Anonyme, vraiment ?

Données personnelles, IA et anonymat

Mise en situation. Les IA apprennent à partir d'énormes quantités de données, et beaucoup de ces données parlent de nous : nos publications, nos photos, nos achats, nos trajets, notre santé... Pour protéger la vie privée, on dit souvent que les données sont « anonymisées » : on a retiré les noms.
Mais un fichier sans nom est-il vraiment anonyme ? Dans cette activité, vous allez jouer les détectives, compter des combinaisons pour comprendre pourquoi il est si facile de retrouver quelqu'un, puis apprendre à mieux protéger des données.

Remarque

Fin mai 2025, l'entreprise Meta a commencé à utiliser les publications publiques (textes, photos, commentaires) des utilisateurs européens adultes de Facebook et d'Instagram pour entraîner ses IA. La CNIL, l'autorité française qui protège les données personnelles, a rappelé que chacun pouvait s'y opposer grâce à un formulaire.

1. Qu'est-ce qu'une donnée personnelle ?

Définition

Selon le règlement européen sur la protection des données (RGPD, 2018), une **donnée personnelle** est une information qui se rapporte à une personne **identifiée** ou **identifiable**, directement (son nom, sa photo) ou indirectement (en **croisant** plusieurs informations).
Des données sont **anonymes** s'il est impossible, même en les croisant avec d'autres informations, de retrouver les personnes. Les données vraiment anonymes ne sont plus soumises au RGPD : c'est pourquoi il est important de savoir si une « anonymisation » est réussie.

▶ À vous de jouer !

Parmi les informations suivantes, lesquelles sont des données personnelles ? Justifier.

a) Le nom et le prénom d'un élève.

b) Une photo de visage.

c) L'adresse IP d'un ordinateur à la maison.

d) La température moyenne à Dijon en juillet.

e) Le nombre d'élèves d'un lycée.

f) Le sexe, la date de naissance et la commune d'un élève, sans son nom.

2. Le fichier « anonyme »

Pour une étude sur la santé des lycéens, l'infirmerie d'un lycée publie le fichier suivant. Les noms ont été retirés : le fichier est présenté comme « anonyme ».

Ligne	Sexe	Date de naissance	Commune	Motif du passage
1	F	12/03/2010	Saint-Ambre	Migraine
2	G	05/07/2009	Montclair	Entorse
3	F	12/03/2010	Montclair	Allergie
4	G	23/11/2010	Saint-Ambre	Asthme
5	F	30/01/2009	Saint-Ambre	Migraine
6	G	05/07/2009	Valbrune	Mal de ventre
7	F	18/09/2010	Valbrune	Allergie
8	G	23/11/2010	Saint-Ambre	Stress

Par ailleurs, le site du lycée publie la liste des élèves du club théâtre, avec leur date d'anniversaire pour pouvoir les fêter :

Nom	Sexe	Date de naissance	Commune
Chloé Martin	F	12/03/2010	Montclair
Emma Petit	F	30/01/2009	Saint-Ambre
Lucas Bernard	G	05/07/2009	Montclair
Yanis Robert	G	23/11/2010	Saint-Ambre

Les personnes, le lycée et les communes de cette activité sont inventés.

▶ À vous de jouer !

1. Le fichier de l'infirmerie contient-il des noms ? Contient-il des données personnelles ?
2. En croisant les deux tableaux, retrouver le motif du passage à l'infirmerie de Chloé, d'Emma et de Lucas.
3. Que peut-on apprendre sur Yanis ? Pourquoi ne peut-on pas être certain de son motif ?
4. Dans le fichier de l'infirmerie, combien de lignes sont **uniques**, c'est-à-dire qu'aucune autre ligne n'a le même sexe, la même date de naissance et la même commune ?

Coup de pouce

Chloé est une fille née le 12/03/2010 à Montclair : cherchez la ligne qui a exactement ces trois valeurs. Attention, la ligne 1 a la même date mais pas la même commune.

Remarque

Ce n'est pas un cas d'école. En 1997, la chercheuse Latanya Sweeney a retrouvé le dossier médical du gouverneur de l'État du Massachusetts dans des données de santé « anonymisées », en les croisant avec une liste électorale qui contenait, elle aussi, le code postal, la date de naissance et le sexe.

3. Pourquoi est-on si facile à retrouver ?

Définition

Chaque personne correspond à une **combinaison** de caractéristiques (par exemple : sexe et date de naissance). Si le nombre de combinaisons possibles est beaucoup plus grand que le nombre de personnes, la plupart des combinaisons sont vides, et la plupart des personnes sont **seules** dans leur combinaison : elles sont identifiables.

▶ À vous de jouer !

On s'intéresse aux habitants d'une commune de 6 000 habitants, âgés de 0 à 79 ans. On compte 365 jours par an.

1. Combien y a-t-il de dates de naissance possibles sur 80 ans ? En tenant compte du sexe (2 possibilités), combien y a-t-il de combinaisons (sexe, date de naissance) ?
2. En moyenne, combien d'habitants de la commune y a-t-il par combinaison ? Qu'en déduire ?
3. Même question si l'on ne publie que le sexe et l'**année** de naissance.
4. Même question avec le sexe, le **mois** et l'année de naissance.

