# Anonymisation et partage des données:
# Enjeux et pratiques

Cédric Eichler, Christophe Sauvion

Forum « IA et entreprises »

16 octobre 2019

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES CENTRE VAL DE LOIRE

LIFO

Merci à Benjamin Nguyen pour avoir fourni l'inspiration à certains slides

# IA & données

- La donnée est une ressource
- En particulier dans l'IA:
  - Exploitation
  - Entraînement
- Premier point rapport Villani:
  - Se doter d'un réelle politique de data governance
  - Favoriser l'open data
- Un intérêt pour le partage

# Le partage de données

- En open data
  - Publier des données de transport public pour favoriser l'émergence de nouvelles solutions privées
- Entre partenaires
  - Faire analyser les habitudes d'achats de ses clients
  - Faire analyser les données de tests médicaux
- Problème: Transmission de données à caractère personnel

⟹ Des obligations légales qui peuvent être lourdes (GDPR)

…sauf si l'on anonymise!

# GDPR et données anonymes

## Considérant 26

> (…) *Il n'y a dès lors* ***pas lieu d'appliquer les principes relatifs à la protection des données aux informations anonymes****, à savoir les informations ne concernant pas une personne physique identifiée ou identifiable, ni aux données à caractère personnel rendues anonymes de telle manière que la personne concernée ne soit pas ou plus identifiable. Le présent règlement ne s'applique, par conséquent, pas au traitement de telles informations anonymes, y compris à des fins statistiques ou de recherche.*

## Droits de traitement des données anonymes:

- Le RGPD s'applique aux données personnelles
- Les données anonymes ne sont pas personnelles par définition

# Des données anonymes?

## 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 du traitement ou par toute autre personne pour identifier la personne physique directement ou indirectement, tels que le ciblage. 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.* (…)

- Définition de « l'identifiabilité »
- Précision de la portée des techniques de réidentification: obligation de moyens

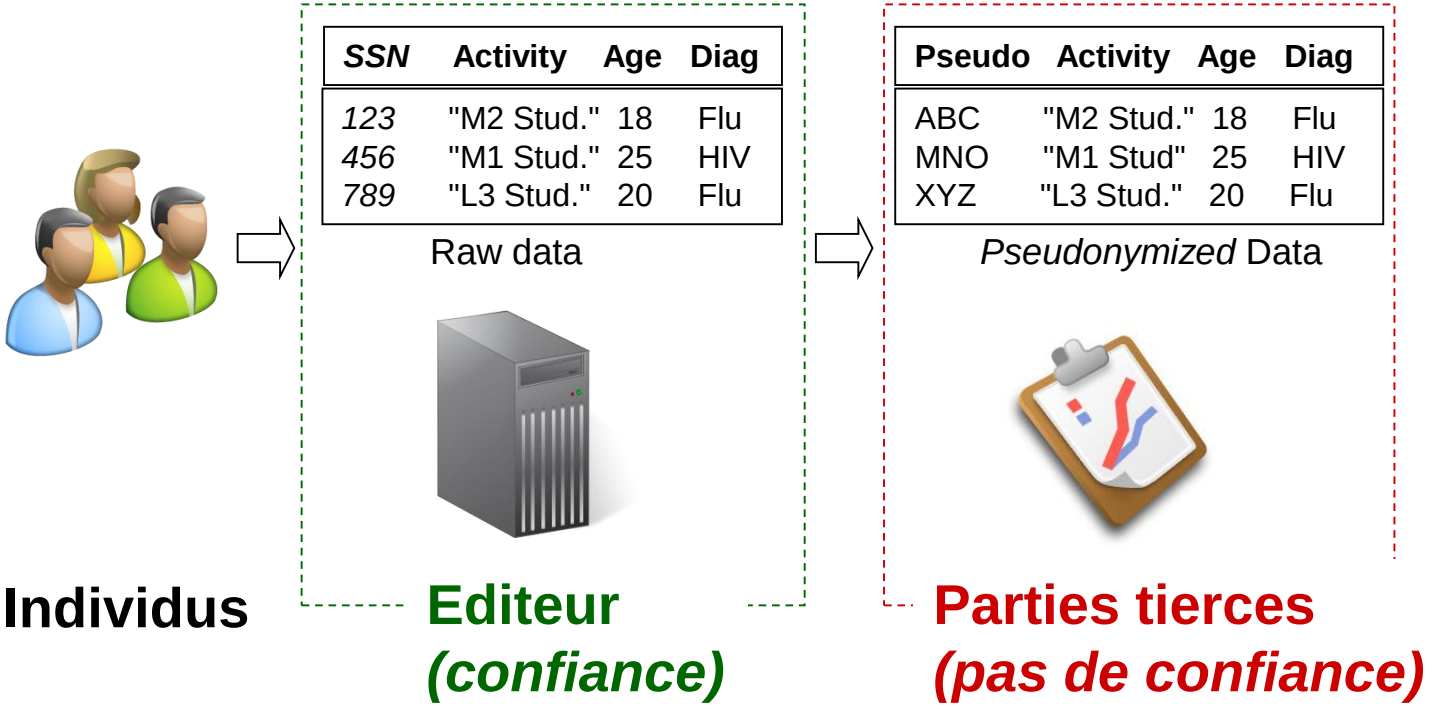
# Comment anonymiser?

