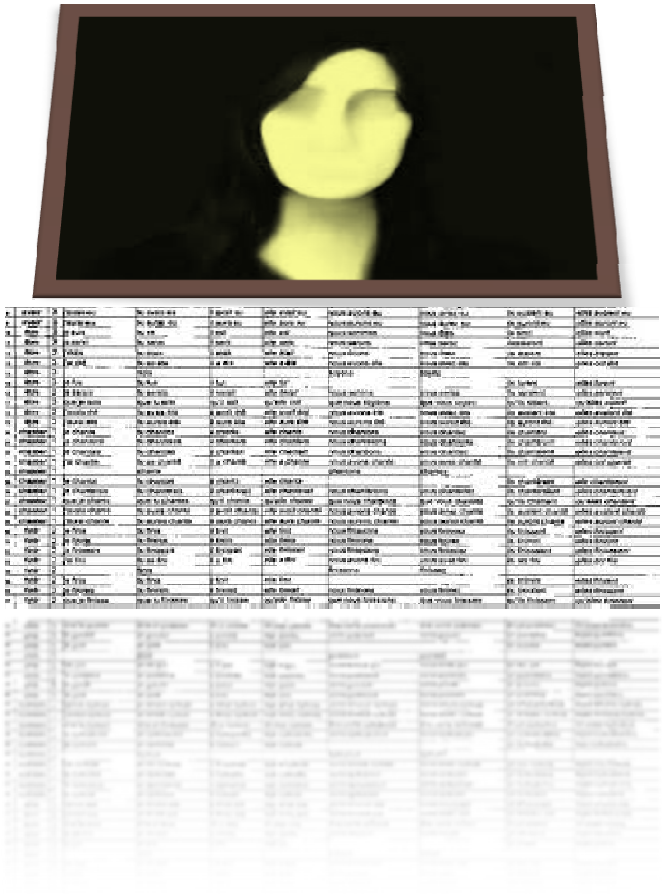


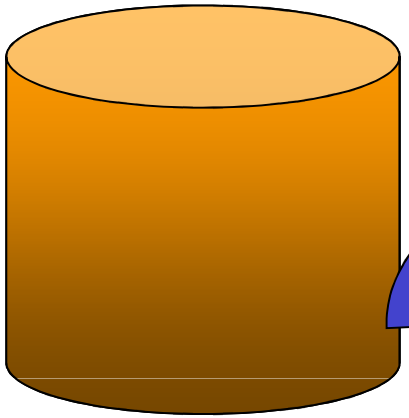
Anonymisation : de la théorie à la pratique

Nora Cuppens-Boulahia et Frédéric Cuppens
Institut Mines-Télécom / Télécom Bretagne
CNRS Lab-STICC

Sommaire

- Objectifs
- Evaluation, quantification du risque
- Techniques d’anonymisation par floutage
- Limites des techniques
- Conclusion

Objectifs



Base de
données sensibles

Transformation

Open
data

La transformation doit garantir :

- ❑ **Les données ne sont plus sensibles**
- ❑ **L'utilité des données est préservée**

Objectifs

	Non sensible		Sensible	
	Code postal	Age	Nationalité	Etat
1	13053	28	Russe	Maladie cardiovasculaire
2	13068	29	Américain	Maladie cardiovasculaire
3				
4				
5				
6				
7				
8				
9				
10	13053	37	Indien	
11	13068	36	Japonais	Cancer
12	13068	35	Américain	Cancer

- **Prévenir les risques**

  Risque de réidentification

  Risque de divulgation d'attributs

Transformation

L'utilisation des données sous forme d'Open Data doit permettre la réalisation de scénarios tels que