Coup de pouce

$$\text{Nombre moyen d'habitants par combinaison} = \frac{\text{nombre d'habitants}}{\text{nombre de combinaisons}}.$$

Remarque

En 2000, Latanya Sweeney a estimé, avec les données du recensement des États-Unis de 1990, que 87 % des habitants (216 millions sur 248) pouvaient être identifiés par seulement trois informations : le code postal, le sexe et la date de naissance. En 2019, une étude publiée dans *Nature Communications* a estimé que 99,98 % des Américains pourraient être ré-identifiés dans n'importe quel fichier contenant 15 caractéristiques (âge, sexe, situation familiale…), même incomplet.

4. Mieux protéger : la généralisation

Définition

Pour protéger un fichier, on peut **généraliser** les informations : remplacer la date de naissance par l'année, la commune par un secteur plus large, ou supprimer une colonne.
Un fichier est dit k**-anonyme** si chaque combinaison publiée est partagée par au moins k personnes : chacun se cache alors parmi au moins k personnes.

▶ À vous de jouer !

On reprend le fichier de l'infirmerie.

1. On publie seulement le sexe, l'**année** de naissance et la commune. Combien de lignes sont uniques ?
2. On publie seulement le sexe et l'année de naissance. Combien de lignes sont uniques ? Le fichier est-il 2-anonyme ?
3. On publie seulement le sexe, et on remplace l'année par « 2009-2010 » pour tout le monde. Le fichier est-il 2-anonyme ? Quelle est la plus grande valeur de k ?
4. Quel est l'inconvénient de trop généraliser ?

5. Compter les lignes uniques en Python

Exemple

Chaque ligne du fichier est une liste `[sexe, date, commune, motif]`.

```
T = [["F", "12/03/2010", "Saint-Ambre", "Migraine"],
     ["G", "05/07/2009", "Montclair", "Entorse"],
     ["F", "12/03/2010", "Montclair", "Allergie"],
     ["G", "23/11/2010", "Saint-Ambre", "Asthme"],
     ["F", "30/01/2009", "Saint-Ambre", "Migraine"],
     ["G", "05/07/2009", "Valbrune", "Mal de ventre"],
     ["F", "18/09/2010", "Valbrune", "Allergie"],
     ["G", "23/11/2010", "Saint-Ambre", "Stress"]]

def identiques(l1, l2, colonnes):
    for c in colonnes:
        if l1[c] != l2[c]:
            return False
    return True

def nb_uniques(table, colonnes):
    n = 0
    for l1 in table:
        compte = 0
        for l2 in table:
            if identiques(l1, l2, colonnes):
                compte = compte + 1
        if compte == 1:
            n = n + 1
    return n
```

La liste `colonnes` indique les colonnes publiées : 0 pour le sexe, 1 pour la date, 2 pour la commune.

▶ À vous de jouer !

1. Que renvoie `identiques(T[3], T[7], [0, 1, 2])` ? Et `identiques(T[0], T[2], [0, 1, 2])` ?
2. Dans la fonction `nb_uniques`, pourquoi une ligne est-elle unique quand `compte` vaut 1 ?
3. Que renvoie `nb_uniques(T, [0, 1, 2])` ? Et `nb_uniques(T, [0, 2])` ? Et `nb_uniques(T, [1])` ?
4. Comment utiliser `nb_uniques` pour savoir si un fichier est 2-anonyme ?

6. L'IA et nos données

Remarque

- Les grands modèles d'IA sont entraînés sur d'énormes quantités de textes et d'images collectés sur Internet, qui contiennent des données personnelles. Des chercheurs ont montré qu'on peut parfois faire « recracher » à un modèle de langage des extraits de ses données d'entraînement, y compris des noms et des coordonnées.
- Une fois qu'une donnée a servi à entraîner un modèle, il est très difficile de l'en retirer : c'est pourquoi il était conseillé de s'opposer **avant** le début de l'entraînement des IA de Meta, fin mai 2025.
- Le RGPD donne des droits à chacun : droit d'accès à ses données, droit de les faire rectifier ou effacer, droit de s'opposer à certains usages.

▶ À vous de jouer !

1. Pourquoi est-il difficile de retirer une donnée d'un modèle d'IA déjà entraîné ?
2. Proposer trois gestes pour protéger ses données personnelles en ligne.

7. Pour aller plus loin

★ Question étoilée

Un fichier « anonyme » contient le sexe et la date de naissance des 1 000 élèves d'un lycée, tous nés sur une période de 3 ans.

1. Combien y a-t-il de combinaisons (sexe, date de naissance) possibles ? Combien d'élèves par combinaison en moyenne ?
2. Alice est une élève du lycée. Pour qu'elle soit seule dans sa combinaison, chacun des 999 autres élèves doit avoir une combinaison différente de la sienne. On admet que la probabilité que cela arrive vaut $\left(\frac{2\,189}{2\,190}\right)^{999}$. Calculer cette probabilité, arrondie au centième.
3. Environ combien d'élèves du lycée sont identifiables par ce fichier ?

À retenir

- Une **donnée personnelle** permet d'identifier une personne, directement ou en **croisant** plusieurs informations.
- Retirer les noms ne suffit pas : quelques informations banales (sexe, date de naissance, commune) suffisent souvent à retrouver quelqu'un, car il y a beaucoup plus de **combinaisons** possibles que de personnes.
- On protège mieux un fichier en **généralisant** ses informations ; il est k**-anonyme** si chacun se cache parmi au moins k personnes. Mais trop généraliser rend les données moins utiles.
- Les IA sont entraînées sur nos données, et il est difficile de les en retirer : le RGPD nous donne des droits (accès, effacement, opposition) qu'il faut connaître et utiliser.