Déployer une procédure d'anonymisation :

se poser les bonnes questions

LIVRE BLANC

| septembre 2020 |

DATA STORM

Table des matières

1 Introduction

Depuis la promulgation du Règlement Général sur la Protection des données (RGPD[1]) en mai 2018, le cadre concernant l'utilisation des données à caractère personnel a été clarifié et les droits de leurs propriétaires renforcés.

Avant de stocker, diffuser ou publier des données à caractère personnel, il est nécessaire de s'assurer que le traitement respecte les principes du RGPD (limitation des finalités et de la conservation, licéité, loyauté et transparence, minimisation des données, exactitude, intégrité et confidentialité). Le règlement prévoit, de surcroît, dans son article 9, des règles spécifiques aux données "sensibles" (e.g. les données médicales, ou traitant des origines ethniques ou raciales, de la religion, des opinions politiques).

La notion de consentement occupe alors une place centrale dans le règlement. Un accord libre, spécifique, éclairé et univoque permet, dans de nombreux cas, de légitimer le traitement des données à caractère personnel. Par la suite, les données identifiantes doivent être conservées dans une durée légitimée par le traitement et énoncées clairement aux personnes concernées. Les données archivées doivent également bénéficier d'une démarche garantissant la confidentialité. Globalement, le responsable du traitement doit s'organiser et mettre en place des actions pour garantir une adéquation au RGPD et un niveau de sécurité adapté au risque.

Si le RGPD a donc contribué à une plus grande protection des individus, le cadre dans lequel un acteur peut exploiter des données à caractère personnel a cependant été fortement contraint et amène les entreprises à trouver de nouveaux moyens pour réaliser leurs traitements.

Des approches fondées sur la conception même de l'architecture des systèmes collectant, stockant et exploitant les données à caractère personnel (« *privacy-by-design* ») permettent de répondre à ces obligations et notamment de pouvoir assurer le calcul d'indicateurs, l'entraînement et l'application de modèles sur les équipements des personnes concernées ou dans une architecture fédérée[2], ce qui permet une Data Science compatible avec la non-collecte des données.

Néanmoins, une telle approche est souvent excessivement complexe par rapport à de nombreux cas d'usage. De plus, elle est incompatible avec certains usages de la donnée. Par exemple, il est légitime de vouloir collecter des données à caractère personnel, sous réserve de l'obtention du consentement, pour permettre d'améliorer les traitements. Tout comme il est important de pouvoir archiver pour des besoins ultérieurs indéfinis. Que ce soit pour obtenir le consentement dans un climat de confiance, comme pour archiver des données à l'origine personnelle, ou des données impersonnelles sensibles, une anonymisation des données est nécessaire.

[1] https://www.cnil.fr/fr/reglement-europeen-protection-donnees
[2] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

En pratique, l'anonymisation doit prévenir plusieurs risques :

- **Le risque d'isolement** : il ne doit pas être possible d'isoler une personne à partir de ses données, car dans ce cas toutes ses informations sont connues avec certitude.

- **Le risque d'inférence** : il ne doit pas être possible d'estimer avec une forte probabilité des informations sur une personne à partir de ses données, de celles des autres individus ou d'informations exogènes (d'une autre base de données, par exemple).

- **Le risque de corrélation** : si plusieurs bases de données sont liées aux mêmes individus, il ne doit pas être possible de faire le lien entre leurs observations. Par exemple, si des identifiants directs (nom, prénom…) ont été pseudonymisés dans plusieurs bases, ils ne doivent pas correspondre.

Une étape d'anonymisation des données peut avoir du sens à plusieurs stades du flux de données d'un système d'information.

Avant le stockage :

- En amont du projet, durant la collecte de la donnée. *Exemple d'un opérateur téléphonique qui met à jour des indicateurs globaux sur l'ensemble de ses clients (évolution du nombre d'utilisateurs journaliers, des habitudes d'utilisation, des remontées d'informations liées à la domotique) : avant d'être stockées dans les bases de données de l'entreprise, les données sont agrégées, généralisées et/ou bruitées (tranche d'âge, zone géographique, date d'achat bruitée …).*

Après le stockage :

- En cas d'extraction ponctuelle, par une personne accréditée, d'une série d'observations sur quelques variables. *Exemple* : *une entreprise possède des données RH et a recours à un prestataire externe pour produire un bilan social complet. Afin de minimiser les risques encourus en faisant sortir les données du cercle RH, cette entreprise décide de mettre en œuvre un processus d'extraction de données avec dé-identification.*

- Via une API, ou un tableau de bord, ouverts à un plus grand nombre de personnes, permettant d'effectuer des interrogations de la base de données. *Exemple : la même entreprise, par souci de transparence, souhaite ouvrir un portail à l'attention de ses collaborateurs. Cet outil leur permettant de réaliser des comparaisons à profil similaire, sans qu'il soit possible de cibler une personne et de retrouver sa rémunération ou toute autre information à caractère personnel.*

L'anonymisation consiste en la recherche d'un équilibre entre des besoins en analyses statistiques nécessitant un bon niveau de précision et des risques en matière de protection des données.

Prenons l'exemple de données issues du recensement : un niveau de détail élevé permettrait de réaliser des analyses très approfondies, mais rendrait possible l'identification de certains ménages recensés.

