La confidentialité différentielle : quelle quantification de la privacy dans le monde de l'apprentissage automatique ?

Edwige Cyffers

Université Paris 1 UFR 10
12, place du Panthéon
75231 - Paris cedex 05

Thèse de master en philosophie

La Confidentialité Différentielle
Quelle quantification de la privacy dans le monde de l'apprentissage automatique ?

Edwige Cyffers

2021

Encadré par Marco Panza et Alberto Naibo

À mon chat

Abstract

The recent advances of technology, that make the storage and the processing of large amount of data affordable, drove up the collection of sensitive data. For instance, a smartphone can track its owner position, her sport activity, her messages, her photos, her queries and browsing. When considering the trend of Big Data or the Covid outbreak, one cannot deny the rise of digitization. Data leakage, malicious or not, is thus a burning issue of the digital era. How can we guarantee privacy, this slippery concept on the fringe of obfuscation, unlinkability, anonymity, confidentiality and data minimization ?

Differential privacy is currently the gold standard both in research and industry for machine learning applications. It quantifies the privacy loss occurring during the use of a record, by synthetizing its impact in a scalar. This work addresses how the definition was introduced and how it relies on statistical learning hypothesis. We see how the context of digitization induces a shift in the privacy protection and test the limit of differential privacy through its variants and real-world implications.

Keywords : Differential privacy, Privacy, Big Data, quantification, machine learning

Résumé

L'effondrement des prix de stockage de l'information, la couverture croissante des usages informatiques et des collectes de données qui y sont associées ainsi que l'accroissement des capacités de traitement de l'information sont autant de bouleversements techniques dans le domaine de l'information. Que l'on parle de *Big Data*, ou que l'on considère simplement les conséquences de la numérisation lors de la crise sanitaire ces deux dernières années, la collecte généralisée de données sensibles est un nouvel enjeu de notre société. À titre d'exemple, un téléphone récolte le généralement la position instantanée, les relations, les heures de sommeil, les questions et autres données de santé de son utilisateur.

La nécessité de sécuriser et d'éviter les fuites de données, qu'elles soient malicieuses ou non, est donc un enjeu clé de la transition numérique. Mais comment peut-on garantir la *privacy* ? Ce concept a de nombreuses facettes : offuscation, droit à l'oubli, anonymat, confidentialité, minimisation des données. Dans le cadre de l'apprentissage automatique (*Machine learning*), une métrique s'est imposée au sein de la recherche et des applications des GAFAM pour quantifier le niveau de *privacy* d'un procédé donné.

La confidentialité différentielle (*differential privacy*) est en effet une définition mathématique qui réduit à un nombre réel le niveau de persistance d'une donnée dans les sorties d'un algorithme. Ce mémoire décrit l'émergence et les facteurs qui ont contribué au succès de cette quantification, ainsi que les conséquences implicites de cette définition sur les attentes de l'apprentissage automatique et le rapport entre l'individu et ses données. Nous abordons donc l'évolution de la notion de *privacy* face aux nouvelles réalités techniques, nous mettons en contexte la définition de confidentialité différentielle comme une technique de quantification et nous analysons ses variantes comme limites de la définition originelle.

Mots-clés : Confidentialité différentielle, protection des données, Big Data, apprentissage automatique, identité numérique, quantification

Table des matières

Introduction

Le droit à la vie privée est un droit reconnu par la déclaration universelle des droits de l'Homme [Nat1948], il y est ainsi formulé :

> *« Nul ne sera l'objet d'immixtions arbitraires dans sa vie privée, sa famille, son domicile ou sa correspondance, ni d'atteintes à son honneur et à sa réputation. Toute personne a droit à la protection de la loi contre de telles immixtions ou de telles atteintes.* »

Cet article 12 s'inscrit dans une reconnaissance de ce droit comme composante de la liberté. Il s'agit de définir les champs d'application les plus emblématiques de la spère privée, telle la famille ou la correspondance. Derrière le terme d'immixtions arbitraires, on place déjà ce droit dans un contexte de défense et d'équilibre : il s'agit de résister à l'oppression que constituerait une transparence absolue, mais il ne s'agit pas non plus d'un droit absolu à garder secret tout ce qui peut relever de la sphère privée.

L'étude de cet équilibre est en soi un sujet d'étude à part entière [SUP2008], tant elle a pu connaitre des géométries variables et des pratiques variées. On peut ainsi mentionner des applications toutes aussi diverses que la nudité, les délibérations judiciaires, les personnalités dites publiques [Nis2010]. Quelle information doit être partagée et avec quel public dans chacun de ces situations ? Quel est le niveau d'intimité qui doit prévaloir dans le rapport au corps ? Faut-il renforcer la transparence de la justice ou laisser les jurés exprimer leur pensée à labri des oreilles indiscrètes ? Une personne peut-elle renoncer à sa vie privée parce qu'elle occupe un rôle spécifique dans la société ? Quelle doit être la quantité d'information qu'un État doit divulguer à sa population, ou collecter [Sno2019] ?

Affirmer que le personnel est politique a été un moyen pour les féministes de repenser les violences contre les femmes comme un problème systémique, et non comme une déviance appartenant au cadre privé [RB2018]. La transparence, voire la revendication, de l'altérité sont également des outils pour impulser des changements sociétaux, ce qui a parfois remis en question le bénéfice de limiter l'intrusion du public dans la sphère privée[SUP2008]. Pourtant, les atouts, voire la nécessité de disposer d'une sphère privée ne saurait être mise en doute de façon radical : chacun maintient son intégrité grâce à cette zone qui lui est strictement personnelle, et choisit selon les contextes de dissimuler ou non certaines informations [Wes1967]. L'offuscation, qui consiste à rendre difficile d'accès une information, est donc un outil de résistance à l'oppression [1], qui a pu être utilisé aussi bien par des aviateurs français lors de la seconde guerre mondiale que par l'opération Vula pour lutter contre l'apartheid [BN2015 - 2015]. De même, le recours à l'anonymat renforce les moyens des lanceurs d'alertes et de la liberté dexpression, mais est aussi souvent pointée du doigt comme un moyen déloyal pour proférer des propos diffamatoires [Tor2017].

1. la résistance à l'oppression est citées comme droits naturels et imprescriptibles de l'Homme dans la d'Article 2 de la Déclaration de l'Homme et du Citoyen

Une multitude de concepts et de techniques se croisent : intimité, droit à la vie privée, anonymat, secret, intégrité, confidentialité, contrôle de l'information... On adoptera le terme anglais « *privacy* » pour couvrir ce large spectre, puisque la langue française manque d'un nom correspondant à l'adjectif « privé ». Dans la lignée de la formalisation de l'intégrité contextuelle [Nis2010], on retrouve plusieurs invariants dans les situations qui relèvent de la *privacy*. En effet, on retrouve tout d'abord le schéma classique de la communication – destinateur, message, destinataire – augmenté de deux paramètres : l'entité auquel se rapporte l'information, et la façon dont l'information est diffusée. Il se tisse le lien entre l'information du message, qui peut être un acte, un fait, une uvre ou même un simple accroissement de probabilité d'une part, et l'entité qui peut avoir une légitimité à restreindre ou conditionner l'accès à cette information d'autre part. Dès lors, la transmission, par cette entité ou par un tiers vers d'autres destinataires peut être soumise à différentes normes et modalités. De ceci découlent différents flots d'informations possibles. Quelles sont les demandes légitimes de l'entité sur cette diffusion d'information ? Quels sont les liens significatifs entre une donnée et une entité ?

Ces questions ont été profondément modifiées par le recours accru à l'informatique [Boe2005]. La croissance exponentielle des capacités de calcul et l'effondrement du prix du stockage de l'information, ainsi que la démocratisation d'Internet ont radicalement transformé les modalités du transfert et du traitement de l'information. Ces dynamiques font apparaitre des résultats contre intuitifs, comme la nécessité de distinguer entre dissimulation effective des informations personnelles et anonymat [Ohm2009], la possibilité de développer des modèles efficients de prédiction malgré l'absence de compréhension de phénomènes sous-jacents [GBC2016], ou encore la suppression des barrières physiques traditionnelles en cas d'une connexion internet suffisante, qui créé des interactions selon des graphes très différents du monde physique [BH2015]. Le code informatique, comme le soulignait déjà Lessig avec la formule « *Code is law* » [Les2009], s'impose donc comme un moteur des relations sociales et politiques du XXI^{e} siècle.

Ce constat a généré plusieurs réponses. Des initiatives législatives comme le Règlement Général de Protection des Données (RGPD) [Con2016] viennent tenter de réinstaurer le respect de la *privacy* dans le cyberspace [Mer ; BM2018]. Des techniques d'offuscation sont aussi monnaies courantes, que ce soit par l'emploi de VPN, de profils correspondant à des identités fictives, ou de plug-in brouillant le traçage publicitaire [BN2015 - 2015]. Le paradoxe entre l'adoption massive d'applications très intrusives en termes de vie privée [HM2016] et la méfiance croissante exprimée par les utilisateurs quant au bien-fondé de ces collectes, et alors même que leurs données est au cœur du système actuel du cyberspace, demande une meilleure compréhension. Si les données sont le nouvel or, les enjeux de traçabilité, de régulation et de transparence sont cruciales [ONe2016 ; Mor2013].

Une partie de la résolution du conflit entre la légitimité des collectes et les attentes

des individus réside peut-être dans la quantification de la *privacy*. La *differential privacy*, traduite par confidentialité différentielle en français, s'est imposée comme la définition mathématique du niveau d'immixtion d'un algorithme dans l'utilisation d'une donnée. Pour cela, cette théorie propose de résumer en un nombre réel, nommé le *budget de privacy*, le niveau d'information que l'on peut extraire de la donnée grâce aux sorties d'un algorithme [DR2014].

Cette mesure a été reconnue par la communauté scientifique, notamment via la remise du prix Gödel en 2017 aux auteurs du papier séminal[Dwo+2006] seulement 11 ans après sa publication. Son utilisation dépasse le cadre seul de la recherche puisque les plus grands acteurs de la *tech*, tel qu'Apple ou Google, l'ont adapté et ont des équipes dédiées à son implémentation. Les administrations peuvent aussi y avoir recours, puisque le US Census[Dwo2019] l'a adopté, soulevant des interrogations voire des recours juridiques [Per2021].

L'objectif de ce mémoire est de comprendre et d'analyser la confidentialité différentielle, son émergence et son utilisation. Pour cela, nous analyserons comment les schémas traditionnels de protection de la vie privée ont été mis à mal par les données massives et les traitements d'apprentissage automatiques, afin de comprendre la nécessité de nouvelles réponses. Nous replacerons donc cette évolution dans le cadre d'une évolution notamment technique, qui a transformé un concept relativement binaire de la *privacy* – est-ce que la donnée privée est rendue publique à mauvais escient – vers un besoin de quantification. Parce que cette évolution est le fruit des changements d'utilisation du numérique, on intitule cette partie Privacy 2.0, en référence au Web 2.0.

Nous nous attacherons ensuite à la définition de la confidentialité différentielle. Ceci nécessite de se placer dans un cadre probabiliste et algorithmique que nous introduirons, dans la perspective de justifier son introduction comme une conséquence des hypothèses de l'apprentissage automatique. Après l'étude de la définition, nous nous intéressons à ses propriétés : comment comprendre les garanties mathématiques qui sont établies en terme d'applications, et pourquoi ces propriétés sont désirées.

Enfin, nous repassons du monde mathématique aux applications en nous intéressant à linstanciation pratique du budget de *privacy*. Les différents paramètres de la définition, confrontés aux flux d'informations utilisés en apprentissage automatique, ont fait émerger diverses variantes. La façon dont un enregistrement se relie à un individu dénote particulièrement cette évolution, non seulement vers une vision quantifiée et probabiliste de la *privacy*, mais aussi de l'information de façon plus générale, et donc ultimement de l'individu. Enfin, si la confidentialité différentielle garantie un type de quantification, nous verrons qu'elle laisse hors champ d'autres flux de données qui peuvent aussi légitimement être vus comme des attaques de la *privacy*.

Première partie

Privacy 2.0

1. Pourquoi étudier la *privacy*

En commençant ce travail, la considération de la *privacy* me semblait acquise. Il s'agit de ces nécessités, de ces droits fondamentaux qui s'imposent d'eux-mêmes tant on sent intimement qu'il serait dommageable et faux de vivre sans. Cette évidence pourtant, si elle n'a jamais été remise en cause lors des nombreuses discussions avec des inconnus, a été questionnée sous trois angles différents, notamment dans le cadre scolaire et celui du Y20 [1]. À titre de motivation, il me semble donc important de parcourir ces questionnements, qui sont de plusieurs types.

Une première attaque, que je qualifierais de relativisme absolu, consiste à dire que la *privacy*, comme d'ailleurs tout droit de l'Homme abstrait, n'est pas nécessaire, que l'on vit très bien sans et qu'il s'agit donc d'un non sujet. Cette position était également utilisée de façon plus générale par certains délégués des pays du Sud sur toute régulation des nouvelles technologies ou du travail.

« Nous ne voulons pas de contraintes sur les données ni sur leurs usages, même s'il y a des risques avérés de discrimination, car nous voulons donner le maximum d'avantage aux entreprises, pour les convaincre d'apporter leurs services », telle est la position récurrente rencontrée. Que les droits de l'Homme soient une lubie occidentale qui ne nourrit pas n'est pas un discours nouveau. Que les besoins fondamentaux, l'accès à l'eau, à la nourriture, aux infrastructures soient des prérequis pour s'intéresser au respect d'abstraction telle la liberté, l'égalité ou encore la *privacy* n'est pas un point de vue que je souhaite discuter ici. Je me place donc dans un cadre où défendre ces concepts à un sens, ce qui me semble légitime dans le « pays des Droits de l'Homme » et dont je suis convaincue qu'il admet aussi un sens universel.

La seconde attaque est plus spécifique à la privacy. « Si on n'a rien à cacher, l'absence de privacy n'est pas un problème » est le leitmotiv de l'apologie de la société de réputation, qui transforme le désir de *privacy* en désir de dissimuler des méfaits [2]. Tout d'abord cette affirmation repose une hypothèse « Si on n'a rien à cacher ». Celle-ci est fausse en général, ce qui invalide alors l'intégralité du raisonnement. De façon plus générale, avoir quelque chose à cacher est la norme : code secret, rideaux, secret médical, secret des délibérations, copies anonymisés, vêtements sont autant d'exemple de

1. Groupe d'engagement consacré à la jeunesse auprès du G20. Rassemblant des délégués des 20 membres du G20, il produit un communiqué produit par consensus dans l'objectif de formuler des recommandations bénéficiant à l'ensemble de la jeunesse. Le communiqué en question est accessible https://www.youngambassadorssociety.it/Y20_2021_Communique.pdf

2. On peut également se tourner vers la formule lapidaire d'Edward Snowden : « saying that you don't care about privacy because you have nothing to hide is no different from saying you don't care about freedom of speech because you have nothing to say » [Sno2019]

dispositifs qui montrent la banalité de ce besoin.

De plus, déduire la conclusion de la prémisse est également faux : ne puis-je pas simplement désirer qu'autrui puisse cacher quelque chose également ? Il est parfois bien plus confortable d'ignorer certains aspects de la vie d'autrui, voire plus équitable de ne pouvoir accéder à certaines informations. Ce point de vue exclut aussi les changements temporels : comment savoir si demain je n'aurai pas quelque chose à cacher parmi les informations que j'ai révélée aujourd'hui ? Si je divulgue tout aujourd'hui, mon changement de comportement, un jour, sera détectable. Renoncer à la *privacy* aujourd'hui est donc y renoncer définitivement. Par ailleurs, pourquoi souhaiter renoncer à la *privacy* par défaut ? Cette dénonciation a d'ailleurs déjà été développée de longue date : nous modifions notre comportements quand nous sommes surveillés, et ceci est déjà à la base de l'analyse du panoptique par Foucault [Fou1975].

Enfin, on mentionne parfois l'abandon de cette valeur par ma génération. La multiplication des réseaux sociaux, qui encouragent les détails intimes jusqu'au scabreux, prouverait que nous ne souhaitons plus avoir de *privacy*. Et effectivement, de nombreuses pratiques largement adoptées y sont contraires. Il faut pourtant voire à la fois quelles sont les origines de ces pratiques et voir à quel point elles créent un paradoxe avec les attentes de *privacy* d'une part, et d'autre part il y a de même de nombreuses pratiques contraire au respect de l'égalité, sans que l'on en déduise que ce concept est *has been*.

Ces réticences, voire oppositions à la *privacy*, si elles n'ont pas pour vocation d'occuper la place d'un mémoire sur la quantification de la privacy, demande toutefois une justification et sont autant d'éléments de contexte. Aurait-on osé questionner la liberté de même ? Ou la sécurité ? Nous reviendrons donc tout d'abord sur quelques points de définitions, avant de voir des exemples de dispositifs techniques plus anciens pour garantir la *privacy*. Puis, nous verrons comment la transposition dans l'espace numérique a changé les possibilités et donc rendu obsolète certaines protections efficaces dans un monde non numérique. Notamment, le recours à des données massives fait disparaitre la différence entre données privées et données publiques, c'est-à-dire rend l'anonymat inefficace pour faire disparaitre le lien entre une entité et la donnée qui lui est reliée. On passe d'une conception de la *privacy* relativement binaire – où l'information est révélée de façon inadéquate ou non – à une quantification continue.

Le droit d'être laissé à soi-même, en anglais *the right to be left alone*, est vu comme une nécessité par les juristes Warren et Brandeis en 1890 [WB1890]. Le besoin d'expliciter ce droit est alors une réaction face à l'essor de journaux qui empiètent sur la vie privée des individus, via des rubriques consacrés à la vie des célébrités. Il s'agit donc déjà d'une évolution technique dans la diffusion de l'information qui change les flux d'information traditionnels. Se faisant, d'après les auteurs, cette avancée technologique

– journaux peu chers et avec une logistique plus fluide – poussent à des évolutions dangereuses de la société, alors même que la société est en mesure, et devrait donc, fournir de plus grandes protections. Les rumeurs de village ne sont bien sûr pas un fait nouveau au XIX^e^ siècle, mais les pratiques intrusives journalistiques, et la large diffusion des journaux qui encouragent les lecteurs à s'intéresser à ces contenus provoquent un changement d'échelle qui rend nécessaire de nouvelles protections juridiques, dépassant la simple propriété intellectuelle ou l'interdiction de diffamation.

Cet article ne propose pas de définition positive de la *privacy*. Cependant, la liste d'exemples problématiques soulevés permet de cartographier le spectre de son application. Il ne s'agit pas simplement de se protéger des vols d'idées qui ont une valeur intrinsèque, comme dans le cas de la propriété intellectuelle, ni de s'arrêter à l'autre extrémité avec la diffusion de faits particulièrement embarrassants ou répréhensibles, mais de protéger l'ensemble des faits, anodins lorsqu'ils sont pris individuellement, constituent néanmoins notre liberté d'action et la définition de nous-mêmes. Ce qu'on doit protéger va de la photographie d'un défunt à la correspondance privée des individus, qui n'a pourtant rien de répréhensible ou de surprenant pour l'immense majorité des cas.

La nécessité sociale de garantir la *privacy* a été reprise ensuite par de multiples philosophes [SUP2008, pages 79-100], même si elle a pu aussi être interprétée comme une menace pour la collectivité. L'absence de *privacy* est en effet un moyen de forcer, par le regard des autres un certain comportement, qui correspond aux normes sociales en vigueur, tandis que sa garantie permet les déviances, les réflexions individuelles et les contre-pouvoirs.

Des dispositifs techniques paradigmatiques de cette tension sont les dispositifs de vote. Dans le cadre de vote d'élus, le vote est public et consigné par écrit, afin que chaque citoyen, s'il le souhaite, puisse vérifier que son représentant a voté en cohérence avec le programme sur lequel il a été élu. Cette transparence signifie que l'élu doit pouvoir rendre des comptes, et ses faits pourront par exemple être pointés par ses opposants dans l'élection suivante, pour une éventuelle sanction par les urnes. Ici, on privilégie donc l'absence de *privacy*, car l'acte doit être public pour permettre la représentativité des élus.

Au contraire, pour l'individu, on recourt à l'isoloir, qui garantit physiquement la *privacy*. Le citoyen accomplit son vote selon un cadre prédéfini (choix des modalités d'élections, choix des candidats) mais son choix personnel est protégé des regards d'autrui par un rideau. Il n'y a pourtant rien de répréhensible dans son acte, dans une démocratie chacun est *a priori* libre de choisir de soutenir un candidat sans pression, et certains citoyens choisissent même de communiquer largement à propos de leur candidat préféré, par exemple pour persuader des connaissances de voter pour lui également. Pourtant, cet anonymat est vu comme un rempart de la démocratie, car il permet à chacun, au moment où il se trouve dans l'isoloir, de prendre une décision sans avoir

à se justifier, ni à se conformer à l'avis qu'il perçoit comme dominant. Seul sa propre conscience devrait normalement guider son acte.