1. Analyse statistique par tranche d'âge quinquennale

2. Analyse statistique sur la fréquentation des hôpitaux

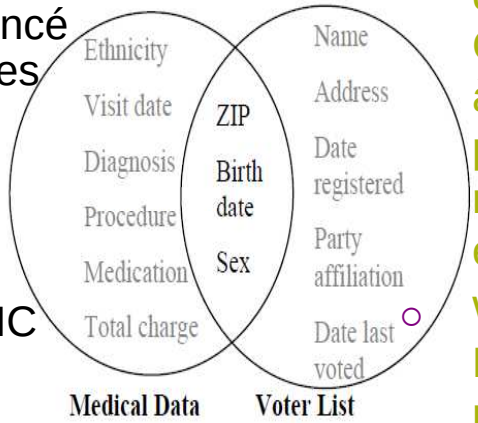
Le risque de réidentification :

au sens remonter jusqu'à l'identité de quelqu'un et non la valeur du dommage engendré

Déjà en 1996…
Le gouverneur William Weld

- L'Association nationale des organismes de données sur la santé (NAHDO) : 37 États américains ont des mandats législatifs pour recueillir des données au niveau des hôpitaux,17 états ont commencé à recueillir des données sur les soins ambulatoires dans les hôpitaux, les cabinets de médecins, les cliniques, etc.

- Dans le Massachusetts, la GIC est responsable de l'achat d'assurances santé pour les employés de l'état. GIC a recueilli des données spécifiques aux patients avec près d'une centaine d'attributs pour environ 135 000 employés de l'État et de leurs familles



- Comme les données ont été considérées comme étant anonymes, le GIC a donné une copie des données aux chercheurs et vendu une copie à l'industrie. Pour 20$, LATANYA SWEENEY achète la liste des inscriptions aux listes électorales. Ces informations peuvent être corrélées à celles médicales en utilisant le code postal, date de naissance et le sexe, reliant ainsi le diagnostic, les procédures et notamment des médicaments
- William Weld était gouverneur du Massachusetts à l'époque et ses dossiers médicaux étaient dans les données GIC. Il vivait à Cambridge Massachusetts
- Dans la liste des électeurs de Cambridge, 6 personnes avaient la même date de naissance que lui, mais seulement 3 d'entre eux étaient des hommes, et il était le seul dans son code postal à 5 chiffres

