » fondateur [68]. La Chine a également amplement communiqué sur sa vision [69].

Avant de conclure cette partie sur la place de l'intelligence artificielle dans la souveraineté numérique et ses enjeux stratégiques associés, prenons un dernier exemple pour illustrer les risques associés au domaine dans son ensemble, vis-à-vis de décisions qui dépassent les acteurs scientifiques et techniques du domaine. En 2018 aux Etats-Unis[11], des directives ont été publiées afin de revoir la liste de technologies régulées par les lois de contrôle à l'exportation (Export Controls). Ces lois à destination des matériels stratégiques

[10] Terme entendu dans ce contexte dans un sens large, comme l'ensemble des outils, algorithmes et techniques permettant de produire un système autonome de décision, majoritairement dans sa version basée sur des données (apprentissage statistique).

[11] https://www.govinfo.gov/app/details/PLAW-115publ232

pour la sécurité nationale des Etats-Unis visent à imposer des contraintes sur l'exportation à l'étranger des technologies figurant dans cette liste. Puis en novembre 2018, le BIS (Industry and Security Bureau) a précisé une liste potentielle de technologies[12] « emerging and foundational » à considérer. Cette liste incluait de nombreuses technologies, dont un ensemble de technologies relatives à l'intelligence artificielle et à l'apprentissage statistique (software et hardware). L'analyse de l'époque faite par les spécialistes américains a été relativement unanime [70], [71], [72], [73], [74] : un contrôle drastique à l'exportation des technologies logicielles d'intelligence artificielle serait contre-productif pour les Etats-Unis (algorithmes, software open-source) mais un contrôle (vis-à-vis notamment de la Chine) mis en place rapidement pour limiter les capacités de production du matériel permettant de faire fonctionner les IA (photolitographie notamment) serait probablement efficace pour permettre aux Etats-Unis de garder leur avance technologique. Les positions stratégiques que nous voyons aujourd'hui sont les conséquences des décisions qui découlèrent de cette analyse.

Quelles conséquences tirer de ce contexte stratégique ?

- S'il n'est pas du rôle d'un chercheur de prendre position sur des enjeux stratégiques, nous pensons qu'il faut prendre en considération le fait que les algorithmes d'intelligence artificielle sont, et resteront, un pilier central de la stratégie des états. Ainsi il est important pour les chercheurs de s'impliquer dans la construction d'une compréhension scientifique des décideurs aussi précise que possible.
- La complexité de l'écosystème numérique mondial doit conduire à une approche pragmatique de la souveraineté numérique. Celle-ci ne peut se réduire à la volonté de maîtriser 100% de la chaine de la valeur d'une technologie. Dans un contexte de compétition scientifique mondiale et de partage par l'intermédiaire des logiciels open-source, nous pensons qu'une question centrale est : à partir de l'état d'un logiciel open-source à un instant t_0, combien de personnes sont capables de continuer à faire avancer ce logiciel ? Plutôt que : quel est le pourcentage de code réécrit et maitrisé de façon fermée par un organisme ou un état ?
- Nous partageons donc l'idée que l'accent doit être mis sur l'excellence scientifique (production des méthodes avancées), la capacité à réduire le temps entre la production de méthodologies et leurs mises à disposition de l'économie, ainsi que la maîtrise technologique nécessaire à leur mise en place efficace.

[12] https://www.federalregister.gov/documents/2018/11/19/2018-25221/review-of-controls-for-certain-emerging-technologies

6.1.2 Economie de l'apprentissage statistique

J'ai intégré le Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) en 2010 avec la volonté de contribuer à des axes de recherche susceptibles d'avoir un impact relativement rapide (1 à 5 ans) sur des secteurs applicatifs variés, et d'être un acteur de ce transfert de technologies. La Direction de la Recherche Technologique (DRT) était donc pour moi un choix relativement naturel. Cette direction a notamment un rôle important dans l'une des missions du CEA[13] qui consiste à « contribuer, au service de la compétitivité de la France, au développement technologique et au transfert de connaissances, de compétences et de technologies vers l'industrie, notamment dans le cadre régional, ainsi qu'à la valorisation des résultats des recherches qu'il mène ». J'ai eu la chance, depuis la fin de ma thèse, d'être impliqué dans des domaines de recherche extrêmement dynamiques de ce point de vue, et il me semble utile à ce stade de chercher à mieux comprendre les dynamiques des cycles de recherche et de valorisation dans le domaine de l'apprentissage statistique.

Tout d'abord il faut noter que les algorithmes d'intelligence artificielle sont considérés par les économistes comme une Technologie à Usage Général (le terme consacré est General Purpose Technology [75], i.e. GPT en anglais). Elles se caractérisent par une utilisation omniprésente dans de nombreux secteurs applicatifs, ont un potentiel important d'améliorations techniques, et des améliorations de performances dans ces technologies génériques résultent en des gains de productivité dans de nombreux domaines. Ces technologies attirent depuis des dizaines d'années beaucoup d'attention et il est acquis que la longueur de ces cycles se raccourcit drastiquement (cf [76] relatant les dizaines d'années nécessaires à l'introduction du moteur par exemple).

La problématique de la valorisation des données, des algorithmes d'apprentissage statistique et des modèles résultant des apprentissages, ont connu de nombreuses évolutions depuis une quinzaine d'années. Les études économiques ayant tenté d'analyser les interactions récentes entre les avancées technologiques associées aux vagues du « Big Data » et de l'« Intelligence Artificielle » et les modèles de valorisation sont assez rares, probablement parce que la vitesse d'évolution de ces modèles rend assez rapidement obsolètes les conclusions. Quelques tentatives récentes [77], [78], [79], [80], [81] fournissent une analyse en s'appuyant sur les modèles de fondation en langage ou vision (comme ChatGPT d'OpenAI). Cette analyse fournit une base de réflexion pour la cristallisation

[13] https://www.legifrance.gouv.fr/loda/id/JORFTEXT000032242719, Article 2. Notons cependant que cet article a été abrogé le 27 Décembre 2023, mais en reprenant cette mission à l'identique pour le CEA dans le décret portant sur la partie réglementaire du code de la recherche https://www.legifrance.gouv.fr/jorf/id/JORFARTI000048709366.

future des approches d'apprentissage statistique dans les modalités plus confidentielles (séries temporelles, données spatiales). Dans [79], les auteurs tirent de nombreuses conclusions sur l'émergence actuelle d'une approche dominante pour la valorisation de l'apprentissage statistique et des modèles entrainés par apprentissage statistique. Les auteurs observent notamment que « without the training data and computing power, no one can refine models ». S'appuyant sur les travaux de [82], les auteurs indiquent que l'accès à des algorithmes permettant d'entraîner des modèles statistiques, est en passe de devenir un service (« utility ») tel que l'électricité ou l'eau. Dans cette optique l'élément différenciant stratégique est alors la donnée. Dans cette optique, la construction d'écosystèmes de plateformes devient alors un élément clé du management de la technologie : une plateforme rassemble un ensemble d'innovations à partir desquelles d'autres innovations peuvent être construites, avec une complexité réduite pour le concepteur. L'émergence d'une plateforme dominante requiert un savant mélange entre une architecture ouverte (attirer la majorité) et un ensemble de mécanismes fermés qui permettent au sponsor de la plateforme de tirer des profits à partir de la plateforme. Après une phase centrée sur le développement de techniques d'apprentissage statistique en tant que service dans le cloud, a lieu une phase visant à porter l'utilisation des modèles résultants sur l'ensemble des périphériques, au plus près des utilisateurs. Les critères clés de cette phase sont alors la confidentialité ou encore la latence. Enfin un dernier point est noté dans [79] : dans le cas des applications utilisant des techniques d'apprentissage statistique dans lesquelles la donnée doit être massive, on doit observer un phénomène appelé « winner-take-all », une position dominante permet à un acteur installé d'acquérir plus de données de qualité, lui permettant ainsi de conforter sa position de leader.

Quelles conclusions tirer de ces observations ?

- L'accès à une donnée de qualité est probablement l'élément le plus important de la chaîne de valeur de l'IA.
- La disponibilité de gigantesques bases de données n'étant pas transposable à l'ensemble des domaines applicatifs, les modèles de valorisation de l'IA pour ces domaines restent largement ouverts. Deux hypothèses semblent possibles aujourd'hui : premièrement des avancées importantes dans la capacité à générer des données réalistes en grande quantité (ou à transposer efficacement entre des domaines d'application variés) pour des applications particulières ; deuxièmement la convergence vers des modèles de valorisation spécifiques aux cas d'application en régime de données limitée.
- Les analyses économiques sont basées sur des règles du marché actuel, laissant une place très importante à la génération de profit. La réglementation ainsi que

les enjeux climatiques pourraient constituer des facteurs plus impactant qu'ils ne le sont aujourd'hui à moyen terme.

6.1.3 Le paysage légal et réglementaire de « l'intelligence artificielle »

Dans son article fondateur de 2020 sur une approche européenne de l'intelligence artificielle [83], la commission européenne a présenté sa feuille de route afin de favoriser le développement d'un écosystème européen orienté vers le bénéfice : 1) des citoyens, en apportant par exemple des améliorations dans le système de santé, la fiabilité des machines, un transport public plus sûr et moins polluant ; 2) les acteurs économiques, en particulier les secteurs dans lesquels l'Europe est forte, comme par exemple les machines-outils, le transport ou la cybersécurité ; 3) au service de l'intérêt public, par exemple en réduisant le coût des services publics ou améliorant la durabilité des produits. L'Europe militait alors pour la mise en place d'un écosystème d'excellence, mettant en garde contre des initiatives nationales. En parallèle, les conclusions d'un groupe d'experts étaient détaillées dans le même article, afin d'agir pour construire un ensemble de mesures afin de viser sept objectifs réglementaires : 1) « Human agency and oversight », 2) « technical robustness and safety », 3) « privacy and data governance », 4) « transparency », 5) « diversity, non-discrimination and fairness », 6) « societal and environmental wellbeing », 7) « accountability ». Cette feuille de route a conduit à plusieurs réglementations ou projets de réglementations. Nous allons parcourir les principales.

Vie privée et règlement pour la protection des données

Le Règlement Général pour la Protection des Données (RGPD) a été proposé en 2012, adopté en 2016 puis est entré en vigueur en 2018. Il s'applique à toutes formes de traitement[14] de données personnelles[15]. Notons que la notion de données personnelles soulève quelques débats technico-juridiques. Les données synthétiques peuvent par exemple être générées en entrainant un modèle statistique pour générer des nouvelles données à partir d'un jeu de données initial. Dans ce cas les avis divergent sur le fait que ces données peuvent être considérées comme non-personnelles, dans la mesure où des recouvrements peuvent apparaître entre le jeu de données initial et les données synthétiques. Ce règlement présente un ensemble de principes qui doivent être vérifiés par tout traitement impliquant des données personnelles, en particulier les principes de limitation du stockage ou encore de la minimisation de la quantité de données. Le règlement note cependant que le traitement de

[14] « any operation or set of operations ... on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction ».

[15] « any information relating to an identified or identifiable natural person ('data subject') »

données personnelles a une base légale si la performance d'une tâche justifie l'utilisation de ces données, et si l'entité procédant au traitement de données se fait dans le cadre d'une tâche pour l'intérêt public ou dans l'exercice d'une autorité officielle.

Artificial Intelligence Act

La Commission européenne prépare actuellement un nouveau cadre réglementaire pour les algorithmes d'intelligence artificielle, appelé AI Act [84]. Dans ce nouveau règlement, il est prévu une approche basée sur le risque, c'est-à-dire que les exigences dépendront du risque potentiel associé au système exploitant un modèle entrainé par apprentissage statistique. Quatre niveaux de risques ont été définis :

- Risque inacceptable : système dangereux qui va à l'encontre des valeurs européennes (par exemple le scoring social). Ces systèmes sont bannis.
- Haut risque : les systèmes entrants dans cette catégorie devront se conformer à des normes exigeantes pour en assurer un haut niveau de confiance, de protection de la sûreté des citoyens et des droits fondamentaux des citoyens.
- Risque limité : ces systèmes reposant sur des algorithmes d'intelligence artificielle seront sujets à un ensemble limité d'obligations (par exemple la transparence).
- Risque minimal : tous les autres systèmes ne feront pas l'objet de règles additionnelles.