Une stratégie d'anonymisation minimale, qui consisterait en la simple pseudonymisation des variables directement identifiantes, n'est pas une méthode viable car les autres variables présentent souvent un risque de réidentification élevé : à titre d'exemple, 87% des américains peuvent être réidentifiés rien qu'avec le code postal, le sexe et la date de naissance.[3]

Si le RGPD motive, pour le traitement des données à caractère personnel, la démarche d'anonymisation, soulignons que celle-ci s'applique tout aussi bien à des sujets sans données à caractère personnel, comme par exemple l'exploitation de données industrielles ou commerciales sensibles. Dans la suite de ce document, nous nous focalisons plus particulièrement sur les données à caractère personnel, mais le lecteur pourra transposer la démarche pour les autres cas.

> Dans ce Livre Blanc, nous établissons les bases permettant de déployer une procédure générique et rigoureuse d'anonymisation.

[3] https://dataprivacylab.org/projects/identifiability/paper1.pdf

2 La définition du besoin

2.1 Identifier les personnes concernées par les données

Il s'agit de déterminer quelles sont les personnes (ou unités statistiques) physiques auxquelles la diffusion des données non-anonymisées pourrait porter préjudice.

S'il y a plusieurs types d'unités statistiques (individu, entreprise), il faut les hiérarchiser et associer à chacun les variables qui le concernent. Ensuite, leurs variables peuvent être anonymisées conjointement ou séparément.

Dans le cas d'une anonymisation conjointe, le nombre de variables est pénalisant, car plus il y a d'informations, plus les modifications doivent être importantes pour dé-identifier les observations. Dans l'autre cas, il peut être nécessaire d'appliquer plusieurs méthodes sur les variables communes afin de contrôler les risques de chaque unité statistique.

Age	Code postal	Type poste	Entreprise	Salaire
22	75000	Type 1	Etp A	30 495
35	78000	Type 2	Etp A	42 405
40	77000	Type 2	Etp A	54 029
34	75000	Type 1	Etp B	24 059
51	93000	Type 3	Etp B	39 404
35	75000	Type 2	Etp B	33 048
62	78000	Type 2	Etp B	40 194

Dans la table ci-dessus, les unités statistiques sont les individus et les entreprises. En effet, la connaissance des salaires moyens par poste ou par âge est une information sensible pour une entreprise. Avant de publier cette base (dont la colonne entreprise sera certainement supprimée), il faut donc s'assurer qu'il n'est possible de récupérer de l'information ni sur les individus, ni sur les entreprises. Les variables sont toutes liées aux individus. En revanche, les variables âge et code postal ne sont pas associées aux entreprises.

2.2 Se renseigner sur la base de données

Les informations sur les modalités de création de la base de données peuvent être utiles lors de l'anonymisation. Il est par exemple important de savoir si la base de données cible une population entière ou seulement un échantillon. Le cas échéant, la démarche d'échantillonnage peut engendrer une sur-représentation ou une sous-représentation de certaines catégories d'individus. Ces informations permettent de dégager un *a priori* sur le niveau de risque associé aux individus ou groupes d'individus.

Il est également important de lister les sources de données annexes (publiques ou non) pouvant être croisées avec la base anonymisée et faciliter la réidentification d'individus. A cet effet, il existe des outils dédiés à la recherche des données publiques[4].

[4] https://www.data.gouv.fr, https://any-api.com/?tag=open%20data

2.3 Identifier le type d'analyse

Puisque les résultats doivent être le moins impactés possibles par les modifications des données, la procédure d'anonymisation dépend largement du type d'analyse envisagé et de la ou des problématiques sous-jacentes. Certaines méthodes d'anonymisation ont d'ailleurs été conçues précisément pour avoir un impact minimal sur certains types d'analyses.

Il est alors important d'avoir une idée précise des attentes, détailler la problématique en une liste de questions et une liste d'analyses associée. A cette étape, il est possible de distinguer les valeurs, les corrélations, les distances que l'on voudra préserver, de celles que l'on pourra perturber pour empêcher une réidentification.

> Selon le cas d'usage et les problématiques associées aux données, plusieurs indicateurs (dispersion, corrélation), modèles ou données graphiques peuvent être utilisés pour vérifier que l'anonymisation ne dégrade pas trop les données.

2.4 Définir le mode de publication / diffusion

Le nombre de personnes qui aura accès – même partiellement – à la base anonymisée est un facteur de risque important. Une diffusion publique sur internet (i.e. *Open Data*) est beaucoup plus risquée qu'une diffusion à faible portée sur un serveur sécurisé. Sans pour autant réfléchir à tous les détails de l'infrastructure, il est nécessaire d'avoir dès le départ une connaissance du niveau de diffusion. Par la suite, il faudra s'assurer que ce niveau de diffusion est effectivement respecté.

2.5 Estimer le besoin en matière de sécurité

Outre le niveau de diffusion, le besoin en matière de sécurité dépend des possibilités de croisement de la base de données avec des informations annexes.

Le niveau de sécurité peut être affiné en raisonnant du point de vue d'un « attaquant ». Comme en cryptologie ou en cybersécurité, il s'agit de définir un « modèle de menace » [5] en dressant des hypothèses sur un attaquant potentiel et en listant les actions possibles d'un individu souhaitant réidentifier les éléments de la base de données. Il s'agit aussi de définir les informations que l'on souhaite garder secrètes et ce qui peut être rendu public (ou du moins, ouvert à un plus large public).

Il ne faut pas avoir peur de trouver des failles : au contraire, il faut en identifier un maximum pour pouvoir les traiter correctement et contrôler leur probabilité d'occurrence. L'essentiel est d'évaluer si le contexte (nature des données, niveau de diffusion) justifie la mise en

[5] https://slides.nothing2hide.org/atelier-modele-de-menace.html
https://ssd.eff.org/fr/module/votre-plan-de-securite

œuvre d'une anonymisation plus forte, afin de réduire les risques associés aux failles identifiées.

