CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

PROTÉGER les données personnelles
ACCOMPAGNER l'innovation
PRÉSERVER les libertés individuelles

Présentation de la doctrine technique de la CNIL concernant l'anonymisation

Congrès de la SIF

Amandine JAMBERT

Définitions RGPD

- Donnée anonyme :

  donnée ne permettant pas d'identifier, **ni directement, ni indirectement** un individu. Cette identification doit être impossible par le détenteur du jeu de données ou toute autre personne.

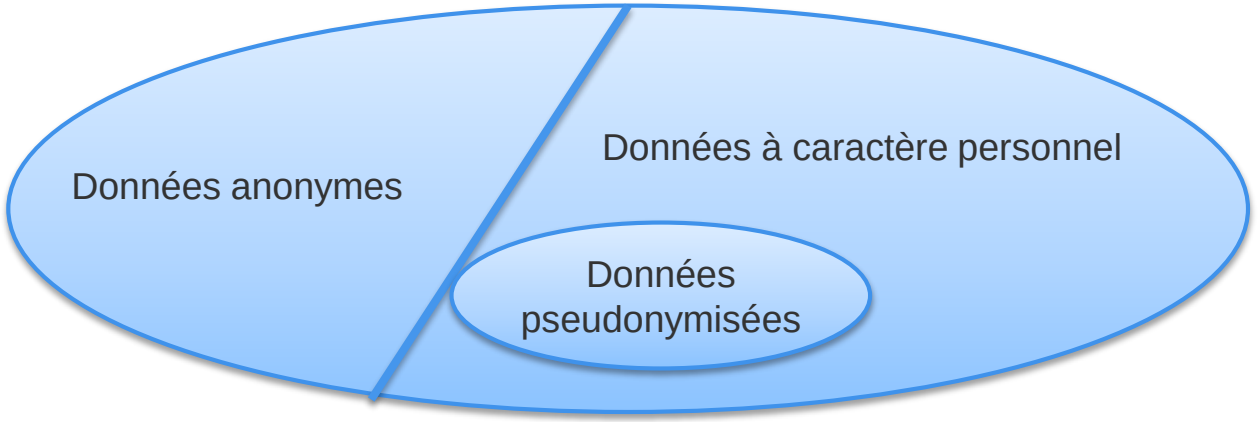
- Donnée pseudonyme :

  donnée pour laquelle il existe **un lien au moins indirect** avec la personne concernée.

  - **dans les deux sens :** s'il existe une table de correspondance (ou une fonction bijective) entre les pseudonymes et les données d'identité.
  - **dans un sens :** si une fonction à sens unique est utilisée pour permettre un suivi des individus sans permettre leur identification directe

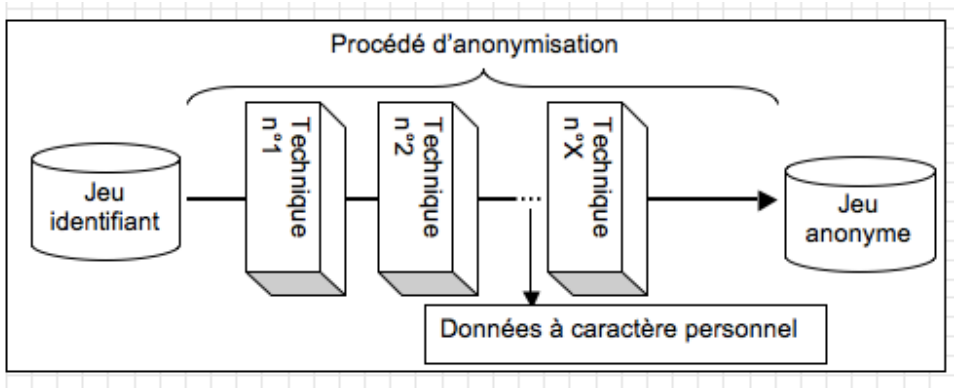
En résumé

- Une donnée anonyme n’est plus une donnée à caractère personnel
- Un jeu de données pseudonymisé n’est pas anonyme



Procédé d'anonymisation

- C'est un ensemble de techniques par lesquelles des données à caractère personnel sont rendues anonymes.



- Si le procédé n'est pas appliqué de façon suffisamment poussée, les données conservent alors leur caractère personnel.

Qu'anonymise-t-on ?

Les données faisant l'objet d'une anonymisation peuvent provenir :

- D'un traitement préexistant répondant à une nécessitait propre
  - Réutilisation de **données préexistantes**
  - Réutilisation de **données publiques** dans certains cas

- Ou être directement collectées dans le but d'être anonymisées
  - **Collecte directe** des données (expressément prévue ou article 7)

Anonymisation – point juridique

- Le processus d'anonymisation est un traitement de données personnelles

- Tout traitement doit* respecter le RGPD, notamment en reposant sur l'une des conditions fixées à l'article 7 de la loi I&L :
  - le consentement préalable des personnes ;
  - le respect d'une obligation légale incombant au responsable de traitement ;
  - l'exécution d'une mission de service public dont ce dernier est investi ;
  - la réalisation de son intérêt légitime.