On voit donc que la limite de la *privacy* dépend du contexte, et de la confiance de chacun en la qualité des normes du collectif par rapport à celles de l'individu. Y aurait-il ici un élément de réponse à ma surprise initiale sur le manque de reconnaissance relatif de l'importance de la privacy dans le monde académique ? Se conformant eux-mêmes naturellement à l'injonction dominante de la société sur la réussite sociale, et ne ressentant pas de honte particulière leurs actions, l'absence de *privacy* n'est peut-être pas autant perçue comme une menace au sein de la classe dominante, et encore mieux, les protègent éventuellement des déviances d'autrui. Au contraire, celui qui se sent imposteur, qui expérimente différentes voies et échoue, hésite, y voit une protection indispensable. Le simple fait d'appartenir au deuxième sexe est une éducation à la nécessité de *privacy* : comment s'en passer quand l'oppression des corps est la norme, s'il est nécessaire d'avoir une chambre à soi, comment méconnaitre ce besoin de *privacy*.

Parce que la norme historique peut être violente et discriminante, la nécessité de se réserver un espace à soi, un lieu de tranquillité qui n'oblige pas à rendre des compte est pour tous ceux qui ne se conforment pas entièrement aux attentes sociétales une composante de la liberté, une protection pour pouvoir être soi-même et aller vers les autres. Qui sort de la norme ? Et ne devons-nous pas préserver la possibilité d'un droit à l'erreur, et donc d'un droit à l'oubli ? Il semble au moins qu'il existe dans la *privacy* quelque chose à protéger, que ce soit comme fin en soi, ou comme moyen d'assurer d'autres protections, telle qu l'égalité ou la liberté.

Permettre à chacun de pouvoir dissimuler certaines zones d'ombre est une façon de lutter contre ce que Victor Hugo dénonçait déjà dans les Misérables. « Tant qu'il existera par le fait des lois et des mœurs, une damnation sociale créant artificiellement, en pleine civilisation, des enfers, et compliquant d'une fatalité humaine la destinée », pouvoir sortir de sa condition en omettant ses erreurs passées est l'unique voie possible pour sortir d'un cercle vicieux. Monsieur Madeleine n'existerait pas si Jean Valjean ne pouvait obtenir la chance de dissimuler son passé de bagnard par hasard. Et pourtant, certaines avancées techniques rendent plus improbables ces échappatoires.

Ces exemples s'appuient directement sur notre société, et montrent que même dans un cadre vu comme respectueux des libertés, l'existence inéluctable de traditions, de normes et de position dominante rend la *privacy* nécessaire. Sa fonction en est d'autant plus évidente que les garanties de liberté et le respect des droits sont défaillants. On peut penser à l'espionnage des journalistes dans le but d'empêcher la révélation d'agissements illégaux [Dam2021], à la surveillance des autorités de contrôle dans le but de corrompre ou d'étouffer les lanceurs d'alertes. C'est aussi le rempart contre le contrôle social dictatoriale, tel qu'il est dénoncé dans le cadre du crédit social en Chine. Entre la multiplication des *nudges* pour inciter certains comportements vertueux, la collecte des données de *tracing* et de santé, la différence avec un crédit social peut se poser.

La résistance directe à l'oppression est certes le cas le plus extrême, mais il semble difficile de nier son existence. À l'échelle individuelle, le droit à l'oubli pour tous est cependant vu comme une nécessité par le Règlement Général de Protection des Données (RGPD) [Con2016]. Offrir la possibilité de se définir différemment dans différents contextes au cour de sa vie fait partie des moyens d'accès aux libertés. Quelque que soit la situation, il existe des normes, qui régissent les *a priori* d'une société, et dont un individu peut avoir envie ou besoin de s'affranchir.

Il s'agit ici de montrer comment cette sensibilité d'écart toléré à la norme est en fait intrinsèquement lié à la possibilité d'analyse, et donc de montrer comment un changement dans la capacité de collecter et traiter l'information change drastiquement le niveau de tolérance d'écart à la norme. Cet aspect technique rend le problème de la *privacy* nouveau aujourd'hui, à cause des nouveaux supports fournis par Internet et les techniques de fouilles de données. Parce que les possibilités de mesure ont changé, il faut repenser l'importance accordée à sa protection.

La motivation initiale de Warren et Brandeis est en effet déjà un changement technique. Si les paroles s'envolent, les écrits restent, et le préjudice du à la publication de faits personnels *via* des journaux, qui ont des conséquences très différentes d'un simple ragot, demande une évolution du concept de *privacy*. Alors que le ragot est vu avec circonspection et meurt chassé par le suivant, une publication peut être ressortie à une data ultérieure, et fournit une précision accrue et qui ne se dissipe pas dans le temps. Le stockage numérique, qui permet les traitements automatisés et les recoupements rapides, demande donc de façon similaire d'être étudié techniquement pour comprendre les implications nouvelles qui y sont liées.

En particulier nous verrons comment la puissance numérique actuelle permet à la fois de changer d'échelle de couverture, c'est-à-dire d'extraire des informations simultanément sur une multitude de personnes, d'échelle de précision des informations, à la fois en fiabilité et en capacité de comparaison, et d'échelle dans l'acceptabilité, c'est-à-dire de l'intrusion perçue par l'individu au niveau de la collecte voire de l'exploitation des données.

2. Internet, de l'anonymat à la ferme à informations

> There are sufficient interests to move the Net95 from a default of anonymity to a default of identification
>
> **Code 2.0, Lessig**

L'espace numérique a tout d'abord été décrit comme un espace ou l'anonymat régnait, offrant un espace virtuel à tous ceux frustrés par le réel, dans lequel ils pouvaient être complètement eux-mêmes. Derrière un commentaire, un post de blog, les handicaps et les différences restaient invisibles. La diversité des secrets est infinie, et n'est pas nécessairement honteuse : il peut s'agir de vouloir être écouté pour ce qu'on dit plutôt que pour son physique pour une très belle femme, ou encore d'avoir plusieurs identités selon ses humeurs et ne pas être relié à sa position publique officielle[Les2009].

Cet anonymat du Web 1.0 permet d'éventuelles dérives que Lawrence Lessig décrit également. Ainsi, de simples commentaires haineux fait sous anonymat peuvent détruire l'ambiance entière d'un forum, comme l'auteur en a fait l'amère expérience. Un élève de son cours avait posté sur le forum du cours des propos avec une agressivité systématique, qui avait poussé l'ensemble de la classe à se désinvestir y compris lors des interactions réelles. Chacun était devenu un suspect potentiel des propos haineux. OBien que dépité, l'auteur n'avait pas voulu utiliser les données techniques du site (logs) pour découvrir qui se cachait derrière le pseudonyme, car il souhaitait ne pas ressentir contre un élève précis dans la vie réelle la répugnance que les propos lui avaient inspirée dans le cadre virtuel. À l'époque, il s'agit donc de différencier le réel où la relation professeur élève doit être conservée, de l'espace dérégulé encore en construction que constitue internet.

Ce sentiment d'impunité, de zone de non droit sans responsabilité, semble en effet dominé les débuts du Web, et influence encore la rhétorique sur la nécessité de lever l'anonymat, de n'autoriser que les « vrais comptes ». En effet, face à l'explosion de cyberharcèlement, avec près de deux adolescents sur cinq ayant subi des propos haineux en ligne, le besoin d'agir est tangible [DoS2021]. Parce que ce n'est pas l'espace réel, parce qu'il n'y pas les mêmes structures que dans la vie réelle mais des relations entre profils qui portent des idées similaires, rien ne vient freiner la violence ni même faire ressentir au profil haineux qu'il agit de façon répréhensible.

Cependant, si la levée de l'anonymat est parfois populaire dans les débats grand public, ce choix n'atteint pas la même popularité pour chez les associations de défense, qui

voient dans l'éducation et la régulation deux approches complémentaires. L'anonymat se révèle aussi être un refuge pour les victimes.

Toutefois, l'anonymat évoqué est celui des pseudonymes des comptes au niveau visible par les utilisateurs, au niveau de l'interface graphique qui se présente à eux lorsqu'ils se connectent *via* leur moteur de recherche favori. Il s'agit donc d'un anonymat affiché par le site, mais qui n'est qu'une infime partie de ce qu'il se passe effectivement lors d'une interaction avec internet.

En effet, l'évolution d'un internet très décentralisé, où les hébergeurs et les contenus étaient de tailles humaines, vers un internet principalement gouverné par quelques plateformes a néanmoins changé la donne. Aujourd'hui, un utilisateur *lambda* est parfaitement identifiable, ses comptes sont reliés à des données réelles, comportant généralement des moyens de paiement et ses différents comptes sont connectés entre eux. Le web est moins une combinaison de sites statiques et de forums gérés par de petits acteurs (Web 1.0) que le lieu de collecte et de partage d'informations ininterrompu (Web 2.0).

Une technique traditionnelle pour délimiter ce qui relève de la *privacy* est de se référer au consentement de l'utilisateur. Dans le cadre d'un portrait par exemple, il ne peut être diffuser qu'avec l'accord de la personne représentée. Parce que l'importance accordée à une information peut dépendre de l'individu, certains seront indifférent à la diffusion de leur image, tandis que d'autres s'y opposeront. Selon les coutumes et la position de l'individu, une tenue particulière peut appartenir à la sphère privée ou non, et cette délégation de pouvoir à l'individu contourne le problème d'expliciter ce qui relève de l'intime ou non de façon générique. On enlève le diktat d'une règle générale et rigide de la définition de l'intime pour la remplacer vers une pratique au cas par cas, puisque c'est l'utilisateur qui fixe la limite.

On comprend donc que le consentement soit une idée particulièrement séduisante lorsqu'il s'agit de légiférer sur la protection des données. On met en place une règle générale forçant le consentement, et le discernement de l'utilisation vient répondre aux cas particuliers. Cependant, cette technique sous-entend certaines hypothèses :

— l'information est personnelle et n'appartient qu'à l'utilisateur concerné ;

— le contenu est clairement circonscrit ;

— les conséquences de la divulgation du contenu sont prévisibles au moment où le consentement est récolté ;

— le consentement n'est pas biaisé ;

— le consentement peut être retiré si l'utilisateur le souhaite.

Or aucun de ces principes ne s'avère vérifié en pratique. La notion de donnée personnelle, qui pourtant est au cœur de la réglementation européenne avec le RGPD, ainsi que la plupart des régulations mondiales, n'est pas adaptée aux résultats empiriques

d'utilisation des données. Détaillons rapidement pourquoi les assertions précédentes sont erronées.

L'idée que l'information est possédée par un unique individu revient à sectionner virtuellement les liens qu'il possède au sein de la société. Un exemple classique est l'analyse d'ADN. *A priori*, rien n'est plus individuel que cette séquence qui encode génétiquement l'individu. Pourtant, cela révèle aussi potentiellement les histoires de toute une famille, des prédisposition génétiques dangereuses pour l'ensemble des parents proches, voire tout simplement des liens de parenté qui étaient inconnus. Tant que le stockage de l'information est coûteux, et la possible d'effectuer des recherches limitée par un procédé manuelle, l'approximation selon laquelle l'information n'appartient qu'à l'individu est pertinente, car l'information est quasi-certainement perdue, rendue inutilisable et non transmise avant qu'elle puisse servir à des recoupements.

On peut donc la considérer personnelle, car elle est difficilement exploitable dans le but d'obtenir des informations sensibles sur d'autres personnes que l'individu qui a décidé de faire le test. Mais lorsque la pratique évolue et qu'au contraire la plus-value du stockage est établie – par exemple dans ce cas pour la recherche scientifique où l'observation du génome est riche d'enseignements, et les algorithmes de comparaison de séquences un domaine actif,– le paradigme « 1 donnée = 1 individu » tombe en défaut. Une compagnie d'assurance peut par exemple en déduire une augmentation des risques pour tous les parents proches des personnes testées, utilisant donc une information pour laquelle les personnes en question n'ont jamais consentie à la collecte.

De même, les recoupements rendent impossibles d'établir en amont la portée d'une donnée. Un exemple certes assez basique mais toutefois emblématique par la portée des conséquences est l'utilisation d'applications gourmandes en données personnelles par des militaires. Un premier scandale a lieu avec Strava en 2017 [Ash2020]. Cette application permet d'optimiser ses performances sportives, et demande l'accès à la géolocalisation du téléphone. On comprend assez facilement que ceci permet d'en déduire des informations sur la vitesse de déplacement du porteur du téléphone, information définitivement judicieuse pour quantifier des performances sportives. Cependant, plutôt que de simplement exploiter ces données pour en extraire les mesures sportives, la localisation était complètement collectée et gardée par l'application. En 2017, une carte du monde montrant les chemins utilisés par les utilisateurs est partagée publiquement. Sur cette carte, on distingue clairement les bases militaires secrètes des États-Unis.

Depuis, la réponse n'a pas été dans une obligation de meilleure information de l'utilisateur ou une limitation des autorisations de collecte mais simplement par l'interdiction au sein de l'armée de recourir à des applications de fitness développées par des tiers. D'autres façons de révéler des secrets d'États à des start-up ont depuis vu le jour. A l'heure de l'écriture de ce mémoire, la dernière affaire en date sont les applications d'aide à la mémorisation, qui ont été utilisés par certains soldats pour s'entrainer à des tests internes. Les cartes mémo sont en fait, dans la version gratuite de la plupart des

sites, visibles par les autres utilisateurs. Une simple recherche de mots-clés a permis de récolter des informations sur les bases nucléaires américaines en Europe [Foe2021]. Il semble raisonnable d'affirmer que les militaires n'avaient pas conscience de la brèche de sécurité qu'ils créaient en consentant aux conditions d'utilisation de cette application.

Pire encore, l'information sous-tirée à un instant donnée peut *a posteriori* être valorisée via des associations (jointures) difficilement imaginables, reliant par exemple l'orientation sexuelle à une simple notation de film sous pseudonyme (3). Le dommage est aussi sans commune mesure par rapport à l'époque pré-moteurs de recherche. Si une information est reliée à un individu à un instant donné, elle peut se diffuser pour un coût absolument négligeable contrairement à l'information sur support papier. Même si la copie originale est désindexée – c'est-à-dire qu'elle n'apparait plus dans les moteurs de recherches, – des copies peuvent indéfiniment refaire surface. Le consentement, qui peut être théoriquement retiré est donc en pratique confisqué lors d'une mise en ligne, car même si la plateforme originale de diffusion respecte celui-ci, la mise hors ligne n'est jamais définitivement acquise, car d'autres acteurs possédant des copies ne sont pas toujours correctement inclus dans la relation de consentement.

Les conditions du consentement éclairé ne sont pas réunies dans les cas d'usages numériques. Les exemples mentionnées, non exhaustifs, montre que les utilisateurs ne circonscrivent pas quelles sont les conséquences de la collecte de données qu'ils acceptent, et la portée dépassent leur cadre individuel, sans leur permettre le contrôle de l'information. Quelques données techniques expliquent les observations précédentes.

En consultant les prix d'AWS[AWS2021], en optant pour un accès d'un délai de l'ordre des quelques millisecondes, le coût du du gigat-octet de stockage est de 0.0125 dollars par mois, sachant qu'un stockage papier d'un giga-octet est de l'ordre de grandeur d'une belle armoire remplie intégralement de papier. N'importe quel particulier peut facilement accéder à des stockages de l'ordre de plusieurs téraoctets. Stocker les moindres détails de la navigation web d'internet est donc devenu non seulement possible mais même rentable : le profilage des individus à grande échelle permet d'optimiser les ventes, les assurances, les recrutements, les profils à risque, de revendre et de croiser ces données sans relâche et de passer à léchelle la quantité d'individus analysés.

Les moyens de collectes se multiplient. Derrière les « cookies » se cachent le traçage des utilisateurs d'un site à l'autre, de nombreux sites demandent la géolocalisation et l'« intérêt légitime » derrière lequel peuvent s'abriter les entreprises pour collecter des données peut être limité à la simple optimisation de l'interface graphique. À titre d'exemple Fullstory propose d'enregistrer complètement le comportement des utilisateurs, en transmettant les questionnaires abandonnés en cours de route aussi bien que les mouvements de souris. Il est possible pour un développeur web d'avoir le même

niveau d'intimité que s'il se positionnait juste au dessus de l'épaule de l'utilisateur [2020]. Ceci est un exemple du niveau de détail que peut prendre la collecte de données de façon banale aujourd'hui.

De l'anonymat d'un terminal aveugle, incapable de distinguer les utilisateurs des uns des autres, Internet s'est transformé en une collecte sans limite d'informations personnelles. D'un type d'entrées, uniquement textuel et codifiée, la prise d'information est devenue protéiforme et invisible pour lutilisateur. Sa frustration est mesurée par les mouvements de sa main tenant son portable, son état nerveux en vitesse de frappe, ses différents écrans et comptes sont reliés entre eux pour une expérience toujours plus personnalisée, et toujours plus invisiblement invasive.

C'est en effet le troisième pilier de cette révolution : l'information n'est pas uniquement collectée et traitée plus facilement, l'utilisateur est également inconscient de la perquisition numérique qu'il subit. On lit sa correspondance, mais ses emails ne sont pas décachetés, on le piste dans tous ses déplacements, mais nulle filature n'est visible, on observe toute sa navigation mais personne ne se penche au dessus de son épaule.

Si la richesse des données pour mettre au point des algorithmes menant une discrimination massive semble évidente, le milieu lui même a tardé à prendre conscience de la valeur des données ainsi collectées. L'échec de l'anonymat en ligne n'a été acté qu'avec la démocratisation de la fouille de données et quelques ré-identifications célèbres, et le potentiel totalitaire qui en découle n'a peut-être été rendu complètement tangible qu'avec l'institution du crédit social en Chine.

3. Un désordre bien ordonné : le miracle de la réidentification

Any structure will necessarily contain an orderly substructure

Ramsey Theory

Le mot ordinateur a une étymologie bien particulière en français. Du latin « ordinat », qui signifie ordonne, il s'agit d'un mot certes désuet, qui signifie alors « Dieu qui met de l'ordre dans le monde ». C'est IBM qui choisit en 1955 ce terme parmi plusieurs propositions. [1]. Ce terme peut sembler moins adéquat que la traduction anglaise *computer* qui traduit bien une machine à calculer. Cependant, ce terme est en fait également très juste. La possibilité de copier du texte, très rapidement est avec très peu d'erreurs et de comparer des contenus est en fait une fonctionnalité tout aussi disruptive que le calcul numérique dans les ordinateurs. De l'amas de données, il est donc possible de créer un ordre, de mettre en forme, de rendre accessible le contenu.

Cette faculté a faire émerger l'ordre à partir de la donnée est souvent assimilée à l'apprentissage automatique voire plus généralement à l'intelligence artificielle. Pourtant, quand on parle de fouille de données et de valorisation, voire plus prosaïquement de nettoyage de données, il s'agit d'un processus bien peu créatif. Mais parce qu'il est possible de l'automatiser et de brasser des quantité inhumaines de données, le résultat peut être très informatif et permettre par exemple de réidentifier des individus dans des données *a priori* non personnelles. La puissance de copie, de calcul et de stockage est alors le cœur de la technique, qui permet de transformer le désordre de la collecte en valeur.

Pour comprendre à quel point cette ré-identification est à la fois facile et contre-intuitive, commençons par reprendre les définitions proposées par le Règlement Général sur la Protection de Données(RGPD) [Con2016]. Ce règlement a été introduit en 2018, législation alors ambitieuse pour protéger l'ensemble des citoyens européens. Elle est pour autant largement décriée, vu comme inapplicable par de nombreux acteurs et surtout obsolète dès son entrée en vigueur. Il s'agit de modérer ces critiques par la propension des acteurs de l'informatique à s'estimer fondamentalement différent de toute autre formes de progrès technique et à ce titre non régulable. De fait, à la fois les nouvelles régulations européennes (AI act) et extra-européennes se sont ins-

1. IBM est alors un acteur crucial qui introduit le PC, et s'illustre déjà par la diversité des applications possibles, et aux conséquences éthiques. https ://goomics.net/302/

pirées de ce règlement [2]. En recherche informatique, cette régulation est régulièrement citée comme motivation, et elle a donc également un effet performatif fort. L'article 4 définit en particulier les acteurs et les objets de la régulation, et a donc introduit une structure pour penser la protection des données.

La première définition du RGPD est celle de données à caractère personnelles, qui est plus simplement traduite par *personal data* en anglais. Est une « données à caractère personnel », « toute information se rapportant à une personne physique identifiée ou identifiable (ci-après dénommée "personne concernée") ; est réputée être une "personne physique identifiable" une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale ; »

Avec cette définition, on se heurte à deux difficultés. La première est qu'elle repose sur une notion de personne physique identifiable, qui essaie donc de définir ce qu'est un individu d'un point de vue du cyberespace, où le terme anglais de *data subject* est peut-être plus parlant.Définir un individu dont l'existence et l'unicité est ici actée d'un point de vue numérique grâce à une liste de facteurs est en partie arbitraire (quels sont les facteurs qui délimitent les traces d'un individu données, par exemple pour un enfant dont les traces numériques sont en grande part la résultante des adultes qui exercice l'autorité sur lui) mais qui ne pose pas nécessairement de questions directement informatique. L'autre partie, qui consiste à définir le lien entre les données et cet individu, qui est pourtant centrale à la définition, est totalement laissé en friche. Il n'y a que la notion de « se rapportant à ».

Ce flou aurait pu conduire à une application très restrictive pour les responsables de traitement (traduction de *data processor*) en considérant que toutes les données collectées à partir d'un terminal individuel sont générées par la personne concernée, et se rapporte donc à cette personne, tombant donc sous le joug de la définition de « données à caractère personnel ». Cependant, l'application effective est beaucoup plus réduite : les données qui ont été en pratique reconnues comme données à caractère personnel sont les données dont il est évident qu'elle identifie uniquement la personne en question. C'est-à-dire les informations valorisables par un être humain directement, par exemple : « cet individu a pour numéro de carte bleue xxx ».