> Pour chacun des risques identifiés, il s'agit de définir des critères à satisfaire pour valider l'anonymisation. Si le risque est trop élevé, l'information doit être supprimée des données.

3 La sélection et le pré-traitement des variables

Dans une base de données, toutes les variables n'ont pas la même importance, tant en matière d'information sur les individus qu'en termes d'intérêt pour le cas d'usage. Dans le contexte de l'anonymisation, on distingue plusieurs catégories de variables, qui sont traitées différemment.

Identifiant direct	Quasi-identifiants			Var inutile	Var sensible
Nom	**Département**	**Age**	**Sexe**	**Couleur pref**	**Salaire**
Albert A.	78200	21	F	Rouge	2100
Bastien B.	78400	28	F	Vert	2800
Claire C.	78600	29	F	Bleu	4500
Damien D.	45100	37	F	Vert	3100
Eloïse E.	45300	31	F	Vert	3300
Fabrice F.	45100	36	F	Rouge	3200
Gabrielle G.	33000	53	M	Bleu	3400

La table ci-dessus illustre les différentes catégories de variables présentes dans les bases de données, auxquelles il est fait référence dans les paragraphes suivants.

3.1 Déterminer et traiter les identifiants directs

Les identifiants directs sont les variables qui, prises isolément, permettent d'identifier des individus (nom, numéro de sécurité sociale, …).

Si une personne peut être concernée par plusieurs observations dans la base de données, il est possible de les conserver pour faire le lien entre leurs différentes observations, mais il est impératif de les pseudonymiser[6]. Sinon, elles doivent être supprimées.

Il existe plusieurs techniques pour pseudonymiser un identifiant. Ces techniques sont présentées en page 22 de l'avis du G29 (groupe des CNILs européennes) sur les techniques d'anonymisation[7].

[6] Voir glossaire en fin de document

[7] https://www.cnil.fr/fr/le-g29-publie-un-avis-sur-les-techniques-danonymisation

3.2 Déterminer les quasi-identifiants

Les quasi-identifiants sont les variables qui, croisées avec d'autres données (de la base ou exogènes), permettent d'identifier des individus (code postal, âge, date de naissance, …). Ce sont ces variables qui vont être anonymisées. Cependant, plus leur nombre est élevé, plus la détérioration des données est importante pour obtenir un même niveau d'anonymisation.

Plusieurs opérations permettent de faciliter l'anonymisation :

- Agréger les quasi-identifiants qui se rapportent à une même notion en une nouvelle variable ;
- Agréger les modalités semblables des quasi-identifiants qui contiennent un grand nombre de modalités ;
- Simplifier les valeurs des quasi-identifiants (par exemple en prenant un arrondi) ;
- Catégoriser/généraliser les variables numériques. Pour gérer les valeurs aberrantes, créer une catégorie « $\geqslant$ » ou « $\leqslant$ ».

3.3 Déterminer les variables sensibles

Les variables sensibles sont des variables différentes des quasi-identifiants, dont la connaissance est importante pour le traitement. Dans certains cas, elles peuvent faciliter la réidentification.

Par exemple, dans une base de données médicales, les affections des patients sont souvent des variables sensibles. Sachant qu'une personne est présente dans la base et souffre d'une maladie rare, il est possible d'inférer avec une forte probabilité ses autres informations présentes dans la base de données, telles que le traitement administré ou le lieu d'hospitalisation. Elles peuvent donc être anonymisées (généralisation, recodage des modalités, ajout de bruit), de préférence individuellement pour limiter la perte d'information.

3.4 Supprimer les autres variables

« La teneur de nos communications en dit rarement autant sur nous que d'autres éléments qui restent tacites. […] Les métadonnées peuvent apprendre à celui qui vous surveille pratiquement tout ce qu'il a envie ou besoin de savoir sur vous, à l'exception de ce qui se passe dans votre tête. » *Edward Snowden, Mémoires Vives, P.200-202.*

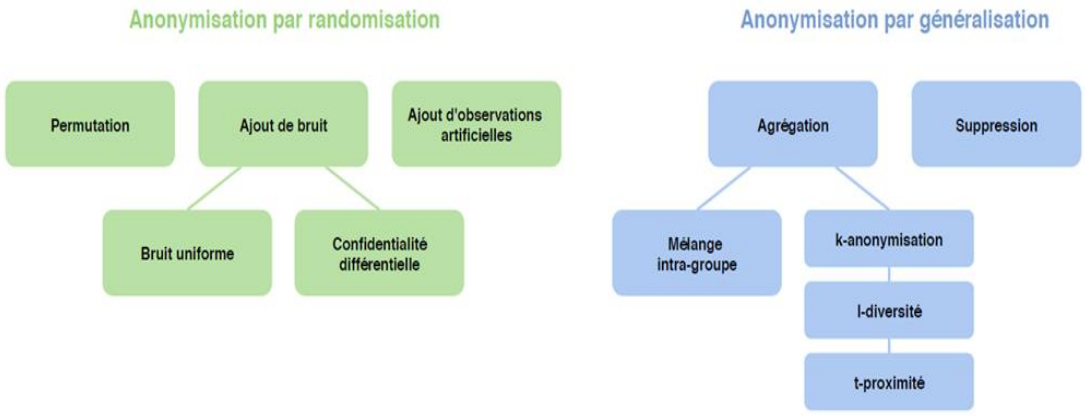
Chaque nouvelle information peut donner lieu à une faille de confidentialité, y compris les métadonnées ou tout autre attribut que l'on jugerait anodin *a priori*.