Les discussions récentes, issues de questionnements franco-germano-italiens sur les modèles fondateurs (comme ChatGPT), ont de plus convergé vers des clauses particulières afin de ne pas bloquer les capacités d'innovation de l'Europe dans ce domaine : les activités de recherche et le développement de composants d'intelligence artificielle libres et ouverts seraient largement exemptés de nombreuses règles de l'AI Act.

Autres réglementations en préparation

D'autres initiatives ont été lancées afin de mieux soutenir la vision européenne à l'heure du numérique. Trois d'entre elles auront nécessairement un impact sur le contexte économique et scientifique associé à l'intelligence artificielle : d'une part les « Cyber-Resilience Act » et « Cybersecurity Act » visant à imposer un haut niveau d'exigence sur la cybersécurité des composants numériques, incluant les composants intégrant des systèmes autonomes de décisions ; d'autre part la nouvelle « Product Liability Directive » [85], qui serait applicable aux produits intégrant des systèmes de décision automatisés, permettrait de mieux protéger les citoyens face au dommages provoqués suite à un produit intégrant de telle techniques, et instituerait une présomption de causalité pour les personnes ayant subis des préjudices causés suite à l'utilisation de systèmes d'intelligence artificielle.

La situation hors Europe

Si l'Europe a manifestement des spécificités quant à sa volonté d'inscrire le droit fondamental des citoyens au cœur de sa stratégie sur l'intelligence artificielle, les autres grandes puissances telles que la Chine et les Etats-Unis développent également leur propre régulation [69], [86]. La législation autour des algorithmes d'intelligence artificielle devient ainsi un nouveau terrain pour l'extraterritorialité des lois.

Quelques éléments d'analyse de ces réglementations

Dans le domaine de la recherche, il est appréciable de constater les efforts de la Commission européenne pour ne pas bloquer l'innovation. Cependant, il faut faire plusieurs constats : d'une part, même si de nombreuses exceptions existent pour favoriser l'innovation, l'accumulation de réglementations (européennes et internationales) font peser sur les laboratoires de recherche le poids de l'analyse. Les chercheurs experts d'apprentissage statistique ne pouvant pas s'improviser juristes, l'empilement des textes impose aux laboratoires de s'appuyer sur des équipes conséquentes de juristes spécialistes, en particulier lorsqu'une partie de l'activité s'oriente vers une valorisation vers des partenaires industriels. Même si des services juridiques existent dans certaines organisations, la situation actuelle nécessite souvent des interactions nombreuses entre les équipes juridiques et techniques afin de déterminer dans quelle mesure des traitements automatisés résultant de composants entrainés par apprentissage statistique peuvent être sereinement entrepris dans de nombreux cas. Mentionnons par ailleurs que les exemptions de recherche posent question, notamment dans [87], où on peut lire « It can be concluded that the normative distinction between research, on the one hand, and other purposes subject to the rules under the AI Act breaks down considering modern research practices which are conducted by public and private actors for commercial and non-commercial reasons alike in a complex, entangled and diaphanous process ». Plusieurs années après sa mise en application, les règles relativement simples du RGPD posent toujours de nombreuses questions, notamment soulevées par [88], qui s'interroge sur les capacités réelles aujourd'hui du « machine unlearning »[16], sous-entendu dans l'article 17 du RGPD. Hors du scope de ce manuscrit, nous préférons omettre quelques sujets qui sont d'importance dans le cas de la manipulation de données potentiellement sous copyright. Ces cas de figure (images et textes) emmènent alors sur des questions liées au risque de génération d'un contenu sous copyright, mais aussi du principe américain, défendu par OpenAI dans son approche plutôt « libérale » de l'utilisation des données publiques, de « fair use ».

[16] Machine unlearning : capacité à supprimer la contribution d'une donnée particulière dans un modèle entrainé par apprentissgae statistique, sans procéder au réentrainement complet du modèle.

Loin d'être anecdotiques, ces considérations réglementaires vont prendre une importance de plus en plus grande dans le futur, et nous devons tenir compte des évolutions prochaines dans les directions de recherche à privilégier. Notons enfin que ces considérations devront être complétées par les exigences d'éthiques qui deviennent de plus en plus importantes dans de nombreux projets (en particulier dans les projets européens). Le lecteur intéressé pourra se référer à [89], [90].

6.1.4 Perspectives épistémologiques de l'intelligence artificielle

Nous n'aurons pas ici la prétention de fournir une analyse épistémologique de la vague de « l'intelligence artificielle » et de la science des données. Quelques éléments précis sont néanmoins utiles pour éclairer certaines lignes directrices de mon projet de recherche. Dans [91], [92], repris récemment par [93], les auteurs expliquent que l'intelligence artificielle a évolué d'un projet de compréhension de l'intelligence humaine, vers un projet purement technologique visant des capacités à copier des performances de calcul et de compréhension propres à l'homme, sans grand rapport avec le fonctionnement du cerveau humain. Ainsi, comme rappelé dans [93], la rupture entre théorie et données conduit à un nouveau paradigme de production de connaissances, qui privilégie « faire » à « comprendre ». Les algorithmes d'intelligence artificielle produisent des résultats technologiques impressionnants et permet d'accélérer la production de connaissances dans tous les domaines scientifiques. Néanmoins il convient de s'interroger sur les quelques alertes levées sur son impact sur la construction de la connaissance. Dans [94], [95], [96], nous voyons des indices qui pointent vers un besoin de synthèse entre la construction de la connaissance basée sur la théorie et l'appui des technologies d'intelligence artificielle au service de la connaissance. Cette synthèse pourrait permettre de réconcilier deux paradigmes, le règne des mathématiques sur la formalisation du monde par la théorie [97], et le règne de la donnée pour agir sur le monde [98].

6.2 Projet de recherche

En tant que chercheur sur les aspects méthodologiques de l'apprentissage statistique, mon rôle est de fournir aux acteurs socio-économiques des connaissances, des outils et des méthodes dans mon champ de compétences. En tant que responsable d'une équipe au CEA, mon rôle est également d'aligner au mieux les objectifs de recherche avec :

- Les forces et contraintes internes,
- L'environnement hyper-compétitif de l'écosystème public et privé dans le domaine,

- Les facteurs stratégiques, économiques, réglementaires et éthiques décrits dans la partie précédente,

afin que les travaux entrepris visent les impacts les plus significatifs possibles. Le CEA est un organisme public dont une des forces est de pouvoir mettre en synergie des équipes aux spécialités variées autour de grands instruments ou d'ambitions supra-laboratoires. Il est de plus organisé pour accompagner efficacement les passages de preuves de concept scientifiques, aux prototypes en contextes contrôlés, et enfin aux transferts visant l'exploitation des technologies résultantes en milieu industriel. L'organisme s'appuie sur des techniciens, ingénieurs et chercheurs qui, avec l'aide d'équipes supports efficaces, ont mené des développements, souvent pendant de nombreuses années, pour conduire à des impacts sur nos vies de tous les jours. Le CEA a été particulièrement prolifique sur des technologies critiques ou orientés par des applications critiques, de par ses missions initiales. Celles-ci impactent encore aujourd'hui son mode d'organisation interne. Dans un tel contexte, l'utilisation de l'apprentissage statistique pose d'énormes défis. Mon projet de recherche s'appuie sur quelques-unes des forces du CEA :

- Ancrage des recherches sur quelques applications clés du CEA[17], notamment le contrôle, le monitoring et le diagnostic d'infrastructures critiques,
- Accès à des données réelles spécifiques exploitant des technologies de capteurs maîtrisées par des experts en interne,
- Accès à certaines filières privilégiées pour la valorisation des travaux (rails, nucléaires, aéronautique),

pour tenter de relever les défis associés à l'utilisation de composants entrainés par apprentissage statistique dans ces domaines. Le type d'applications envisagé en priorité nécessite en effet de nombreuses recherches afin de fournir l'arsenal méthodologique permettant d'identifier, de mesurer, d'atténuer ou d'annuler les risques liés à l'utilisation de l'apprentissage statistique. Nous pensons de plus que cet arsenal pourra bénéficier à de nombreuses applications au-delà des applications visées en priorité. Les exigences des applications visées ont en effet le potentiel de créer une saine émulation entre les aspects applicatifs et méthodologiques.

Parmi les différents verrous envisageables dans le contexte énoncé, nos travaux se positionneront dans un premier temps sur les aspects suivants :

Q1 / 6.2.1	Quels mécanismes de collaboration en apprentissage statistique pour les applications critiques, exigeant des garanties

[17] L'objectif n'est pas de se restreindre absolument à certaines applications, mais plutôt de profiter des applications maîtrisées en local pour identifier les verrous les plus critiques et bénéficier d'une avance pour proposer de nouvelles approches pour ces verrous, en étant capables de mesurer clairement l'impact des méthodes proposées.

	importantes sur la robustesse, la confidentialité des données ou encore la robustesse aux attaques ?
Q2 / 6.2.2	Quelles sont/seraient les conséquences de l'introduction de composants entrainés par apprentissage statistique dans les systèmes critiques, du point de vue de la cybersécurité ?
Q3 / 6.2.3 et 6.2.4	Quelles méthodologies, accompagnées de garanties théoriques ou pratiques, peuvent fournir à des décideurs, législateurs ou utilisateurs des moyens de prioriser ou hiérarchiser des critères de coûts lors de l'entraînement ou de l'exploitation des systèmes de décision autonome basés sur l'apprentissage statistique ?
Q4 / 6.2.5	Comment articuler la personnalisation et la collaboration entre entités dans des régimes où peu de données sont étiquetées et de lourdes contraintes de confidentialité sont présentes ?

6.2.1 Au-delà de l'apprentissage fédéré

Les paradigmes d'apprentissage classiques impliquent généralement une entité unitaire, qui utilise un processus de génération de données ou un ensemble de données statiques pour apprendre un modèle mathématique à travers les étapes successives d'un algorithme d'optimisation spécialisé. Cette approche présente des inconvénients importants, à savoir l'existence d'un propriétaire de modèle unique et la nécessité, pour le propriétaire du modèle, de collecter des données potentiellement sensibles au cours du processus d'apprentissage. Des tendances récentes soutiennent la nécessité de développer des approches d'apprentissage innovantes entièrement décentralisées. En particulier, cela ouvrirait la voie à l'autonomisation des utilisateurs concernant les services cruciaux basés sur l'apprentissage automatique actuellement mis à disposition par quelques entreprises américaines et reposant sur l'exploitation de données sensibles des utilisateurs.

Le concept d'IA fédérée propose une stratégie de centralisation de l'évolution locale des modèles dans un modèle global qui est ensuite redistribué à tous les systèmes. Cette vision permet aux systèmes centralisés de bénéficier des efforts des acteurs locaux et de rester maîtres du modèle final et de sa redistribution. Dans REDEEM[18], projet ciblé du PEPR Intelligence Artificielle[19], une approche plus collaborative est visée, permettant une répartition plus large et plus équitable du pouvoir. Elle consiste en un partage de

[18] https://redeem-pepria.github.io/en/
[19] https://www.pepr-ia.fr/en/accueil-english/

connaissances entre pairs (protocole pair à pair). Cette vision de l'IA distribuée est intéressante pour les utilisateurs, mais elle doit être étudiée en prenant en compte les enjeux majeurs de sécurité, de robustesse et de personnalisation : comment garantir la conformité d'un modèle appris dans un autre contexte ? Comment protéger notre réseau d'IA de l'introduction de connaissances biaisées, malveillantes ou non, voire de fonctions « backdoor » ? Si la mutualisation consiste en une optimisation simultanée, comment s'assurer de la validité d'apports pas toujours explicables ?

Au sein du projet REDEEM, les contributeurs se focaliseront essentiellement sur les approches à base de réseaux de neurones, dont les performances ont été démontrées dans de nombreux domaines. Plusieurs volets sont visés dans le projet : d'une part, de nouvelles approches de distribution pour les réseaux de neurones, à la fois lors des phases d'entraînement et d'inférence. Au fil des différents sujets proposés, le projet vise des approches innovantes dans le cas où les données des individus sont disjointes entre les contributeurs, mais aussi dans le cas où les caractéristiques sont séparées entre les différents participants aux phases d'inférence ou d'apprentissage. D'un point de vue théorique, REDEEM a pour objectif de fournir des bases solides pour les approches proposées, et en particulier dans le cas où des protagonistes malveillants participent aux étapes, et avec l'objectif primordial de respecter au maximum la confidentialité des données utilisées. Au-delà des nouvelles approches de distribution, REDEEM vise des implémentations efficaces, offrant à la communauté des librairies largement accessibles. Afin de valider les approches proposées, REDEEM a enfin prévu des interactions cruciales avec d'autres projets dans des PEPR, notamment Santé Numérique, TASE, Décarbonation de l'industrie, Cloud, cybersécurité et 6G, dans lesquels certains partenaires de REDEEM sont également impliqués. Pour ce faire, ce projet rassemble un consortium d'équipes et de chercheurs complémentaires, avec une expertise principale en apprentissage automatique, en optimisation distribuée, en algorithmes de consensus et en théorie des jeux.