Pourquoi se poser la question?

- Producteur de données: comment faire?
- Consommateur: quel impact sur les données récupérées?

# Une première approche

- Pseudonymisation: suppression d'identifiants directs



# 1ère mini table ronde

- **Cette technique vous semble-t-elle raisonnable?**
- Objectifs:
  - Se présenter
  - Appliquer la technique vue
  - Discuter de ses limitations potentielles
- Support:
  - Extrait d'une DB

# 1ère mini table ronde

**Retexp:**

**La pseudonymisation est-elle satisfaisante?**

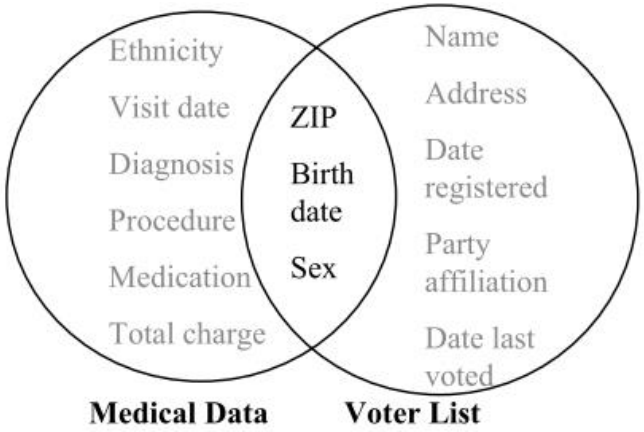
# Attaque sur la pseudonymisation

Sweeney 2002, *k*-anonymity: a model for protecting privacy (IJUFK-BS)

La pseudonymisation rend vulnérables les enregistrements dans lesquels une partie de la donnée permet de réidentifier l'individu concerné.

Sweeney a montré l'existence de **quasi identifiants**:

1. Des données médicales ont été « anonymisées » puis publiées
2. Une liste d'électeur était disponible publiquement




➡ L'identification des enregistrements du gouverneur Weld a été possible en faisant une jointure entre ces deux datasets sur les *quasi-identifiants.*

Recensement US de 1990: « 87% of the population in the US had **characteristics that likely made them unique** based only on {5-digit Zip, gender, date of birth} »

# GDPR et pseudonymisation

**Considérant 26**

*Il y a lieu d'appliquer les principes relatifs à la protection des données à toute information concernant une personne physique identifiée ou identifiable. Les données à caractère personnel qui ont fait l'objet d'une pseudonymisation et qui pourraient être attribuées à une personne physique par le recours à des informations supplémentaires devraient être considérées comme des informations concernant une personne physique identifiable.* (…)

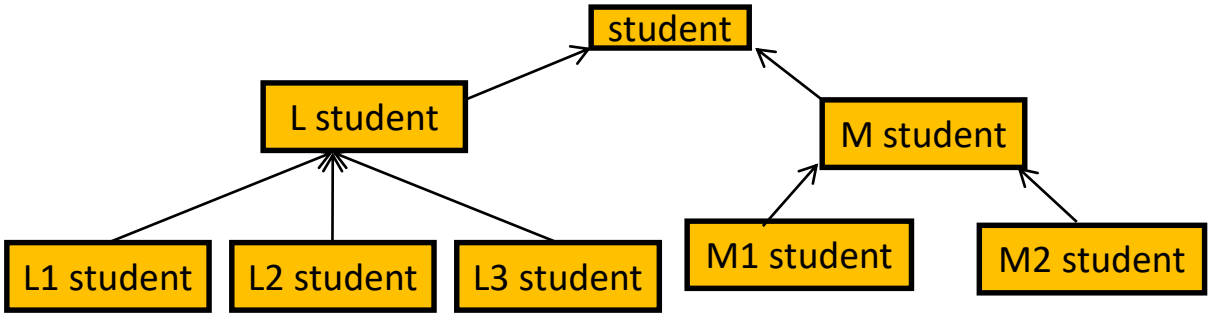
**La pseudonymisation n'est pas de l'anonymisation!**