Une deuxième définition vient confirmer la restriction sous-jacente à la première définition. La « pseudonymisation » est définie par « le traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des

2. par exemple le California Consumer Privacy Act qui a pris effet en 2020 ou encore la législation turque qui s'est aussi globalement alignée sur le traitement des données personnelles

mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable ». Il s'agit donc d'un procédé qui, aux yeux du législateur, permet d'éviter de stocker des données à caractère personnel, en enlevant les parties qui font le lien entre la personne et ses données.

Ces deux définitions donnent donc un découpage des problèmes de protection des données en délimitant l'entité à protéger – la personne physique identifiable – et théorise un moyen de lutte avec pseudonymisation. Il faut donc questionner la pertinence de ce cadre.

Lorsqu'on parle de stockage informatique, on se réfère à des bases de données, c'est-à-dire à des grands tableaux [3]. Chaque colonne correspond à un paramètre, tandis que chaque ligne correspond à une entrée (typiquement la contribution d'un utilisateur par exemple). L'idée de la pseudonymisation, c'est de distinguer colonnes identifiantes et colonnes non identifiantes. Par exemple, imaginons qu'un site internet permette de noter des films. Une base de données de la forme suivante pourrait être extraite 3.1

Id	Name	Age	Pseudo	Blade Runner	Parasite	Gataca
195923	P. Aaron	42	Riley	5 (14/06/2012)	2 (03/03/2020)	-
195924	A. Smith	58	Darcy	3 (05/08/2011)	-	1 (07/10/20017)
195925	J. Brown	19	Dee	-	5 (01/04/2021)	3
195926	C. Doe	27	Emery	1 (26/11/2018)	4 (15/07/2021)	2 (18/02/2019)
195927	E. Garcia	31	Jude	-	-	4 (07/10/2015)

Table 3.1. – Table de données fictive de notations de films

Ici on peut donc décider que les deuxième et troisième colonnes doivent être retirées. Ce très clairement des données permettant d'identifier une personne. Ensuite, par mesure de précaution, il semble aussi judicieux de retirer le pseudonyme, qui est souvent réutilisé par un individu entre différent site internet, et facilement ré-identifiable (par exemple parce que certains comptes contiennent aussi l'identité, voire des photos.) Reste alors uniquement un identifiant unique, qui a été choisi au hasard par le site web pour assurer que chaque utilisateur est identifié de façon unique, et les notes des films ainsi que le moment de leur notation, visible dans la table 3.2.

Ceci correspond donc à la notion de pseudonymisation introduite par le RGPD, et c'est intuitivement une façon performante de protéger la *privacy* des utilisateurs. En effet, on peut encore construire des profils type d'utilisateurs et donc améliorer les systèmes de recommandations, ce qui améliore selon les métriques usuelles l'expérience utilisateur, mais il n'y a plus de données à caractère personnel.

C'est dans cette optique que le jeu de données de Netflix a été ouvert au public dans une compétition où le premier participant réussissant à améliorer les prédictions de

3. comme dans Excel par exemple, mais sans les bugs ou les éléments de mise en forme

Id	Blade Runner	Parasite	Gattaca
195923	5 (14/06/2012)	2 (03/03/2020)	-
195924	3 (05/08/2011)	-	1 (07/10/20017)
195925	-	5 (01/04/2021)	3
195926	1 (26/11/2018)	4 (15/07/2021)	2 (18/02/2019)
195927	-	-	4 (07/10/2015)

Table 3.2. – Table de données fictive de notations de films, pseudonymisée

scores de 10% par rapport à l'algorithme alors en vigueur remportait une mise d'un million de dollars. La foire aux questions comportait alors une question sur la *privacy* : « Is there any customer information in the dataset that should be kept private ? »

La réponse est tout à fait compatible avec le RGPD dans son application actuelle :

> *« No, all customer identifying information has been removed ; all that remains are ratings and dates. This follows our privacy policy, which you can review here. Even if, for example, you knew all your own ratings and their dates you probably couldnt identify them reliably in the data because only a small sample was included (less than one-tenth of our complete dataset) and that data was subject to perturbation. Of course, since you know all your own ratings that really isnt a privacy problem is it ? »*

Deux chercheurs ont pourtant questionné cette évidence [NS2006]. En guise d'exemple, ils collectent une autre base de données directement accessible sur Internet, 6000 utilisateurs rassemblés dans la base de données MovieLens. À partir de cette base auxiliaire, une première tentative est de réaliser ce qu'on appelle une jointure en base de données.

Une jointure consiste a faire se correspondre les valeurs de deux tables en les reliant entre elles. Pour cela, le plus simple est d'avoir une colonne commune : si la base de Netflix contenait encore les pseudonymes, cela permettrait d'aligner facilement les lignes qui se rapportent au même individu, et donc de créer une unique table rassemblant les données des deux tableaux. C'est en premier lieu ce procédé qui est désigné dans la définition de pseudonymisation par « avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles ». Ici la donnée annexe est publique et ne saurait donc être vue comme suffisamment séparée, mais on pourrait croire qu'on a enlevé les éléments permettant de réaliser une jointure.

En effet, les seuls éléments qui permettent a priori de faire le lien entre les deux tables sont les notations elles-mêmes, et on pourrait se dire que noter un film comme des milliers d'autres personnes, ne nous fait par sortir du rang. Cette impression de banalité est en fait fausse. L'article montre que très peu de paramètres créent déjà des schémas uniques. Un utilisateur a plus de 9 chances sur 10 d'être identifié si l'attaquant connait 3 ou 4 notes avec un intervalle d'incertitude sur la date de notation de trois jours (voir figure 3.1).

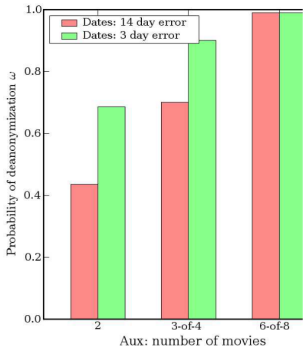


Figure 1: De-anonymization: adversary knows exact ratings and approximate dates.

FIGURE 3.1. – Illustration issue de la publication[NS2006]. On voit que la probabilité de ré-identification augmente rapidement dès que plusieurs données sous partagées, et que même une marge d'erreur sur les dates de notation ne permettent pas de briser l'unicité des profils.

Les auteurs soulignent deux aspects reliant cette ré-identification à une perte substantielle de vie privée. Certains utilisateurs peuvent choisir de noter publiquement des films populaires mais noter en privé des films plus engagés, qui permettent de supposer leur orientation sexuelle ou politique. Relier le profil public aux notes privées peut donc être utilisé à des fins de profilage sans que les utilisateurs aient consenti à divulguer leur données. Cette notion de profilage est également définie dans le RGPD comme

> toute forme de traitement automatisé de données à caractère personnel consistant à utiliser ces données à caractère personnel pour évaluer certains aspects personnels relatifs à une personne physique, notamment pour analyser ou prédire des éléments concernant le rendement au travail, la situation économique, la santé, les préférences personnelles, les intérêts, la fiabilité, le comportement, la localisation ou les déplacements de cette personne physique

Même si beaucoup d'utilisateurs ne révèlent sans doute pas de données sensibles, il semble tout aussi probable qu'au moins une partie des utilisateurs ayant souhaiter garder privée les classements de films les voyaient un élément privé, que ce soit par simple honte de leur gout cinématographiques ou parce qu'ils souhaitaient dissimuler leur engagement politique à leur employeur. Comme le soulignent les auteurs, le cas moyen n'est pas forcément la bonne unité de mesure : ce sont les pires cas, c'est-à-dire ce qui se retrouve fortement exposés malgré leur précaution, dont il faut assurer la protection.

On voit donc qu'une base de données qui semblait particulièrement sécurisée révèle des informations que l'on peut raisonnablement qualifiées de données à caractère personnel. D'autres jeux de données ont connu des assauts similaires de ré-identifications victorieuses. Pour les exemples un peu douteux, on peut mentionner les recherches des utilisateurs sur le moteur de recherche AOL, qui avaient été rendues public également à des fins de recherche en 2006. Outre les recherches en soi choquantes (« Comment tuer sa femme », « photos de personnes décédées »,...) il est apparu qu'il l'anonymisation des données, reposant uniquement sur la suppression des colonnes identifiant clairement l'utilisateur, n'était pas suffisantes. Le New York Times a donc contacté Thelma Arnold, veuve sans histoires, qui a reconnu avoir effectué un certain nombres de recherches compromettantes [4] [Ohm2009].

Ces exemples reposent tous sur le même mécanisme sous-jacent : même si les données sont humainement inexploitables, elles possèdent une structure intrinsèque forte, dont l'exploitation mène à la reconstruction. Ceci a été particulièrement étudié par Latanya Sweeney qui avait fait sensation en 1997 en envoyant son dossier médical au gouverneur du Massachusetts pour prouver le danger de rendre public les données de santé des citoyens sans garantie d'anonymisation suffisante. Elle n'avait utilisé que des données licitement consultable pour réaliser cette ré-identification ciblée. Comment est-ce possible ?

La façon la plus simple de comprendre la ré-identification est de la voir comme la reconstruction de données manquantes avec de nombreuses contraintes. Dans un grille de Sudoku, peu de chiffres sont initialement affichés. Mais ils suffisent déjà à reconstruire l'intégralité de la grille, car d'autres contraintes sont présentes. Ces contraintes peuvent être encodées directement dans un type de programmes particulier, qu'on appelle « solveur linéaire », qui va calculer automatiquement, grâce à des heuristiques très perfectionnées, quelle est la solution. Si les avancées de ces solveurs linéaires sont moins médiatisées que celles en machine learning, elles sont pourtant très marquantes, et permettent de considérer la ré-identification des habitants d'un département comme une tâche de la complexité d'un sodoku pour un humain.

La force de ces techniques est de permettre une résolution de plus en plus rapide, mais aussi de savoir résoudre des problèmes malgré des inexactitudes (petit nombres de données erronées, données imprécises ou manquantes) et reposant sur des relations de plus en plus complexes. Aujourd'hui, il est possible, pour tout organisme de taille raisonnable, de valoriser des données en les croisant efficacement avec des données publiquement disponibles. De plus, chaque donnée réidentifiée donne davantage de prise pour la prochaine attaque. Imaginons que vous fassiez partie des utilisateurs réidentifiés dans le jeu de données de Netflix. Si vous faites des recherche sur AOL sur les films que vous avez aimé peu de temps après les avoir vus, vous êtes d'autant plus

4. des recherches comportant notamment « soixantenaire célibataire » et « chien urinant partout »

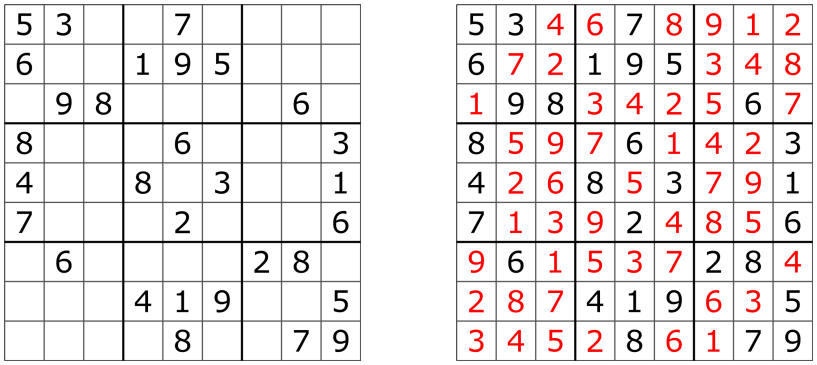


Figure 3.2. – Exemple de reconstruction de données complètes à partir d'information partielle. La vision « pseudonymisation » du monde considère que la grille de gauche protège les cases vides car elles sont vides. Les attaques de ces dernières années illustrent comment la structure des données permet de récréer les données manquantes et donc de supprimer l'anonymat annoncé. L'existence de telles structures dans les données de grandes dimensions ne peut pas être évité.

identifiable dans ce second jeu de données, et toute l'information s'agrège si bien qu'il devient à peu près impossible de poster un fait nouveau sur Internet sans qu'il ne puisse être rapproché du profil déjà existant.

Enfin, il ne s'agit pas seulement de résoudre le problème de la perte de compétitivité due à l'impossibilité de rendre des données publiques. Il s'agit aussi de savoir comment une entreprise peut collecter des données sans devenir la proie d'une cyberattaque. Cette activité est en plein extension, très lucrative et relativement peu risquée, d'autant plus que les défenses des entreprises sont souvent très faibles. Ainsi, IBM estime que les fuites de données ne sont détectées en moyenne que 280 jours après l'intrusion initiale, et 80% des fuites concernent des données personnelles. Avec un coût direct élevé pour l'entreprise, de l'ordre de 100 dollars par utilisateur aux données compromises, entre les frais de pertes de chances pour l'entreprise, les frais de détection puis de réparation, les cyberattaques sont devenues un risque conséquent pour les entreprises [Sec2020]. Le nombre d'attaques et les fuites données sont en augmentation constantes et la dynamique actuelle, vers un accroissement du numérique, laisse penser que la tendance ne devrait pas s'inverser dans les années à venir.

Il y a donc aujourd'hui une nécessité de ne plus se reposer sur une anonymisation factice à base de pseudonymisation. Reconstruire, à grande échelles des bases de données, que se soit en croisant les jeux de données accessibles ou en recourant à des cyberattaques est une réalité qui ne peut plus faire débat. Même en étant particulièrement prudent, la méthode fondée sur la séparation entre donnée personnelles et non personnelles ne donne pas de garantie pour l'avenir. Dès lors que des données intéressantes et exactes sont conservées dans un jeu de données, une nouvelle collecte postérieure

peut venir sagréger sur les données précédentes, et, de jointures en jointures, de résolutions de système linéaires en techniques de ré-identification, la précision des faits se referme inéluctablement autour d'un unique individu. Le mécanisme était impossible à la main, son automatisation est une profession à part entière aujourd'hui.

4. Quand les données personnelles deviennent des flots

Nous avons vu comment la limite entre données personnelles et techniques perdait de sa pertinence en raison de la facilité à croiser et extraire des informations toujours plus intéressantes de bases de données diverses, par agrégation successives. De même, savoir si une information est dévoilée ou non ne peut donc être vue comme une simple donnée binaire, car elle dépend des autres informations disponibles et de comment celles-ci sagrègent au cours du temps. Similairement, Helen Nissenbaum a proposé la théorie de l'*intégrité contextuelle* pour analyser et détecter lorsque des utilisations de données vont à l'encontre de la *privacy*.

On y retrouve la disjonction entre *secrecy* qui se contenterait de réguler quels informations doivent rester non collectées, et celles qui peuvent l'être, pour penser plutôt les problèmes de *privacy* comme une écologie des flots d'informations, où l'analyse ne se restreint pas au type de contenu mais au contraire doit comporter cinq éléments [Nis2010]

— Le destinateur, source du flot d'information ;

— Le destinataire qui reçoit cette information ;

— La personne physique identifiable (comme nous l'avons vu dans la définition du RGPD) ;

— Le message ;

— Les principes de transmission.

Il s'agit donc tout d'abord de faire réapparaitre le schéma classique de la communication – destinateur, message, destinataire – et d'y ajouter la personne physique identifiable dont le rôle apparaissait auparavant. Le destinateur et les destinataire sont souvent dans ce cadre des personnes morales et non physiques : site internet, entreprise, institutions. L'accent est cependant aussi mis ici sur les moyens de la transmission, qui peut être très varié, et qui change l'acceptabilité du flot en question. Le flot peut, à titre d'exemples, résulter d'un achat, de la contrainte, se faire avec le consentement, avec un mandat, avec réciprocité... Il faut donc vérifier la cohérence de ces cinq éléments, et non seulement le message pour détecter la légitimité d'un flot.

L'intégrité contextuelle est conservée quand le flot d'information est conforme aux les normes sociales qui le régissent usuellement. Parce que ces normes peuvent différer selon les coutumes et les époques, l'adéquation d'un flot évolue avec la société et peut

dépendre des individus. Cette approche permet de détecter une violation des normes quand bien même la donnée qui est en cause n'est pas nécessairement entièrement secrète ou personnelle. Ainsi, il ne s'agit pas, par exemple, d'empêcher par principe la collecte des données de santé, mais de garantir que les destinataires peuvent effectivement être restreints aux personnels de santé grâce à un flot d'information suffisamment sécurisé, et dans un but bénéfique au patient, c'est-à-dire que les données collectés sont un moyen d'améliorer les soins qu'il reçoit. De même, il n'y a pas de possession par défaut de ses propres données : il n'est pas possible par exemple de communiquer à un patient les conclusions de ses séances avec son thérapeute sans l'accord du thérapeute, bien que ce soit « ses » données de santé. Il s'agit donc de dépasser la disjonction données personnelles ou non, car elle ne représente qu'un seul des cinq points de la définition. Il est possible pour un même élément d'avoir des conclusions différentes selon le contexte.

De même, réduire la *privacy* au choix individuel de la personne physique identifiable n'est pas suffisamment pour comprendre la dynamique. Westin avait suivi cette voie en proposant de répartir en trois catégories les individus : les fondamentaux, les pragmatiques et les non-concernés [Wes1967]. Les fondamentaux seraient systématiquement inquiets des conséquences des collectes de données, les non-concernés n'en auraient cure tandis que les pragmatiques incarneraient un juste milieu qui évaluent les conséquences au cas par cas. On peut noter que les études ultérieures montrent une part croissante de fondamentalistes et décroissante de non-concernées. D'après cette approche, la régulation doit se concentrer sur le modèle du « pragmatique », qui n'a pas d'*a priori* marqué sur le respect de la confidentialité des données par les entreprises, mais qui juge de façon éclairé en prenant en compte le rapport coût bénéfice à chaque opération. Les comportements des non-concernés et des fondamentalistes sont traités par Westin comme des propensions individuelles par rapport à un comportement pragmatique sain, qui peuvent être jugulés *via* les choix personnels d'autorisation ou de refus de partage des données. Pourtant, on observe bien expérimentalement que les variables liés au contexte, c'est-à-dire au principe de transmission, sont des variables explicatives très marquées en comparaison de l'appartenance aux groupes définis par Westin[MN2017].

Remettre en contexte les flots d'informations pour déterminer quelles sont les attentes en termes de *privacy* permet de comprendre comment le Big Data et les nouvelles technologies, même si elles ne collectent pas forcément des données auparavant jamais collectées, ni ne s'adressent à des individus différents, soulèvent pour autant des problèmes inédits. En effet, les destinataires peuvent en revanche y être fortement élargis, et les formes de transmission sont également changeantes [BGN2017]. La systématisation de la collecte et le déséquilibre permanent entre les individus soumis à la collecte et l'opacité des collecteurs génèrent des modes de transmissions nouveaux.

En effet, si l'information est stockée de façon peu sécurisée, il est légitime de consi-

dérer que la donnée peut *de facto* avoir n'importe quel destinataire, y compris ceux particulièrement à éviter : les conversations et prises de positions personnelles qui sont communiqués aux supérieurs hiérarchiques, les données de santé qui sont revendues aux assurances santé, les données bancaires accessibles aux escrocs les plus divers sont des cas classiques de données dont la collecte n'est ni nouvelle ni ne doit être prohibée, mais dont les destinataires inappropriés sont désormais plus probables.

Les flots d'information ne s'arrêtent pas à la collecte par une entreprise ou un gouvernement, mais les données sont souvent retravaillées pour être revendues à des tierces paries, éventuellement enrichies par différentes bases de données qui permettent de modifier le message initial du flot en lui associant une interprétation plus riche. Pour assurer le respect des normes, une partie de la solution peut être le résultat d'implémentations dans les algorithmes manipulant ces données. Comment assurer mathématiquement, informatiquement que les flots sont adéquats ? Qu'il ne seront pas redirigés vers d'autres destinataires, ni que leur mode de partage sera changé ?

La confidentialité différentielle est un des éléments techniques qui répond à ce formalisme. Le message considéré est celui des sorties d'un algorithme entrainé sur une base de données, et la quantification se fait selon une notion d'adjacence qui reprend l'échelle de la personne physique identifiable, tandis que le destinataire reste libre. Comme nous le verrons dans la partie suivante, un nouveau principe de transmission est proposé, qui se fonde sur une quantification de l'information transmise. Il s'agit de majorer la quantité de corrélation qui peut être extraite à partir du message. On réduit donc, comme l'a déjà fait Shannon [Mac2002], l'information à une quantité (bit d'information) indépendamment de son contenu, et on utilise cette quantification pour mesurer la *privacy*.

La dissymétrie entre le collecteur et le collecté

> Une Intelligence Avenante logée comme une araignée au fond d'une base de données pense à eux, amoureusement, à chaque instant. Elle accueille sans se lasser, le plus infime, le plus intime, le plus insignifiant de leur comportement, l'interprète comme un désir secret, pour un jour pouvoir y répondre, au bon endroit et au bon moment.
>
> **Alain Damasio, Les furtifs**

Cette première partie a donc illustrée les changements significatifs de la *privacy* dans le cadre d'une augmentation de la numérisation. De nouvelles dimensions doivent maintenant être prises en compte pour comprendre la signification des collecte et des traitements de donnes effectués. L'individu se retrouve notamment dans une position très différente et très vulnérable en comparaison de celui qui valorise et exploite la donnée.