6.2.2 Sécurité des systèmes intégrant des composants entrainés par apprentissage statistique

La volonté d'intégrer des systèmes entrainés par apprentissage statistique dans de nombreux outils et le besoin croissant de sécurité ont eu pour conséquence de focaliser l'attention de plusieurs agences chargées de la sécurité des technologies et des systèmes d'information. L'ANSSI[20], l'ENISA[21] ou encore le NIST[22], agences respectivement française, européenne et

[20] Agence Nationale de la Sécurité des Systèmes d'Information : https://cyber.gouv.fr
[21] European Union Agency for Cybersecurity. https://www.enisa.europa.eu
[22] National Institute of Standards and Technology. https://www.nist.gov

américaine, ont lancé de nombreuses initiatives sur le sujet de la cybersécurité des composants entrainés par apprentissage statistique. Les rapports et recommandations qui en ont découlé permettent de disposer aujourd'hui d'une cartographie complète des risques cyber associés aux IA, selon les étapes du cycle de vie de celles-ci [44], [99], [100]. En complément, MITRE ATLAS™[23] tient à jour une base de connaissance détaillée à ce sujet, et l'ANSSI propose une méthode d'analyse des risques, EBIOS RM, conçue pour la cybersécurité générale, qui s'adapte très bien aux enjeux des algorithmes d'intelligence artificielle[24].

A partir des différents éléments exposés jusqu'à présent dans ce manuscrit, nous défendons le besoin de développer une vision unitaire de la cybersécurité des systèmes intégrant des composants entrainés par apprentissage statistique, combinant la prise en compte des failles, en lien avec les capteurs, la sécurité physique et la sécurité numérique. Cette vision a conduit à proposer le projet KINAITICS[25], que je coordonne. Plusieurs pistes de recherche vont être explorées dans ce projet, à partir de l'analyse de cas d'usage précis. Nous avons par exemple au sein du CEA un ensemble de cartes électroniques pour l'instrumentation, qui embarquent des algorithmes entrainés par apprentissage statistique, s'appuyant sur l'émission et la réception de signaux ultrasonores pour le diagnostic automatisé de rails. Ce système de mesure et de diagnostic distribué est un élément clef de l'infrastructure critique de transport pour assurer la sûreté. L'analyse des composants de ce système a mis en valeur, d'une part des surfaces d'attaques au niveau des composants, d'autre part une augmentation de la surface d'attaque lorsque l'attaquant est capable de combiner des attaques multi-vecteurs. Par exemple, un monitoring comportemental du système permet d'inférer des informations sur les modèles de décision, qui ensuite peuvent être utilisées pour lancer des attaques adverses plus efficaces sur les modèles statistiques intégrée dans le système. La problématique illustrée ici englobe les questions associées à la sécurité des systèmes intégrant des composants entrainés par apprentissage statistique, dans un monde dans lequel les interconnexions entre les volets numériques et physiques sont de plus en plus importantes : les attaques sur le monde physique peuvent avoir un impact sur le numérique (par exemple une attaque adverse), et des attaques numériques sont de plus en plus ciblées pour engendrer des conséquences néfastes sur le monde physique.

[23] https://atlas.mitre.org

[24] J'ai participé en 2022-2023 à un groupe de travail du Campus Cyber pour l'application de la méthodologie EBIOS RM au cas d'un système intégrant une IA.

[25] https://kinaitics.eu

6.2.3 Compromis théoriques en apprentissage fédéré et décentralisé

Plusieurs résultats existent aujourd'hui pour comprendre les intrications existant entre différentes propriétés, souvent liées à la confiance des systèmes, de l'apprentissage statistique dans des contextes distribués [101], [102], [103]. Souvent dans le cadre de deux propriétés, parfois trois, une approche plus systématique des compromis théoriques entre des propriétés de systèmes apprenants semble nécessaire pour avancer vers des systèmes dont les exigences augmentent de plus en plus de la part des citoyens et des décideurs. Ayant contribué à la compréhension de différents compromis, par exemple entre performance et robustesse aux attaques adverses, ou encore performance et respect de la vie privée, nous estimons que la recherche de ponts entre ces notions, comme nous l'avons proposé dans [W1] doit être une priorité dans les années à venir.

6.2.4 Compromis pratiques de l'apprentissage décentralisé

Au-delà des ponts théoriques entre différentes notions associées aux systèmes d'apprentissage statistiques, nous voyons émerger aujourd'hui plusieurs enjeux sociétaux majeurs autour de la diffusion des composants entrainés par apprentissage statistique. Parmi les plus importants, nous avons identifié les enjeux énergétiques, les enjeux de contrôle et de maîtrise (pas uniquement par le développeur mais surtout par le citoyen, le décideur ou le législateur), et les enjeux de confiance. Si les résultats théoriques pourront amener une partie des réponses, le chercheur en apprentissage statistique doit aussi prendre en compte le besoin de fournir des moyens intelligibles de fixer les compromis selon les contraintes d'utilisation et les exigences des utilisateurs. Ce sujet n'est pas simple à aborder car nous savons aujourd'hui que ce sont les systèmes les plus simples, dans lesquels la plupart des compromis et des paramètres sont cachés, qui sont les plus susceptibles de rencontrer une forte adoption. Des efforts importants sont donc nécessaires afin de trouver le moyen de satisfaire les exigences mentionnées ci-dessus.

6.2.5 Personnalisation et collaboration

Un des enjeux de l'apprentissage décentralisé est de gérer efficacement l'articulation entre le partage de modèles permettant de bénéficier de connaissances provenant de nombreux contextes et le risque de mélanger des modèles issus de distributions hétérogènes. Ce compromis apparaît de manière directe dans l'apprentissage fédéré, mais il impacte en réalité tous les modèles statistiques et leur pouvoir de généralisation à de nouvelles données. Nous avons commencé à aborder cette question dans la thèse de Fabiola Espinoza Castellon, en collaboration avec les travaux de Eduardo Fernandes Montesuma dans [C1] à partir

d'une approche basée sur des dictionnaires et des techniques de transport optimal. De nombreux autres travaux sont nécessaires pour gérer de manière efficace ce compromis dans des cas plus variés.

6.2.6 La place des applications

Mes ambitions aujourd'hui, en tant que responsable d'un laboratoire d'une trentaine de personnes au CEA, sont de conduire, d'initier ou de susciter des recherches qui sont susceptibles d'avoir un impact applicatif ou sociétal à court ou moyen terme. Mon expérience personnelle m'a convaincu que la recherche totalement guidée n'est pas une approche qui maximise la qualité des résultats obtenus. A l'inverse j'ai observé qu'à l'échelle d'un laboratoire, un cadre de recherche trop large, accompagné d'un rythme effréné des cycles de recherche (grand nombre de publications, obsolescence rapide des sujets de recherche) conduisait à une situation de stress pour un nombre non-négligeable de chercheurs. Ces observations, en complément du contexte réglementaire décrit précédemment, et des modèles de valorisation qui se dessinent pour l'utilisation de l'apprentissage statistique, m'ont conduit à remettre une place importante aux applications dans mes recherches futures, en accord avec la légitimité et les atouts du CEA. Cette volonté s'est accompagnée d'un positionnement suivant les principes suivants :

- Couvrir au maximum la chaine de valeur de la donnée, en particulier par les capacités à produire des systèmes de mesures : nous avons été témoins dans de nombreux domaines d'une « unreasonable effectiveness of data » [98], par exemple en transcription de la parole. Il est fondamental de positionner les actions dans des domaines où la production de la donnée est maitrisée, et dans lesquels une filière industrielle existe pour intégrer et utiliser les résultats des recherches conduites. Le CEA a une position de choix dans cette optique dans certains domaines particuliers, notamment dans les technologies de mesures de rayonnements ionisants ainsi que du contrôle non-destructif. Ces domaines constituent pour le laboratoire un ancrage applicatif tirant les exigences algorithmiques.
- Accès aux experts métiers : il est primordial d'avoir de plus accès aux experts qui maîtrisent le système de production de la donnée, qui savent interpréter les mesures, et sont capables d'identifier les verrous applicatifs actuels.
- Positionnement méthodologique : nous avons de plus fait le choix, au vu des expériences acquises ces dernières années, de focaliser les efforts de développement méthodologique sur les techniques d'apprentissage collaboratif et de mesure distribuée.

J'ai aujourd'hui la chance d'être dans un laboratoire qui combine ces aspects de manière remarquable. Plusieurs ingénieurs-chercheurs en électronique analogique/numérique, et systèmes embarqués ont intégré le laboratoire et développent actuellement nos capacités à produire nos propres systèmes de mesure. Le laboratoire est intégré dans un service ancré historiquement dans le contrôle industriel. Nous évoluons donc au contact des experts de ces techniques et pouvons aiguiller nos recherches en étant à l'écoute de leurs problématiques scientifiques. Nous pensons qu'avec la concurrence actuelle dans le domaine de l'apprentissage statistique, il est nécessaire de réduire la portée de nos recherches pour maximiser leur impact. Néanmoins, cette approche ne signifie pas que nos ambitions en apprentissage statistique sont réduites. Il convient de maintenir une excellence dans ce domaine, en disposant de barrières à l'entrée pour l'accès aux données, et en projetant plus facilement les résultats de nos recherches vers des technologies et des filières maitrisées.

6.3 Conclusion

Nous conclurons ce manuscrit en rendant plus explicite la raison de la présence d'un paragraphe sur l'épistémologie de l'intelligence artificielle. Dans son volet tourné vers un projet technologique, l'intelligence artificielle s'est écartée d'une approche classique de la construction de la connaissance basée sur les mesures et les corpus méthodologiques. Nous souhaitons orienter une partie de nos recherches afin de tenter de faire une synthèse entre la rigueur de l'approche classique, et la précision apportée par des techniques de plus en plus avancées issues de la vague la plus récente de l'apprentissage statistique. A court terme, cette synthèse sera tentée en mettant une partie de nos travaux en apprentissage statistique au service de l'instrumentation.

7 Parcours académique

7.1 Présentation du profil

De formation pluridisciplinaire (école d'ingénieur en électronique, traitement du signal et des images, master recherche en sciences cognitives), j'ai commencé mon parcours académique en 2006 avec une thèse sur les interfaces cerveau-machines sous la direction de Christian Jutten et Marco Congedo. Ma thèse s'est focalisée sur la conception d'algorithmes asynchrones [Th1], *i.e.* en l'absence d'un indicateur spécifiant l'instant exact de l'activité cérébrale recherchée, pour le traitement des signaux électroencéphalographiques (EEG) afin d'en extraire des commandes à destination d'un robot ou d'un ordinateur. Après un postdoctorat au Laboratoire de Recherche en Informatique (LRI)[26] sous la direction de Michèle Sebag et Anthony Larue, j'ai rejoint le Commissariat à l'énergie atomique et aux énergies alternatives (CEA) en 2010 dans le but de développer des algorithmes de traitement du signal et d'apprentissage statistique pour des applications variées, dans le domaine de l'énergie, du manufacturing, de la cybersécurité ou encore du traitement de données biomédicales. Visant des approches robustes et efficaces dans divers cas d'applications réels, mes activités de recherche se sont largement orientées vers l'exploration de compromis sous-jacents à la conception et l'utilisation d'algorithmes d'apprentissage statistique. J'ai principalement exploré les axes suivants : la parcimonie grâce à l'utilisation de dictionnaires surcomplets, les statistiques robustes pour la détection de ruptures dans les signaux multivarié, les algorithmes approximatifs pour le traitement de flux de données ou la robustesse aux attaques adverses, et plus récemment différentes pistes pour concilier l'apprentissage statistique avec la protection de la confidentialité des données et des modèles.