> Toutes les variables qui n'ont pas d'intérêt pour le traitement prévu doivent être supprimées. Elles constituent un risque de réidentification inutile.

4 Le choix de la méthode d'anonymisation

4.1 Choisir un algorithme d'anonymisation

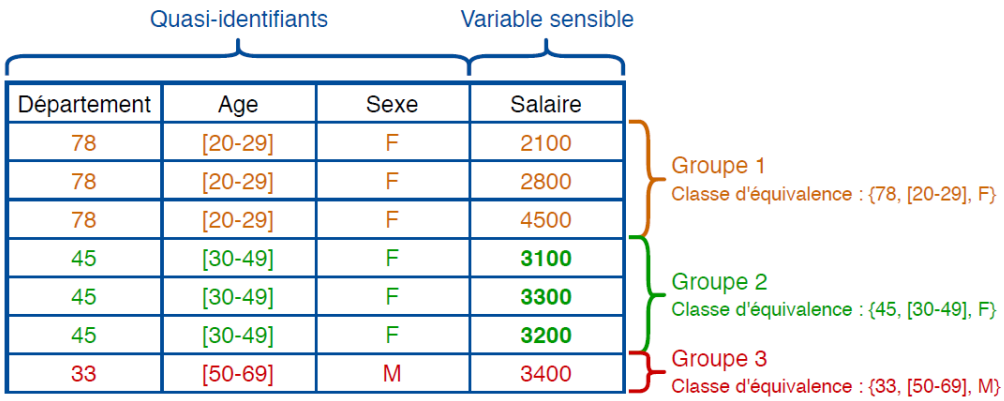
Il existe deux grandes familles de techniques d'anonymisation : les **techniques de randomisation** qui consistent à altérer les données par ajout de bruit par ajout de données artificielles à la base ; et les **techniques de généralisation** qui consistent à créer des groupes d'individus dont les valeurs sont harmonisées (par création d'un intervalle ou d'un ensemble de valeurs, par affectation de la moyenne ou du mode, …).



Parmi les techniques de randomisation, **la confidentialité différentielle** est une méthode d'anonymisation de requête par ajout de bruit. Elle ne permet pas d'anonymiser directement une base de données, mais le résultat d'une requête (numérique) exécutée sur la base (moyenne d'une variable, nombre d'occurrences d'une modalité). Elle présente l'avantage d'offrir des garanties statistiques sur son niveau de confidentialité.

La randomisation inclut également les techniques de **construction d'un nouveau jeu de données** à partir du jeu de données initial. Pour cela, il est possible de permuter les informations issues de plusieurs observations, ou d'ajouter et de supprimer aléatoirement des observations.

La k-Anonymisation, méthode d'anonymisation par généralisation la plus courante, consiste à généraliser les modalités des quasi-identifiants afin de rendre chaque individu indistinguable d'au moins k-1 autres individus. Il existe plusieurs algorithmes permettant de satisfaire cette propriété. Il est également possible de lui ajouter des contraintes (cf. *l-Diversité*, *t-Proximité*), si le risque de réidentification demeure trop élevé.



La table ci-dessus est un exemple de généralisation (à partir de la table des différents types de variables). L'identifiant direct et la variable inutile pour le cas d'usage ont été supprimés. Les autres variables ont été modifiées de manière à former des groupes de k = 3 individus. La dernière observation est un outlier : l'associer à l'un des autres groupes détériorerait significativement ses informations, il vaut donc mieux la supprimer.

Une excellente compréhension de ces techniques est requise pour les appliquer mais la k-Anonymisation et la confidentialité différentielle méritant chacune une série d'articles, l'objet de ce Livre Blanc n'est pas de détailler ces différentes méthodes. L'avis du G29 de la Commission européenne les introduit de manière concise et claire. De plus, la littérature présente des applications de ces méthodes à différents types d'analyse :

- k-anonymisation pour la classification[8]
- k-anonymisation pour la régression[9]
- Confidentialité différentielle pour des comptages [10]
- Confidentialité différentielle pour les séries temporelles[11]
- Confidentialité différentielle pour une régression linéaire [12]
- Confidentialité différentielle pour la descente de gradient [13]

Pour aller plus loin : *Domingo-Ferrer, J., Torra, V. Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation | Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," | Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. Technical report, 2013.*

[8] K. Wang, P. S. Yu and S. Chakraborty - Bottom-up generalization: A data mining solution to privacy protection, 2004

[9] K. LeFevre, DJ. DeWitt, R. Ramakrishnan – Workload-aware anonymization, 2006

[10]Differentially Private Mechanisms for Count Queries, Parastoo Sadeghi, Shahab Asoodeh, Flavio du Pin Calmon

[11]Rastogi, Vibhor & Nath, Suman. (2010). Differentially private aggregation of distributed time-series with transformation and encryption

[12] Differentially Private Simple Linear Regression, Daniel Alabi, Audra McMillan, Jayshree Sarathy, Adam Smith, Salil Vadhan

[13] Differentially Private Stochastic Coordinate Descent, Georgios Damaskinos, Celestine Mendler-Dünner, Rachid Guerraoui, Nikolaos Papandreou, Thomas Parnell

> Le premier critère de sélection d'une technique est le type de données à anonymiser.

La confidentialité différentielle est pertinente dans le cas d'une API ou d'une application de visualisation de données avec maintien d'une cohérence entre plusieurs analyses. Inversement, la k-Anonymisation cible une base de données constituée de quasi-identifiants et d'attributs sensibles.

Il est aussi possible d'appliquer des procédés moins contraignants, tels que l'ajout de bruit, la généralisation de modalités, ou la suppression d'observations trop extrêmes.