On relie encore souvent le numérique à un monde sans loi et anonyme, mais cet imaginaire ne résiste pas à la réalité de la collecte massive. La perception du risque d'utilisation déloyale en est donc souvent sous-estimée, ce qui fausse d'autant la possibilité de consentir. Ces facteurs construisent une situation où l'on peut donc surveiller bien plus d'individus qu'avec les méthodes manuelles avec un coût abordable, où les informations que l'ont peut obtenir sont beaucoup plus précises que ce dont on pourrait rêver auparavant, et où le tout se fait sans que les individus puissent de façon claire et transparente savoir quel est le niveau de surveillance qu'ils subissent.

En effet, il s'agit de considérer que l'information dévoilée l'est de façon irréversible, puisque la copie est quasi instantanée et sans coût, et que ce qui la rend expressive est particulièrement son agrégation avec les données d'autrui, limitant donc la possibilité de lutte à l'échelle individuelle. Les techniques de ré-identification peuvent être vues en elle-même comme un bouleversement technologique.

Ainsi, le changement d'échelle du *Big Data* induit un changement sémantique : mettre en ligne une donnée aujourd'hui n'a pas le même sens que la partager dans un monde analogique. Le niveau d'intrusion possible, le niveau de personnes visées, la vitesse d'application est sans commune mesure avec le passé. Si je souhaite protéger une information, l'anonymiser n'est plus une réponse valable. Tout peut se révéler données personnelles pour peu d'être en quantité suffisante, et savoir quelles informations ont été collectées sur lui-même ne permet pas à l'individu de savoir quels sont les torts qui peuvent en dérouler, quels bénéfices en extrait l'entreprise ou le gouvernement.

Dans ce cadre où la *privacy* est repensée pour correspondre à la réalité technologique,

plusieurs outils émergent pour garantir des protections aux utilisateurs. Que ce soit motivé par des raisons éthiques, légales et commerciales, savoir comment garantir que l'utilisation d'une donnée est respectueuse de la *privacy* est une demande croissante. Dans ce cadre la confidentialité différentielle, propose une démarche particulièrement emblématique.

Deuxième partie

Privacy pour l'apprentissage automatique, la confidentialité différentielle

5. Apprentissage automatique ou généralisation automatique ?

> Differential privacy is a definition of privacy tailored to the problem of privacy-preserving data analysis
>
> **Cynthia Dwork, ACM**

L'apprentissage automatique, plus connu sous la traduction anglaise de *Machine Learning*, est un des champs de l'intelligence artificielle. Il se consacre à des méthodes où le cœur de l'algorithme est occupé par les données, et non par des règles pré-établies. On peut en première approche voir les algorithmes traditionnels comme des boites noires prenant des données, appliquant un traitement, qui a été défini à l'avance et suffisamment général pour convenir pour l'ensemble des entrées possibles. On peut donc y voir une certaine forme d'explicabilité : n'importe qui ne pourra pas expliquer comment marche le programme, mais on peut espérer que les développeurs pourront justifier les lignes de code écrites, soit en y reliant directement une sémantique précise, soit en redirigeant vers d'autres sources qui relie ces codes avec des propriétés théoriques connues.

On peut même souvent relier les règles implémentées par l'algorithme à des règles « métiers » c'est-à-dire à un savoir issu d'un autre domaine de connaissances. Si je souhaite que mon algorithme retourne la position d'un solide tombant en chute libre depuis une altitude donnée h_0, la physique vient à mon secours pour me donner la formule :

$$h(t) = h_0 - gt^2$$

De ce savoir issu des lois de la physique et codifiée mathématiquement, je peux en déduire un algorithme qui prend en entrée le temps et retourne la position, et facilement encoder celui-ci dans un langage de programmation. On transfère donc un savoir déjà existant, indépendant des données reçues, vers un algorithme puis un code.

En opposition, l'apprentissage est dit *automatique* parce qu'il doit créer, grâce à une propre structure interne, ses propres règles. Il n'hérite pas de connaissances pré-établies ; il doit se contenter au contraire de discerner par l'apprentissage quelles sont les règles pertinentes.

On peut donc dire que l'on inverse la place de l'algorithme et des données (voir fi-

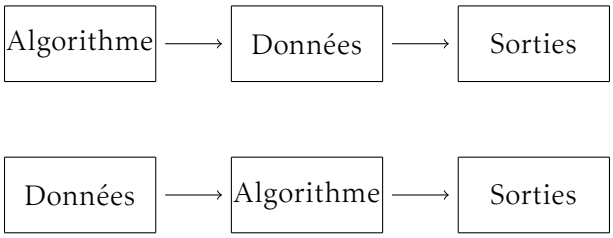


FIGURE 5.1. – Schéma des deux paradigmes : en algorithmie traditionnelle, les données sont traitées par l'algorithme, tandis qu'en apprentissage automatique l'algorithme est lui-même défini par les données. L'algorithme a toujours des *a priori* sur les données, mais les paramètres sont inconnus en amont de l'exécution, donc l'algorithme n'est plus antérieur aux données mais au contraire produit par celui-ci.

gure 5.1. Alors que l'algorithme pré-existe et est indépendant des données rencontrées dans un schéma de programmation traditionnel, il est conditionné par l'exposition aux données dans le cadre de l'apprentissage automatique. Plus qu'un « apprentissage » automatique, il s'agit donc d'une généralisation automatique à partir d'un jeu de données.

Détaillons donc les hypothèses à l'œuvre dans l'apprentissage automatique. On considère, de façon générale, que les données sont des éléments d'un ensemble $\mathcal{X}$ et qu'il faut fournir une sortie pour chaque entrée dans un élément d'un ensemble $\mathcal{Y}$ de sorte à minimiser (ou maximiser) un objectif préétabli. Une représentation des données, le choix de l'objectif et la méthode d'optimisation composent alors l'apprentissage d'un modèle statistique. Dans un modèle dit *supervisé*, on suppose disposer de paires $(x, y) \in \mathcal{X} \times \mathcal{Y}$ [1]. Ces couples sont appelés le *jeu d'entrainement*. On va présenter des entrées x à l'algorithme, et lui demander de se corriger si la sortie qu'il a donné diffère du y correct. Ce procédé de rectification au fur à mesure que l'algorithme tente d'apprendre les réponses correctes est l'*entraînement* du modèle. L'objectif à minimiser est alors une fonction proche de celle du nombre d'erreurs commises.

Pourquoi seulement utiliser une fonction proche et non pas le score parfait pour tous les exemples dont on dispose ? Le but final n'est pas de savoir la réponse qui a été définie comme correcte pour chacune des paires connues, tâche pour laquelle un simple dictionnaire, qui mémorise les sorties correctes une fois pour toute est optimal. Le but ici, et c'est le pari qui n'a rien d'anodin de l'apprentissage automatique, est de généraliser à partir des exemples d'entrainement la situation réelle dans son entière complexité [GBC2016]. De même qu'un bébé à qui on a désigné quelques chats va pouvoir pointer du doigt un nouveau représentant de l'espèce féline, on souhaite qu'un modèle d'apprentissage supervisé puisse catégoriser correctement des exemples qu'il

1. le symbole $\in$ signifie l'appartenance et $\times$ le produit cartésien. On lit donc une paire (x, y) dont la première composante est un élément de l'ensemble $\mathcal{X}$ et la seconde un élément appartenant à l'ensemble $\mathcal{Y}$

n'a jamais vu auparavant.

Cette capacité de généralisation correspond à ce que l'on évalue dans un premier temps au moment de test du modèle, puis lors son déploiement dans un cadre réel. La vraie fonction à minimiser est le nombre d'erreurs que le modèle fait sur les prédictions qu'il génère à partir du modèle appris, c'est-à-dire le nombres de mauvaises réponses qu'il fera sur des données qu'il n'a jamais vues auparavant. Ceci n'est par définition pas mesurable en amont, puisque l'on cherche à résoudre des exemples dont on ne dispose pas encore, mais on suppose que réduire l'erreur sur le jeu d'entrainement permet d'approcher suffisamment bien cette résolution.

Pour que ceci soit possible, il faut donc que deux conditions soient réunies. D'une part, il faut que cette catégorisation existe et soit accessible *via* l'entrée qui est fournie. D'autre part, il faut que le modèle utilisé soit suffisamment expressif pour pourvoir discriminer les différentes catégories et suffisamment simple pour pourvoir être efficacement entrainé sur le peu de données disponibles.[Mac2002 ; GBC2016]

La première condition peut être assurée de différentes façons : dans l'exemple précédent, puisqu'un être humain est capable dé savoir si une photo représente un chat ou un chien, la tâche est soluble. L'information présente sur l'image, même s'il est difficile de formuler des règles métiers pour la codifier via un algorithme traditionnel, est suffisante pour trancher. Parfois on peut connaitre la réponse y par d'autres méthodes, et ne pas avoir d'alternatives pour résoudre la tâche. On peut simplement avoir une forte présomption qu'il doit exister un moyen de discriminer les différentes catégories.

C'est le cas de la détection de cellules cancéreuse par exemple. Il est difficile pour un humain, même qualifié, de savoir catégoriser cellules saines et cancéreuses dans un stade précoce. Le résultat est par contre enregistrable *a posteriori*. On peut donc avoir un jeu d'entrainement qui catégorise les cellules, et espérer une détection plus fine grâce à des modèles d'apprentissage plus perfectionnés que ce que peut faire un être humain. Cette capacité sur-humaine va souvent demander avec un certain prix. Pour que le modèle soit performant, il faudra suffisamment de données, et d'une bonne qualité, alors même qu'il s'agit de données sensibles. Parfois, on tente aussi d'effectuer un apprentissage sans avoir la certitude que le jeu d'entrainement possède suffisamment de richesse pour permettre un apprentissage fonctionnel. Par exemple, peut-être que dans un stade trop précoce, il n'existe aucun marqueur visuel de la cellule cancéreuse, et que le modèle, aussi bien pensé soit-il ne peut faire mieux que l'aléatoire.

La deuxième condition, à savoir choisir un modèle adapté à la tâche considérée, est encore plus complexe à remplir. Il faut donc définir ce que l'on sait sur les données à venir. Plus exactement, l'apprentissage automatique utilise ici le monde des probabilités pour formaliser le raisonnement. On suppose que l'ensemble des données qui vont être observées sont issues de tirages aléatoires à partir d'une même loi de probabilité. On parle également de *réalisations*. Cette loi est décrite mathématiquement par des paramètres qui sont appris lors de l'entrainement. Illustrons cette démarche sur un

exemple très simple unidimensionnel.

Imaginons que l'on souhaite prédire la position d'un solide qui se déplace à vitesse constante inconnue en ligne droite au cours du temps. Pour cela, on dispose d'un certaine nombre de points de mesure à différents instants qui forment donc des paires $(t, x) \in \mathbb{R}^+ \times \mathbb{R}^+$ de réels positifs. La mesure elle-même est sans doute bruitée en raison de l'imprécision des capteurs. Cependant, une hypothèse forte mais souvent nécessaire va être de supposer que les capteurs sont non biaisés, c'est-à-dire qu'il se trompe aussi souvent en mesurant au-dessus ou en dessous de la valeur réelle. On possède alors un ensemble de points que peuvent s'exprimer comme la somme des deux éléments :

$$\hat{x}(t) = v_0 t + z_t$$

où la valeur mesurée à l'instant t est notée $\hat{x}(t)$, et où z_t est le bruit du capteurs que l'on va généralement supposer suivre une loi Gaussienne, dont la variance, c'est-à-dire la propension à s'écarter de sa moyenne est un paramètre dépendant du capteur. Dans ce cadre, à partir de multiple paires (t_i, x_i) il est possible de calculer la valeur la plus probable de v_0 (cf 5.2). Les données d'entrainement permettent donc d'en déduire de façon générale la réponse qu'on obtiendrait avec le capteur, à l'exception du bruit. Si on nous demande de prédire l'emplacement à un nouvel instant t, on peut alors répondre

$$x(t) = v_0 t$$

Où v_0 a été calculé expérimentalement. On notera que pour faire une prédiction correcte ici, il fallait que le modèle de prédiction soit suffisamment riche pour admettre une dépendance de la réponse avec t : on ne peut pas bien approximer le résultat avec un simple modèle constant. Au contraire, le modèle est plus simple à résoudre (un unique point permet déjà une approximation) que si on avait enlevé l'hypothèse de vitesse continue. si par exemple on ne pouvait garantir qu'une accélération constante, mais pas forcément nulle, il aurait fallu au minimum deux exemples pour fixer les deux paramètres :

$$x(t) = \frac{1}{2} a_0 t^2 + v_0$$

La seule différence est qu'en apprentissage automatique, les modèles sont généralement beaucoup plus compliqués qu'une régression linéaire, et comportent facilement des millions de paramètres, ce qui signifie qu'il y a un changement d'échelle. Des modèles différents vont permettre d'apprendre à distinguer des critères de classification différents, et doivent donc être choisis en fonction des connaissances que l'on possède sur les données (cf 5.3). La grande taille des modèles utilisés ne permet de plus pas forcément de tracer efficacement les dépendances possibles entre un élément du jeu d'entrainement et le modèle qui a été appris. Les choix sont donc plus difficiles à effectuer, et ne correspondent pas toujours à l'intention en petites dimensions.

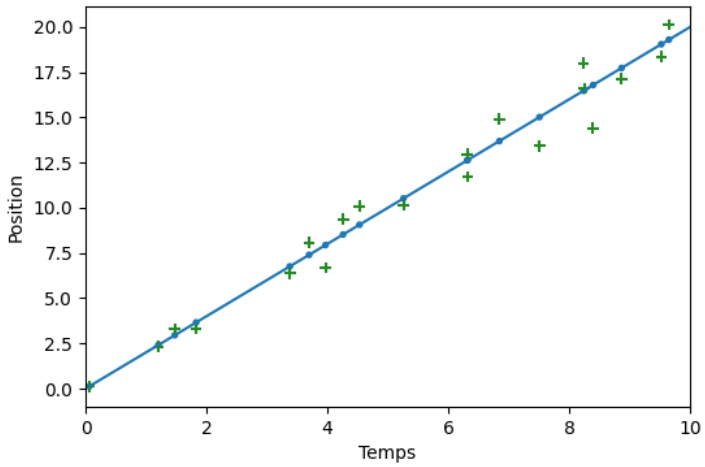


FIGURE 5.2. – Exemple d'apprentissage à partir des données. Les points d'entrainement (verts) sont des pairs où l'entrée x est le temps et où l'on observe y la position du solide. On ajoute le savoir *a priori* que la vitesse de déplacement est constante, ce qui garantie que le modèle appris est une droite. On prend la droite qui minimise les erreurs par rapports aux points mesurés, qui contiennent une part de bruit. Pour un point donnée, la prédiction qui serait faite est le point à l'intersection de l'abscisse et de la courbe apprise (points bleus)

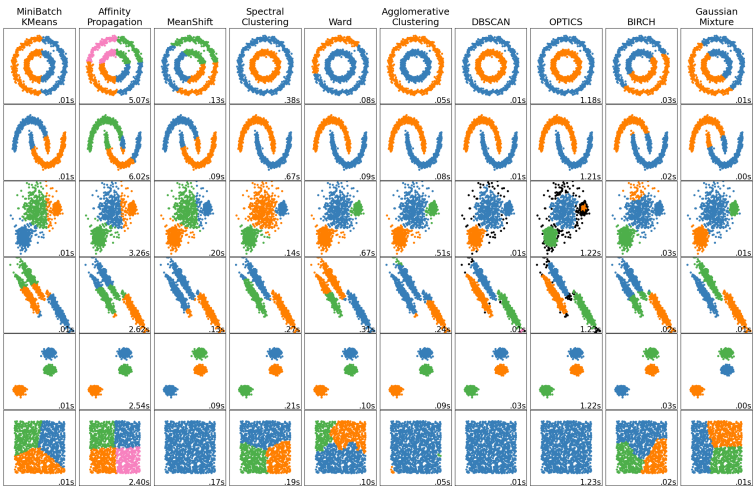


FIGURE 5.3. – Image tiré de `scikit-learn`, une bibliothèque classique d'apprentissage automatique. Pour différents jeux de données, les modèles les plus performants ne sont pas toujours les mêmes, et leur choix constitue un enjeu phare du domaine

Au cœur du processus de généralisation se situe donc l'invariance entre la source des données permettant l'entrainement, et des données que l'ont cherche à prédire ensuite. Avoir ajusté les paramètres sur le jeu d'entrainement doit permettre d'avoir les bons paramètres pour les données inconnues. En termes probabilistes, il faut donc que les données suivent la même loi de probabilité indépendamment de leur appartenance au jeu d'entrainement. Par exemple, dans notre étude de cas précédente, il faut sans doute que notre solide soit le même, ou du moins ait exactement les mêmes caractéristiques de déplacement, entre jeu de test et entrainement pour que les prédictions soient précises.

Une autre façon de voir cette invariance, et qui est particulièrement utile dans le cadre de la confidentialité différentielle, est qu'on ne souhaite pas mémoriser les exemples de jeu d'entrainement, mais uniquement leur essence. Il ne faut que la règle générale, mais pas ces incarnations particulières. Comment dans ce cadre, garantir mathématiquement que le modèle appris n'a pas gardé en mémoire des artefacts du au hasard de l'apprentissage, mais uniquement les invariants que l'on cherchait à capturer ?

En effet, le risque de mémorisation d'exemples spécifiques est un risque bien documenté dans le cadre de l'apprentissage automatique et en particulier pour les réseaux de neurones profonds. À titre d'exemple, Copilot [Zie2021], qui propose de générer automatiquement du code à partir d'une spécification en langage naturel a soulevé de nombreux doutes, y compris juridiques, chez certains utilisateurs. Ce modèle a été entrainé sur l'ensemble des codes publiés sur Github [2], qui peuvent donc être disponibles sous des licences variées. Il s'avère que si la spécification du code demandée est trop spécifique à un unique exemple présent sur Github, le modèle peut retourner non pas un code généré à l'occasion de cette demande, mais un morceau de code « appris par cœur ». Les concepteurs pensent même éventuellement ajouter une option permettant de détecter ces réponses de « par cœur » pour les signaler à l'utilisateur qui devra alors décider s'il peut réutiliser ce code ou s'il doit créer sans l'aide de Copilot sa propre version.

De la même façon, il est possible que ce soit en ayant accès au modèle entrainé ou juste à ces prédictions, de reconstruire le jeu d'entrainement, ce qui peu donc poser d'importants problèmes de *privacy*, puisque les données d'entrainement sont donc reconstruites et disponibles à partir du modèle.[Pap+2017 ; ZLH2019 ; Sho+2017 ; ACW2018]

On peut voir la démarche de l'apprentissage automatique à la lumière de la théorie de l'information [Mac2002 ; CT2006]. Cette théorie initiée par Shannon se donne pour objectif de quantifier l'information, ce qui en fait un aspect centrale des théories de codage, de cryptographie, et dans une moindre mesure actuellement, de l'apprentissage automatique [TZ2015]. Au sein de cette théorie, on peut définir l'information mutuelle de deux variables aléatoires A et B comme le niveau de renseignement que l'on peut

2. Github est un service web d'hébergement et de gestion de développement de logiciels, très populaire : c'est l'endroit de référence pour trouver du code sur internet.

extraire réciproquement l'un de l'autre, que l'on note $I(A : B)$.

On peut alors voir le modèle d'apprentissage comme une façon T d'encoder le signal d'entrée X de sorte à pouvoir sauvegarder la sortie voulue Y (voir figure 5.4). Pour saisir les composantes les plus générales et non celles propres au jeu d'entrainement, il s'agit donc de maximiser l'information mutuelle entre l'encodage et la sortie $I(T : Y)$ tout en minimisant la capacité de remonter de l'encodage à l'entrée $I(X : T)$.

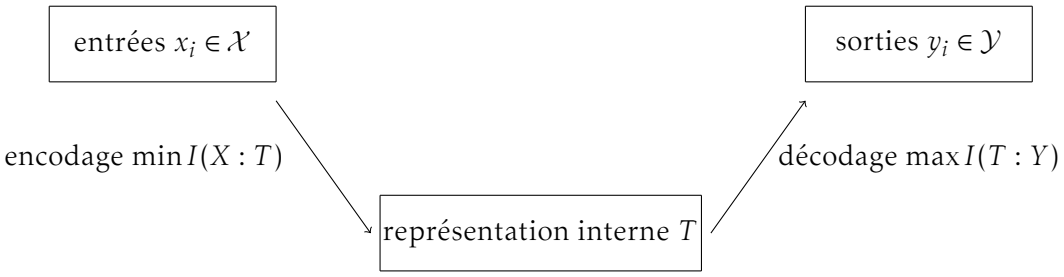


FIGURE 5.4. – Schéma de l'apprentissage automatique vue sous le prisme de la théorie de l'information. Ce compromis entre minimisation et maximisation est l'*information bottleneck*

6. La confidentialité différentielle

> La quantification ne fournit pas seulement un reflet du monde, mais elle le transforme, en le reconfigurant autrement
>
> **Alain Desrosières et Sandrine Kott** [DK2005]

La confidentialité différentielle a été introduite en 2006 dans le papier [Dwo+2006]. Le titre *Calibrating noise to Sensitivity in Private Data Analysis* décrit un exemple particulier de mécanisme qui permet de garantir un certain niveau de la quantification qu'ils définissent à cette occasion. La recherche postérieure a montré que cette définition avait un potentiel bien plus large que l'application initiale, et le champ de recherche associé s'est développé, valant notamment aux auteurs le prix Gödel en 2017.