Avis du G29 [WP 216, Avis 05/2014]

Pour prouver que des données sont anonymes :

1) faire une analyse de risques de ré-identification ;

- en France ➔ risques résiduels quasi nuls

2) démontrer qu'il n'est pas possible :

- **d'isoler/d'individualiser** des informations relatives à un seul individu
- **de relier/corréler** les données d'un même individu ou groupe d'individus
- **de déduire/d'inférer** d'un ensemble d'attributs la valeur d'un autre attribut

Avis du G29

L'avis pose les critères et examine les principales techniques d'anonymisation : principes, points forts et points faibles, erreurs courantes et échecs de chaque technique.

Elles peuvent apporter des garanties (..) mais ***uniquement si leur application est correctement conçue*** *– ce qui suppose que les conditions préalables (le contexte) et les objectif(s) du processus d'anonymisation soient clairement définis.*

Le choix de la solution optimale devrait s'opérer ***au cas par cas****, en utilisant éventuellement* ***une combinaison de techniques différentes***.

Quelques familles de techniques

- **Ajout de bruit :** Altérer la justesse de l'information en ajoutant de l'aléa
- **Permutation :** Mélanger les valeurs d'attributs au sein du jeu de données
- **Généralisation :** Changer la granularité des valeurs pour former des groupes
  - k-anonymat : au moins k personnes ont le même profil
  - l-diversité : au moins l valeurs pour chaque attribut
  - t-proximité : la répartition des valeurs est proche de la distribution initiale

Notre pré-analyse

	Reste-t-il un risque d'individualisation ?	Reste-t-il un risque de corrélation ?	Reste-t-il un risque d'inférence ?
Pseudonymisation	Oui	Oui	Oui
Ajout de bruit	Oui	Peut-être pas	Peut-être pas
Permutation	Oui	Oui	Peut-être pas
Agrégation ou K-anonymat	Non	Oui	Oui
L-diversité	Non	Oui	Peut-être pas

Il n'existe pas de technique "miracle"

En pratique

1) Avant de commencer

- Déterminer l'objectif et les usages prévus pour les données anonymes
- Évaluer si l'anonymisation est réellement l'objectif à atteindre

2) Préparer les données

- Supprimer les éléments d'identification directe et les valeurs rares
- Définir les attributs importants, secondaires et inutiles (i.e. supprimables)
- Définir la granularité optimale et acceptable pour chaque attribut à conserver
- Définir les priorités

En pratique

3) Appliquer un ensemble de techniques

- Choisir les techniques en fonction des cas d'usages prévus
- Evaluer le résultat, en fonction repartir à l'étape 2

4) Après l'anonymisation

- Documenter les techniques (ou types de techniques) utilisées
- Rendre ces techniques aussi publiques que les données
- Effectuer une veille des techniques d'anonymisation et de ré-identification

Merci de votre attention

Et ça marche : quelques exemples

- **Données de santé pour *challenge open data***
  - **Projet “Epidémium” (La Paillasse / Roche)**
- **Données d’opérateurs téléphoniques**
  - **Flux Vision (Orange)**
- **Wifi-tracking**
  - **Suivi de parcours client dans les centres commerciaux ou les lieux publics**
- **Consommation énergétique pour *open data***
  - **Art 179 de la loi “Transition énergétique pour la croissance verte”**

Mais ce n’est pas toujours un objectif atteignable :

“Désidentification” de textes (et pas anonymisation)

- **Comptes rendus médicaux et textes de jurisprudence**

Annexe : Les « données personnelles »

Règlement Général sur la Protection des Données (GDPR)

Article 4 :

- ***données à caractère personnel***, toute information se rapportant à une **personne physique identifiée ou identifiable** ; est réputée être une **«personne physique identifiable» une personne physique qui peut être identifiée, directement ou indirectement,** notamment par référence à un identifiant […]

Considérant (26) :

- Pour déterminer si une personne physique est identifiable, il convient de prendre en considération l'ensemble des moyens **raisonnablement susceptibles** d'être utilisés **par le responsable de traitement ou par toute autre personne** pour identifier la personne physique directement ou indirectement […]
- Pour établir si des moyens sont raisonnablement susceptibles d'être utilisés pour identifier une personne physique, il convient de prendre en considération **l'ensemble des facteurs objectifs,** tels que le coût de l'identification et le temps nécessaire à celle-ci, en tenant compte des technologies disponibles au moment du traitement et de l'évolution de celles-ci.