Les techniques d'anonymisation sont paramétrables. Dans le cas de la confidentialité différentielle et de la k-anonymisation, les paramètres peuvent être considérés comme des curseurs de l'intensité de l'anonymisation.

Le paramétrage doit être réfléchi en estimant des probabilités de réidentification et/ou en cherchant des cas d'usage similaires dans la littérature. Il peut – si le cas d'usage le permet – être affiné empiriquement.

Ces méthodologies sont implémentées maintenant dans une pléthore d'outils gratuits à code source ouvert :

- R : sdcMicro
- Java : muArgus, ArX
- Python: OpenDifferentialPrivacy, TensorFlow Privacy & TensorFlow Federated, PySyft

4.2 Déterminer une méthode d'évaluation de la qualité d'anonymisation

> La qualité d'une anonymisation doit être retranscrite par des indicateurs et son évaluation peut être en partie automatisée, mais l'appréciation humaine reste indispensable.

Un contrôle automatisé peut vérifier l'état d'indicateurs globaux, comparer via certaines métriques les données avant/après anonymisation et porter une attention particulière à certains aspects précis :

- La présence d'*outliers* parmi les individus ou groupes d'individus (risque de réidentification élevé) ;
- L'homogénéité des variables sensibles dans les groupes, particulièrement si les groupes sont disjoints (risque d'inférence élevé) ;
- La confidentialité des individus sous-représentés (risque de réidentification élevé) ;
- La représentativité de l'échantillon dans un contexte d'échantillonnage ;
- L'adéquation d'indicateurs présents dans la littérature.

Cependant, il est compliqué – si ce n'est impossible – de rendre compte totalement de ces risques par des indicateurs de *monitoring*. Les précédents points devraient également faire l'objet d'une évaluation humaine, ainsi que :

- Les données publiques qui pourraient être utilisées pour faciliter la réidentification ;
- Les risques identifiés et non transcriptibles en indicateurs ;
- Les risques non-identifiés et par conséquent non couverts par les indicateurs mis en place.

Aussi, des tentatives de réidentification, à l'instar des tests de pénétration en cybersécurité, doivent être menées pour découvrir des angles morts. Cette approche est détaillée dans la partie 5.4 de ce Livre Blanc comme méthode d'évaluation à long terme du système. Des tests de réidentification ont bien sûr tout autant de sens au moment de son élaboration.

4.3 Paramétrer l'algorithme d'anonymisation

> Une anonymisation est une recherche d'équilibre. En ce sens, il convient de trouver le meilleur compromis entre confidentialité et qualité des données.

La création d'indicateurs et l'automatisation de l'évaluation de la qualité sont importantes car elles permettent d'élaborer une démarche fondée sur des méthodes classiques d'optimisation et de paramétrage :

- Optimisation avec contrainte d'une fonction de coût, ou d'un score d'anonymisation (lagrangien, descente de gradient, …) ;
- Recherche algorithmique du « meilleur » paramétrage (Grid Search, Random Search).

5 Et après ?

5.1 Etablir un compte rendu de la démarche

La CNIL considère la documentation comme un élément clé pour attester de la conformité d'une organisation au RGPD. Un compte rendu de la démarche a toute sa place dans les documents décrivant les procédures internes (registre des activités de traitement). Il convient également de référencer les éléments externes utilisés lors de l'élaboration de la démarche (papiers scientifiques, documentation technique, etc.).

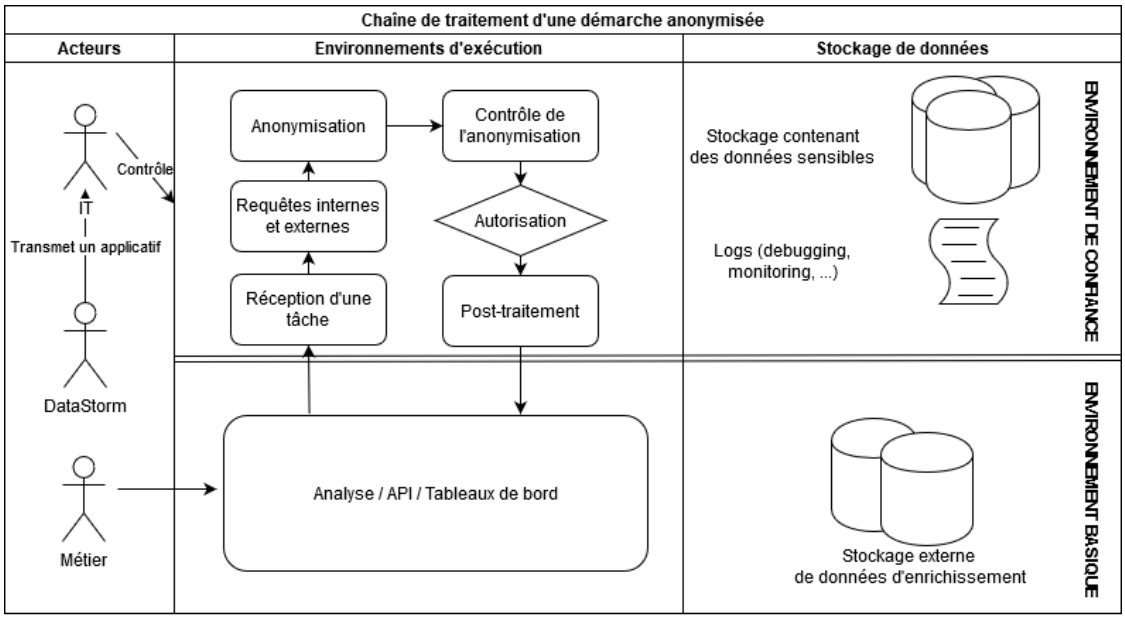
Tout élément permettant d'avoir une vision claire sur les traitements mis en œuvre sera un atout incontestable pour expliquer, reprendre, auditer la démarche de dé-identification mise en place.