Comme introduit précédemment, on se place ici dans un cadre probabiliste où la sortie d'un modèle d'apprentissage automatique est, non pas une réponse unique et déterministe, mais au contraire une distribution de probabilité, dont on va observer un tirage aléatoire particulier. Dans la section précédente, on supposait ainsi que le point était le résultat d'un tirage sur une variable aléatoire gaussienne centrée en la véritable valeur de la position, et on observait en entrée la valeur bruitée. Le bruit était le résultat d'un tirage suivant une distribution Gaussienne. On définit de façon générique un *mécanisme* pour traduire le passage d'une entrée en une sortie probabiliste. Partant d'une base de données X dans un ensemble $\mathcal{X}$, un mécanisme $\mathcal{M}$ retourne un tirage suivant une loi de probabilité sur l'espace des sorties $\mathcal{Y}$. On peut donc noter :

$$\mathcal{M} : \mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$$

Ici, la longue flèche désigne une fonction qui va donc associée à un élément de l'ensemble de gauche un élément de l'ensemble de droite. Le $\mathbb{P}$ désigne les probabilités sur l'ensemble $\mathcal{Y}$, dont la définition formelle sort du cadre de ce mémoire.

La confidentialité différentielle quantifie comment cette distribution de probabilité va être modifiée en fonction de la contribution d'un utilisateur. Au sein d'une base de données, un des enregistrements, qui correspond à l'entité dont on cherche à assurer la *privacy*, peut virtuellement être modifié. Deux bases de données X et X' sont dites *adjacentes* lorsqu'on peut passer de l'une à l'autre en modifiant un unique enregistrement, et on note alors $X \sim X'$. Assurer la *privacy*, nous dit la confidentialité différentielle, revient à garantir un certain niveau d'indiscernabilité entre les bases de données

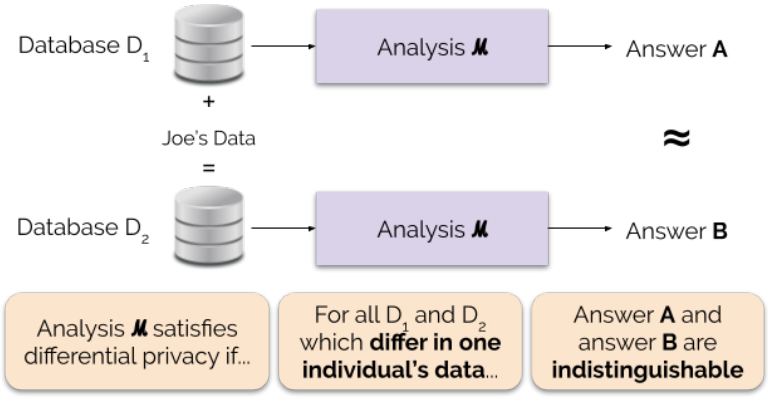


FIGURE 6.1. – Visualisation graphique de l'intuition de la confidentialité différentielle. La réponse observée ne doit pas être fortement modifiée en fonction de la donnée d'un utilisateur fixée. Illustration tirée du blog de l'« US Census » [JB2020]

adjacentes, car cela signifie que les sorties du mécanisme ne permettent pas de remonter à un des enregistrements qui a été utilisé pour produire l'algorithme (cf 6.1). La proximité des sorties est assurée par un ensemble de contraintes sur les sorties. On a, formellement, la définition suivante :

Définition 1 (Confidentialité différentielle). *Soit $\varepsilon \geq 0$. Un mécanisme $\mathcal{M}$ est ε-confidentiellement privé par rapport à la relation d'adjacence $\sim$ lorsque pour toute paire $X \sim X'$, pour tout ensemble mesurable $\mathcal{S}$,*

$$\mathbb{P}(\mathcal{M}(X) \in \mathcal{S}) \leq e^{\varepsilon}\mathbb{P}(\mathcal{M}(X') \in \mathcal{S})$$

où e est la base du logarithme naturel.

Ce jeu de contraintes peut être interprété ainsi : quelles que soient les sorties du mécanismes, la probabilité d'observer cette sortie si la contribution d'un utilisateur était modifiée ne changerait pas qu'un certain ratio, qui est quantifié par le ε. Si on prend un ε très grand, cela signifie que la contrainte est très peu difficile à satisfaire, car le facteur devant la probabilité de droite devient très grand, et donc satisfaire l'inégalité est facile.

À l'autre extrême, si $\varepsilon = 0$, l'inégalité devient en fait une égalité ($\mathbb{P}(\mathcal{M}(X) \in \mathcal{S}) = \mathbb{P}(\mathcal{M}(X') \in \mathcal{S})$) et cela signifie que l'enregistrement dont diffèrent X et X' ne peut pas avoir d'influence sur le résultat : on ne peut donc pas utiliser l'enregistrement un question pour modifier l'apprentissage. En raison du quantificateur universel, on comprend qu'il est impossible de faire un algorithme d'apprentissage automatique $\varepsilon = 0$-confidentiellement privé, car cela revient à dire que les sorties ne sont pas dépendantes de la base de données, ce qui contredit la notion même d'apprentissage à partir des données.

Le paramètre ε va donc quantifier numériquement un compromis entre la quantité

d'information qui peut subsister dans les sorties du mécanisme – ce qui est le fondement de l'apprentissage – et le niveau d'indiscernabilité que l'on requiert. Pour cette raison, on appelle cette variable le *budget de privacy*, qui est donc un nombre réel. Plus il est élevé, plus on dépense de la *privacy via* le mécanisme, c'est-à-dire plus un unique enregistrement risque de produire un biais significatif et facilement observable sur la sortie.

Le papier originel introduit un premier exemple de mécanisme confidentiellement privé. Pour simplifier les notations et la compréhension, nous nous réduisons à nouveau à un exemple unidimensionnel. Considérons une base de données où chaque enregistrement est uniquement composé d'un nombre réel $x \in \mathbb{R}$ et d'une réponse $y \in \mathbb{R}$ qui est comprise entre 0 et 1. Ceci nous permet d'apprendre une fonction f telle que $|f(x) - y|$ soit suffisamment petit, c'est-à-dire que f nous permet de prédire les sorties $f(x) \sim y$ à partir de x. Dans un cadre non privé, on dévoilerait complètement la fonction f. Le mécanisme proposé est l'ajout de bruit à cette prédiction initiale :

$$\mathcal{M}(x) = f(x) + z \text{ avec } z \sim Lap(1/\varepsilon)$$

Cela signifie qu'au lieu d'avoir potentiellement un moyen de revenir à une contribution unique *via* les prédictions du mécanisme, la prédiction est perturbée par l'ajout d'un bruit. Ce bruit est assuré par une variable aléatoire, qui doit bien sûr restée cachée pour qu'on ne puisse pas connaitre le résultat sans ce bruit. L'ampleur de ce bruit dépend du budget de *privacy* : plus le ε est grand, plus la variable aléatoire risque de s'éloigner de zéro, et donc de fournir un résultat éloigné de la version non bruitée. En représentant graphiquement la distribution de sortie, on a donc un pic de probabilité sur la valeur non bruitée, et une décroissance symétrique de part et d'autre. Si on compare les deux distributions, on a donc un ratio qui reste borné.

La *privacy* est donc ici assurée par l'impossibilité de revenir à un enregistrement à partir de l'unique connaissance de la sortie, car exactement le même comportement pourrait être observé avec un enregistrement différent. Imaginons que les enregistrement contiennent une valeur binaire, par exemple « est porteur d'un cancer ou non », il est alors possible d'observer la même sortie globale de l'algorithme, que l'individu soit malade ou non. Même en connaissant parfaitement tous les autres enregistrements, c'est-à-dire en ayant un moyen de voir toute la base de données sauf cette unique ligne, on ne peut garantir la valeur de l'enregistrement. On ne peut donc avoir qu'une certaine probabilité d'accéder à la vraie valeur, qui ne dépasse pas e^{ε}.

La condition d'adjacence donne une garantie sur la *privacy* d'un enregistrement même en connaissant tous les autres éléments de la base de données. Cette condition peut sembler particulièrement forte. Cependant, elle est classique dans le domaine de la cryptographie, où on cherche des garanties contre des attaquants ayant aussi la possibilité de maitriser les autres participants d'un protocole. Ceci permet d'éviter des at-

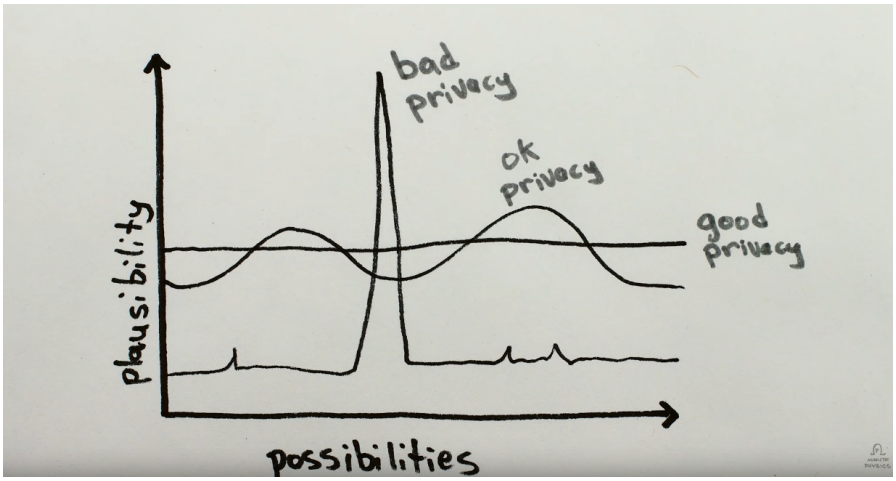


FIGURE 6.2. – Illustration sur le type de distribution de probabilité obtenue avec la confidentialité différentielle. Si la distribution de probabilité comporte un pic trop important qui correspond à la valeur initiale, un attaquant peut facilement inférer la vraie valeur. Au contraire, plus la distribution est uniforme quelle que soit la participation, moins un attaquant peut déduire d'information. Illustration tiré de Minute Physics, en collaboration avec le *US Census*

taques en cascade, où la corruption de quelques utilisateurs entraine progressivement la perte de garantie pour l'ensemble des participants. D'un point de vue strictement pragmatique, il est également souvent plus simple de prouver cette condition plutôt que des hypothèses intermédiaires. Il est également possible de raffiner la notion d'adjacence de base de données pour correspondre exactement à la notion d'attaque souhaitée.

La confidentialité différentielle impose des contraintes fortes sur l'ensemble des sorties possibles : même les sorties très peu probables, qui n'ont qu'une chance sur un million de survenir, ne doivent pas varier de plus d'un facteur e^{ε}. Ces contraintes fortes sont décriées par certains comme étant excessives, pour un bénéfice somme toute limité. Est-ce judicieux par exemple d'ajouter du bruit comme dans l'exemple, ce qui diminue la précision des données, pour une garantie mathématique peu explicite ?

Il faut donc aussi étudier quels sont les bénéfices de cette métrique pour comprendre à la fois son adoption et l'absence de métrique alternative en termes de quantification de la *privacy*.

7. Quelles propriétés mathématiques pour quelle *privacy* ?

> Differential privacy describes a promise, made by a data holder, or curator, to a data subject : « You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available. »
>
> **Cynthia Dwork, ACM**

Pour garantir mathématiquement la *privacy*, il faut définir le type d'attaque et les limites du cadre où la protection doit s'appliquer. L'hypothèse de la confidentialité différentielle est d'autoriser n'importe quel traitement sur les sorties du mécanismes. Tout ce qui peut arriver à partir de ces sorties doit être protégé, sans restriction de capacité de calcul ou de connaissance annexe.

Ceci est souvent formulé en termes de résistance au post-traitement (*post-processing* en anglais). C'est-à-dire que quels que soient les mécanismes ultérieurs appliqués sur les sorties, la confidentialité différentielle reste vérifiée. Cette universalité de la garantie découle de l'approche utilisée. On borne en effet la quantité d'information *via* les contraintes sur la distribution, donc le signal n'est pas simplement offusqué au code, mais bel et bien détruit au-delà du seuil quantifié par le budget de *privacy*. On peut mathématiquement facilement exprimer cette robustesse par : « pour toute fonction f, pour tout mécanisme de qui soit ε-confidentiellement privé, $f \circ \mathcal{M}$ est également ε-confidentiellement privé », où $\circ$ dénote la composition. Il s'agit donc d'une promesse particulièrement forte, comme le décrit Cynthia Dwork, qui ne se limite pas à la réidentification ou à la détection de participation au jeux de données [1], mais à n'importe quel type de traitement possible. Cela peut aussi être vue comme une exigence de *forward privacy* :

> One may state the « forward privacy » property : if someone's privacy is breached (e.g., her anonymous online records have been linked to her real identity), future privacy breaches should not become easier.[NS2006]

La seconde propriété particulièrement désirable est la composition. On peut avoir envie d'utiliser une même donnée de multiples fois, par exemple pour ajuster à plusieurs reprises le modèle que l'on entraine, ou même pour entrainer plusieurs modèles indépendants. Il est donc crucial, pour une quantification de la *privacy*, de pouvoir quantifier le résultat de plusieurs mécanismes, qu'ils soient indépendants ou non, car

1. *membership attack* en anglais

COMPOSE$(\mathcal{A}, \mathcal{M}, k, b)$

Input: $\mathcal{A}$, $\mathcal{M}$, k, b
Output: V^b
  for $i = 1$ to k **do**
    $\mathcal{A}$ requests $(D^{i,0}, D^{i,1}, q_i, M_i)$ for some $M_i \in \mathcal{M}$;
    $\mathcal{A}$ receives $y_i = M_i(D^{i,b}, q_i)$;
  end for
  Output the view of the adversary $V^b = (R^b, Y_1^b, \ldots, Y_k^b)$.

FIGURE 7.1. – Algorithme détaillant le scénario de la composition de différents mécanismes dans [KOV2015]

cela correspond aux situations les plus courantes d'usages de données. Un deuxième enjeu est d'avoir une dégradation la plus lente possible du budget de *privacy* lors des compositions.

La composition est déjà introduite dans l'article fondateur [Dwo+2006]. Cependant, les bornes ont été raffinées dans des cas spécifiques [KOV2015]. La composition naïve prouve que les budgets de *privacy* se somment. Cela signifie que si on considère deux mécanismes $\mathcal{M}_1$ et $\mathcal{M}_2$, définis sur les mêmes ensembles de départ (c'est-à-dire prenant les mêmes types d'entrées), et respectivement ε_1 et ε_2-différentiellement privés, alors dévoiler le mécanisme résultant de n'importe quelle combinaison de ces deux mécanismes est, dans le pire des cas, $\varepsilon_1 + \varepsilon_2$-confidentiellement privé.

La force de cette garantie repose à nouveau dans l'universalité de cette propriété. En effet, le résultat n'est pas uniquement valable dans le cas d'une application indépendante de deux mécanismes, dont on joint ensuite les résultats, mais pour toute construction dans l'organisation de l'utilisation de ces mécanismes. Pour se représenter la force de la garantie, il est coutume d'en appeler à la théorie des jeux et de s'imaginer un combat entre deux adversaires.

Le premier détient la base de données et la possibilité d'y effectuer les calculs tandis que l'adversaire essaie d'« attaquer » en diminuant le plus possible le budget de *privacy* d'un individu donné. Le jeu s'effectue en k manches (cf 7.1). Le jeu commence avec le détenteur des données qui fournit un ensemble de mécanismes à l'attaquant. À chaque tour, l'attaquant choisit le mécanisme qu'il souhaite utiliser parmi l'ensemble des mécanismes disponibles, ainsi que la base de données à l'exception de l'individu cible. Il reçoit en échange la sortie de ce mécanisme. Le but est de maximiser l'information qu'il peut alors extraire de la vue rassemblant l'ensemble des k réponses obtenues.

La confidentialité différentielle garantit alors que l'attaquant perdra toujours si le but est d'attendre un budget de *privacy* supérieur à la somme des ε des mécanismes choisis. Autrement dit, il n'existe pas de façon astucieuse de composer les mécanismes de sorte à diminuer leur protection.

Se contenter d'une garantie se détériorant avec la somme des ε reste cependant trop contraignant en pratique pour les applications réelles. C'est de plus un résultat qui correspond au pire cas possible sur ce qui arrive lors de la composition. En effet, cela revient à dire que les sorties qui ont été observées sont les plus complémentaires possible, et réduisent au maximum l'incertitude observée, ce qui n'a pas de raison d'être dès lors qu'on utilise des mécanismes non pas choisit pour optimiser la surface d'attaque mais l'apprentissage des données.

Pour faire fi de la rigidité de la confidentialité différentielle pure, une variante (ε, δ) a été introduite, et c'est la définition qui est utilisé dans la majeur partie des recherches et des implémentations. Il s'agit d'admettre que dans certains cas, de probabilité suffisamment faible, les contraintes d'inégalités peuvent être violées. La probabilité qu'un tel événement arrive doit alors resté borné par δ. Mathématiquement, cela conduit donc à la définition suivante :

Définition 2 (Confidentialité différentielle avec delta). *Soit $\varepsilon \geq 0$ et $\delta \in [0,1]$. Un mécanisme $\mathcal{M}$ est (ε, δ)-confidentiellement privé par rapport à la relation d'adjacence $\sim$ lorsque pour toute paire $X \sim X'$, pour tout ensemble mesurable $\mathcal{S}$,*

$$\mathbb{P}(\mathcal{M}(X) \in \mathcal{S}) \leq e^{\varepsilon}\mathbb{P}(\mathcal{M}(X') \in \mathcal{S}) + \delta$$

Cette modification change radicalement le type de garantie qui est fournie, puis qu'on peut vérifier cette condition tout en dévoilant complètement les données de certains utilisateurs : il suffit que ce dévoilement est lieu avec une probabilité inférieure à δ. Le gain en flexibilité avec cette définition est cependant tel qu'il semble préférable pragmatiquement d'avoir un ε bien plus petit, et un δ strictement positif, que de rester cantonné à la définition pure. En effet, si un pire cas se manifeste avec un millionième de chance ou moins, et dévoile alors presque surement un point, mais que le reste du temps le budget est cent fois plus petit, connaitre cette garantie est intéressante. Nous verrons par la suite que d'autres compromis ont également été développés (cf 8), qui crée *grosso modo* un continuum : il y a telle probabilité d'avoir tel niveau de *privacy*. L'attente est cependant de se placer dans un régime où le δ reste de l'ordre de la faille théorique uniquement, et n'advient pas en pratique. Il faut que ce pire cas soit vu comme si inédit, pour que le risque reste acceptable par les utilisateurs.

Pour cette raison, il est préconisé en général de ne jamais avoir $\delta > 10^{-4}$, mais aussi de faire décroitre le δ en fonction du nombre d'utilisateurs, par exemple $\delta < \frac{1}{100n}$, voire $\delta < \frac{1}{n^2}$. Ces choix sont aussi des choix éthiques : faut-il choisir une bonne protection moyenne, et autoriser quelques cas pathologiques, ou au contraire se concentrer sur la certitude absolue, même si cela revient à préférer des solutions moins bonnes en moyenne ? Ces choix ont été assez peu discutés dans la littérature pour le moment[LC2011]. Le choix est donc délégué principalement aux ingénieurs qui implémentent les applications concrètes de confidentialité différentielle. La recherche se tourne également vers des métriques alternatives pour éviter d'avoir cet arbitrage au

cours des preuves de mécanismes. Garder variable les valeurs de ε et δ n'est pas nécessairement le signe d'une désaffection des chercheurs des applications réelles dans ce cas précis, mais aussi une simple commodité calculatoire, où tout se fait en toute généralité pour les valeurs de ε et de δ, sans qu'il faille donc restreindre l'analyse à des valeurs précises.

Grâce à la souplesse obtenue *via* le δ, la composition devient beaucoup moins contraignante, puisqu'il est alors possible d'obtenir une dégradation du budget non pas proportionnelle à la somme, mais uniquement à la racine carrée de celle ci. Cela est une diminution suffisante du coût pour étendre assez largement le nombre d'algorithmes qui permettent encore de capter le signal dominant tout en fournissant des garanties de *privacy*.

La résistance au post-traitement et la composition sont donc deux propriétés fortes qui permettent pour un algorithme complexe de déterminer le budget de *privacy* qu'il requiert, et d'assurer que le niveau de celui-ci ne sera pas revu à la baisse ultérieurement. Il s'agit cependant toujours d'une protection relative à la participation, c'est-à-dire comment l'individu qui a participé ne peut avoir perdu plus d'un certain budget de *privacy* par rapport à s'il s'était abstenu. Cela ne borne donc pas le préjudice que peut subir l'individu en termes de conséquences apprises par l'algorithme, mais bien uniquement l'appartenance au jeu de données.

Imaginons qu'on entraine avec confidentialité différentielle un algorithme capable de prédire si un individu va gagner ou non son procès, et que celui-ci fonctionne suffisamment bien. La perte de chance des individus désormais catégorisés comme « perdants » est énorme, car leurs adversaires n'ont plus aucun intérêt à trouver un accord à l'amiable, puisque l'incertitude du procès n'existe plus. Il y a donc une conséquence négative très forte à l'issue de l'apprentissage, mais cela n'est pas du tout vu comme problématique pour la confidentialité différentielle. La protection ne touche que les individus qui ont participé à l'entrainement, dont on ne pourra pas affirmer avec certitude leur participation, ni si eux-mêmes ont gagné ou perdu leur procès avec plus d'assurance que pour les autres individus. On note que si l'algorithme fonctionne bien, accéder aux entrées d'un individu permet bien d'avoir une plus grande confiance en la sortie de l'algorithme qu'auparavant : on ne borne donc pas la quantité d'information qui peut être révélée sur un individu, mais uniquement le gain relatif de connaissance par rapport à l'univers où cet individu n'avait pas fait partie de la base de données.