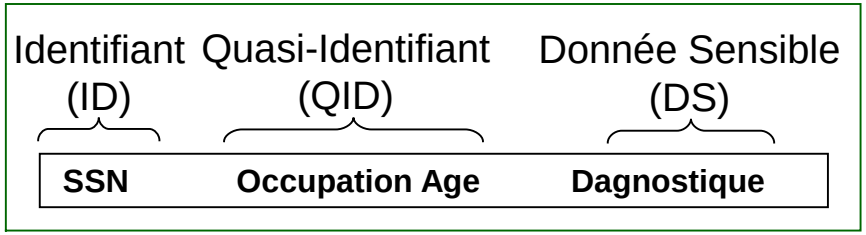
L'UE recommande toutefois de l'appliquer comme mesure preventive complémentaire à l'anonymisation

# Une vraie technique d’anonymisation

# Le *k*-anonymat

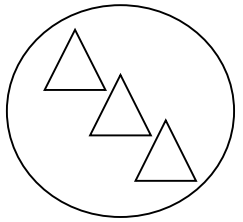
- Pour chaque nuplet:
  - Les identifiants sont retirés
  - Le lien entre QIDs et DS est *obfusqué* mais doit rester globalement correcte
  - Chaque ensemble de QIDs est similaire à  au moins **k-1** autres




| *SSN* | Activity | Age | Diag |
|---|---|---|---|
| *123* | L2 Stud. | 19 | Flu |
| *456* | M1 Stud. | 20 | HIV |
| *789* | L3 Stud. | 22 | N/A |
| 490 | M1 Stud. | 24 | HIV |

| Activity | Age | Diag |
|---|---|---|
| L2 Stud. | 18-25 | Flu |
| M1 Stud. | 20-25 | HIV |
| L3 Stud. | 18-25 | N/A |
| M1 Stud. | 20-25 | HIV |



# 2ème mini table ronde

- **Cette technique est-elle raisonnable?**
- Objectifs:
  - Appliquer la technique vue
  - Identifier des limitations ou problèmes
- Support:
  - Le même extrait

# 2ème mini table ronde

**Retexp:**

**Des problèmes? Des questions?**

# Limitation du k-anonymat

- Garantie du k-anonymat: association d'un individu à une ligne avec probabilité au pire 1/k
- Et si toutes les DS de ces lignes sont les mêmes?

| Activity | Age | Diag |
|---|---|---|
| L Stud. | 18-25 | Flu |
| M1 Stud. | 20-25 | HIV |
| L Stud. | 18-25 | N/A |
| M1 Stud. | 20-25 | HIV |

- Pour un étudiant de master
  - Certes, pas de lien avec une ligne possible
  - Mais, HIV

# La l-diversité

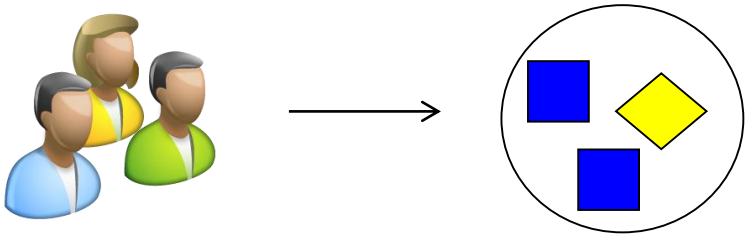
- Contrainte sur les classes d'équivalence
- Chaque classe est liée à au moins l DS différentes
- Garantie: association individu/valeur sensible avec probabilité au pire 1/l

| Activity | Age | Diag |
|---|---|---|
| L Stud. | 18-25 | Flu |
| M1 Stud. | 20-25 | HIV |
| L Stud. | 18-25 | N/A |
| M1 Stud. | 20-25 | HIV |

2-anonyme 1-diverse

| Activity | Age | Diag |
|---|---|---|
| Stud. | 18-20 | Flu |
| Stud. | 18-20 | HIV |
| Stud. | 21-25 | N/A |
| Stud. | 21-25 | HIV |

2-anonyme 2-diverse



# La notion de qualité

- Plusieurs possibilités

| Activity | Age | Diag |
|---|---|---|
| L2 Stud. | 19 | Flu |
| M1 Stud. | 20 | HIV |
| L3 Stud. | 22 | N/A |
| M1 Stud. | 24 | HIV |

| Activity | Age | Diag |
|---|---|---|
| L Stud. | 18-25 | Flu |
| M1 Stud. | 20-25 | HIV |
| L Stud. | 18-25 | N/A |
| M1 Stud. | 20-25 | HIV |

| Activity | Age | Diag |
|---|---|---|
| Stud. | 18-20 | Flu |
| Stud. | 18-20 | HIV |
| Stud. | 21-25 | N/A |
| Stud. | 21-25 | HIV |

- Laquelle est la meilleure?

⇨ Quels traitements sur les données anonymes?

# Discussion