7.2 Situation professionnelle actuelle

Je suis actuellement ingénieur-chercheur en contrat à durée indéterminée (CDI) au sein du CEA Paris-Saclay. Je suis responsable du Laboratoire Instrumentation Intelligente Distribuée et Embarquée (LIIDE), qui est un laboratoire de l'institut LIST[27] de la Direction de la Recherche Technologique (DRT) du CEA. La quotité de temps de travail consacrée à la recherche est d'environ 80 %, répartie de manière variable entre l'encadrement de

[26] Le Laboratoire de Recherche en Informatique est devenu en 2021 le Laboratoire Interdisciplinaire des Sciences du Numérique (LISN).

[27] Laboratoire d'Intégration des Systèmes et des Technologies.

doctorants, de post-doctorants, la participation à des projets de recherche (avec des partenaires académiques ou des industriels), des activités d'animation scientifique au sein du laboratoire que je dirige, des activités liées à mon rôle d'expert senior du CEA, et enfin des expertises scientifiques et techniques hors du CEA. Dans les 20 % de temps résiduel, mes activités portent sur des aspects administratifs, juridiques, financiers et RH de la vie d'un laboratoire. Ces activités ne sont donc pas strictement orientées vers la recherche mais sont nécessaires pour la mise en place de l'environnement de recherche du laboratoire. Ces activités non comptabilisées en recherche sont : recrutement, gestion des carrières, des compétences et plus généralement des problématiques associées aux ressources humaines ; contributions aux questions liées aux contrats et négociations juridiques, en lien avec les juristes CEA ; suivi financier du laboratoire en collaboration avec les équipes dédiées du CEA LIST ; actions de communication et de valorisation.

7.3 Parcours professionnel depuis le doctorat

Depuis Jan 2023	**Chef de laboratoire** CEA-List, Saclay, France. Laboratoire Instrumentation Intelligente Distribuée et Embarquée (LIIDE), 32 personnes au 1er décembre 2024. *Animation scientifique et technique, veille, valorisation, stratégie scientifique, management.*
Jan 2021 – déc 2022	**Chef de laboratoire** CEA-List, Saclay, France. Laboratoire Intelligence Artificielle et Apprentissage Automatique (LI3A), 15 personnes au 31 décembre 2022. *Animation scientifique et technique, veille, valorisation, stratégie scientifique, management.*
Sep 2021 – août 2025	**Expert senior CEA** Spécialité mathématiques, informatique scientifique, logiciel Intelligence artificielle (IA) systèmes experts, méthodes statistiques IA et traitement du signal, sécurité et confidentialité des IA. *Encadrement, choix technologiques, veille stratégique.*
Sep 2015 – août 2021	**Expert du CEA** Spécialité intelligence artificielle, machine learning et traitement du signal. *Encadrement, choix technologiques, veille stratégique.*
Oct 2010 – déc 2020	**Ingénieur-chercheur CEA**

	CEA-List, Saclay, France. Laboratoire d'Analyse de Données et Intelligence des Systèmes (LADIS). Contributeur et responsable technique de plusieurs projets (académiques ou industriels). *Recherche, contributions techniques, encadrement, chef de projet.*
Oct 2009 – oct 2010	**Post-doctorant** LRI, Université Paris-Sud, France. *Prédiction de crises épileptiques et approches non-supervisées pour les interfaces cerveau-machines sous la responsabilité de Michèle Sebag et Anthony Larue.*

7.4 Formation

Oct 2006 – Oct 2009	**Doctorat** GIPSA-lab, Grenoble, France. Titre : Approches asynchrones pour les interfaces cerveau-machines. Encadrement : Christian Jutten (Directeur de thèse) et Marco Congedo (encadrant). Thèse soutenue le 1er octobre 2009 devant le jury : • Maureen Clerc, *rapportrice* • François Cabestaing, *rapporteur* • Olivier Bertrand, *examinateur* • Jean-Philippe Lachaux, *président* • Marco Congedo, *encadrant* • Christian Jutten, *directeur de thèse* https://www.theses.fr/2009GRE10179
Sept 2005 – juin 2006	**Master 2 recherche en Sciences cognitives** Institut National Polytechnique de Grenoble, France.
Sept 2003 – juin 2006	**Diplôme d'ingénieur** ENSERG (PHELMA), INPG, Grenoble, France.

7.5 Projets et collaborations de recherche

J'ai participé à plusieurs projets de recherche, dont les financements provenaient de sources variées (région, France, Europe). Je résume ci-dessous les plus importants, en précisant mon rôle dans le projet :

- 2010-2011 : projet **DIGIBRAIN**, projet financé dans le cadre du programme digiteo de la région Ile-de-France. Contributeur technique principal. Partenaires LRI, CEA.

- 2011-2013 : projet **subénergie**, différents projets IA pour l'énergie, financement interne CEA. Contributeur technique. Partenariat scientifique entre CEA LIST et CEA LITEN (Chambéry).
- 2012-2015 : projet **eco-fev**, financement européen global de 4.3 millions d'euros. Contributeur technique. 13 partenaires, dont Hitachi, Politecnico di Torino, Technical University of Berlin et CEA. https://www.2zeroemission.eu/research-project/eco-fev/.
- 2014-2018 : projet **SCE (smart city for energy)**, financement ANR via l'IRT SystemX. Responsable technique contribution CEA de 2016 à 2018. Partenaires industriels GE, Alstom notamment, et académiques tels que Centrale-Supélec, IRT, CEA.
- 2018-2021 : projet **StreamOps**, financement DataIA. Coordinateur. Partenaires : UVSQ, hôpital Foch et CEA.
- 2019-2021 : projet **mastermind**, financement DGA. Rôle de chef de projet.
- 2020-2022 : projet **confiance.ai**, financement ANR via IRT SystemX. Montage et coordination technique pour l'un des quatre départements du CEA LIST. https://www.confiance.ai. Au-delà de l'implication technique pour un département et le suivi des travaux réalisés au sein du laboratoire, j'ai aussi assuré de 2020 à 2023 le rôle de suppléant pour le rôle de représentant du CEA LIST au comité de pilotage de confiance.ai.
- 2021-2025 : projet européen **STARLIGHT** (fiabilité de l'intelligence artificielle pour les acteurs de la sécurité). Responsable de lot (WP8 leader: AI-based cybersecurity and protection of Law Enforcement Agency AI solutions). 51 partenaires, budget du projet 17 M€. https://www.starlight-h2020.eu.
- 2022-2025 : projet européen **KINAITICS** (attaques et défenses dans les systèmes cyber-physiques intégrant des composants entrainés par apprentissage statistique). Coordinateur du projet. 7 partenaires, budget du projet 4 M€. https://kinaitics.eu.
- 2023-2027 : projet du PEPR IA, **REDEEM** (Apprentissage machine résilient, décentralisé et respectueux de la vie privée). Co-coordinateur du projet. Partenaires : CEA, INRIA, CNRS, X. Budget du projet : 8 M€.
- 2022-2028 : projet du PEPR Cybersécurité, **SUPERVIZ** (supervision de la sécurité). Responsable de workpackage. https://superviz.inria.fr.
- 2024-2026 : projet européen **SAFE4SOC** (Standard Alert Format Exchange for SOCs). Responsable de lot (composants entrainés par apprentissage statistique & IDMEFv2). 10 partenaires, budget projet 7 M€. IDMEFv2.

Au-delà de ces projets collaboratifs regroupant de nombreux partenaires, j'ai obtenu plusieurs financements de thèses grâce à des appels compétitifs, en particulier un

financement DIGITEO, un demi-financement DGA et 3 financements internes CEA (thèses sélectionnées par le Haut-Commissaire du CEA).

7.6 Expertises et animation de la communauté scientifique

Je suis également impliqué dans des activités d'animation de la communauté scientifique :

- DataIA : expertise de projets de thèses (appel à projets UDOPIA).
- ANR : expertises pour la sélection des projets de recherche.
- Center for Data Science : représentant du CEA LIST au sein du comité exécutif du CDS (réunions de pilotage, reviews, financement d'actions).
- Membre du comité de programme : EGC (2019, 2020, 2021, 2022), ECML PKDD (2020).
- Activités d'expertise d'articles scientifiques : Neural processing letters, Eusipco, journal of neuroscience methods, sensors, Plos One, IEEE Journal of Selected Topics in Signal Processing, IEEE Transactions on signal processing, Clinical neurophysiology, Neural networks, signal processing (elsevier).
- Expertise de projets : Digicosme (appel 2019), ANR (preps 2018), Programme Transverse CEA -- PTC (2018), projets Carnot CEA LETI (2016).
- ICDCS 2024. Membre du comité de programme pour la session « Federated Learning, Analytics, and Deployment ».
- IEEE CSF 2024, Workshop on Security, Privacy and Information Theory, Protect-IT'24 : membre du comité de programme.
- BPI France : expert pour la sélection et les auditions suite à l'appel à projets lancé par BPI France intitulé « Usages de l'intelligence artificielle générative ».

7.7 Activités de recherche interdisciplinaires

J'ai été à l'origine de quelques actions/projets avec des acteurs du monde hospitalier :

- Projet IA et transplantation (2020) : financement par l'association Vaincre la Mucoviscidose, partenariat CEA, hôpital Foch. Rôle : responsable et intervenant technique côté CEA.
- Projet SEPSIS (2020-2025) : FHU impliquant le CEA. Rôle : montage et responsable pour le CEA dans le FHU.

De plus mes activités au CEA se sont aussi largement ouvertes aux collaborations avec les acteurs du monde économique, en particulier dans les secteurs secondaire et tertiaire. Dans ce cadre, mon rôle a évolué au fil des années de contributeur technique sur des projets modestes, à chef de projet/responsable technique sur des projets d'envergure. J'ai eu ainsi l'occasion de participer aux projets suivants :

- Acteur du domaine de l'énergie, 2011-2013 : apprentissage statistique pour la robustesse des capteurs. Contributeur technique.
- **Schneider**, 2014 : veille dans le domaine de la détection d'anomalies. Contributeur technique.
- **Efluid / UEM**, 2016-2018 : apprentissage statistique pour la prédiction de consommation BT et HTA (électrique/gaz). Responsable technique CEA.
- **RATP**, 2019-2020 : apprentissage statistique et monitoring de systèmes critiques industriels. Chef de projet et responsable technique CEA.

7.8 Communications grand public

- Article 2017 pour Clefs du CEA : algorithmes prédictifs, une efficacité déraisonnable. http://www.cea.fr/multimedia/Pages/editions/clefs-cea/big-data.aspx.
- Article 2017 pour The Conversation : les algorithmes prédictifs, enjeux de l'interprétation. https://theconversation.com/les-algorithmes-predictifs-enjeux-de-linterpretation-78411.
- Juin 2018 : participation à l'émission de France culture « la méthode scientifique », 1h co-invité avec Claire Mathieu. https://www.franceculture.fr/emissions/la-methode-scientifique/la-methode-scientifique-du-mercredi-27-juin-2018.
- Juillet 2018 : film France culture sur les algorithmes prédictifs. https://www.youtube.com/watch?v=U_YSGsO5k_0.
- Septembre 2019 : Risques et limites de l'IA. Article pour le magazine préventique co-écrit avec Rafaël Pinot.
- Septembre 2019 : conférence invitée pour le magazine Workplace, introduction à l'IA. https://soundcloud.com/workplace-magazine/conference-inaugurale-etes-vous-sures-de-tout-savoirsur-la-data-par-cedric-gouy-pailler.
- Novembre 2019 : article Clefs du CEA : Attaques adverses, Atténuer les risques. http://www.cea.fr/multimedia/Pages/editions/clefs-cea/intelligence-artificielle.aspx.
- Novembre 2019 : masterclass X/CEA sur l'intelligence artificielle (conférence invitée devant 3 classes de terminales). https://www.youtube.com/watch?v=rLiUNh98m6g.

- Janvier 2020 : film pédagogique sur les algorithmes pour le jeu vidéo “le prisonnier quantique”. https://prisonnier-quantique.fr/index.html.
- Janvier 2020 : AI versus Wild. Démonstration CES 2020 à Las Vegas.
- Juillet 2022 : Cryptographie homomorphe, l’art de partager sans divulguer, article paru dans les cahiers techniques d’usine nouvelle. https://t.co/mzodcpdDnt.
- Novembre 2023 : tout savoir sur le big data. Conférence invitée à la Bibliothèque Publique d’Information (BPI). https://agenda.bpi.fr/evenement/tout-savoir-sur-big-data/.