Enfin, la confidentialité différentielle a un certain niveau d'universalité, c'est-à-dire que si on utilise des mécanismes qui ne génèrent pas de distorsion dans l'information révélée, alors on peut reconstruire l'information. Plus exactement, il est possible de calculer un niveau de distorsion minimum sans lequel l'ensemble des réponses est « manifestement non privé » [2], c'est-à-dire que l'on peut alors reconstruire toute la base

2. traduction libre de *blatantly non-private* [DR2014, chapitre 8]

de données avec un nombre négligeable d'erreurs. « Négligeable » est à prendre ici au sens mathématique : pour une base de données de taille n, il y aura au plus $o(n)$ erreurs, où la notation de Landau signifie que le ratio $\frac{\text{nombre d'erreurs}}{n}$ tend vers 0 quand le n augmente.

Autrement dit, quelles que soient les innovations techniques à venir et les alternatives à la confidentialité différentielle exprimables, on ne peut pas s'affranchir du compromis entre introduction de bruit et d'imprécisions, et l'impossibilité de reconstruction de la base de données [DN2003]. Certains raffinements du calcul du niveau de bruit nécessaire ont été étudiés, dans le cas de grandes bases de données [DN2004] et sur les bornes inférieurs nécessaires pour pallier le risque de reconstruction *via* la confidentialité différentielle [De2011]. Même s'il n'y a pas d'équivalence formelle entre la possibilité de reconstruction et la confidentialité différentielle – on sait juste que l'on est protégé en fonction du budget si on choisit de recourir à la confidentialité différentielle – le principal désavantage de celle-ci est prouvé comme étant nécessaire. On sait donc que toute technique protectrice doit être aléatoire et introduire de la dispersion, ce qui est donc très lié à la notion de confidentialité différentielle, et explique sans doute pourquoi cette définition s'est imposée.

Une quantification par et pour le traitement statistique

> Les équations sont une des nombreuses catégories de traduction, et c'est à la suite de toutes les autres traductions qu'elles doivent être étudiées.
>
> **Bruno Latour, La science en action**

La confidentialité différentielle, même si elle ne fait appel qu'à des notions très simples de probabilité, est bel et bien reliée à l'émergence de l'apprentissage statistique. Il s'agit en effet de tracer le flot d'information en le remontant, des sorties de l'algorithme vers ses entrées.

Un premier prérequis est donc de se placer dans un cadre d'exploitation des données. Contrairement au cadre de l'algorithmie traditionnelle ou des procédés cryptographiques, la donnée possède ici un rôle déterminant. Le point phare de la définition de la confidentialité différentielle est de traiter la donnée comme étant le moyen d'accéder à une vérité, dont on veut supprimer un effet indésirable, qui est la subsistance de la donnée dans les sorties. On traduit donc, *via* cette quantification de la *privacy*, une hypothèse fondamentale de l'apprentissage automatique, qui n'est pas nécessairement vérifiée ou vérifiable, à savoir qu'il existe une structure sous-jacente que l'on peut apprendre grâce à l'analyse de données qui sont représentatives de cette source. Nous avons discuté quelques exemples de ces sources possibles, et de la possibilité d'avoir des données qui respectent cette représentativité et donc d'avoir accès à tout le modèle, et non à un simple morceau de celui-ci, ou à une vision déformée.

Par définition, puisque les sorties du mécanismes ne doivent pas dépendre trop fortement d'un enregistrement spécifique, il faut avoir un nombre d'enregistrements important, ce qui correspond au cadre de l'apprentissage automatique. C'est dans ce cadre que la magie opère, c'est-à-dire qu'il est possible d'extraire un signal sans que le bruit nécessaire à la dissimulation du signal individuel ne soit un obstacle définitif. Quand on met ensemble toutes les contributions, le bruit, qui n'a pas de structure propre, s'annihile. À l'opposé, les signaux présents dans chacun des enregistrements se renforcent mutuellement, et prennent donc le pas sur le bruit.

Borner l'influence, la résurgence d'un individu dans la boite noire construite, c'est la quantification proposée de la *privacy*. La mesure de la différence entre le scénario

incluant l'individu dont on mesure le budget, et le scénario l'excluant est effectuée au travers du calcul de la distance entre les distributions des deux scénarios. Ce calcul peut être accompli de différente façons, avec la version pure ou en acceptant un δ par exemple.

Dans tout la cas, la quantification effectuée de la perte de *privacy* n'est pas une quantification absolue de l'information qui est générée, mais un garantie relative entre deux scenarii, celui où l'individu participe et celui où il s'abstient. Pour cette raison, on peut dire que la confidentialité différentielle protège des inconvénients de la participation à un jeu de données.

Troisième partie

L'individu probabiliste et sa *privacy* quantifiable

8. Interpréter un budget de *privacy*

> Finally, the question of how to intuitively interpret and convey the privacy guarantees and limitations of differential privacy at various privacy loss levels to the public, remains open.
>
> **Jun Tang et al,** [Tan+2017]

La confidentialité différentielle met au cœur de sa définition la notion d'enregistrement, puisque les contraintes sont sur toutes les bases de données qui diffèrent d'un enregistrement. Comment cependant interpréter la garantie qui est obtenue par l'individu, pour un budget de *privacy* fixé ? Le budget mentionné peut avoir des conséquences très différentes selon les données qui sont contenues, selon le vecteur qui est utilisé pour stocker l'information et selon la taille de la base de données.

La compréhension de ce que signifie ce budget à l'échelle de l'individu est cruciale dans la perspective d'un choix éclairé par l'utilisateur lui-même. Pourtant, l'interprétation du simple nombre réel du budget de *privacy* peut varier selon de nombreux paramètres. Ceci a d'ailleurs été un des arguments retenus pour introduire la confidentialité différentielle locale, ainsi que d'autres variantes.

En effet, dans le modèle classique de la confidentialité différentielle, la garantie dépend de l'ensemble du jeu de données. Supposons donc que l'on souhaite calculer une statistique très simple, à savoir le pourcentage d'individus qui sont des femmes dans le jeu de données. Chaque enregistrement contient un booléen, qui est vrai si l'individu est une femme et faux sinon. Quitte à interpréter vrai par 1 et faux par 0, il suffit donc d'additionner toutes les lignes de la bases de données. Si on révèle le résultat tel quel, le résultat n'est pas confidentiellement privé. En effet, il suffit que la base de données ne contiennent que des femmes pour qu'on puisse en déduire que l'individu cible est également une femme.

En revanche, si on ajoute au hasard un entier relatif, on est différentiellement privé. La question est alors de savoir quel est l'amplitude que doit avoir cet entier relatif : peut-on se contenter de $+1, 0, -1$ avec une grande probabilité, ou au contraire prendre des valeurs plus grandes ? On comprend intuitivement que si la base de données est très grande, on espère avoir un effet protecteur, car notre signal est mélangé avec plus d'autres signaux. Cependant, quantifier exactement quelle quantité de bruit est nécessaire dans quelle situation, c'est-à-dire calculer le budget de *privacy*, mais pour une question aussi basique que celle-ci, n'est pas évidente. Comment cette valeur est mo-

	Naïve Composition				Advanced Composition				zCDP				RDP			
ϵ	Loss	1%	2%	5%	Loss	1%	2%	5%	Loss	1%	2%	5%	Loss	1%	2%	5%
0.01	.93	0	0	0	.94	0	0	0	.93	0	0	0	.92	0	0	0
0.05	.92	0	0	0	.93	0	0	0	.92	0	0	0	.94	0	0	0
0.1	.94	0	0	0	.92	0	0	0	.94	0	0	0	.91	0	0	1
0.5	.92	0	0	0	.94	0	0	0	.90	0	0	0	.68	0	3	27
1.0	.93	0	0	0	.93	0	0	0	.88	0	0	0	.51	4	21	122
5.0	.91	0	0	0	.89	0	0	0	.62	2	11	45	.16	39	95	304
10.0	.90	0	0	0	.87	0	0	0	.47	15	38	137	.09	55	109	329
50.0	.65	0	2	16	.64	19	31	73	.15	44	102	291	.02	70	142	445
100.0	.48	6	29	152	.53	18	47	138	.08	58	121	362	.00	76	158	456
500.0	.10	53	112	328	.29	42	88	256	.00	80	159	487	.00	86	166	516
1,000.0	.01	65	138	413	.20	57	111	301	.00	86	172	514	.00	93	185	530

Table 5: Number of individuals (out of 10,000) exposed by Yeom et al. membership inference attack on logistic regression (CIFAR-100). The non-private ($\epsilon = \infty$) model leaks 129, 240 and 704 members for 1%, 2% and 5% FPR respectively.

FIGURE 8.1. – Exemple de quantification du budget via la ré-identification selon différentes attaques et selon les métriques utilisées [JE2019]. La tâche est ici fictive au sens où les images ré-identifiées à partir du modèle appartiennent à un jeu de données très connu (CIFAR) et standard en apprentissage automatique, qui classifie des images d'objets et d'animaux.

difiée, si on prend désormais en compte une troisième catégorie non-binaire ? Faut-il alors bouger les différents compteurs avec plus de bruit, ou avec le même niveau ? La dynamique à l'échelle d'une base de données entière peut être difficile à appréhender, notamment parce qu'au moment où la personne choisit de partager ses données, le reste du jeu de données n'est pas entièrement stabilisé.

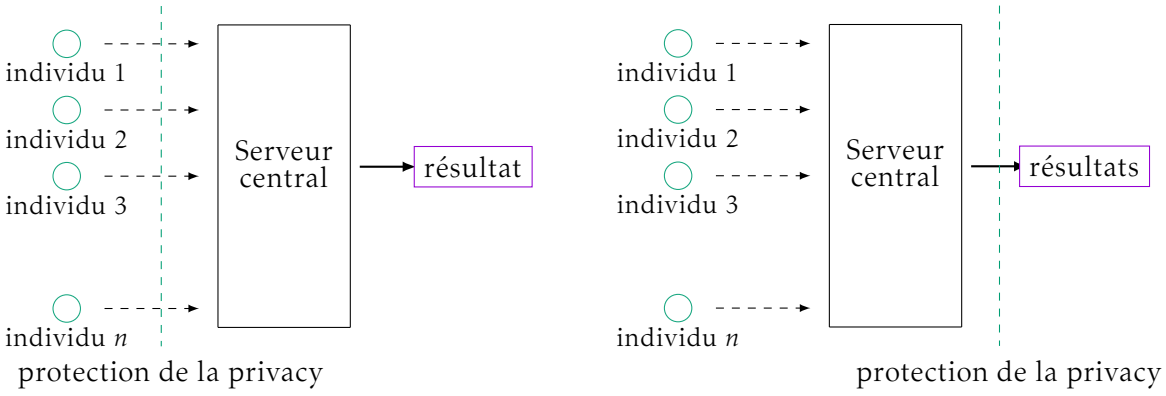
Une approche possible est de présenter une quantification du budget en fonction des attaques connues, ce qui ramène la confidentialité différentielle à une mesure empirique, alors même qu'on souhaite y exprimer un garantie théorique qui soit robuste aux progrès techniques ultérieurs.(cf le tableau 8.1)

Avoir un unique paramètre *via* le budget de *privacy* pallie la multiplicité des situations en proposant un budget unifié pour tous les cas possibles : même si on ne sait pas encore comment on protègera les données, on peut garantir quel niveau de traçabilité sera possible. Mais ce raffinement peut aussi être vu comme un facteur d'illisibilité : un même nombre correspond à des situations très différentes. De plus, il suppose une confiance totale en la personne, l'institution ou le dispositif qui va ajouter le mécanisme nécessaire. Ne pourrait-on pas laisser l'individu assurer lui-même sa *privacy* ?

Ajouter du bruit à l'échelle locale, et non globale, est en fait un procédé plus ancien que la confidentialité différentielle à l'échelle des bases de données, et a donné lieu à la définition de confidentialité différentielle locale (cf 8.2). Son origine viendrait de travaux en sociologie. Comment convaincre les participants de répondre honnêtement aux questions embarrassantes ? Lorsqu'un chercheur souhaite mesurer la proportion d'individus ayant une pratique prohibée, les résultats peuvent souffrir des mensonges

des individus. Plutôt que de révéler une consommation de drogue, un individu peut préférer mentir, malgré le contexte protecteur de l'enquête. En termes statistiques, les échantillons collectés permettent de construire un *estimateur*, mais celui-ci est biaisé. Cela signifie que l'espérance – c'est-à-dire *grosso modo* la moyenne de tous les cas possibles – de l'estimateur est différente de l'espérance réelle des données. Comme il n'y a pas de moyen d'estimer facilement l'amplitude de ce biais [1], cela pose un problème de fiabilité des résultats obtenus qui est difficilement évitable.

En 1965, Warner propose un protocole pour contourner ce problème [War1965]. Chaque individu est invité à lancer une pièce, puis à répondre à la question de façon honnête si la pièce est sur pile, et répondre selon un second lancer de pièce dans le cas contraire (par exemple « oui » si pile, « non » si face). L'idée sous-jacente est que désormais tout le monde est susceptible de répondre « Oui » pour un fait répréhensible, même ceux qui n'ont rien à se reprocher. La réponse est donc toujours liée à l'individu, mais parce qu'elle est en partie aléatoire, l'individu peut toujours prétendre que la réponse transmise correspond à la partie aléatoire.



(a) Avec la version locale de la confidentialité différentielle

(b) Avec la version globale et la confiance dans le serveur qu'elle implique

FIGURE 8.2. – Schéma des flots d'information et de l'emplacement de la protection donnée par ces instances de la confidentialité différentielle

On fait donc porter la responsabilité de la *privacy* sur l'individu lui-même, qui peut plus facilement contrôler le mécanisme, et comprendre ce qu'il signifie. On peut toujours extraire l'information au niveau global. Supposons qu'on récolte 100 réponses *via* ce procédé. Environ 50 sont dues aux lancers de pièce, avec environ 25 « oui » et 25 « non ». Le reste est composé uniquement de réponses qu'on suppose non biaisées grâce au protocole. On peut donc enlever le bruit qui a été ajouté, car on connait son comportement moyen, et avec quelle proportion on peut avoir un tirage déséquilibré,

1. Il semblerait que poser récursivement la question « Avez-vous menti à la question précédente ? » soit une méthode inefficace

donc un peut aussi facilement calculer le niveau de confiance dans la donnée. Par la loi des grands nombres, on sait que le bruit va décroitre proportionnellement au signal en la racine du nombre de participants.

La confidentialité locale est donc l'application de la confidentialité différentielle au cas particulier de bases de données à un élément, réduite à la contribution d'un individu. On déplace la protection de la *privacy* de l'échelle de l'agrégation des données (au niveau de l'entreprise, de l'état, du serveur) à celle de l'individu. Cela a cependant un coût : avec la multiplication des mécanismes locaux, la quantité d'information qui peut être extraite des données est beaucoup plus limitée. On voit donc apparaitre un nouveau niveau de compromis : après un premier compromis entre un budget de *privacy* élevé mais permettant une exploitation plus complète des données, et un budget de *privacy* exigeant mais rendant l'apprentissage complexe, on a de façon similaire un compromis entre mettre sa confiance dans un tiers capable de protéger simultanément de nombreuses données et avoir donc une bonne rentabilité des données, ou au contraire vouloir assurer une protection individuelle et en payer le prix en termes d'utilité des données très bruitées qui vont alors être récoltées.

D'autres paradigmes ont même proposé des compromis plus subtils entre une approche purement locale ou purement globale. Certaines applications spécifiques demandent en effet des petits ajustements. On peut citer par exemple le cadre des algorithmes en ligne, où la base de données n'est pas conservée en entier mais uniquement l'état courant, et qui demande une protection uniquement contre le dévoilement d'une unique étape de calcul, et non contre l'ensemble des états parcourus, qui a donné la *pan-privacy* [Dwo+2010].

Une autre variante, intéressante car elle modifie la structure des flots d'information, est l'exemple de la confidentialité différentielle jointe [2]. En effet, une partie des algorithmes d'apprentissage sont utilisés dans des buts de fournir un service personnalisé. L'algorithme apprend au cours du temps par plusieurs itérations : tout le monde interagit localement avec une version de l'algorithme et y trouve des erreurs et des comportements sous-optimaux. Beaucoup d'erreurs sont cependant des erreurs globales, dont la correction bénéficie à tous. D'autres sont et doivent rester locales, car elles sont spécifiques à l'utilisateur, et relève donc de la personnalisation.

On comprend donc dans ce cadre que l'on souhaite garder le cadre de la confidentialité différentielle au niveau global, c'est-à-dire que le modèle sans personnalisation doit laisser transparaitre le moins d'information possible sur les utilisateurs participant à l'entrainement. Mais on voit aussi que cela limite de façon absurde la qualité du modèle pour l'individu : il ne sert à rien de dissimuler à un individu ses propres données, puisqu'il les possède déjà. Il est donc intéressant d'avoir une protection sur le modèle global qui s'applique pour tous les autres utilisateurs sauf lui-même. Alors

2. *Joint Differential Privacy* en anglais, il n'existe pas de document français connu employant ce terme.

la personnalisation est possible, sans changer les garanties vis-à-vis des autres. On a donc dans cette définition admis un flot dinformation plus complexe que la participation puis la révélation du modèle, pour adopter une métrique adaptée aux aller-retours entre modèle global et local, et on différencie donc le flux qui revient à soi des autres flots.

Un même budget de *privacy* n'a pas donc la même interprétation selon le cadre dans lequel il est appliqué : une protection locale sera vue comme un gage supérieur à une garantie uniquement centrale. De même, les différents choix pour calculer la divergence entre les deux scénarios, avec ou sans l'individu, donnent des garanties un peu différentes, comme on l'a vu entre la version pure avec uniquement ε ou celle avec (ε, δ). D'autres recherches ont mis en avant la confidentialité différentielle de Rényi [Mir2017], qui utilise un calcul entre les distributions de probabilités un peu différent. Il est intéressant de noter que si cela rend l'interprétation du budget complexe, il existe pour toutes ses définitions des résultats permettant le passage d'un formalisme à l'autre.

Dans le cadre de l'introduction de *Randomised Response*, le petit algorithme avec deux lancers de pièces, l'interprétation qui est donnée de ce mécanisme log(3)-confidentiellement privée, est de fournir un déni plausible [3]. En effet, dans le cadre d'un rapport d'humain à humain, en cas de reproches directs, l'individu peut toujours communiquer sur le fait que la réponse était factice. Quel est le niveau d'incertitude pour que le déni soit suffisamment plausible pour ne pas nuire à l'individu ? Cette réponse définit la valeur que doit prendre le budget de *privacy*. Mais on voit à nouveau que le budget acceptable pour une application n'est pas le même pour un autre : peut-être vais-je décider que l'information n'est pas assez fiable dans certains cas, même que la légère hausse de probabilité suffira à discriminer l'individu dans d'autres situations.

À travers ces différentes définitions, on retrouve donc la question des hypothèses sur la force de l'attaquant. Celui-ci discriminera-t-il à partir d'un certain niveau de confiance uniquement ? Ou est-ce que le niveau d'information qui lui parvient est trop faible, trop appauvri pour qu'il ait intérêt à renoncer ? La théorie de l'information nous garantit qu'avec un ε nul, il ne peut pas avoir d'avantage à exploiter l'algorithme, car on ne peut pas recréer l'information à partir de rien. Ensuite, les bits d'information deviennent plus nombreux avec l'augmentation du budget de *privacy*, jusqu'à un niveau proche de la donnée brute que l'on souhaite protéger. On peut donc voir la confidentialité différentielle comme une interprétation sous forme d'optimisation sous contraintes dans le cadre de la théorie de l'information [Mac2002 ; CT2006].

En effet, comme vu précédemment (5), l'apprentissage automatique est déjà aux prises avec cette volonté de minimiser l'information propre au jeu de données tout en maximisant la richesse du modèle entrainé. Quand on ajoute la confidentialité dif-

3. *plausible deniability* est également un concept juridique aux États-Unis, on entend ici ce terme dans son sens littéral, la possibilité de nier de façon plausible que l'on a commis un acte

férentielle, on peut donc interpréter cette condition comme étant une contrainte stricte sur le flot d'information allant du jeu de données vers la sortie. Ce n'est pas exactement l'information mutuelle vis-à-vis des sorties, mais c'est une garantie qui relève de la même dynamique. La question est alors de savoir quantifier quand l'intérêt du modèle appris est suffisant pour compenser le coût de subsistance de la donnée. C'est dans ce cadre qu'ont été proposées des analyses économiques du budget de *privacy*, ce qui nous demande aussi de revenir au lien entre l'individu et la version numérique capturée par l'enregistrement présent dans la base de données.

9. L'utilisateur, l'individu, et l'enregistrement

And now you see another me,
I have been reloaded
Don't shut me down, ABBA

La confidentialité différentielle est une protection fondée sur la notion d'enregistrement, qui définit la notion d'adjacence $X \sim X'$ entre deux bases de données X et X'. Ceci repose sur l'idée d'une bijection entre le monde réel et son pendant informatisé : chaque individu de la population cible pourrait être cantonné à une ligne précise, et réciproquement chaque ligne apporterait de la connaissance sur un individu.