5.2 Sécuriser les infrastructures

L'anonymisation n'a de sens que si les données restent effectivement accessibles uniquement aux personnes censées y avoir accès. L'effort d'anonymisation ne doit pas être mis en défaut par une faiblesse dans la sécurité des infrastructures. Cela est d'autant plus vrai dans le contexte de la confidentialité différentielle, où chaque utilisateur peut effectuer un certain nombre de requêtes sur la base de données. Un *monitoring* des requêtes est alors nécessaire, par exemple pour qu'il ne soit pas possible d'obtenir plusieurs versions anonymisées de la même requête (dont la moyenne donnerait une information plus précise).

Ainsi, les serveurs, librairies et programmes utilisés doivent être mis à niveau. Tout comme les pratiques de sécurité qui évoluent sans cesse : à l'image des pratiques de stockage d'un mot de passe dans une base de données d'authentification.

Les accès aux infrastructures depuis le monde extérieur (HTTPS, SSH, ports ouverts) doivent être monitorés. La gestion des accès aux bases de données doit être stricte et réfléchie. Il est même possible de chiffrer certaines tables afin de s'assurer que seules les personnes accréditées puissent accéder à l'information.



Les infrastructures doivent être classées, *a minima*, en deux catégories :

- Un environnement dit de confiance, où sont stockées les données sensibles, le système d'authentification, l'applicatif d'anonymisation ;

- un environnement basique, moins sécurisé, accessible par un plus large public pour interagir avec les utilisateurs et restituer les résultats (visuellement ou non). Faute d'avoir un contrôle suffisant sur ce dernier, il sera considéré comme corrompu.

La sécurité d'une anonymisation repose sur un prérequis fort : celui de contrôler l'accès aux données anonymisées. Dans l'exemple du schéma ci-dessus, seul l'IT d'une entreprise cliente de DataStorm a accès à l'environnement de confiance. L'autorisation d'un flux sortant de ce périmètre va au-delà du simple contrôle d'identité et suppose de savoir valider efficacement une anonymisation, et de savoir déceler des comportements suspicieux.

Pour surveiller dans le temps les accès aux données, toute exécution de la chaîne de traitement doit générer un historique analysable par la machine et par un œil humain. Ainsi, des « logs » sont enregistrés au sein de l'environnement sécurisé. Ces détails d'exécution,

traçage des actions de lecture, écriture, alimentent le volet *monitoring* de l'applicatif d'anonymisation.

5.3 Monitorer l'anonymisation

En cybersécurité, une infrastructure doit être surveillée au jour le jour. En Data Science, l'évolution d'un modèle doit être étudiée au cours du temps. Il en va de même de l'anonymisation.

Une fois l'application ou l'API anonymisée publiée, il convient de garder en tête qu'aucun système n'est infaillible et que l'histoire est là pour en attester :

- En 2006, AOL a publié une base de données « anonymisée » de plusieurs millions de requêtes internet de plusieurs centaines de milliers d'utilisateurs. AOL pensait avoir anonymisé tous les quasi-identifiants, mais c'était sans compter les requêtes elles-mêmes, qui contenaient des informations très identifiantes.
- En 2007, Netflix a publié une base de données « anonymisée » de plusieurs millions de critiques de films de quelques centaines de milliers d'utilisateurs. Leur anonymisation était fondée sur la suppression de certaines informations et la pseudonymisation des identifiants, mais c'était sans compter les sources de données exogènes présentes sur internet. En croisant les informations de la base anonymisée avec celles de l'Internet Movie Database, il était possible de réidentifier des personnes. Or, les informations des critiques, supposées anonymes, peuvent être sensibles et traduire des orientations politiques ou autres.

La plus grande erreur serait de considérer une sécurité comme acquise. Auditer régulièrement un processus d'anonymisation permet de déceler au plus tôt une faille non identifiée, et permet de réagir à l'évolution du contexte extérieur (évolution des données exogènes, des connaissances de l'attaquant, du modèle de menace précédemment établi). A titre de garantie, il est possible de formaliser une procédure d'audit et de désigner un responsable (service chargé de la mise en œuvre du traitement, délégué à la protection des données, etc.).

Il est également possible de programmer des alertes en cas de comportements anormaux ou inhabituels, et de bloquer par mesure de sécurité l'accès à la donnée. Le *monitoring* ne s'improvise pas. Des architectures types existent, des outils spécialisés ont été développés pour s'adapter à ce cas particulier (exemple : la croissance rapide du coût de stockage d'un *monitoring*).

> Comme pour la surveillance des infrastructures, l'anonymisation peut être monitorée afin de tracer l'évolution des indicateurs utilisés pendant le paramétrage de la méthode d'anonymisation. Le *monitoring* permet d'anticiper un simple recalibrage ou des modifications plus profondes.

5.4 Eprouver les infrastructures et la méthode de dé-identification

Les sportifs, les cryptologues, les responsables informatiques s'accordent pour dire que la meilleure façon de tester la robustesse d'une défense, c'est de la mettre à l'épreuve. Déceler ses failles afin de corriger ce qui est perfectible, dresser des parades et surveiller les points faibles. Dans bien des domaines, se forme une dualité entre l'attaque et la défense : le cryptographe contre le cryptanalyste, les ingénieurs en cybersécurité contre les concepteurs de virus, etc.