Annexe

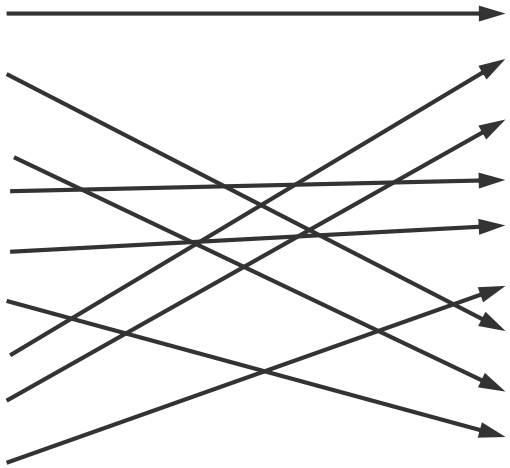
- Anonymisation d'un jeu de donnée par *k*-anonymisation et *l*-diversité
- But : Etudier les affections dont souffrent les personnes d'une Université X en fonction de leur âge et de leur activité

Nom	Activité	Age	Diagnostic
Jean	Master 2	21	Grippe
Pierre	Maître de Conférences	27	Cancer
Anne	Doctorant	26	Cancer
Jérôme	Licence 1	21	VIH
Sophie	Licence 3	20	Grippe
Elisabeth	Doctorant	24	Cancer
Tiphaine	Master 1	22	Rhume
Frédéric	Master 2	23	Rhume
Christophe	Licence 2	21	Allergie

Annexe

- **Application de la k-anonymisation :** technique qui permet d'éviter qu'un individu puisse être isolé dans un jeu de données.
- Pour cela, on effectue des groupements de k individus dans une même classe d'équivalence (ici k = 3).

Nom	Activité	Age	Diagnostic
Jean	Master 2	21	Grippe
Pierre	Maître de Conférences	27	Cancer
Anne	Doctorant	26	Cancer
Jérôme	Licence 1	21	VIH
Sophie	Licence 3	20	Grippe
Elisabeth	Doctorant	24	Cancer
Tiphaine	Master 1	22	Rhume
Frédéric	Master 2	23	Rhume
Christophe	Licence 2	21	Allergie



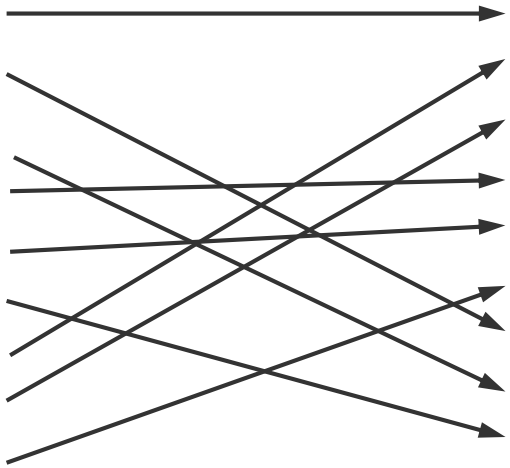
Activité	Age	Diagnostic
Master	[21, 23]	Grippe
Master	[21, 23]	Rhume
Master	[21, 23]	Rhume
Licence	[20, 21]	VIH
Licence	[20, 21]	Grippe
Licence	[20, 21]	Allergie
Enseignant	[24, 27]	Cancer
Enseignant	[24, 27]	Cancer
Enseignant	[24, 27]	Cancer

Données *k*-anonymisées avec *k* = 3

Annexe

- **Problème :** si tous les individus d'une classe d'équivalence présentent les mêmes valeurs, alors il y a un risque de ré-identification. Par exemple, en considérant les données k-anonymisées on peut déduire qu'un enseignant ayant un âge compris entre 24 et 27 ans souffre forcément du cancer. Donc, si on sait qu'Elisabeth est une doctorante, alors on peut en déduire qu'elle a le cancer.

Nom	Activité	Age	Diagnostic
Jean	Master 2	21	Grippe
Pierre	Maître de Conférences	27	Cancer
Anne	Doctorant	26	Cancer
Jérôme	Licence 1	21	VIH
Sophie	Licence 3	20	Grippe
Elisabeth	Doctorant	24	Cancer
Tiphaine	Master 1	22	Rhume
Frédéric	Master 2	23	Rhume
Christophe	Licence 2	21	Allergie