L'anonymat, c'est-à-dire la possibilité de garder les données telles quelles mais en supprimant les liens vers les individus est, comme on l'a vu au début de ce mémoire (chapitre 3), impossible en grande dimension, car il est impossible de couper les liens entre un individu et ses données, et de bout en bout les propriétés d'une personne permettent de converger vers son identité. Mais si ces liens sont impossibles à dissoudre, établissent-ils pour autant une relation simple entre enregistrement et individu ? Sont-ils notamment limités à une unique individu ? Cette problématique se situe souvent en dehors du champ fixé par l'informatique, où le chercheur va le plus souvent considérer un jeu de données qui est déjà fixé. Par conséquent, les articles traitent ce passage entre individu et enregistrement comme une des données fixées du problème. « each row is a record associated with some individual »[NS2006] se contentent de signaler les auteurs dans leur description de l'attaque. Pourtant combien de comptes Netflix sont en réalité partagés par plusieurs personnes physiques identifiables ?

Une autre désignation est celle de « contributeur individuel », qui garde donc l'idée d'un individu, d'une entité bien définie, et l'idée clé de la participation, en affirmant que le but est de quantifier « the privacy loss of an individual contributor to the input dataset »[Fel+2018]. Dans le papier phare qui introduit l'analyse de l'utilisation de la confidentialité différentielle pour l'apprentissage profond, la définition est encore plus elliptique : « a single entry, that is, if one image-label pair is present in one set and absent in the other. »[Aba+2016].

Les papiers consacrés à des attaques de modèles, où les auteurs montrent comment des modèles permettent une ré-identification du jeu d'entrainement se doivent de formuler les conditions de succès de l'attaque, et sont donc un peu plus prolixes. Mais là encore, le rapport entre enregistrement et individu est le plus souvent *ad hoc*

Ainsi, pour un modèle fournissant un pré-entrainement pour faire de la reconnais-

sance de visage, les auteurs montrent que les visages des participants au jeu d'entrainement sont vulnérables :

> This attack violates the privacy of an individual who is willing to provide images of themselves as training data, as the adversary can potentially reconstruct images of every individual in the training set. [FJR2015]

Pour une étude des données spatiales, c'est la révélation de quatre positions avec leurs horaires qui jouent le rôle du représentant réel [Mon+2013].

Ces définitions montrent cependant la diversité de l'entité privée que l'on souhaite protéger : image à reconstruire, données personnelles en fonction du nom [Car+2020], trajectoire complète à partir de quelques points, paire de pseudonymes désignant un même utilisateur. Si on se réfère à la théorie de l'intégrité contextuelle (cf 4), on voit donc que la protection est dépendante du message dans le cadre de confidentialité différentielle, plus que directement de la personne physique identifiable. Les liens vers la personne physique ne sont donc pas forcément l'alpha et l'oméga du schéma de protection, ils peuvent être même considérés comme du bruit par rapport à d'autres mesures.

Ainsi, peu importe que l'algorithme ne sache pas qui est la personne physique identifiable derrière l'enregistrement 325486541 tant que cette donnée peut être exploitée pour calculer le prix maximal auquel la personne est encore prête à acheter : le flou de la définition d'adjacence dans la confidentialité différentielle protège sans s'intéresser au retour au monde réel de la donnée protégée.

En rester à cette abstraction est pourtant limitant sur au moins deux aspects : si c'est l'individu qui choisit le budget de *privacy*, la définition de l'individu reste partie intégrante du problème. D'autre part, comment passe-t-on de l'individu, qui est censé être l'entité à protéger, au contributeur, qui est en pratique l'entité qui décide ou non de partager l'enregistrement ?

On voit donc que si la confidentialité différentielle permet de quantifier le coût de l'utilisation d'un enregistrement, elle réduit trois éléments de l'intégrité contextuelle – le message, la personne physique identifiable et le destinateur – à une unique notion d'adjacence de base de données. L'entité commune qui en découle n'est donc pas nécessairement bien définie comme une personne physique identifiable.

Plus généralement, on va donc réduire la complexité du problème à une discrétisation arbitraire, celle de l'enregistrement. Si la photo comporte plusieurs personnes, est l'œuvre d'un photographe, contient un arrière plan spécifique, tout va être résumé à un seul acteur, celui qui choisit d'inclure ou non la photo dans le jeu d'entrainement. De même, derrière un utilisateur peut se cacher une famille entière, plusieurs amis, schémas qui s'écartent de la simplicité assumée par la confidentialité différentielle.

Certains ajustements sont possibles. Par exemple, la *group privacy* propose de considérer plusieurs enregistrements simultanément jusqu'à k d'entre eux dans la notion d'adjacence, avec à nouveau un passage simple de la définition traditionnelle à celle-

ci. Un mécanisme ε-différentiellement privé est au moins $k\varepsilon$-confidentiellement privée par un groupe d'au plus k éléments, et cette borne est optimale et peut être atteinte en pratique.

On voit cependant qu'on laisse au sein de la définition de la relation d'équivalence un vrai degré de liberté. Ainsi, pour des données génétiques, faut-il rester sur l'échelle de l'individu, ou considérer l'échelle du groupe de la famille, et avec quel degré de parenté ? La quantification de la *privacy* est-elle alors adaptée quand la découpe du flot en enregistrement montre son arbitraire ? Si un enregistrement peut relier plusieurs personnes physiques identifiables, de façon symétrique plusieurs enregistrements peuvent se rapporter à un unique individu. Comment articule-t-on le budget de *privacy* dans le cadre d'utilisateur utilisant plusieurs comptes, justement dans l'objectif de protéger leur vie privée ? Toutes ces questions doivent être arbitrées en amont de l'utilisation des mécanismes de confidentialité différentielle.

Cette discrétisation *via* le découpage en enregistrements est donc, au même titre que le choix du budget de *privacy*, une quantification. Celle-ci se fait en fonction du découpage le plus propice à l'algorithme d'apprentissage, de l'unité qui sera retenue pour le jeu d'entrainement, et non nécessairement pour la cohérence dans le monde réel qu'elle engendre. Ainsi, de façon très terre-à-terre, les auteurs de [Car+2020] conseillent d'éviter les données dupliquées dans les jeux de données pour limiter les risques de fuites : plus l'information est redondante, plus elle risque d'être mémorisée telle quelle par le modèle, même avec un usage de la confidentialité différentielle, car le budget de *privacy* augmente avec la répétition de la donnée, alors même que la qualité de l'apprentissage n'est pas forcément supérieure.

Symétriquement, on peut aussi voir que la valeur apportée par un enregistrement dans un modèle peut grandement varier. Certains chercheurs ont donc proposé une autre quantification, celle de la plus-value apportée à l'apprentissage global. La perspective est alors beaucoup plus économique : comment répartir la richesse créée par un jeu de données entre les différents contributeurs ? Plutôt que de se contenter d'une répartition uniforme, on peut voir certaines contributions comme étant plus structurantes et pertinentes que d'autres.

L'évaluation de données – *data valuation* en anglais – propose une quantification presque complémentaire, dont les méthodes de calculs varient mais sont plus tournées vers l'estimation *a posteriori*. On quantifie l'amélioration spécifique due à une donnée précise, *via* notamment la valeur de Shapley. On a donc exactement le complémentaire de l'évaluation du budget de *privacy*. Dans un cas, on tente de mesurer quel est le niveau de persistance des spécificités individuelles dans le modèle, tandis que dans la *data valuation* on quantifie le niveau de généralité qui est extrait du cas spécifique en question.

Ceci montre à nouveau que la nature de l'information qui est retenue est une information quantifiable, mais non énumérable : ce sont les réels qui sont conviés pour

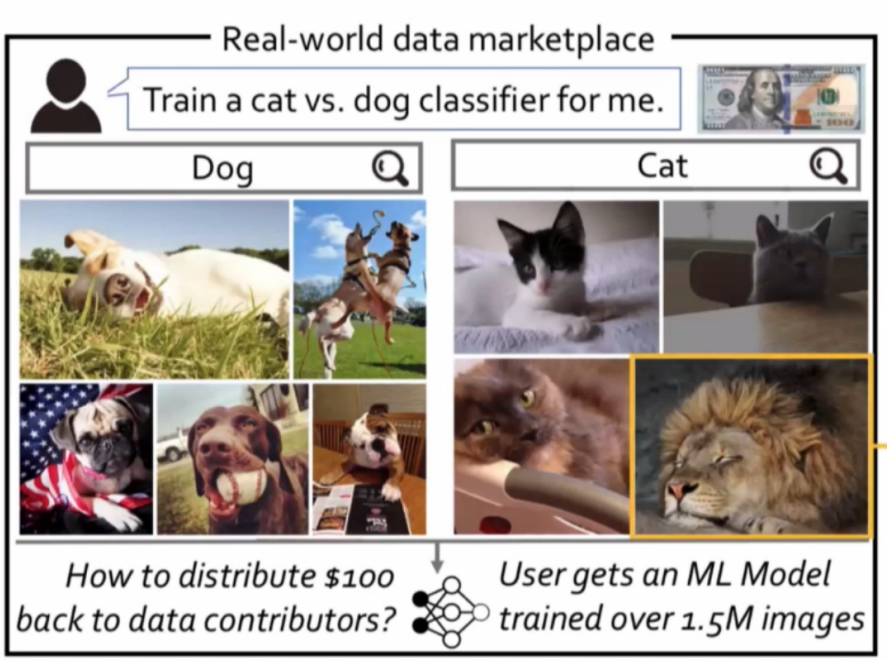


FIGURE 9.1. – Extrait de la présentation de Dawn Song, lors de NeurIPS Workshop SpicyFL. Pour répartir un budget fixé, la proposition est d'évaluer la pertinence de chaque image et de la répartir proportionnellement au gain que constitue l'image pour la tâche. Ici, par exemple, l'image de lion ne reçoit rien du tout, alors qu'avoir un chat sur un chaise est peut-être plus original qu'un chat sur un lit, et « mérite » donc une plus grande incitation financière.

définir la quantité, et non des entiers. On passe donc d'une base de données, où chaque enregistrement est un entité distincte, à un continuum où chacun peut être n'importe quelle fraction. L'information est pourtant totalement identifiée (car reliée à un enregistrement spécifique) mais le niveau de certitude transmis est modulable car on se place dans un cadre probabiliste : doubler le bruit divise par deux l'information par exemple et on peut appliquer n'importe quelle facteur de réduction au niveau d'information contenu dans les données.

Cette approche probabiliste tend donc à inclure la donnée discrète comme un artefact d'un signal continu, du flot d'information. Si on constitue un budget de *privacy*, ou que l'on répartit par fraction la valeur de la donnée, on passe implicitement d'une vision déterministe de l'enregistrement – est-ce que telle ou telle propriété est vérifiée pour cet individu – à une vision probabiliste des données – à quelle point la masse de probabilité se concentre dans cette zone de propriétés ?

On a donc *via* le formalisme de la confidentialité différentielle une présentation de la donnée, et de la décision autour de son utilisation, sous un prisme entièrement probabiliste. C'est cohérent avec les modèles d'apprentissage automatique, qui traitent déjà de toute façon chaque point comme si c'était un tirage d'une loi de probabilité. En revanche, c'est une déplacement de la notion déterministe de l'individu vers une interprétation probabiliste qui est inhabituelle au niveau humain. Cela modifie-t-il la perception de l'individu pour en faire un individu probabiliste ? Est-ce que le budget pourra être interprété à terme comme une vision quantifiée du consentement ? Est-ce qu'un juge acceptera la notion d'un protection dont l'individu est malheureusement été un des cas du δ qui avait été prévu par le protocole et dont les données ont été dévoilées ?

10. Discriminer, mais en toute confidentialité

> This, to my thinking, actually represented the great nexus of the Intelligence Community and the tech industry : both are entrenched and unelected powers that pride themselves on maintaining absolute secrecy about their developments. Both believe that they have the solutions for everything, which they never hesitate to unilaterally impose. Above all, they both believe that these solutions are inherently apolitical, because they're based on data, whose prerogatives are regarded as preferable to the chaotic whims of the common citizen.
>
> **Edward Snowden,** [Sno2019]

La confidentialité propose une quantification au sein d'un schéma certes adaptable (8) mais qui reste centré autour de l'idée de participation à une base de données dans le but de faire émerger une vérité sous-jacente (5). Un des axiomes de la protection est donc cette invariance qu'aurait un modèle appris parfait par rapport aux données d'entrainements. On va même plus loin en supposant qu'on peut alors utiliser le modèle appris sans problème de *privacy*, car on n'utilise alors qu'une vérité établie, une vérité scientifique pourrait-on dire. Nul ne peut s'estimer être lésé par la simple application de la vérité à son cas.

On protège donc le flux qui correspond à la construction de modèle, c'est-à-dire que l'on se restreint à une notion très classique du jeu d'entrainement comme un flux séparé de la mise en pratique. Or, maintes applications reposent sur un second flux, bien plus massif, lors du déploiement du modèle. La *privacy* est alors souvent inexistante dans le sens où l'opération de personnalisation peut se faire en local, et donc ne sera pas renvoyée vers une personne tierce.

Par exemple dans le cadre de la prédiction de contenu, le jeu d'entrainement sera une certaine partie des utilisateurs, puis un modèle est établi, qui va calculer en fonction des paramètres d'un individu donné quel affichage est le plus efficace pour lui, selon la « vérité » établie avec le jeu d'entrainement. Cette étape n'invoque généralement pas de *privacy*, cela est d'ailleurs formalisé par la notion de confidentialité différentielle, car on ne pourrait d'ailleurs pas vraiment parler de personnalisation si avec une grande probabilité, le contenu n'est pas calculé en fonction des vraies données de l'utilisateur.

On parle donc, dans un tel scénario, d'apprentissage respectueux de la *privacy* car le service calculant le modèle et toute personne ayant accès au modèle ne peut pas extraire l'information personnelle des individus concernées. Cependant, on discrimine bien l'individu suivant ses caractéristiques, simplement on ne le sait pas nous-mêmes, ce n'est que l'application d'un modèle général discriminant par essence. On n'observe

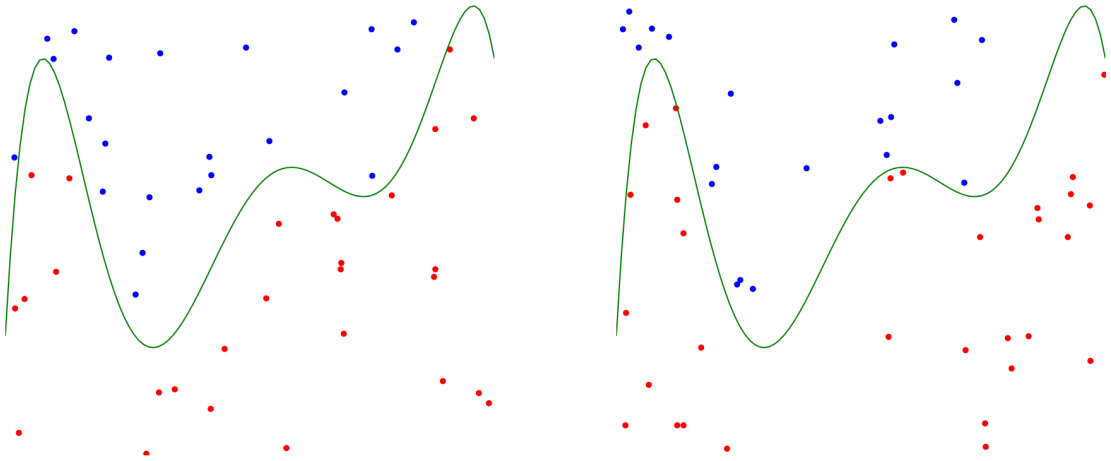


FIGURE 10.1. – Illustration de l'indépendance supposée des données : pour apprendre la courbe, peu importe qu'on se fonde sur les données de la figure de gauche ou de la figure de droite : ce ne sont que des exemples pour paramétrer correctement la courbe, mais aucun point n'est irremplaçable d'après la théorie de l'apprentissage automatique

pas la donnée, on ne la connait pas, on ne la vérifie pas explicitement, mais on l'utilise tout autant, éventuellement avec un bénéfice financier important.

On peut donc totalement respecter de bout en bout la confidentialité différentielle, mais utiliser des informations très personnelles et apprendre un modèle qui va agir différemment selon ces données personnelles. Ce faux sentiment de sécurité, qui semble garantir que tout est sous contrôle, est d'ailleurs également dénoncé par certains. La chercheuse Carmela Troncoso dénonce ainsi la persistance de l'existence même de points aberrants dans les données synthétiques générées en respectant la confidentialité différentielle, dans un séminaire intitulé « Privacy mirages in Machine learning ».

Elle y dénonce que le gain obtenu par la quantification de la confidentialité reste finalement faible car on apprend toujours avec suffisamment de données des caractéristiques qui sont par essence du domaine privé. On peut aller encore plus loin. Au lieu de prétendre pouvoir tracer une limite entre un modèle neutre car dévoilant uniquement un savoir scientifique – une simple découverte par une promenade dans les grandes dimensions, similaire à la découverte de l'Amérique dans le monde physique – et des enregistrements liés aux personnes et qui devraient rester dans leur zone d'intimité, certaines applications de l'apprentissage automatique sont un résumé de contingences spécifiques à la base de données actuelle, et la généralisation ne peut être qu'incomplète. Il n'y a pas d'universel à capturer dans certains modèles, la prémisse de l'apprentissage automatique qui suppose cette généralisation possible est invalidée. Il ne s'agit que de tendances, qui peuvent faire mieux que le hasard uniquement en se fon-

dant justement sur des discriminations illicites, reproduisant des *statu quo* et l'inertie de la société.

La confidentialité différentielle n'est pas invalidée par cette limite, car elle quantifie un aspect contre-intuitif de l'apprentissage automatique qui est aussi une de ses faiblesses : par défaut certains éléments risquent d'être mémorisés tels que dans le modèle, alors qu'il suffit pour apprendre d'avoir un signal suffisamment fort dans un nuée bruitée. On enlève donc les cas où l'on viole la *privacy* pour rien, car le bruit devrait être compatible avec l'apprentissage jusqu'à un certain point. Apprendre avec confidentialité différentielle, c'est donc finalement peut-être plus proche d'apprendre de façon robuste, que de garantir une quantification facile de la *privacy*.

En effet, une autre quantification importante en apprentissage automatique, surtout pour les applications critiques, est cette notion de robustesse. Si la notion est complexe à définir précisément, on la relie souvent à l'absence de discontinuité ou de surspécialisation du modèle, et plus généralement au maintien de la stabilité vis-à-vis d'entrées incorrectes ou de défaillance. Dans ce cadre, la confidentialité différentielle peut aussi être une alliée. En traçant le flot de l'information de l'enregistrement aux sorties, le calcul de la confidentialité différentielle demande une opération proche d'une forme d'audit, et peut être dans le meilleur des cas des garde fous supplémentaires.

On peut donc y voir également un lien avec le champ de la *fairness*, où une métrique proche de la confidentialité différentielle a été proposée par Cynthia Dwork également. Sous le nom d'*individual fairness*, il s'agit d'imposer cette fois-ci des contraintes sur la différence possible des sorties en fonction d'une métrique représentant la *fairness* en fonction des similitudes des individus [Dwo+2011]. Cependant, cette métrique présente une difficulté similaire à la définition d'adjacence dans le cas de la confidentialité différentielle, mais cette fois sur la définition de la métrique de similitude des individus.

Enfin, on peut voir une connexion avec la recherche pour la quantification de la transparence, puisque borner la contribution d'un enregistrement signifie également qu'on sait quantifier non seulement la persistance des signaux propres à l'individu, mais aussi son poids maximal dans l'apprentissage des paramètres du modèle.

De même que dans ces champs connexes, la confidentialité différentielle ne quantifie qu'un type de protection, qui ne saurait rendre la complexité de la chaine de l'information. Selon ce qu'on considère comme le mécanisme, la relation d'adjacence, la précision de l'analyse du mécanisme pour définir finement le budget de *privacy*, l'interprétation varie fortement. Ainsi Apple avait-il été attaqué dans sa première implémentation de confidentialité différentielle. En particulier, ils utilisaient l'hypothèse selon laquelle on pouvait remettre à zéro l'ensemble des budgets de *privacy* des utilisateurs chaque jour [Tan+2017].

On conçoit aisément que si la beauté technique d'avoir déployé une analyse entièrement confidentiellement privée était au rendez-vous, la protection de l'utilisateur, elle,

restait encore à améliorer. L'avancée technique était donc déjà là, même si elle continue de se développer. En revanche, l'analyse du flux global et l'interprétation pourtant assez évidente – nos informations personnelles évoluent peu d'un jour à l'autre – avaient été négligées. Si ce problème spécifique a été réglé, d'autres restent encore en jachère.

Comment la personnalisation du service peut être un problème de *privacy per se*, puisque même si Apple ne récupère jamais les données personnelles il peut néanmoins les exploiter directement par le calcul local semble encore rester un impensé. Avec un risque alors tout particulier, puisqu'on peut récréer, sans même le souhaiter ou le coder, les discriminations réelles, sans même accéder aux données qui nous font créer ces discriminations. Ces effets de bulle, où chaque utilisateur n'a accès qu'à un monde restreint par ses actions antérieures et un certain déterminisme social, est pourtant bien documenté [BH2015 ; ONe2016].