Avoir une meilleure maîtrise des risques de fuite d'information passe par une expertise de la réidentification. A l'instar de la cryptanalyse, perpétuel domaine de recherche. A l'instar de Kali Linux, outil de cybersécurité pour réaliser des tests de pénétration. Théoriser et se doter d'outils pour faciliter la réidentification permet, à terme, de bâtir de meilleures anonymisations.

Comment raisonner comme un attaquant ? La partie qui suit présente sous une forme ludique des axes stratégiques – non exhaustifs – qu'un attaquant peut suivre pour mettre à mal une anonymisation.

Règle 1 (dite de Murphy) : si une donnée annexe peut fuiter, alors elle fuitera

Aujourd'hui, personne n'est à l'abri d'une attaque sur des données à caractère personnel et/ou confidentielles. Les entreprises publiques ou privées, les associations sont régulièrement la cible de cyberattaques concernant les données des milliers d'utilisateurs/clients.

La possession par l'attaquant de logs ou d'une table qui ont fuité change le paradigme prévu dans le cahier des charges et offre à ce dernier la possibilité de croiser cette information annexe, parfaitement identifiante, avec une variable de la table anonymisée.

Règle 2 : être ou ne pas être, c'est toujours de l'information

Les fuites d'informations peuvent être classées en deux catégories : les divulgations positives, et les divulgations négatives.

Exemple de divulgation négative : soit une base de données d'une agence de voyage recensant les destinations de ses clients. Si l'attaquant sait qu'une personne fait partie d'un groupe d'individus et si aucun des individus du groupe ne s'est rendu à une destination D, alors l'attaquant sait que cette personne ne s'est pas rendue là-bas.

Règle 3 : les variables sensibles n'ont pas été traitées comme quasi-identifiantes mais peuvent servir à réidentifier

Si les variables sensibles n'ont pas (ou peu) été traitées, c'est parce que le cas d'usage nécessitait la plus grande précision. Imaginons que k lignes d'une table soient semblables au vu des quasi-identifiants, mais qu'une observation sorte du lot sur une variable considérée comme sensible. Cette distinction pourrait aider un attaquant à réidentifier un individu qu'il sait dans la table. Réidentifier lui permettra potentiellement d'apprendre de nouvelles informations grâce aux autres variables sensibles.

Règle 4 : une fuite d'information peut avoir lieu grâce à la réidentification de plusieurs personnes

Les connaissances d'un attaquant sur certaines personnes peuvent lui permettre de dégager de l'information sur d'autres personnes. Reprenons l'exemple de l'agence de voyage. Si l'attaquant sait que plusieurs personnes font partie d'un groupe d'individus (par exemple à cause de leur âge), et s'il est capable d'identifier certaines personnes de ce groupe, alors il obtient une information plus précise sur les autres membres du groupe. Si les personnes identifiées par l'attaquant étaient les seules associées à certaines destinations, alors il en conclut que les autres personnes du groupe ne s'y sont pas rendues.

Règle 5 : l'attaquant peut effectuer le nombre de requêtes qu'il veut, quand il veut pour étudier différentes réponses à une même question, ou à plusieurs questions légèrement différentes

L'attaquant saura tirer profit des autorisations qu'il possède pour faire parler la donnée. Il pourra utiliser autant d'allers-retours avec la source de données qu'il le souhaite et pourra même les répartir dans le temps. L'idée est de comparer un ensemble de résultats pour effectuer des traitements statistiques.

Dans le cas d'une anonymisation par ajout de bruit, la moyenne de plusieurs résultats d'une même requête constitue un estimateur plus précis. Si l'attaquant ne peut pas extraire plusieurs fois la même requête, il peut encore extraire plusieurs requêtes similaires en faisant varier le périmètre d'un delta.

Il est possible de mettre en place des restrictions afin de réduire le degré de liberté des utilisateurs, mais il est peu vraisemblable de parer à toutes les situations.

> En matière d'anonymisation le risque zéro n'existe pas, mais il est possible de le maîtriser. Des audits récurrents, une organisation autour de la recherche de failles (hackathon, bug bounty) permettent de garantir la confiance des usagers envers le système mis en place.

6 Conclusion

L'apparition massive des données numériques a multiplié les opportunités de traitement et leur utilisation dans tous les secteurs d'activité. Et la demande continue de croître. Pour répondre à cette demande de manière responsable, il faut réfléchir à de nouveaux modes de gestion des données respectueux des personnes qu'elles exposent, ainsi que leur intégration dans un système d'information (potentiellement basé sur les solutions cloud d'Amazon, Google et Microsoft).

Sur la base du RGPD, les CNILs européennes ont publié une liste de recommandations concernant l'exploitation et la diffusion de données à caractère personnel, dans lesquelles elles encouragent l'utilisation des méthodes d'anonymisation. Ces recommandations ont d'autant plus de sens que le potentiel offert par l'intelligence artificielle sur des données très personnelles est considérable. Le déploiement d'une organisation, de processus, et de moyens de suivi de l'anonymisation permet de multiplier les leviers de performance en exploitant l'ensemble du patrimoine informationnel tout en respectant le caractère personnel ou sensible des données.

Vous souhaitez obtenir des informations sur les cas d'usages et les méthodes d'anonymisation qui leur correspondent ?