7.9 Encadrement de doctorants, stagiaires et post-doctorants

Encadrement de doctorants

- **Imane MEDDOUR** (nov. 2024 –), *directeur de thèse*, implication de **30%** Etalonnage fédéré de capteurs pour l’instrumentation. Thèse co-encadrée avec Andréa MACARIO BARROS.
- **Fabiola ESPINOZA CASTELLON** (nov. 2020 – fév. 2024), *directeur de thèse*, implication de **50%** Apprentissage fédéré : personnalisation et collaboration. Thèse co-encadrée avec Aurélien MAYOUE. J’ai obtenu une dérogation de l’université Paris-Saclay en 2020 pour diriger cette thèse sans HDR. https://www.theses.fr/s259235. Soutenue le 6 février 2024.
- **Pierre-Emmanuel CLET** (nov. 2020 – jan 2024), implication de **25%.** Co-conception de réseaux de neurones profonds adaptés au FHE (chiffrement homomorphe). Thèse co-encadrée avec Renaud SIRDEY et Aymen BOUDGUIGA. https://www.theses.fr/s263414. Soutenue le 15 janvier 2024. Ingénieur-chercheur au CEA Paris-Saclay.
- **Arnaud GRIVET-SEBERT** (mar. 2020 – juin 2023), implication de **50%.** Combining differential privacy and homomorphic encryption for privacy-preserving collaborative machine learning. Soutenue le 12 juin 2023. Thèse co-encadrée avec Renaud SIRDEY. https://www.theses.fr/s253047. Post-doctorant à l’école polytechnique.
- **Rafaël PINOT** (oct. 2017 – déc. 2020), implication de **33%.** On the impact of randomization on robustness in machine learning. Soutenue le 2 décembre 2020. Thèse co-encadrée avec Florian YGER et Jamal ATIF (Université Paris-Dauphine). **Prix de thèse de la fondation Dauphine 2021**. Titulaire d’une chaire mathématiques et IA à Sorbonne Université (the mathematical foundation of computer and data science). https://www.theses.fr/2020UPSLD038.

- **Anne MORVAN** (oct. 2015 – nov. 2018), implication de **50%.** Contributions to unsupervised learning from massive high-dimensional data streams : structuring, hashing and clustering. Thèse co-encadrée avec Jamal ATIF (Université Paris-Dauphine). Soutenue le 12 novembre 2018. **Prix de thèse de la fondation Dauphine 2019** et **Prix de thèse de la DGA 2020**. Chercheuse chez Expedia (Genève). https://www.theses.fr/2018PSLED033.
- **Flore HARLÉ** (nov. 2012 – juin 2016), implication de **33%.** Bayesian multiple change-point detection in multivariate time series, soutenue le 21 juin 2016. Thèse co-encadrée avec Sophie ACHARD (CNRS, LJK-INRIA) et Florent CHATELAIN (Grenoble INP, GIPSA-lab). Ingénieure de recherche Carestream. https://www.theses.fr/2016GREAT043.
- **Yoann ISAAC** (nov 2011 – mai 2015), implication de **33%.** Représentations redondantes pour les signaux d'électroencéphalographie, soutenue le 29 mai 2015. Thèse co-encadré avec Michèle SEBAG (Université Paris-Saclay, LRI) et Jamal ATIF (Université Paris-Dauphine). Ingénieur de recherche chez Vidal Group. https://www.theses.fr/2015PA112072.

Stagiaires M2 Recherche

- **Thomas LEBRUN** (2020), Apprentissage fédératif en contexte non-iid. Stage co-encadré avec Aurélien MAYOUE.
- **Rafaël PINOT** (2017), Confidentialité différentielle dans les données structurées par des graphes. Stage co-encadré avec Anne MORVAN, Florian YGER et Jamal ATIF (Université Paris-Dauphine).
- **Dialecti VALSAMOU** (2011), apprentissage non supervisé pour les interfaces cerveau-machines. Stage co-encadré avec Michèle SEBAG.

Post-doctorants

- **Mohammad AL SHAER** (2019–2021) : algorithmes pour le traitement de flux de données massifs. Data scientist chez Linkfluence.
- **Yohan PETETIN** (2013–2015) : on-line bayesian data assimilation for photovoltaic systems. Maître de conférence Télécom SudParis.
- **Xavier ARTUSI** (2013–2014) : leaks and contaminations detection in water distribution networks. Ingénieur-chercheur CEA.
- **Olaf KOUAMO** (2011–2013) : estimation and detection in stochastic processes with applications in electrical vehicles. Chief Data Scientist chez Stellantis.

- **Boujemaa AIT EL FQUIH** (2011–2013) : anomaly detection using online recursive filtering. Chercheur chez King Abdullah University of Science and Technology (KAUST).
- **Anthony MOURAUD** (2010–2012) : spike-based metrics for sparse representations. Responsable Software Engineering & Innovation chez Lhyfe.

7.10 Accompagnement de la formation doctorale

- Intervenant à la journée "Carrières en Signal, Image & Vision" à destination des doctorants (7 mars 2019). J'ai témoigné de mon expérience dans un EPIC auprès de futurs docteurs. La journée était organisée par le GDR ISIS http://www.gdr-isis.fr/index.php?page=reunion&idreunion=34.
- Examinateur dans le jury de thèse de Maroua Bahri (5 juin 2020). http://www.theses.fr/2020IPPAT017.
- Examinateur dans le jury de thèse de Mainak Jas (12 avril 2018). https://www.theses.fr/2018ENST0021.
- Membre du comité de suivi de thèse de Nicolas Aussel (juin 2017), Maroua Bahri (mars 2019), Thibault Allenet (mai 2020), Manon Césaire (juillet 2021), Bastien Vuillod (juin 2024) et Sébastien Patté (juillet 2024).

7.11 Enseignements & interventions pédagogiques

- 2015 – 2023 : Web and Social network analysis. Télécom SudParis, niveau M2. 12 heures par an.
- 2016 – 2017 : Fouille de données. Université Paris-Dauphine, niveau M2. 9 heures.
- 2019 – 2020 : Graph Neural Networks. Institut Polytechnique de Paris, M2. 4 heures.
- 2020 – 2023 : Robustesse des réseaux de neurones face aux attaques adverses. Master 2 SETI. 6 heures par an.
- 2024 : Cybersécurité de l'intelligence artificielle dans le domaine de la santé. Diplôme Universitaire IA et Santé. 1 heure.
- 2024 : MLOps. Certification en intelligence artificielle pour une société de services informatiques. 3h.
- 2024 : MLOps. Chef de projet IA en formation continue. 3h.

- 2024 : Intelligence artificielle et cybersécurité : opportunités et menaces pour la sécurité intérieure. Institut des Hautes Etudes du Ministère de l'Intérieur (IHEMI). Cycle supérieur d'intelligence artificielle. 1h.

7.12 Bibliographie de l'auteur du manuscrit

Les publications sont classées par type de document, puis dans l'ordre chronologique décroissant. La bibliographie de cette partie présente uniquement les articles auxquels j'ai contribué. La bibliographie générale du document est accessible au chapitre 8.

7.12.1 Articles de journaux

[J1] Fessler, J., Gouy-Pailler, C., Finet, M., Zuber, B., Messika, J., Glorion, M., Sage, E., De Wolf, J., Roux, A., Brugière, O., et al. (2024). Prediction of primary graft dysfunction after double-lung transplantation: a machine learning approach. Soumission à Journal of Heart and Lung Transplantation.

[J2] Pinot, R., Meunier, L., Yger, F., Gouy-Pailler, C., Chevaleyre, Y., and Atif, J. (2022). On the robustness of randomized classifiers to adversarial examples. Mach Learn 111 (9), 3425-3457. Lien url.

[J3] Grivet Sébert, A., Pinot, R., Zuber, M., Gouy-Pailler, C., and Sirdey, R. (2021). SPEED: secure, PrivatE, and efficient deep learning. Mach Learn 110, 675–694. Lien url.

[J4] Isaac, Y., Barthélemy, Q., Gouy-Pailler, C., Sebag, M. & Atif, J (2017). Multi-dimensional signal approximation with sparse structured priors using split Bregman iterations. *Signal Processing* **130**, 389–402. Lien url.

[J5] Harlé, F., Chatelain, F., Gouy-Pailler, C. & Achard, S. (2016) Bayesian Model for Multiple Change-Points Detection in Multivariate Time Series. *IEEE Transactions on Signal Processing* **64**, 4351–4362. Lien url.

[J6] Sameni, R. & Gouy-Pailler, C. (2014). An iterative subspace denoising algorithm for removing electroencephalogram ocular artifacts. *Journal of Neuroscience Methods* **225**, 97–105. Lien url.

[J7] Barthélemy, Q., Gouy-Pailler, C., Isaac, Y., Souloumiac, A., Larue, A., and Mars, J.I. (2013). Multivariate temporal dictionary learning for EEG. Journal of Neuroscience Methods 215, 19–28. Lien url.

[J8] Gouy-Pailler, C., Sebag, M., Larue, A. & Souloumiac, A. (2011). Single trial variability in brain–computer interfaces based on motor imagery: Learning in the presence of labeling noise. *Int. J. Imaging Syst. Technol.* **21**, 148–157. Lien url.

[J9] Gouy-Pailler, C., Congedo, M., Brunner, C., Jutten, C. & Pfurtscheller, G. (2010). Nonstationary Brain Source Separation for Multiclass Motor Imagery. *IEEE Transactions on Biomedical Engineering* **57**, 469–478. Lien url.

[J10] Congedo, M., Gouy-Pailler, C. & Jutten, C. (2008). On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics. *Clinical Neurophysiology* **119**, 2677–2686. Lien url.

7.12.2 Conférences internationales avec actes et comité de lecture

[C1] Espinoza Castellon, F., Fernandes Montesuma, E., Ngolè Mboula, F., Mayoue, A., Souloumiac, A., and Gouy-Pailler, C. (2024). Federated Dataset Dictionary Learning for Multi-Source Domain Adaptation. In proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024).

[C2] Grivet-Sébert, A., Checri, M., Stan, O., and Sirdey, R. (2023). Combining homomorphic encryption and differential privacy in federated learning. In Proceedings of the 20th Annual International Conference on Privacy, Security & Trust (PST 2023).

[C3] Grivet Sébert, A., Zuber, M., Stan, O., Sirdey, R., and Gouy-Pailler, C. (2023). A Probabilistic Design for Practical Homomorphic Majority Voting with Intrinsic Differential Privacy. In Proceedings of the 11th Workshop on Encrypted Computing & Applied Homomorphic Cryptography WAHC'23. (Association for Computing Machinery).

[C4] Espinoza Castellon, F., Singh, D., Mayoue, A., and Gouy-Pailler, C. (2023). FUBA: Federated Uncovering of Backdoor Attacks for Heterogeneous Data. In Proceedings of the Fifth IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications.

[C5] Espinoza Castellon, F., Mayoue, A., Sublemontier, J.-H., and Gouy-Pailler, C. (2022). Federated learning with incremental clustering for heterogeneous data. In 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. 10.1109/IJCNN55064.2022.9892653.

[C6] Madi, A., Stan, O., Sirdey, R., and Gouy-Pailler, C. (2022). SecTL: Secure and Verifiable Transfer Learning-based inference. In 8th International Conference on Information Systems Security and Privacy (ICISSP 2022), pp. 220--229.

[C7] Madi, A., Stan, O., Mayoue, A., Grivet-Sébert, A., Gouy-Pailler, C., and Sirdey, R. (2021). A Secure Federated Learning framework using Homomorphic Encryption and Verifiable Computing. In 2021 Reconciling Data Analytics, Automation, Privacy, and Security: A Big Data Challenge (RDAAPS), pp. 1–8.

[C8] Al Shaer, M., Garcia Rodriguez, S., and Gouy-Pailler, C. (2020). Detecting Anomalies from Streaming Time Series using Matrix Profile and Shapelets Learning. In 32nd International Conference on Tools with Artificial Intelligence (ICTAI'20), virtual event.

[C9] Garcia Rodriguez, S., Al Shaer, S., and Gouy-Pailler, C. (2020). STREAMER: A Powerful Framework for Continuous Learning in Data Streams. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), virtual event.

[C10] Azzabi, R., Gouy-Pailler, C., Valley, F., and Dubois, H. (2020). Visualization and machine learning for interactive cyber threats analysis in critical infrastructures. International Conference on Nuclear Security, ICONS 2020. Vienna, Austria, March 2020 (2020).

