Protection des données personnelles

Mehdi Haddad

mehdi.haddad@u-pec.fr

Introduction

- Génération massive des données personnelles
  - Créées par les individus
  - Les équipements digitaux (smartphones, équipements d'auto mesure, compteurs électriques intelligents)
  - Mises à disposition par les organisations (banques, administrations, centres médicaux, etc.)
- L'ensemble de ces données constitue la vie numérique (souvent privée) de l'individu
  - Décrivant ses déplacements, sa consommation, ses relations, son état médical, social, financier, ses comportements, ses préoccupations, etc

→ Nécessité de protéger ces données

Nuances entre sécurité et respect de la vie privée (privacy)

- Similitudes
  - Protection des données sensibles contre les accès non autorisés
  - Empêcher la modification frauduleuse des données sensibles

- Différences
  - Le respect de la vie privée (privacy) concerne des données liés à des individus. On peut accéder aux données tant qu'on ne peut pas les relier à un individu en particulier
  - Droit de d'accès des individus aux données collectées

Exemples d'exigences légales sur les SI

- Chiffrement des communications
- Mise en œuvre de politique d'accès assurant la confidentialité des données
- Mise en œuvre de mécanismes d'audit permettant de vérifier, à postériori, la mise en œuvre de la politique
- **Anonymisation des données avant de les publier**
- Limiter la durée de rétention des données
- Droit à l'oubli
- etc

Impact des législations sur les SI

- Impact des exigences non fonctionnelles (e.g., sécurité) sur les exigences fonctionnelles
- La sécurité est souvent moins prioritaire que les exigences fonctionnelles dans le cycle de développement
  - Ajout des modules conçu spécifiquement pour se conformer aux législations
  - Augmentation du coût de développement

Vue global

Cahier de charge spécifier en langage naturel

Niveau abstrait: description formelle des législations

Niveau concret qui illustre l'impact des des exigences non fonctionnelles

Attaques contre les bases de données

- Attaques liées à l'environnement de la BD
  - Application tierces non sécurisées accédant à la BD
  - Exemple : Injections SQL
- Attaque liées à mise en place de la politique
  - Un utilise parvient à inférer/déduire une information confidentielle
  - Exemples : BD statistique, recoupement d'information

Données personnelles

- Une donnée personnelle est reliée à un individu
- Une donnée est dite anonyme si elle ne peut pas, de quelque manière que ce soit, être reliée à un individu donné
- Approches pour protéger les données personnelles
  - Base de données statistiques
  - Anonymisation des données

Attaques sur les base de données statistiques

- Base de données statistique est une base de données qui autorisent l'accès à des données agrégées mais pas aux valeurs individuelles
- Exemple : on peut accéder à la moyenne des salaires d'une entreprise mais pas au salaire d'un employé
- Objectif : permettre les traitement statistique sans porter atteinte à la vie privée des individus

Exemple attaque statistique

- On considère la relation Analyse (Patient, H/F, Age, Mutuelle, Leucocyte)

Patient	H/F	Age	Mutuelle	Leucocyte
Dupont	H	30	MMA	6000
Durand	F	25	LMDE	3000
Dulac	F	35	MMA	7000
Duval	H	45	IPECA	5500
Dubois	H	55	MGEN	3500
Dumont	H	38	MMA	7500
Dupré	F	32	IPECA	7200
Dupuis	F	50	MGEN	6800
Dufour	H	45	MAAF	4000
Dumas	H	40	Rempart	3800

Exemple de requêtes

- La requête « quelle est la moyenne du taux de leucocytes des patients ayant plus de 30 ans ? » est autorisé

```
Select Leucovyte
From Analyse
Where Age >30
```

- La requête « quel est le taux de leucocytes de Dupont ? » est interdite

```
Select Leucovyte
From Analyse
Where Patient = 'Dupont'
```

Exemple d'attaque

- Un utilisateur U veut connaitre le taux de Leucocyte de Dubois
- U sait que Dubois est un adhérant masculin à la MGEN
- Comment U peut procéder (en utilisant uniquement des requêtes statistiques) pour deviner le taux de Leucocyte de Dubois ?