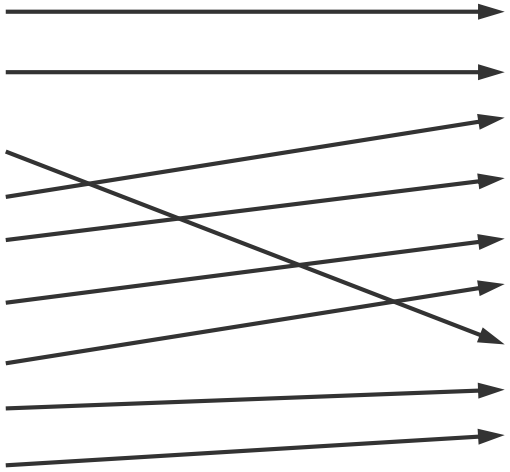
Activité	Age	Diagnostic
Master	[21, 23]	Grippe
Master	[21, 23]	Rhume
Master	[21, 23]	Rhume
Licence	[20, 21]	VIH
Licence	[20, 21]	Grippe
Licence	[20, 21]	Allergie
Enseignant	[24, 27]	Cancer
Enseignant	[24, 27]	Cancer
Enseignant	[24, 27]	Cancer

Données *k*-anonymisées avec $k = 3$

Annexe

- **Application de la l-diversité :** technique permettant de rajouter une contrainte sur les classes d'équivalence obtenues par k-anonymat. Non seulement k profils d'individus doivent être regroupés dans une même classe, mais en plus le champ sensible (ici Diagnostic) doit prendre au minimum l valeurs distinctes (ici l = 3) dans chacune des classes d'équivalence.

Nom	Activité	Age	Diagnostic
Jean	Master 2	21	Grippe
Pierre	Maître de Conférences	27	Cancer
Anne	Doctorant	26	Cancer
Jérôme	Licence 1	21	VIH
Sophie	Licence 3	20	Grippe
Elisabeth	Doctorant	24	Cancer
Tiphaine	Master 1	22	Rhume
Frédéric	Master 2	23	Rhume
Christophe	Licence 2	21	Allergie



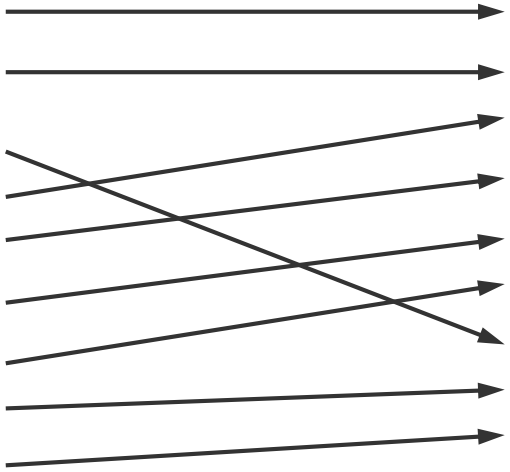
Activité	Age	Diagnostic
Etudiant/Enseignant	[21, 27]	Grippe
Etudiant/Enseignant	[21, 27]	Cancer
Etudiant/Enseignant	[21, 27]	VIH
Etudiant/Doctorant	[20, 24]	Grippe
Etudiant/Doctorant	[20, 24]	Cancer
Etudiant/Doctorant	[20, 24]	Rhume
Etudiant/Doctorant	[21, 26]	Cancer
Etudiant/Doctorant	[21, 26]	Rhume
Etudiant/Doctorant	[21, 26]	Allergie

Données *k*-anonymisées avec *k*= 3 et *l*-diverses avec *l* = 3

Annexe

- **Cependant,** il faut noter qu'il reste possible dans certains cas de procéder à de l'inférence. Par exemple, on peut déduire qu'un étudiant/doctorant de 20 ans aura une probabilité de 33% (soit 1/l) d'avoir la grippe, une probabilité de 33% d'avoir le cancer et une probabilité de 33% d'avoir un rhume… Et surtout aucune chance de souffrir d'une autre pathologie !

Nom	Activité	Age	Diagnostic
Jean	Master 2	21	Grippe
Pierre	Maître de Conférences	27	Cancer
Anne	Doctorant	26	Cancer
Jérôme	Licence 1	21	VIH
Sophie	Licence 3	20	Grippe
Elisabeth	Doctorant	24	Cancer
Tiphaine	Master 1	22	Rhume
Frédéric	Master 2	23	Rhume
Christophe	Licence 2	21	Allergie



Activité	Age	Diagnostic
Etudiant/Enseignant	[21, 27]	Grippe
Etudiant/Enseignant	[21, 27]	Cancer
Etudiant/Enseignant	[21, 27]	VIH
Etudiant/Doctorant	[20, 24]	Grippe
Etudiant/Doctorant	[20, 24]	Cancer
Etudiant/Doctorant	[20, 24]	Rhume
Etudiant/Doctorant	[21, 26]	Cancer
Etudiant/Doctorant	[21, 26]	Rhume
Etudiant/Doctorant	[21, 26]	Allergie

Données *k*-anonymisées avec *k*= 3
et *l*-diverses avec *l* = 3