[C11] Pinot, R., Meunier, L., Araujo, A., Hisashi K., Yger F., Gouy-Pailler, C. & Atif, J. Theoretical evidence for adversarial robustness through randomization. in Proceedings of the thirty-third Conference on Neural Information Processing Systems, NIPS 2019, Vancouver, Canada, December 2019 (2019).

[C12] Fessler, J., Vallee, A., Gouy-Pailler, C., Davignon, M., Fischler, M., and Le Guen, M. (2021). Machine-Learning for Primary Graft Dysfunction in Lung Transplantation. The Journal of Heart and Lung Transplantation 40, S380.

[C13] Stan, O., Sirdey, R., Gouy-Pailler, C., Blanchart, P., BenHamida, A., and Zayani, M.-H. (2018). Privacy-Preserving Tax Calculations in Smart Cities by Means of Inner-Product

Functional Encryption. In 2018 2nd Cyber Security in Networking Conference (CSNet), pp. 1–8.

[C14] Pinot, R., Morvan, A., Yger, F., Gouy-Pailler, C. & Atif, J. Graph-based clustering under differential privacy. In Proceedings of the thirty-fourth conference on uncertainty in artificial intelligence, UAI 2018, monterey, california, USA, august 2018 329–338 (2018).

[C15] Morvan, A., Souloumiac, A., Gouy-Pailler, C. & Atif, J. Streaming binary sketching based on subspace tracking and diagonal uniformization. In 2018 IEEE international conference on acoustics, speech and signal processing, ICASSP 2018, 2421–2425 (2018).

[C16] Morvan, A., Choromanski, K., Gouy-Pailler, C. & Atif, J. Graph sketching-based space-efficient data clustering. In SIAM international conference on data mining, SDM 2018, 10–18 (2018).

[C17] Bojarski, M. et al. Structured adaptive and random spinners for fast machine learning computations. In Proceedings of the 20th international conference on artificial intelligence and statistics, AISTATS 2017, 54, 1020–1029 (2017).

[C18] Blanchart, P. & Gouy-Pailler, C. WHODID: web-based interface for human-assisted factory operations in fault detection, identification and diagnosis. In Joint european conference on machine learning and knowledge discovery in databases, ECML PKDD 2017, 437–441 (2017).

[C19] Harlé, F., Chatelain, F., Gouy-Pailler, C. & Achard, S. Rank-based multiple change-point detection in multivariate time series. In 2014 22nd european signal processing conference, EUSIPCO 2014, 1337–1341 (2014).

[C20] Kouamo, O. & Gouy-Pailler, C. Multi-scale test procedure for non-stationarity in short and long memory time series. In 2013 IEEE international conference on acoustics, speech and signal processing, ICASSP 2013, 5368–5372 (2013).

[C21] Isaac, Y., Barthelemy, Q., Atif, J., Gouy-Pailler, C. & Sebag, M. Multi-dimensional sparse structured signal approximation using split bregman iterations. In IEEE international conference on acoustics, speech and signal processing, ICASSP 2013, 3826–3830 (2013).

[C22] Ait-El-Fquih, B. & Gouy-Pailler, C. Backward hidden Markov chain for outlier-robust filtering and fixed-interval smoothing. In IEEE international conference on acoustics, speech and signal processing, ICASSP 2013, 5504–5508 (2013).

[C23] Mouraud, A. et al. From Neuronal cost-based metrics towards sparse coded signals classification. In proceedings of the 20th European Symposium on Artificial Neural Networks, ESANN 2012, Bruges, Belgium (2012).

[C24] Gouy-Pailler, C. et al. Distance and similarity measures for sensors selection in heavily instrumented buildings: application to the INCAS platform. In 28th international conference of CIB w78 (2011).

[C25] Chatelain, F., Achard, S., Michel, O. & Gouy-Pailler, C. Multivariate approach for brain decomposable connectivity networks. In IEEE statistical signal processing workshop, SSP 2011, 817–820 (2011).

[C26] Gouy-Pailler, C., Sebag, M., Larue, A. & Souloumiac, A. SABIN: a resampling-based learning algorithm for idle state identification in asynchronous brain-computer interfaces. In first workshop on brain decoding: pattern recognition challenges in neuroimaging 1–4 (2010).

[C27] Gouy-Pailler, C., Mattout, J., Congedo, M. & Jutten, C. Uncued brain-computer interfaces: a variational hidden markov model of mental state dynamics. In 17th european symposium on artificial neural networks, ESANN 2009, Bruges, Belgium, april 22-24, 2009, proceedings (2009).

[C28] Gouy-Pailler, C., Sameni, R., Congedo, M. & Jutten, C. Iterative subspace decomposition for ocular artifact removal from EEG recordings. In International conference on independent component analysis and signal separation, ICA/LVA 2009, 419–426 (2009).

[C29] Gouy-Pailler, C., Congedo, M., Jutten, C., Brunner, C. & Pfurtscheller, G. Model-based source separation for multi-class motor imagery. In 16th european signal processing conference, EUSIPCO 2008, 1–5 (2008).

[C30] Gouy-Pailler, C., Zijp-Rouzier, S., Vidal, S. & Chêne, D. A haptic based interface to ease visually impaired pupils' inclusion in geometry lessons. In International conference on universal access in human-computer interaction, HCI 2007, 598–606 (2007).

[C31] Gouy-Pailler, C. et al. Topographical dynamics of brain connections for the design of asynchronous brain-computer interfaces. In 29th annual international conference of the IEEE Engineering in medicine and biology society, EMBC 2007, 2520–2523 (2007).

7.12.3 Conférences nationales avec actes et comité de lecture

[Cn1] Espinoza Castellon, F., Singh, D., Mayoue, A., and Gouy-Pailler, C. (2023). Défense contre les attaques par porte dérobée en apprentissage fédéré par estimation du motif d'attaque et élagage. In XXIXème Colloque Francophone de Traitement du Signal et des Images.

[Cn2] Pinot, R., Morvan, A., Yger, F., Gouy-Pailler, C. & Atif, J. Graph-based Clustering under Differential Privacy. in *Actes de la conférence sur l'apprentissage automatique (Cap 2019)* (2019).

[Cn3] Isaac, Y., Barthélemy, Q., Gouy-Pailler, C., Atif, J., and Sebag, M. (2015). Généralisation des micro-états EEG par apprentissage régularisé temporellement de dictionnaires topographiques. In XXVème Colloque Francophone de Traitement du Signal et des Images.

[Cn4] Harlé, F., Chatelain, F., Gouy-Pailler, C., and Achard, S. (2015). Utilisation de la vraisemblance empirique pour un test d'homogénéité. In XXVème Colloque Francophone de Traitement du Signal et des Images.

[Cn5] Barachant, A., Cycon, R., and Gouy-Pailler, C. (2015). P300-speller: Géométrie Riemannienne pour la détection multi-sujets de potentiels d'erreur. In XXVème Colloque Francophone de Traitement du Signal et des Images.

[Cn6] Isaac, Y., Barthélemy, Q., Atif, J., Gouy-Pailler, C., and Sebag, M. (2013). Régularisations spatiales pour la décomposition de signaux EEG sur un dictionnaire temps-fréquence. In XXIVème Colloque Francophone de Traitement du Signal et des Images.

[Cn7] Chatelain, F., Achard, S., Gouy-Pailler, C., Michel, O.J.J., and Amblard, P.-O. (2011). Graphe de connectivité cérébrale et longue dépendance. In XXIIIème Colloque Francophone de Traitement du Signal et des Images.

[Cn8] Lemoine, J., Gouy-Pailler, C., Achard, S., and Amblard, P.-O. (2009). Recherche de la connectivité de réseaux complexes. Application en fMRI. In XXIIème Colloque Francophone de Traitement du Signal et des Images, pp. 1–4.

[Cn9] Gouy-Pailler, C., Rivet, B., Achard, S., Souloumiac, A., Jutten, C., Maby, E., and Congedo, M. (2007). Théorie des graphes et dynamique des connexions cérébrales pour la conception d'interfaces cerveau-machines asynchrones. In XXIème Colloque Francophone de Traitement du Signal et des Images, pp. 1–4.

7.12.4 Brevets

[B1] Ait El Fquih, B., Gouy-Pailler, C., and Guillemin, S. (2017). Management of the recharging of the battery of an electric vehicle.

[B2] Chaintreuil, N., Gouy-Pailler, C., Lespinats, S., and Plissonnier, A. (2016). Method and device for detecting an electric arc in a photovoltaic installation.

[B3] Gouy-Pailler, C., and Chaintreuil, N. (2014). Method and device for detecting electric arc in a photovoltaic installation.

7.12.5 Manuscrit de thèse

[Th1] Gouy-Pailler, C., « Vers une modélisation dynamique de l'activité cérébrale pour la conception d'interfaces cerveau-machines asynchrones », PhD Thesis, University of Grenoble, 2009.

7.12.6 Divers

[W1] Pinot, R., Yger, F., Gouy-Pailler, C., et Atif, J., « A unified view on differential privacy and robustness to adversarial examples », in *Workshop on Machine Learning for CyberSecurity at ECMLPKDD 2019*, 2019.

[W2] Morvan, A., Souloumiac, A., Choromanski, K., Gouy-Pailler, C., et Atif, J., « On the Needs for Rotations in Hypercubic Quantization Hashing », ArXiv180203936 Cs, févr. 2018. Disponible sur: http://arxiv.org/abs/1802.03936

[W3] Sirdey, R., Grivet Sébert, A., et Gouy-Pailler, C., « [Cahier technique] Cryptographie homomorphe : l'art de partager sans divulguer », *Florent Robert - Ind. Technol.*, 2022.

8 Bibliographie générale

[1] L. Clissa, « Survey of Big Data sizes in 2021 », *Front. Big Data*, vol. 6, p. 1271639, oct. 2023, doi: 10.3389/fdata.2023.1271639.

[2] F. Pedregosa *et al.*, « Scikit-learn: Machine learning in Python », *J. Mach. Learn. Res.*, vol. 12, p. 2825-2830, 2011.

[3] R. Kitchin et G. McArdle, « The Diverse Nature of Big Data », 18 septembre 2015, *Rochester, NY*: 2662462. doi: 10.2139/ssrn.2662462.

[4] R. Pinot, « On the impact of randomization on robustness in machine learning », Thèse de doctorat, Université Paris sciences et lettres, 2020. Consulté le: 4 janvier 2024. [En ligne]. Disponible sur: https://hal.science/tel-03121555

[5] A. Grivet Sébert, « Combining differential privacy and homomorphic encryption for privacy-preserving collaborative machine learning », Thèse de doctorat, Université Paris-Saclay, 2023. Consulté le: 4 janvier 2024. [En ligne]. Disponible sur: https://theses.hal.science/tel-04223076

[6] P.-E. Clet, « Contributions to the optimization of TFHE's functional bootstrapping for the evaluation of non-polynomial operators », Thèse de doctorat, Université Paris-Saclay, 2024. Consulté le: 25 février 2024. [En ligne]. Disponible sur: https://theses.hal.science/tel-04431993

[7] F. Espinoza Castellon, « Contributions à un apprentissage fédéré efficace et sécurisé avec des données client hétérogènes », Thèse de doctorat, université Paris-Saclay, 2024. Consulté le: 25 février 2024. [En ligne]. Disponible sur: https://www.theses.fr/s259235

[8] A. Morvan, « Contributions to unsupervised learning from massive high-dimensional data streams : structuring, hashing and clustering », Thèse de doctorat, Université Paris sciences et lettres, 2018. Consulté le: 3 janvier 2024. [En ligne]. Disponible sur: https://theses.hal.science/tel-01982476

[9] M. K. Islam, A. Rastegarnia, et Z. Yang, « Methods for artifact detection and removal from scalp EEG: A review », *Neurophysiol. Clin. Neurophysiol.*, vol. 46, n° 4, p. 287-305, nov. 2016, doi: 10.1016/j.neucli.2016.07.002.

[10] M. Fatourechi, A. Bashashati, R. K. Ward, et G. E. Birch, « EMG and EOG artifacts in brain computer interface systems: A survey », *Clin. Neurophysiol.*, vol. 118, n° 3, p. 480-494, mars 2007, doi: 10.1016/j.clinph.2006.10.019.

[11] T. P. Jung *et al.*, « Removing electroencephalographic artifacts by blind source separation », *Psychophysiology*, vol. 37, n° 2, p. 163-178, mars 2000.