Exemple d'attaque

- Un utilisateur U veut connaitre le taux de Leucocyte de Dubois
- U sait que Dubois est un adhérant masculin à la MGEN
- U effectue les deux requêtes suivantes :

Requête 1 :	Requête 2 :
SELECT COUNT (Patient) FROM Analyse WHERE H/F = 'H' AND Mutuelle = 'MGEN' ; Résultat : 1	SELECT SUM (Leucocyte) FROM Analyse WHERE H/F = 'H' AND Mutuelle = 'MGEN' ; Résultat : **3500**

Pourquoi l'attaque a réussi?

- L'attaquant posséder une information externe sur les caractéristique de l'individu attaqué (masculin, MGEN)
- La cardinalité du résultat de la première requête est 1
- Le système doit refuser de réponse aux requêtes ayant une cardinalité de résultat inférieur à une certaine borne k
- Une approche pour éviter l'attaque précédente consiste à refuser les requêtes statistiques n'utilisant qu'une seule ligne

Exemple attaque statistique 2

Requête 3 SELECT COUNT (Patient) FROM Analyse Résultat : 10	Requête 4 SELECT COUNT (Patient) FROM Analyse WHERE NOT (H/F = 'H' AND Mutuelle = 'MGEN') ; Résultat: 9
Requête 5 SELECT SUM (Leucocyte) FROM Analyse Résultat : 54300	Requête 6 SELECT SUM (Leucocyte) FROM Analyse WHERE NOT (H/F = 'H' AND Mutuelle = 'MGEN') ; Résultat : 50800 ;

- L'attaquant n'a plus qu'à calculer la différence entre le résultat 5 et le résultat 6 : 54300 – 50800 = **3500**

Limitation de la cardinalité du résultat

- Le système doit refuser de répondre aux requêtes dont la cardinalité est supérieur à N-k, N étant le nombre total de ligne de la relation
- Au final le système ne doit répondre qu'aux requêtes ayant une cardinalité c tel que : $k \leqslant c \leqslant N-k$
  - En choisissant une valeur de k adéquate

Importance du choix de la valeur de k

- Une valeur de k trop petite
  - Dans l'exemple précédent k=3 ne permet pas de prévenir des attaques
- Une valeur de k trop grande limite les requêtes accepter et donc la disponibilité des donnée
- Equilibre entre confidentialité et disponibilité

Choix de la valeur k

Requêtes avec une cardinalité c avec $3 \leqslant c \leqslant 7$ (N=10, k =3)

Requête 7 SELECT COUNT (Patient) FROM Analyse WHERE H/F = 'H' ; Résultat : 6	Requête 8 SELECT COUNT (Patient) FROM Analyse WHERE H/F = 'H' AND NOT (Mutuelle = 'MGEN') ; Résultat : 5
Requête 9 SELECT SUM (Leucocyte) FROM Analyse WHERE H/F = 'H' ; Résultat : 30300	Requête 10 SELECT SUM (Leucocyte) FROM Analyse WHERE H/F = 'H' AND NOT (Mutuelle = 'MGEN') ; Résultat : 26800 ;