Pour fixer les idées, on peut considérer le cas de l'apprentissage de données spatiales, par exemple pour faire de la personnalisation de suggestion à partir de la position de l'utilisateur. On peut alors recréer des ghettos virtuels de façon entièrement confidentiellement privée. Chaque utilisateur contribue au modèle avec une position géographique bruitée. La carte des préférences générée reconstruit donc, grâce à l'annulation des bruits entre eux, quels sont les contenus préférés de chaque quartier. Un individu interagissant avec le modèle appris se verra donc proposer du contenu adapté à sa vraie localisation, puisque c'est la vérité apprise. Cependant, si la différentiation spatiale peut sans doute être acceptable pour certains contenus (par exemple quelqu'un cherchant quelles plantes faire pousser sur son balcon sera sans doute plus content avec des plantes adaptées aux conditions météorologiques de sa région), ce choix est beaucoup plus discutable dans d'autres contextes. Est-il éthique de personnaliser les offrent d'emploi en utilisant les données GPS ? Même si le modèle ne permet pas de remonter à la position de l'individu, la perte de chance est maintenu, et ce genre de pratique renforce les boucles de rétro-action qui poussent à la radicalisation des positions politiques, à la reproduction des inégalités et autres effets des *filter bubbles*[ONe2016 ; BH2015].

Ceci est autant de questions qui sont soulevées par les flux de données utilisés par l'apprentissage automatique, et qui ne peuvent pas être contrôlés par la confidentialité différentielle. Cette quantification est donc un outil parmi toute une panoplie pour contrôler les flux d'information en s'attaquant à un type de problèmes rencontrés avec l'apprentissage statistique.

Conclusion

We will treat code-based environmental disasters –like the loss of privacy, like the censorship of censorware filters, like the disappearance of an intellectual commons– as if they were produced by gods, not by Mans. We will watch as important aspects of privacy and free speech are erased by the emerging architecure of the panopticon, and we will speak, like modern Jeffersons, about the nature making it so–forgetting that here, we are nature.

Code 2.0, Lessig

Il y a vingt ans, les tours du World Trade Center s'effondraient. Le *Patrioct Act* était adopté en réaction, dans l'espoir que la large collecte de données serait gage de sécurité pour les Américains, en permettant la détection des criminels avant leur passage à l'acte. Sacrifier la *privacy* pour la sécurité de tous était la position morale dominante. Les mathématiques et la loi de Bayes, ont pourtant continué de s'appliquer, et assurent toujours que détecter des événements très peu probables dans une large population mène à un nombre de faux positifs. Les résultats restent donc inexploitables pour cet objectif de détection de signal très faible, et seul un traçage ciblé peut fonctionner. Si les rapports ont montré l'inefficacité des dispositifs [PB2014], ceci n'a pas arrêté le procédé, comme si la mise en place des infrastructures avait rendu leur révocation impossible.

Aujourd'hui, un nouveau péril sanitaire fait revivre ce mythe de l'information d'utilité publique et de la technologie combinant le meilleur des deux mondes [MIL+2021]. Collecte de données de santé peu sécurisées pour hâter le rythme de la recherche, traçage permanent des individus, contrôle d'identité systématique avec des outils d'une vétusté informatique patente, rupture du secret médical sont difficilement dénoncés sans passer pour un ennemi de la santé publique. Déjà l'Australie récupère les données collectées pour les enquêtes policières [Ram2021], déjà les prix des services numériques sont suspectés d'irrégularité [Ant2021], déjà la permanence ou l'élargissement de ce traçage est suspecté, sans que l'équilibre entre la dérive sécuritaire et les respect de libertés individuelles ne soient questionné.

La définition de la confidentialité différentielle est intrinsèquement reliée à l'émergence de la collecte de données massives, et à l'acceptation de la donnée comme vecteur principal de connaissance dans les applications numériques. Alors que l'anonymat promettait intuitivement le meilleur des deux mondes – récupérer les données pour étendre le savoir de tous, tout en ne nuisant à personne – la persistance de structures en grande dimension et les capacités de calcul ont anéanti ce rêve.

Il y a donc un prix à payer pour l'exploitation des données, elles subsistent dans les sorties aux dépends de ceux qui les fournissent. Chaque implémentation propose un compromis entre le niveau de fidélité des données et le respect de la *privacy*. La confidentialité différentielle quantifie en un unique nombre réel ce compromis, comme étant la magnitude du flot de données qui peut être extrait des sorties de l'algorithme à propos d'une des entrées de celui-ci.

En se fondant sur une protection de la participation, cette métrique s'inscrit donc dans la lignée d'une information pour le Bien commun, dont la *privacy* est un tiraillement, une concession faite dans une perspective pragmatique pour rassurer celui qui cède ses données. Parce que le modèle est censé généraliser, il se doit d'exister un paramétrage qui mémorise peu les données et offre une bonne confidentialité différentielle tout en maintenant une précision suffisante. Les difficultés de mise en œuvre pratiques prouvent cependant l'inexactitude de cette hypothèse.

Parce qu'on se repose sur une notion d'adjacence qui simplifie en un enregistrement la multiplicité des causes – messages, personne physique identifiable, destinateur –, et simplifie les déformations des sorties en un unique nombre, la confidentialité différentielle ne permet de théoriser qu'une partie du flot de données de l'apprentissage. Le reste, non garanti mathématiquement, résulte pourtant aussi de choix cruciaux pour l'utilisateur.

Ainsi, imaginer un consentement éclairé quantifiable par un budget de *privacy* semble inatteignable à moyen terme. Cela n'enlève cependant rien à l'étude d'algorithmes confidentiellement privés, car cette métrique fournit une quantification de notre capacité à extraire le signal de la masse de données en étant robuste aux incertitudes et aux données inexactes. Elle permet aussi de penser l'utilisation des bases de données dans le cadre de l'apprentissage automatique, sous réserve de ne pas confondre garantie mathématique et garantie effective de protection de l'utilisateur. En quantifiant la *privacy*, on rend l'optimisation selon ce paramètre possible. On l'inscrit donc comme une des dimensions qui importent, et on ouvre la voie au débat sur l'équilibre à choisir entre la précision souhaitée et la *privacy*.

Bibliographie

[WB1890] Samuel D. Warren et Louis D. Brandeis. « The right to privacy ». In : *Harvard Law Review* 4.5 (1890), p. 193-220. URL : http://faculty.uml.edu/sgallagher/Brandeisprivacy.htm.

[Nat1948] Assemblée générale dees Nations Unies. *Déclaration Universelle des Droits de l'Homme*. 1948.

[War1965] Stanley L. Warner. « Randomized Response : A Survey Technique for Eliminating Evasive Answer Bias ». In : *Journal of the American Statistical Association* 60.309 (mars 1965), p. 63-69. DOI : 10.1080/01621459.1965.10480775. URL : https://doi.org/10.1080/01621459.1965.10480775.

[Wes1967] Alan F. Westin. *Privacy and Freedom*. New York : Atheneum, 1967.

[Fou1975] Michel Foucault. *Surveiller et punir : naissance de la prison*. Gallimard, 1975.

[Mac2002] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. USA : Cambridge University Press, 2002. ISBN : 0521642981. URL : https://www.cambridge.org/fr/academic/subjects/computer-science/pattern-recognition-and-machine-learning/information-theory-inference-and-learning-algorithms.

[DN2003] Irit Dinur et Kobbi Nissim. « Revealing Information while Preserving Privacy ». In : jan. 2003, p. 202-210. DOI : 10.1145/773153.773173.

[DN2004] Cynthia Dwork et Kobbi Nissim. « Privacy-Preserving Datamining on Vertically Partitioned Databases ». In : *Advances in Cryptology – CRYPTO 2004*. Sous la dir. de Matt Franklin. Berlin, Heidelberg : Springer Berlin Heidelberg, 2004, p. 528-544. ISBN : 978-3-540-28628-8.

[Boe2005] Pieter Boeder. « Habermas heritage : The future of the public sphere in the network society ». In : (2005). URL : https://firstmonday.org/ojs/index.php/fm/article/download/1280/1200.

[DK2005] Alain Desrosières et Sandrine Kott. « Quantifier ». In : *Genèses* 58.1 (2005), p. 2. DOI : 10.3917/gen.058.0002. URL : https://doi.org/10.3917/gen.058.0002.

[CT2006] Thomas M. Cover et Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA : Wiley-Interscience, 2006. ISBN : 0471241954. URL : https://www.amazon.fr/Elements-Information-Theory-Thomas-Cover/dp/0471241954.

[Dwo+2006] Cynthia Dwork et al. « Calibrating Noise to Sensitivity in Private Data Analysis ». In : *Theory of Cryptography*. Sous la dir. de Shai Halevi et Tal Rabin. Berlin, Heidelberg : Springer Berlin Heidelberg, 2006, p. 265-284. ISBN : 978-3-540-32732-5.

[NS2006] Arvind Narayanan et Vitaly Shmatikov. « How To Break Anonymity of the Netflix Prize Dataset ». In : *CoRR* abs/cs/0610105 (2006). arXiv : cs/0610105. URL : http://arxiv.org/abs/cs/0610105.

[SUP2008] D.J. Solove, Harvard University et Harvard University Press. *Understanding Privacy*. Understanding privacy vol. 10. Harvard University Press, 2008. ISBN : 9780674027725. URL : https://www.amazon.fr/Understanding-Privacy-Daniel-J-Solove/dp/0674035070.

[Les2009] Lawrence Lessig. *Code 2.0*. 2nd. Scotts Valley, CA : CreateSpace, 2009. ISBN : 1441437649. URL : https://www.decitre.fr/livres/code-and-other-laws-of-cyberspace-version-2-0-9780465039142.html.

[Ohm2009] Paul Ohm. « Broken Promises of Privacy : Responding to the Surprising Failure of Anonymization ». In : (2009). URL : https://ssrn.com/abstract=1450006.

[Dwo+2010] Cynthia Dwork et al. « Pan-Private Streaming Algorithms ». In : *Innovations in Computer Science - ICS 2010, Tsinghua University, Beijing, China, January 5-7, 2010. Proceedings*. Sous la dir. d'Andrew ChiChih Yao. Tsinghua University Press, 2010, p. 66-80. URL : http://conference.iiis.tsinghua.edu.cn/ICS2010/content/papers/6.html.

[Nis2010] H. Nissenbaum. *Privacy in Context : Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2010. ISBN : 9780804752374. URL : https://books.google.fr/books?id=cxb15VjOzCYC.

[De2011] Anindya De. *Lower bounds in differential privacy*. 2011. arXiv : 1107.2183 [cs.CR].

[Dwo+2011] Cynthia Dwork et al. *Fairness Through Awareness*. 2011. arXiv : 1104.3913 [cs.CC].

[LC2011] Jaewoo Lee et Chris Clifton. « How Much Is Enough ? Choosing for Differential Privacy ». In : oct. 2011, p. 325-340. ISBN : 978-3-642-24860-3. DOI : 10.1007/978-3-642-24861-0_22.

[Mon+2013] Yves-Alexandre de Montjoye et al. « Unique in the Crowd : The privacy bounds of human mobility ». In : *Scientific Reports* 3.1 (mars 2013), p. 1376. ISSN : 2045-2322. DOI : 10.1038/srep01376. URL : https://doi.org/10.1038/srep01376.

[Mor2013] E. Morozov. *To Save Everything, Click Here : The Folly of Technological Solutionism*. PublicAffairs, 2013. ISBN : 9781610391399. URL : https://www.amazon.fr/Save-Everything-Click-Here-Technological/dp/1610393708.

[DR2014] C. Dwork et A. Roth. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science Series. Now Publishers, 2014. ISBN : 9781601988188. URL : https://books.google.fr/books?id=J3PUoQEACAAJ.

[PB2014] PRIVACY et CIVIL LIBERTIES OVERSIGHT BOARD. *Report on the Telephone Records Program Conducted under Section 215 of the USA PATRIOT Act and on the Operations of the Foreign Intelligence Surveillance Court*. Jan. 2014. URL : https://documents.pclob.gov/prod/Documents/OversightReport/ec542143-1079-424a-84b3-acc354698560/215-Report_on_the_Telephone_Records_Program.pdf.

[BH2015] Engin Bozdag et Jeroen van den Hoven. « Breaking the filter bubble : democracy and design ». In : *Ethics and Information Technology* 17.4 (déc. 2015), p. 249-265. DOI : 10.1007/s10676-015-9380-y. URL : https://doi.org/10.1007/s10676-015-9380-y.

[FJR2015] Matt Fredrikson, Somesh Jha et Thomas Ristenpart. « Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures ». In : *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. CCS '15. Denver, Colorado, USA : Association for Computing Machinery, 2015, p. 1322-1333. ISBN : 9781450338325. DOI : 10.1145/2810103.2813677. URL : https://doi.org/10.1145/2810103.2813677.

[KOV2015] Peter Kairouz, Sewoong Oh et Pramod Viswanath. *The Composition Theorem for Differential Privacy*. 2015. arXiv : 1311.0776 [cs.DS].

[TZ2015] Naftali Tishby et Noga Zaslavsky. *Deep Learning and the Information Bottleneck Principle*. 2015. arXiv : 1503.02406 [cs.LG].

[Aba+2016] Martin Abadi et al. « Deep Learning with Differential Privacy ». In : *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (oct. 2016). DOI : 10.1145/2976749.2978318. URL : http://dx.doi.org/10.1145/2976749.2978318.

[Con2016] Parlement européen et Conseil. *Réglement relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données*. Avr. 2016. URL : http://data.europa.eu/eli/reg/2016/679/oj.

[GBC2016] Ian Goodfellow, Yoshua Bengio et Aaron Courville. *Deep Learning*. MIT Press, 2016. URL : https://www.amazon.com/Deep-Learning-Adaptive-Computation-Machine/dp/0262035618/ref=sr_1_1?ie=UTF8&qid=1472485235&sr=8-1&keywords=deep+learning+book.

[HM2016] Eszter Hargittai et Alice Marwick. « "What Can I Really Do?" : Explaining the Privacy Paradox with Online Apathy ». Anglais. In : *International Journal of Communication* 10 (jan. 2016), p. 3737-3757. ISSN : 1932-8036. URL : https://doi.org/10.5167/uzh-148157.

[ONe2016] C. O'Neil. *Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016. ISBN : 9780553418828. URL : https://books.google.fr/books?id=NgEwCwAAQBAJ.

[BGN2017] S. Benthall, S. Gürses et H. Nissenbaum. *Contextual Integrity Through the Lens of Computer Science*. Foundations and Trendső in Privacy and Security Series. Now Publishers, 2017. ISBN : 9781680833843. URL : https://books.google.fr/books?id=CaNzswEACAAJ.

[MN2017] Kirsten Martin et Helen Nissenbaum. « Measuring Privacy : An Empirical Test Using Context to Expose Confounding Variables ». In : *Columbia Science & Technology Law Review* 18 (jan. 2017), p. 176-218.

[Mir2017] Ilya Mironov. « Renyi Differential Privacy ». In : *CoRR* abs/1702.07476 (2017). arXiv : 1702.07476. URL : http://arxiv.org/abs/1702.07476.

[Pap+2017] Nicolas Papernot et al. « Practical Black-Box Attacks against Machine Learning ». In : *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (2017).

[Sho+2017] R. Shokri et al. « Membership Inference Attacks Against Machine Learning Models ». In : *2017 IEEE Symposium on Security and Privacy (SP)* (2017), p. 3-18.

[Tan+2017] Jun Tang et al. « Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12 ». In : *CoRR* abs/1709.02753 (2017). arXiv : 1709.02753. URL : http://arxiv.org/abs/1709.02753.

[Tor2017] Vicenç Torra. *Data Privacy : Foundations, New Developments and the Big Data Challenge*. Springer International Publishing, 2017. DOI : 10.1007/978-3-319-57358-8. URL : https://www.springer.com/gp/book/9783319573564.

[ACW2018] Anish Athalye, Nicholas Carlini et David A. Wagner. « Obfuscated Gradients Give a False Sense of Security : Circumventing Defenses to Adversarial Examples ». In : *CoRR* abs/1802.00420 (2018). arXiv : 1802.00420. URL : http://arxiv.org/abs/1802.00420.

[BM2018] A. Basdevant et J. Mignard. *L'Empire des données. Essai sur la société, les algorithmes et la loi*. Non fiction. Éditions Don Quichotte, 2018. ISBN : 9782359496369. URL : https://books.google.fr/books?id=Y95PDwAAQBAJ.

[Fel+2018] Vitaly Feldman et al. « Privacy Amplification by Iteration ». In : *CoRR* abs/1808.06651 (2018). arXiv : 1808.06651. URL : http://arxiv.org/abs/1808.06651.

[RB2018] Frances Rogan et Shelley Budgeon. « The Personal is Political : Assessing Feminist Fundamentals in the Digital Age ». In : *Social Sciences* 7 (août 2018), p. 132. DOI : 10.3390/socsci7080132.

[Dwo2019] Cynthia Dwork. « Differential Privacy and the US Census ». In : *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. PODS '19. Amsterdam, Netherlands : Association for Computing Machinery, 2019, p. 1. ISBN : 9781450362276. DOI : 10.1145/3294052.3322188. URL : https://doi.org/10.1145/3294052.3322188.

[JE2019] Bargav Jayaraman et David Evans. « Evaluating Differentially Private Machine Learning in Practice ». In : *28th USENIX Security Symposium (USENIX Security 19)*. Santa Clara, CA : USENIX Association, 2019, p. 1895-1912. ISBN : 978-1-939133-06-9. URL : https://www.usenix.org/conference/usenixsecurity19/presentation/jayaraman.

[Sno2019] E. Snowden. *Permanent Record*. Henry Holt et Company, 2019. ISBN : 9781250237248. URL : https://books.google.fr/books?id=0XCcDwAAQBAJ.

[ZLH2019] Ligeng Zhu, Zhijian Liu et Song Han. « Deep Leakage from Gradients ». In : *NeurIPS*. 2019.

[Ash2020] Thomas Ashley. *No Place to Hide : Privacy Implications of Geolocation Tracking and Geofencing*. Jan. 2020. URL : https://www.americanbar.org/groups/science_technology/publications/scitech_lawyer/2020/winter/no-place-hide-privacy-implications-geolocation-tracking-and-geofencing.

[Car+2020] Nicholas Carlini et al. « Extracting Training Data from Large Language Models ». In : *CoRR* abs/2012.07805 (2020). arXiv : 2012.07805. URL : https://arxiv.org/abs/2012.07805.

[2020] *FullStory : un outil danalyse du parcours des utilisateurs*. 2020. URL : https://junto.fr/blog/fullstory/.

[JB2020] David Darais Joseph Near et Kaitlin Boeckl. *Differential Privacy for Privacy-Preserving Data Analysis : An Introduction to our Blog Series*. 2020. URL : https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-privacy-preserving-data-analysis-introduction-our.

[Sec2020] IMB Security. *Cost of a data breach Report*. 2020.

[Ant2021] Anticor. *Application StopCovid : Anticor saisit la Cour de Justice de la République*. Mars 2021. URL : https://www.anticor.org/2021/03/22/application-stopcovid-anticor-saisit-la-cour-de-justice-de-la-republique/.

[AWS2021] AWS. *AWS S3 Pricing*. 2021. URL : https://aws.amazon.com/fr/s3/pricing/.

[Dam2021] Martin Untersinger Damien Leloup. *Projet Pegasus : révélations sur un système mondial despionnage de téléphones*. Juill. 2021. URL : https://www.lemonde.fr/projet-pegasus/article/2021/07/18/projet-pegasus-revelations-sur-un-systeme-mondial-despionnage-de-telephones_6088652_6088648.html.

[DoS2021] DoSomething. *11-facts-about-cyber-bullying*. 2021. URL : https://www.dosomething.org/us/facts/11-facts-about-cyber-bullying.

[Foe2021] Postma Foeke. *US Soldiers Expose Nuclear Weapons Secrets Via Flashcard Apps*. Juin 2021. URL : https://www.bellingcat.com/news/2021/05/28/us-soldiers-expose-nuclear-weapons-secrets-via-flashcard-apps/.

[MIL+2021] Stefania MILAN et al. « Promises Made to Be Broken : Performance and Performativity in Digital Vaccine and Immunity Certification ». In : *European Journal of Risk Regulation* 12.2 (2021), p. 382-392. DOI : 10.1017/err.2021.26.

[Per2021] Kelly Percival. *Court Rejects Alabama Challenge to Census Plans for Redistricting and Privacy*. Juin 2021. URL : https://www.brennancenter.org/our-work/analysis-opinion/court-rejects-alabama-challenge-census-plans-redistricting-and-privacy.

[Ram2021] Michael Ramsey. *Privacy infringement fears after police access data from SafeWA contact tracing app*. Juin 2021. URL : https://7news.com.au/news/western-australia-police/wa-police-accessed-contact-tracing-data-c-3118713.

[Zie2021] Albert Ziegler. *Research recitation*. 2021. URL : https://docs.github.com/en/github/copilot/research-recitation.

[Mer] Samir Merabet. *Vers un droit de l'intelligence artificielle*. Dalloz. URL : https://www.boutique-dalloz.fr/vers-un-droit-de-l-intelligence-artificielle-volume-197-p.html.

[BN2015 - 2015] Finn Brunton et Helen Nissenbaum. *Obfuscation : a user's guide for privacy and protest*. eng. Cambridge, Massachusetts : The MIT Press, 2015 - 2015. ISBN : 9780262029735.