[12] A. Delorme, T. Sejnowski, et S. Makeig, « Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis », *Neuroimage*, vol. 34, n° 4, p. 1443-1449, févr. 2007, doi: 10.1016/j.neuroimage.2006.11.004.

[13] T. Goldstein et S. Osher, « The Split Bregman Method for L1-Regularized Problems », *SIAM J. Imaging Sci.*, vol. 2, n° 2, p. 323-343, janv. 2009, doi: 10.1137/080725891.

[14] A. Gramfort, T. Papadopoulo, E. Olivi, et M. Clerc, « OpenMEEG: opensource software for quasistatic bioelectromagnetics », *Biomed. Eng. OnLine*, vol. 9, n° 1, p. 45, sept. 2010, doi: 10.1186/1475-925X-9-45.

[15] M. Elad, P. Milanfar, et R. Rubinstein, « Analysis versus synthesis in signal priors », *Inverse Probl.*, vol. 23, n° 3, p. 947, avr. 2007, doi: 10.1088/0266-5611/23/3/007.

[16] Y. Isaac, « Représentations redondantes pour les signaux d'électroencéphalographie », These de doctorat, Paris 11, 2015. Consulté le: 1 janvier 2024. [En ligne]. Disponible sur: https://www.theses.fr/2015PA112072

[17] F. Harlé, « Détection de ruptures multiples dans des séries temporelles multivariées : application à l'inférence de réseaux de dépendance », These de doctorat, Université Grenoble Alpes (ComUE), 2016. Consulté le: 25 février 2024. [En ligne]. Disponible sur: https://www.theses.fr/2016GREAT043

[18] W. Johnson et J. Lindenstrauss, « Extensions of Lipschitz mappings into a Hilbert space », in *Conference in modern analysis and probability (New Haven, Conn., 1982)*, vol. 26, in Contemporary mathematics, vol. 26. , American Mathematical Society, 1984, p. 189-206.

[19] M. Talagrand, « New concentration inequalities in product spaces », *Invent. Math.*, vol. 126, n° 3, p. 505-563, nov. 1996, doi: 10.1007/s002220050108.

[20] A. Andoni, P. Indyk, T. Laarhoven, I. Razenshteyn, et L. Schmidt, « Practical and optimal LSH for angular distance », in *Proceedings of the 28th international conference on neural information processing systems*, in NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 1225-1233.

[21] V. Bentkus, « On the dependence of the Berry–Esseen bound on dimension », *J. Stat. Plan. Inference*, vol. 113, n° 2, p. 385-402, 2003.

[22] P. Indyk et R. Motwani, « Approximate nearest neighbors: Towards removing the curse of dimensionality », in *Proceedings of the thirtieth annual ACM symposium on theory of computing*, in STOC '98. New York, NY, USA: ACM, 1998, p. 604-613.

[23] M. Charikar, « Similarity estimation techniques from rounding algorithms », in *STOC*, 2002.

[24] K. Terasawa et Y. Tanaka, « Spherical LSH for approximate nearest neighbor search on unit hypersphere », in *WADS*, 2007, p. 27-38.

[25] N. Sundaram *et al.*, « Streaming similarity search over one billion tweets using parallel locality-sensitive hashing », *Proc VLDB Endow*, vol. 6, n° 14, p. 1930-1941, sept. 2013.

[26] A. Andoni, P. Indyk, T. Laarhoven, I. Razenshteyn, et L. Schmidt, « Practical and optimal LSH for angular distance », in *NIPS*, 2015, p. 1225-1233.

[27] K. Abed-Meraim, A. Chkeif, et Y. Hua, « Fast orthonormal PAST algorithm », *IEEE Signal Process. Lett.*, n° 3, p. 60-62, 2000.

[28] T. Falkowski, A. Barth, et M. Spiliopoulou, « DENGRAPH: A density-based community detection algorithm », in *Proceedings of the IEEE/WIC/ACM international conference on web intelligence*, in WI '07. Washington, DC, USA: IEEE Computer Society, 2007, p. 112-115.

[29] P. W. Holland, K. B. Laskey, et S. Leinhardt, « Stochastic blockmodels: First steps », *Soc. Netw.*, vol. 5, n° 2, p. 109-137, juin 1983.

[30] A. Condon et R. M. Karp, « Algorithms for graph partitioning on the planted partition model », *Random Struct Algorithms*, vol. 18, n° 2, p. 116-140, mars 2001.

[31] K. Rohe, S. Chatterjee, et B. Yu, « Spectral clustering and the high-dimensional stochastic blockmodel », *Ann. Stat.*, vol. 39, n° 4, p. 1878-1915, août 2011.

[32] K. J. Ahn, S. Guha, et A. McGregor, « Analyzing Graph Structure via Linear Measurements », in *Proceedings of the Twenty-Third Annual ACM-SIAM*

Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, 2012, p. 459-467.

[33] T. Asano, B. Bhattacharya, M. Keil, et F. Yao, « Clustering Algorithms Based on Minimum and Maximum Spanning Trees », in *Proceedings of the Fourth Annual Symposium on Computational Geometry*, in SCG '88. New York, NY, USA: ACM, 1988, p. 252-257. doi: 10.1145/73393.73419.

[34] Y. Xu, V. Olman, et D. Xu, « Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees », *Bioinformatics*, vol. 18, nº 4, p. 536-545, avr. 2002, doi: 10.1093/bioinformatics/18.4.536.

[35] M. Ester, H. Kriegel, J. Sander, et X. Xu, « A density-based algorithm for discovering clusters in large spatial databases with noise », AAAI Press, 1996, p. 226-231.

[36] N. R. Adam et J. C. Worthmann, « Security-control methods for statistical databases: a comparative study », *ACM Comput. Surv. CSUR*, vol. 21, nº 4, p. 515-556, 1989.

[37] D. E. Denning, « Secure Statistical Databases with Random Sample Queries », *ACM Trans Database Syst*, vol. 5, nº 3, p. 291-315, sept. 1980.

[38] S. Goldwasser et S. Micali, « Probabilistic encryption », *J. Comput. Syst. Sci.*, vol. 28, nº 2, p. 270-299, 1984.

[39] A. Narayanan et V. Shmatikov, « Robust de-anonymization of large sparse datasets », in *IEEE Symposium on Security and Privacy*, 2008.

[40] European Parliament et European Council, « Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC », European Parliament and European Council, 2016.

[41] B. Fung, K. Wang, R. Cheng, et P. Yu, « Privacy-preserving data publishing: A survey of recent developments », *Acm Comput. Surv.*, vol. 42, nº 4, p. 14:1-14:53, juin 2010.

[42] C. Dwork, F. McSherry, K. Nissim, et A. Smith, « Calibrating Noise to Sensitivity in Private Data Analysis », in *Theory of Cryptography*, Springer Berlin Heidelberg, 2006, p. 265-284.

[43] C. Dwork et A. Roth, « The algorithmic foundations of differential privacy », *Found. Trends® Theor. Comput. Sci.*, vol. 9, nº 3–4, p. 211-407, 2014.

[44] ENISA, « Securing Machine Learning Algorithms », ENISA. Consulté le: 4 janvier 2024. [En ligne]. Disponible sur: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

[45] R. Pinot, « Minimum spanning tree release under differential privacy constraints », *ArXiv180106423 Cs Math Stat*, janv. 2018, Consulté le: 8 février 2018. [En ligne]. Disponible sur: http://arxiv.org/abs/1801.06423

[46] O. Michel, A. Hero, et P. Flandrin, « Graphes de représentation minimaux, entropies et divergences: applications », *Trait. Signal*, vol. 17, nº 4, p. 287-297, 2000.

[47] I. Csiszár et J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[48] I. Mironov, « Rényi differential privacy », in *30th IEEE computer security foundations symposium, CSF 2017, santa barbara, CA, USA, august 21-25, 2017*, 2017, p. 263-275.

[49] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, et C. Palamidessi, « Broadening the scope of differential privacy using metrics », in *Privacy enhancing technologies*, E. De Cristofaro et M. Wright, Éd., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, p. 82-102.

[50] B. Biggio *et al.*, « Evasion Attacks Against Machine Learning at Test Time », in *Proceedings of the 2013th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III*, in ECMLPKDD'13. Berlin, Heidelberg: Springer-Verlag, 2013, p. 387-402. doi: 10.1007/978-3-642-40994-3_25.

[51] C. Szegedy *et al.*, « Intriguing properties of neural networks », in *International conference on learning representations*, 2014.

[52] B. McMahan, E. Moore, D. Ramage, S. Hampson, et B. A. y Arcas, « Communication-efficient learning of deep networks from decentralized data », in *Artificial intelligence and statistics*, PMLR, 2017, p. 1273-1282.

[53] T. Gu, K. Liu, B. Dolan-Gavitt, et S. Garg, « Badnets: Evaluating backdooring attacks on deep neural networks », *IEEE Access Pract. Innov. Open Solut.*, vol. 7, p. 47230-47244, 2019.

[54] M. Fredrikson, S. Jha, et T. Ristenpart, « Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures », in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, ACM Press, 2015.

[55] B. Wang *et al.*, « Neural cleanse: Identifying and mitigating backdoor attacks in neural networks », in *2019 IEEE symposium on security and privacy (SP)*, IEEE, 2019, p. 707-723.

[56] S. Kotz, T. J. Kozubowski, et K. Podgórski, *The Laplace Distribution and Generalizations*. Boston, MA: Birkhäuser, 2001. doi: 10.1007/978-1-4612-0173-1.

[57] M. Zuber, S. Carpov, et R. Sirdey, « Towards real-time hidden speaker recognition by means of fully homomorphic encryption ». 2019.

[58] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, et B. McMahan, « cpsgd: Communication-efficient and differentially-private distributed sgd », *NeurIPS*, vol. 31, p. 7564-7575, 2018.

[59] A. Koskela, J. Jälkö, L. Prediger, et A. Honkela, « Tight differential privacy for discrete-valued mechanisms and for the subsampled gaussian mechanism using FFT », in *International conference on artificial intelligence and statistics*, PMLR, 2021, p. 3358-3366.

[60] C. Canonne, G. Kamath, et T. Steinke, « The discrete gaussian for differential privacy », *ArXiv Prepr. ArXiv200400010*, 2020.

[61] N. Agarwal, P. Kairouz, et Z. Liu, « The skellam mechanism for differentially private federated learning », *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.

[62] P. Kairouz, Z. Liu, et T. Steinke, « The distributed discrete gaussian mechanism for federated learning with secure aggregation », in *International conference on machine learning*, PMLR, 2021, p. 5201-5212.

[63] P. Türk, « Définitions et enjeux de la souveraineté numérique », *Cah. Fr. Doc. Actual.*, nº 415, juin 2020, Consulté le: 6 janvier 2024. [En ligne]. Disponible sur: https://hal.univ-cotedazur.fr/hal-02493091

[64] P. Hérault, « Comment renforcer la souveraineté à l'heure des chaînes de valeur mondiales? », *Études L'Ifri Déc. Www Ifri Org*, 2021, Consulté le: 31 décembre 2023. [En ligne]. Disponible sur:

https://www.ifri.org/sites/default/files/atoms/files/herault_chaines_de_valeur_2021.pdf

[65] M. Leonard, J. Pisani-Ferry, E. Ribakova, J. Shapiro, et G. B. Wolff, *Redefining Europe's economic sovereignty*. JSTOR, 2019. Consulté le: 6 janvier 2024. [En ligne]. Disponible sur: https://www.jstor.org/stable/pdf/resrep28498.pdf

[66] R. Gauvain, C. d'Urso, A. Damais, et S. Jemai, « Rétablir la souveraineté de la France et de l'Europe et protéger nos entreprises des lois et mesures à portée extraterritoriale », *Rep. Prime Minist. June*, vol. 26, 2019.

[67] C. Villani *et al.*, *Donner un sens à l'intelligence artificielle: pour une stratégie nationale et européenne*. Conseil national du numérique, 2018. Consulté le: 31 décembre 2023. [En ligne]. Disponible sur: https://books.google.fr/books?hl=en&lr=&id=Q7lUDwAAQBAJ&oi=fnd&pg=PP1&dq=villani+donner+un+sens+%C3%A0+l%27intelligence+artificielle&ots=0ICkR9XyTZ&sig=Ldi55_QpNm-YBHe44Rmui2lkmUY

[68] European Parliament, « White Paper on Artificial Intelligence including follow-up | Legislative Train Schedule », European Parliament. Consulté le: 1 janvier 2024. [En ligne]. Disponible sur: https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-white-paper-artificial-intelligence-and-follow-up

[69] H. Roberts, J. Cowls, J. Morley, M. Taddeo, V. Wang, et L. Floridi, « The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation », *AI Soc.*, vol. 36, nº 1, p. 59-77, mars 2021, doi: 10.1007/s00146-020-00992-2.