Résultat 10 – Résulat 9 = 30300 – 26800 = **3500**

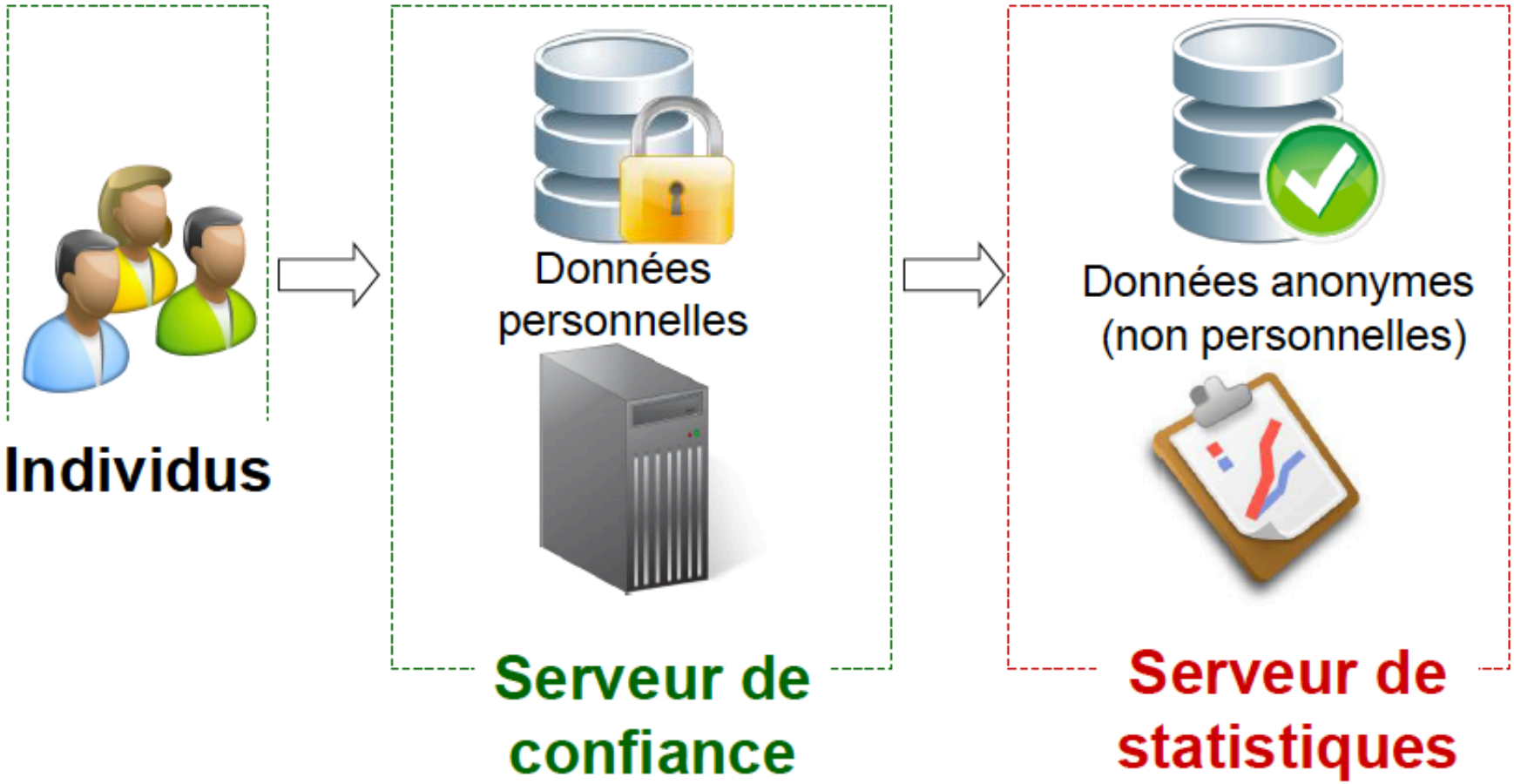
Attaques BD statistiques

- Les attaques par inférence permettent d'illustrer la difficulté d'implémenter correctement une politique d'accès
- Affecté « naïvement » des autorisations aux utilisateurs n'est pas suffisant
- Considérer les accès indirect que peuvent effectuer les utilisateur pour déduire une information interdite

Anonymisation des données

- Permettre l'analyse des données tout en respectant la vie privée
- Approche naïve : supprimer les identifiants des individus avant de publier les données
  - Exemple identifiants : nom, prénom, numéro de sécurité social
- Différentes affaires ont montré que l'approche naïve n'était pas suffisante
  - Des individus ont pu être reconnu même après la suppression des attributs identifiants