Contactez-nous : contact@datastorm.fr

7 Glossaire

> *Ce glossaire reprend, en partie, des définitions présentées par la CNIL dans le document « Guide pratique : les durées de conservations », ou sur leur site internet https://www.cnil.fr/fr/glossaire et des définitions issues de l'avis de 05/2014 sur les techniques d'anonymisation, réalisé par le groupe de travail Article 29.*

API : Interface normalisée d'un logiciel facilitant les communications avec un second logiciel pour offrir un service précis.

Archivage : procédé qui vise à conserver des données qui ont cessé d'être d'utilité courante. L'archivage peut être intermédiaire lorsqu'à l'issue de la phase d'utilisation courante les données présentent encore un intérêt administratif (par exemple pour constituer des éléments de preuve dans le cas d'un contentieux). Il peut également être définitif si, au terme de l'archivage intermédiaire, les données présentent un intérêt particulier.

Anonymiser : rendre impossible l'identification et l'obtention d'informations sur les individus d'une base de données, par la lecture des informations qu'elle renferme mais également par l'ajout d'informations externes à ces dernières.

Attaque par inférence : utilisation de techniques statistiques afin d'estimer certaines caractéristiques d'un individu, à défaut de pouvoir le réidentifier.

Confidentialité différentielle : méthode d'anonymisation de requête par ajout de bruit (moyenne d'une variable, nombre d'occurrences d'une modalité). Elle présente l'avantage d'offrir des garanties statistiques sur son niveau de confidentialité.

Connaissance annexe **:** toute information publiquement accessible qui n'est pas contenue dans la base de données (culture générale, internet, autre base de données, etc.).

Corrélation : capacité de relier plusieurs enregistrements issus d'une ou plusieurs tables, qui correspondent à un même individu ou groupe d'individus.

CNIL : Commission nationale de l'informatique et des libertés. La CNIL est une autorité administrative indépendante. Elle est le régulateur des données à caractère personnel. Elle accompagne et contrôle les professionnels dans leur mise en conformité et aide les particuliers à maîtriser leurs données à caractère personnel et exercer leurs droits. (https://www.cnil.fr/fr/les-missions-de-la-cnil)

Délégué à la protection des données : le délégué à la protection des données (DPO) est chargé d'assurer la conformité au RGPD au sein de l'organisme. Son rôle concerne l'ensemble des traitements concernant des données à caractère personnel réalisés par cet organisme.

Anonymisation par généralisation : généraliser ou diluer les attributs des personnes concernées en modifiant leur échelle ou leur ordre de grandeur respectif (par exemple, une région plutôt qu'une ville, un mois plutôt qu'une semaine).

Finalité d'un traitement : objectif principal de l'utilisation de données personnelles. Les données sont collectées dans un but bien déterminé et légitime. Elles ne sont pas traitées ultérieurement de

façon incompatible avec cet objectif initial. Ce principe de finalité limite la manière dont le responsable de traitement peut utiliser ou réutiliser ces données dans le futur.

K-Anonymité : propriété d'une base de données dans laquelle toute observation a ses quasi-identifiants identiques à au moins k-1 autres observations.

Logs : historique des traitements réalisés au cours du temps par une application. Les logs sont particulièrement utiles pour suivre les activités d'un processus, régler des défauts de fonctionnement.

Modèle de menace : modélisation d'une réalité afin de mettre en évidence des entités à protéger, des forces et faiblesses de sécurité, des attaquants potentiels.

Monitoring : ensemble des activités visant à montrer qu'un système est opérationnel. Collecte et visualisation d'indicateurs sur le long terme, alerte et potentiellement action après une détection d'anomalie.

Open data : un mouvement, né en Grande-Bretagne et aux États-Unis, d'ouverture et de mise à disposition des données produites et collectées par les services publics (administrations, collectivités locales...).

Pseudonymisation : remplacement d'un attribut (généralement un attribut unique) par un autre dans un enregistrement.

Randomisation : une famille de techniques qui altèrent la véracité des données afin d'affaiblir le lien entre les données et l'individu. Si les données sont suffisamment incertaines, elles ne peuvent plus être rattachées à un individu en particulier.

Registre des activités de traitement : le registre des activités de traitement permet de recenser vos traitements de données et de disposer d'une vue d'ensemble de ce que le responsable de traitement fait avec les données à caractère personnel.

Respect de la vie privée par conception : certification en vertu de l'article 42 du RGPD, garantissant qu'un traitement est conçu pour respecter en tout point les principes du RGPD (minimisation, confidentialité, transparence, limitation de la conservation, etc.).

Variable quasi-identifiante : variable ou ensemble de variables d'une base qui, si elles sont combinées avec d'autres variables ou informations externes, peuvent permettre d'obtenir des informations sur des individus de la base.

Variable sensible : variable ciblée par l'étude, contenant des informations confidentielles préjudiciables si elles devaient être reliées à un individu. Ces variables sont en nombre restreint et ne sont généralement pas modifiées lors de l'anonymisation.

La filiale d'expertise et de conseil du Groupe ENSAE-ENSAI

DataStorm adresse la chaîne complète de valeur Data des entreprises

De l'application des résultats des derniers travaux des chercheurs du Groupe ENSAE-ENSAI et de l'Institut polytechnique de Paris à la mise en production et à la maintenance de solutions permettant aux entreprises d'innover grâce à l'exploitation des données.

Think

Nous mobilisons les laboratoires de recherche du Groupe et du plateau de Saclay au service de votre stratégie data.

Model

Nous vous aidons à développer les modèles et algorithmes dédiés à votre métier, avec des leviers opérationnels directs.

Build

Nous construisons les outils de déploiement des algorithmes au sein de vos systèmes d'information.

DATASTORM.FR contact@datastorm.fr


DATA
STORM