[70] S. Ezell et C. Foote, « How Stringent Export Controls on Emerging Technologies Would Harm the U.S. Economy ». Consulté le: 6 janvier 2024. [En ligne]. Disponible sur: https://itif.org/publications/2019/05/20/how-stringent-export-controls-emerging-technologies-would-harm-us-economy/

[71] C. Flynn, « Recommendations on Export Controls for Artificial Intelligence », Center for Security and Emerging Technology. Consulté le: 6 janvier 2024. [En ligne]. Disponible sur: https://cset.georgetown.edu/publication/recommendations-on-export-controls-for-artificial-intelligence/

[72] F. Sevini *et al.*, « Emerging dual-use technologies and global supply chain compliance », in *IAEA Symposium on International Safeguards*, 2018. Consulté le: 31 décembre 2023. [En ligne]. Disponible sur: https://media.superevent.com/documents/20181109/8f69660c6877a27468249829713be243/id324-sevini_paper.pdf

[73] R. C. Thomsen II, « Artificial Intelligence and Export Controls: Conceivable, but Counterproductive? », *J. Internet Law*, vol. 22, nº 5, p. 14-24, 2018.

[74] A. Viski, S. Jones, L. Rand, T. Boyce, et J. Siegel, *Artificial intelligence and strategic trade controls*. Center for International & Security Studies, U. Maryland., 2020.

[75] T. F. Bresnahan et M. Trajtenberg, « General purpose technologies 'Engines of growth'? », *J. Econom.*, vol. 65, nº 1, p. 83-108, janv. 1995, doi: 10.1016/0304-4076(94)01598-T.

[76] P. A. David et G. Wright, « General Purpose Technologies and Productivity Surges: Historical Reflections on the Future of the ICT Revolution », *Econ. Hist.*, Art. nº 0502002, févr. 2005, Consulté le: 31 décembre 2023. [En ligne]. Disponible sur: https://ideas.repec.org//p/wpa/wuwpeh/0502002.html

[77] J. Vipra et A. Korinek, « Market Concentration Implications of Foundation Models », 2 novembre 2023, *arXiv*: arXiv:2311.01550. doi: 10.48550/arXiv.2311.01550.

[78] D. K. Kanbach, L. Heiduk, G. Blueher, M. Schreiter, et A. Lahmann, « The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective », *Rev. Manag. Sci.*, sept. 2023, doi: 10.1007/s11846-023-00696-z.

[79] X. Ferràs-Hernández, P. A. Nylund, et A. Brem, « The Emergence of Dominant Designs in Artificial Intelligence », *Calif. Manage. Rev.*, vol. 65, nº 3, p. 73-91, mai 2023, doi: 10.1177/00081256231164362.

[80] J. E. Bessen, S. M. Impink, L. Reichensperger, et R. Seamans, « The Business of AI Startups », 25 juillet 2023, *Rochester, NY*: 3293275. doi: 10.2139/ssrn.3293275.

[81] I. M. Enholm, E. Papagiannidis, P. Mikalef, et J. Krogstie, « Artificial Intelligence and Business Value: a Literature Review », *Inf. Syst. Front.*, vol. 24, nº 5, p. 1709-1734, oct. 2022, doi: 10.1007/s10796-021-10186-w.

[82] M. Ford, *Rule of the robots: How artificial intelligence will transform everything*. Hachette UK, 2021. Consulté le: 10 janvier 2024. [En ligne]. Disponible sur: https://books.google.com/books?hl=en&lr=&id=eF82EAAAQBAJ&oi=fnd&pg=PT7&dq=ford+utility+AI+2021&ots=IFwuamENaM&sig=vKVUlg2yniKUTuMvBzlFIrqd-TY

[83] European Commission, « White Paper on Artificial Intelligence: a European approach to excellence and trust ». Consulté le: 1 janvier 2024. [En ligne]. Disponible sur: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

[84] European Parliament, « Artificial intelligence act | Legislative Train Schedule », European Parliament. Consulté le: 1 janvier 2024. [En ligne]. Disponible sur: https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence

[85] European Parliament, « AI liability directive | Legislative Train Schedule », European Parliament. Consulté le: 1 janvier 2024. [En ligne]. Disponible sur: https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-ai-liability-directive

[86] B. Benbouzid, Y. Meneceur, et N. A. Smuha, « Quatre nuances de régulation de l'intelligence artificielle. Une cartographie des conflits de définition », *Réseaux*, vol. 232-233, nº 2-3, p. 29-64, 2022, doi: 10.3917/res.232.0029.

[87] L. Colonna et S. L. Submitter, « The AI Act's Research Exemption: A Mechanism for Regulatory Arbitrage? », 19 septembre 2023, *Rochester, NY*: 4575971. doi: 10.2139/ssrn.4575971.

[88] B. A. Juliussen, J. P. Rui, et D. Johansen, « Algorithms that forget: Machine unlearning and the right to erasure », *Comput. Law Secur. Rev.*, vol. 51, p. 105885, nov. 2023, doi: 10.1016/j.clsr.2023.105885.

[89] MSI-NET, « Algorithms and human rights - Study on the human rights dimensions of automated data processing techniques and possible regulatory implications », Council of Europe Publishing. Consulté le: 6 janvier 2024. [En ligne]. Disponible sur: https://edoc.coe.int/en/internet/7589-algorithms-and-human-rights-study-on-the-human-rights-dimensions-of-automated-data-processing-techniques-and-possible-regulatory-implications.html

[90] European Parliament, « Framework of ethical aspects of artificial intelligence, robotics and related technologies | Legislative Train Schedule », European Parliament. Consulté le: 1 janvier 2024. [En ligne]. Disponible sur: https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-ai-ethical-framework

[91] B. Bachimont, « Herméneutique matérielle et Artéfacture: des machines qui pensent aux machines qui donnent à penser; Critique du formalisme en intelligence artificielle », PhD Thesis, Thèse de doctorat d'épistémologie, École Polytechnique, 1996.

[92] J. Lassègue, « La méthode expérimentale, la modélisation informatique et l'intelligence artificielle », 1996, Consulté le: 2 août 2019. [En ligne]. Disponible sur: https://halshs.archives-ouvertes.fr/halshs-00008861

[93] E. Schmitt, « Explorer, visualiser, décider : un paradigme méthodologique pour la production de connaissance à partir des Big Data », Université de technologie de Compiègne, 2018.

[94] G. Varoquaux et V. Cheplygina, « Lessons from shortcomings in machine learning for medical imaging », OECD, Paris, juin 2023. doi: 10.1787/b885eecd-en.

[95] A. Cockburn, P. Dragicevic, L. Besançon, et C. Gutwin, « Threats of a replication crisis in empirical computer science », *Commun. ACM*, vol. 63, n° 8, p. 70-79, juill. 2020, doi: 10.1145/3360311.

[96] C. S. Calude et G. Longo, « The Deluge of Spurious Correlations in Big Data », in *Lois des dieux, des hommes et de la nature*, Nantes, France, oct. 2015, p. 1-18. doi: 10.1007/s10699-016-9489-4.

[97] E. P. Wigner, « The unreasonable effectiveness of mathematics in the natural sciences. », *Commun. Pure Appl. Math.*, vol. 13, n° 1, p. 1-14, 1960, doi: 10.1002/cpa.3160130102.

[98] A. Halevy, P. Norvig, et F. Pereira, « The Unreasonable Effectiveness of Data », *IEEE Intell. Syst.*, vol. 24, n° 2, p. 8-12, mars 2009, doi: 10.1109/MIS.2009.36.

[99] A. Vassilev, A. Oprea, A. Fordyce, et H. Anderson, « Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations », National Institute of Standards and Technology, NIST Artificial Intelligence (AI) 100-2 E2023, janv. 2024. doi: 10.6028/NIST.AI.100-2e2023.

[100] J. Near, D. Darais, N. Lefkovitz, et G. Howarth, « Guidelines for Evaluating Differential Privacy Guarantees », National Institute of Standards and Technology, NIST Special Publication (SP) 800-226 (Draft), déc. 2023. doi: 10.6028/NIST.SP.800-226.ipd.

[101] E.-M. El-Mhamdi *et al.*, « On the Impossible Safety of Large AI Models », 9 mai 2023, *arXiv*: arXiv:2209.15259. doi: 10.48550/arXiv.2209.15259.

[102] Y. Allouah, R. Guerraoui, N. Gupta, R. Pinot, et J. Stephan, Éd., « On the Privacy-Robustness-Utility Trilemma in Distributed Learning », *Proc. 40th Int. Conf. Mach. Learn. - Hawaii*, 2023.

[103] R. Guerraoui, N. Gupta, R. Pinot, S. Rouault, et J. Stephan, « Differential Privacy and Byzantine Resilience in SGD: Do They Add Up? », in *Proceedings of the 2021 ACM Symposium on Principles of Distributed Computing*, in PODC'21. New York, NY, USA: Association for Computing Machinery, juill. 2021, p. 391-401. doi: 10.1145/3465084.3467919.

université
PARIS-SACLAY

Titre : Contributions à la robustesse, la sécurité et la confidentialité en apprentissage statistique

Mots clés : apprentissage statistique ; traitement statistique de signaux multivariés ; attaques adverses ; confidentialité et apprentissage statistique ; apprentissage statistique décentralisé.

Résumé : Quinze ans se sont écoulés depuis la soutenance de ma thèse dédiée aux interfaces cerveau-machines asynchrones. Au-delà d'une formation technique poussée au spectre relativement large (de la séparation de sources en contexte très bruité à la classification statistique), celle-ci m'avait initié à un contexte de recherche compétitif, associant une grande quantité d'équipes actives dans le domaine, des constantes de temps de recherche réduites et des attentes publiques et médiatiques importantes. Le recul de ces quinze dernières années permet de regarder ces années de thèse avec relativisme !

Ce document retrace les pistes de recherche suivies au cours de ces quinze années, façonnées par les résultats positifs et négatifs, les questionnements intérieurs et quelques modes d'un monde qui semblent s'accélérer…

Nos travaux couvrent de nombreux domaines, du traitement de signaux multivariés à la sécurité de l'apprentissage statistique décentralisé

Nous commençons par décrire les travaux en continuité avec la thèse, portant sur l'utilisation de représentations parcimonieuses pour les signaux électroencéphalographiques, puis la segmentation de séries temporelles multivariées. Nous décrivons ensuite les contributions portant sur la gestion des compromis en apprentissage statistique pour le traitement de flux de données massives, avant de décrire celles dédiées à la sécurité et la confidentialité de l'apprentissage statistique en contextes centralisé et décentralisé.

Nous concluons ce document avec une analyse du contexte actuel associé au domaine de l'apprentissage statistique, afin d'expliquer les directions futures de mes recherches.

Title : Contributions to robustness, security, and privacy in statistical machine learning

Keywords : statistical machine learning; statistical signal processing; adversarial robustness; privacy in machine learning; decentralized statistical learning.

Abstract : Fifteen years have passed since I defended my thesis dedicated to asynchronous brain-computer interfaces. Beyond providing a technical training with a relatively broad spectrum (from source separation in a noisy context to classification), it introduced me to a competitive research environment, involving a large number of active teams in the field, reduced research time constants, and significant public and media expectations. The perspective of these last fifteen years allows us to look back at those thesis years with relativism!

This document retraces the research paths followed over these fifteen years, shaped by both positive and negative results, internal questioning, and some trends of a world that seems to be accelerating…

Our work spans numerous fields, from multivariate signal processing to the security of decentralized statistical learning. We begin by describing work in continuity with the thesis, focusing on the use of sparse representations for electroencephalographic signals, followed by the segmentation of multivariate time series. We then describe contributions related to trade-offs in machine learning for processing massive data streams, before detailing those dedicated to the security and privacy of statistical learning in centralized and decentralized contexts.

We conclude this document with an analysis of the current context associated with the field of artificial intelligence, in order to explain future research directions.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France