Anonymisation : le principe

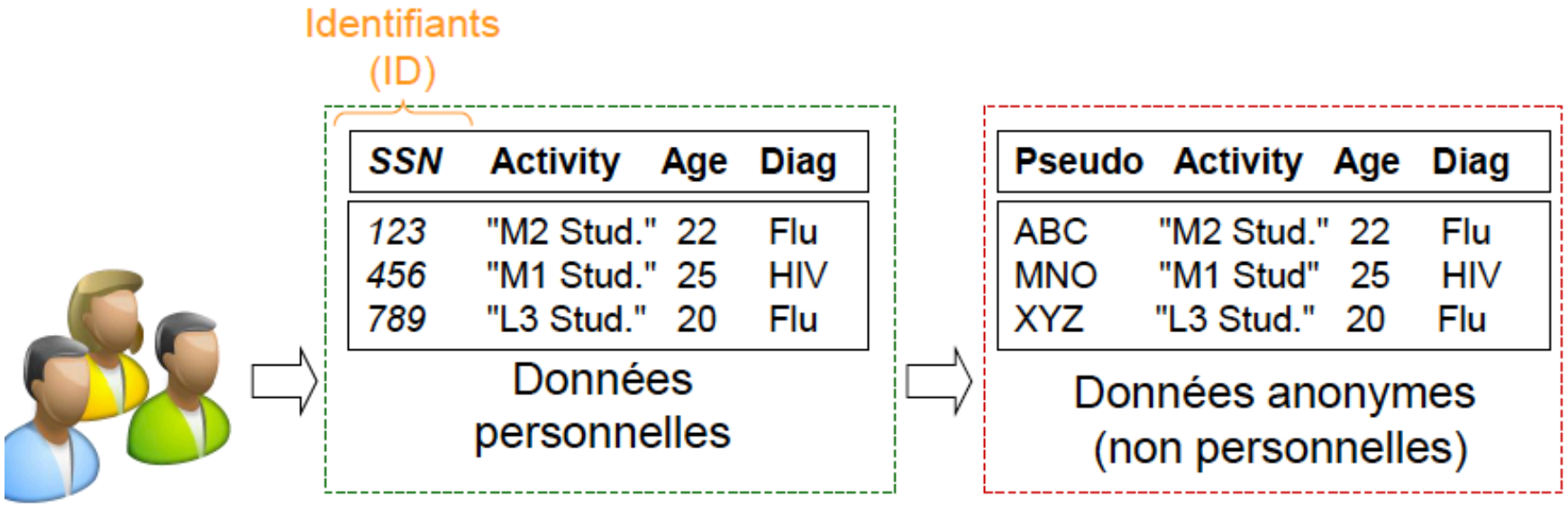


Approches d'anonymisation

- Pseudo-anonymisation
- k-anonymat
- L-diversité

Le pseudonymat

- Base du pseudonymat
  - Retirer les identifiants et les remplacer par un pseudo



…ne garantie pas l'anonymat !

Etude de L. Sweeney – 2002 (1)

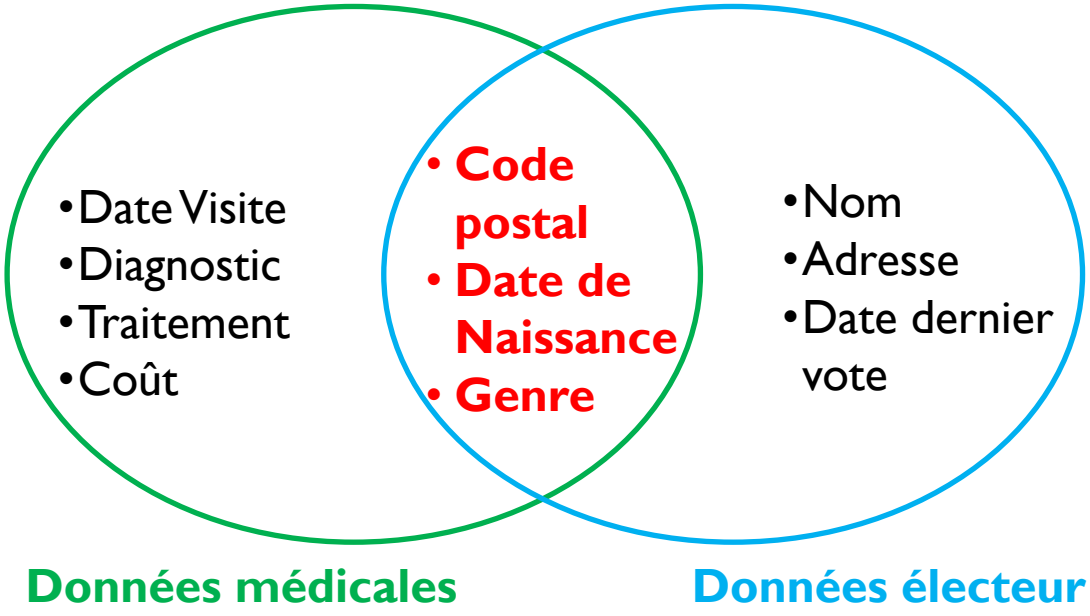
- Un fichier anomyme produit par une compagnie d'assurance
  - Sans d'identifiant (ni nom, ni numéros de sécu, etc.)
  - Avec des données sensibles (traitement médical, diagnostique, etc.)
  - Et d'autres non sensibles (code postal, genre, etc.)
- Un fichier nominatif (liste de grands électeurs)
  - Des identifiants (nom, adresse, parti politiques, etc.)
  - Des champs non sensibles (code postal, genre, etc.)
- Ces deux fichiers étant publics…
  - L'identité de certaines personnes ne peut pas être préservée
  - Sweeney retrouve facilement le dossier médical du gouverneur W. Weld

Comment dé-anonymiser les données ?

Quelle est la proportion de personnes qui reste protégée?

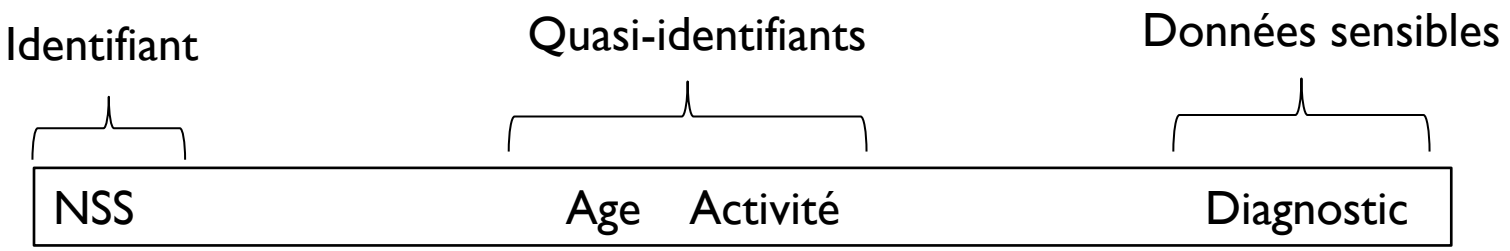
Pseudo anonymisation n'est pas suffisante

- Jointure des deux fichiers sur les données non sensibles



- Aux Etats-Unis, 87% de la population peut être identifiée en basant sur le Code postal, le date de naissance, et le genre

Classification des données



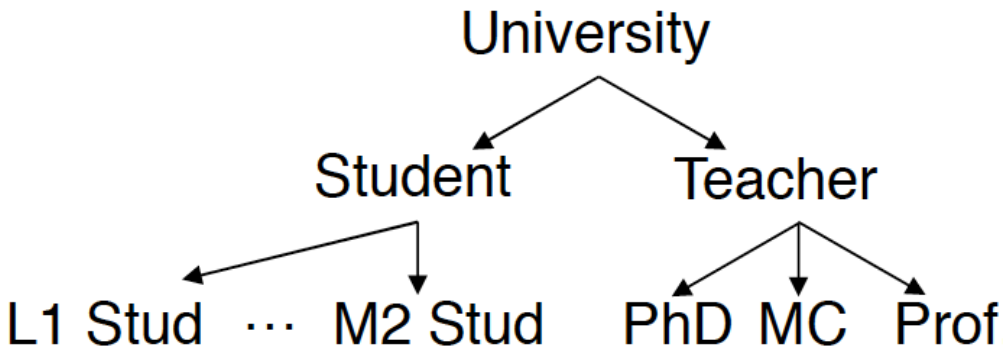
- Pour chaque tuple (enregistrement) :
  - Les identifiants doivent être supprimés
  - Les liens entre les quasi-identifiants et les données sensibles doivent être masqués

k-anonymat : intuition

- Créer des groupes de k individus
- Associer à chaque groupe un ensemble de valeurs sensibles
- Un groupe peut être vu comme une classe d'équivalence
- Méthodes
  - Généralisation
  - Suppression

K-anonymat par généralisation

- Principe : remplacer les quasi-identifiants par des valeurs plus générales suivant une certaine hiérarchie
- Exemple : hiérarchie pour l'attribut activité

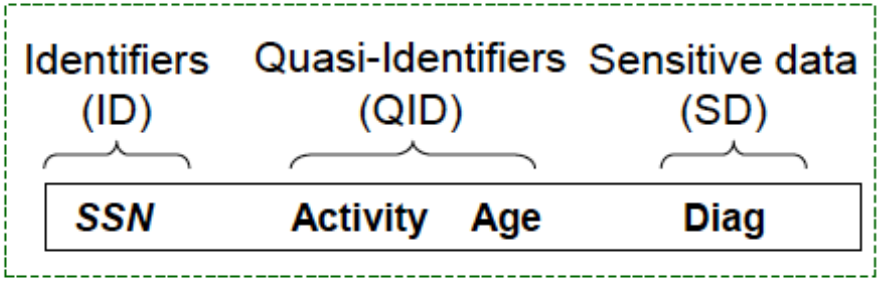


- Les données numériques sont généralisées par des intervalles

k-anonymat

- Le k-anomynat répond au problème du pseudonymat
- Base du k-anonymat

  1) classifier les données



  2) retirer les identifiants (comme pour le pseudonymat)

  3) supprimer et/ou généraliser les quasi-identifiants **(restent vrais!)** …

  … de manière à former des classes d'individus équivalents de taille k

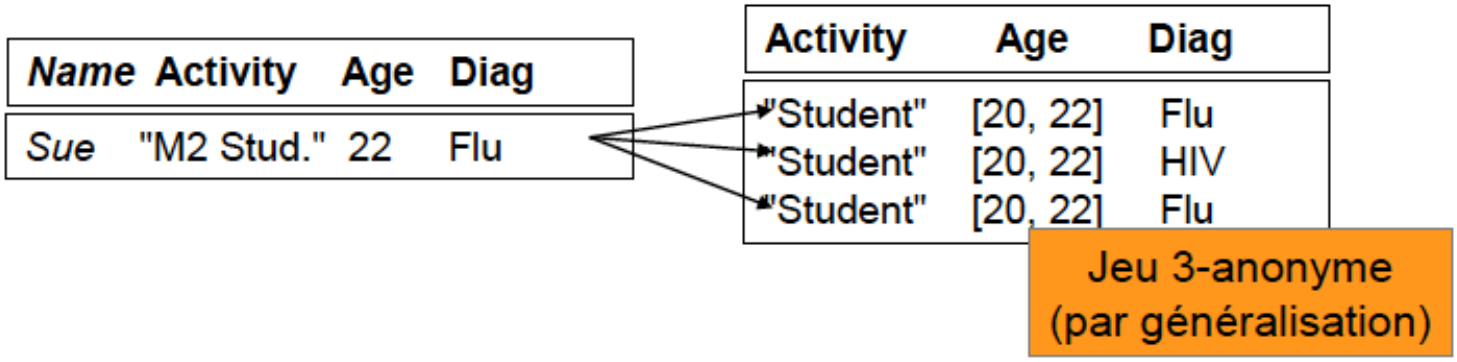
Name	Activity	Age	Diag
Sue	"M2 Stud."	22	Flu
Pat	"MC"	27	Cancer
Dan	"PhD"	26	Cancer
Bob	"M1 Stud."	21	HIV
Bill	"L3 Stud."	20	Flu
San	"PhD"	24	Cancer

Activity	Age	Diag
"Student"	[20, 22]	Flu
"Student"	[20, 22]	HIV
"Student"	[20, 22]	Flu
"Teacher"	[24, 27]	Cancer
"Teacher"	[24, 27]	Cancer
"Teacher"	[24, 27]	Cancer

Jeu 3-anonymes (par généralisation)

Le *k*-anonymat garantit que…

- Un individu donné est toujours associé à au moins *k* individus participants au jeu anonyme
  - C'est-à-dire à tous ceux appartenant à une même classe
  - Par exemple: « Sue » est associée à au moins 3 tuples du jeu 3-anonyme



… mais ne garantit pas tout

- Il n'y a pas de contrôle sur les valeurs des attributs sensibles associées dans une même classe de taille *k*
  - On peut donc avoir moins de *k* valeurs sensibles par classe
  - Voire même une seule valeur sensible !
- Exemple:
  - L'individu « Pat » est bien relié à une classe de taille 3…

    … mais tous les individus de cette classe ont le même *Diag* !

Name	Activity	Age	Diag
Pat	"MC"	27	Cancer

Activity	Age	Diag
"Teacher"	[24, 27]	Cancer
"Teacher"	[24, 27]	Cancer
"Teacher"	[24, 27]	Cancer

➔ Le *k*-anonymat ne protège pas contre la dé-anonymisation des attributs sensibles !

La l-Diversité

Complète le *k*-anonymat

– Afin d'éviter la dé-anonymisation des attributs sensibles

Assure que chaque classe contient au moins ℓ valeurs sensibles différentes et « représentatives »

– « représentatives » peut signifier différentes choses

Name	Activity	Age	Diag
Sue	"M2 Stud."	22	Flu
Pat	"MC"	27	Cancer
Dan	"PhD"	26	Cancer
Bob	"M1 Stud."	21	HIV
Bill	"L3 Stud."	20	Flu
San	"PhD"	24	Cancer
John	"M2 Stud"	22	Cold
Jim	"M2 Stud"	23	Cancer

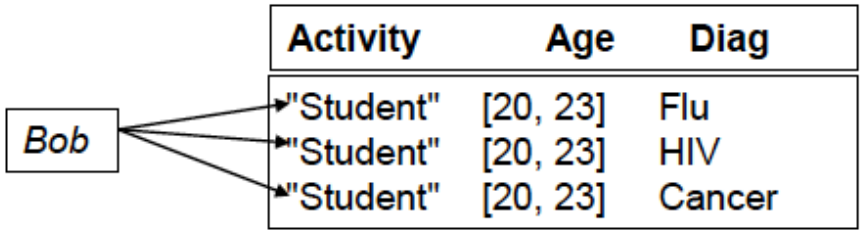
Jeu 3-anonyme et 3-divers

Activity	Age	Diag
"Student"	[20, 23]	Flu
"Student"	[20, 23]	HIV
"Student"	[20, 23]	Cancer
"University"	[22, 24]	Flu
"University"	[22, 24]	Cold
"University"	[22, 24]	Cancer

La l-Diversité garantie que…

Un individu donné est toujours associé à
au moins ℓ valeurs d'attributs sensibles différentes
parmi les plus représentatives

- Par exemple: l'individu « Bob » est bien associé à trois valeurs sensibles représentatives {Flu, HIV, Cancer}



- Mais elle ne garantit pas que la classe contienne des valeurs sensibles reflétant la distribution globale
- Donc: possibilité de discriminer un individu s'il appartient à une classe « à risque »…

Anonymisation et utilité

- Plus les données sont rendues anonymes plus leurs utilité diminue
- Equilibre difficile entre respect de la vie et privée et utilité des données publiées
  - Définition d'une métrique pour l'utilité
- Autre problème : gestion des mises à jour
  - Propagation des mises à jour aux données rendues anonymes