Métrique de risque

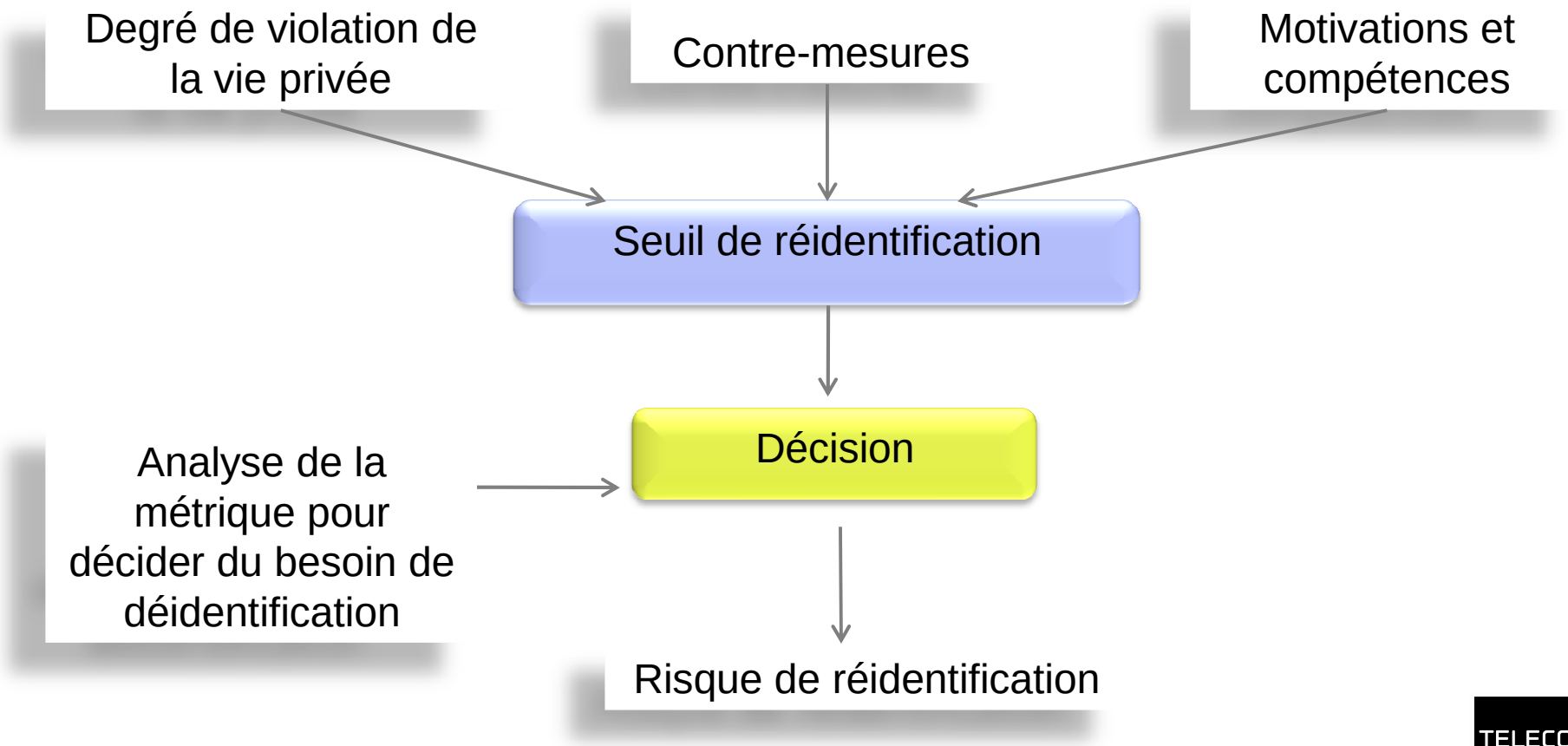
- **Connaissance de l'utilisateur**
  - Aline a été hospitalisée
  - Présente dans la base rendue publique des prescriptions
  - Elle est née en 1987
- **Procédure de réidentification**
  - Trouver les patientes nées en 1987
  - Généralisation $\Rightarrow$ Trouver les patientes nées en 1980 et 1989
  - Si f enregistrements correspondent à la requête alors probabilité =1/f=0,5
- **Les enregistrements ayant la même combinaison de valeurs pour les attributs choisis constitue une classe d'équivalence**
  - Elle est égale à 2 pour Aline
  - Elle est égale à 3 pour Joan

	Genre	Date de naissance	Id- Med	Med
Joan	Mâle	1970 -1979	2046059	-6059
Alain	Mâle	1980 -1989	716839	6839
Hen	Mâle	1970 -1979	2241497	-1497
Gina	Femelle	1990 -1999	2046059	-6059
Mari	Femelle	1980 -1989	392537	2537
Willi	Mâle	1990 -1999	363766	3766
Rob	Mâle	1990 -1999	544981	4981
Aline	Femelle	1980 -1989	293512	3512
Dam	Mâle	1970 -1979	544981	4981
Fréd	Mâle	1990 -1999	596612	6612
Eme	Mâle	1980 -1989	725765	5765

Base de réidentifiée Publique ***Base originale***

Evaluation du risque de réidentification : Vue globale

- **Affectation d'une probabilité d'une réidentification réussie à chacun des enregistrements**



Evaluation du risque de ré-identification

$$R_a = \frac{1}{n} \sum_{j \in J} f_j \times I(\theta_j > \tau)$$

Estimation normalisée du risque de ré-identification

Nombre d'enregistrements à diffuser

Taille de la classe d'équivalence j (les enregistrements partageant la même propriété)

=1 si la probabilité de ré-identification est supérieure au seuil τ.
= 0 sinon

τ est la probabilité maximale autorisée de ré-identification d'un enregistrement

La valeur de la probabilité de ré-identification dans le cas du persécuteur :

$$_p\theta_j = 1/f_j$$

Si la valeur de la ré-identification est supérieure à un certain seuil, le risque de ré-identification sera considéré comme élevé.

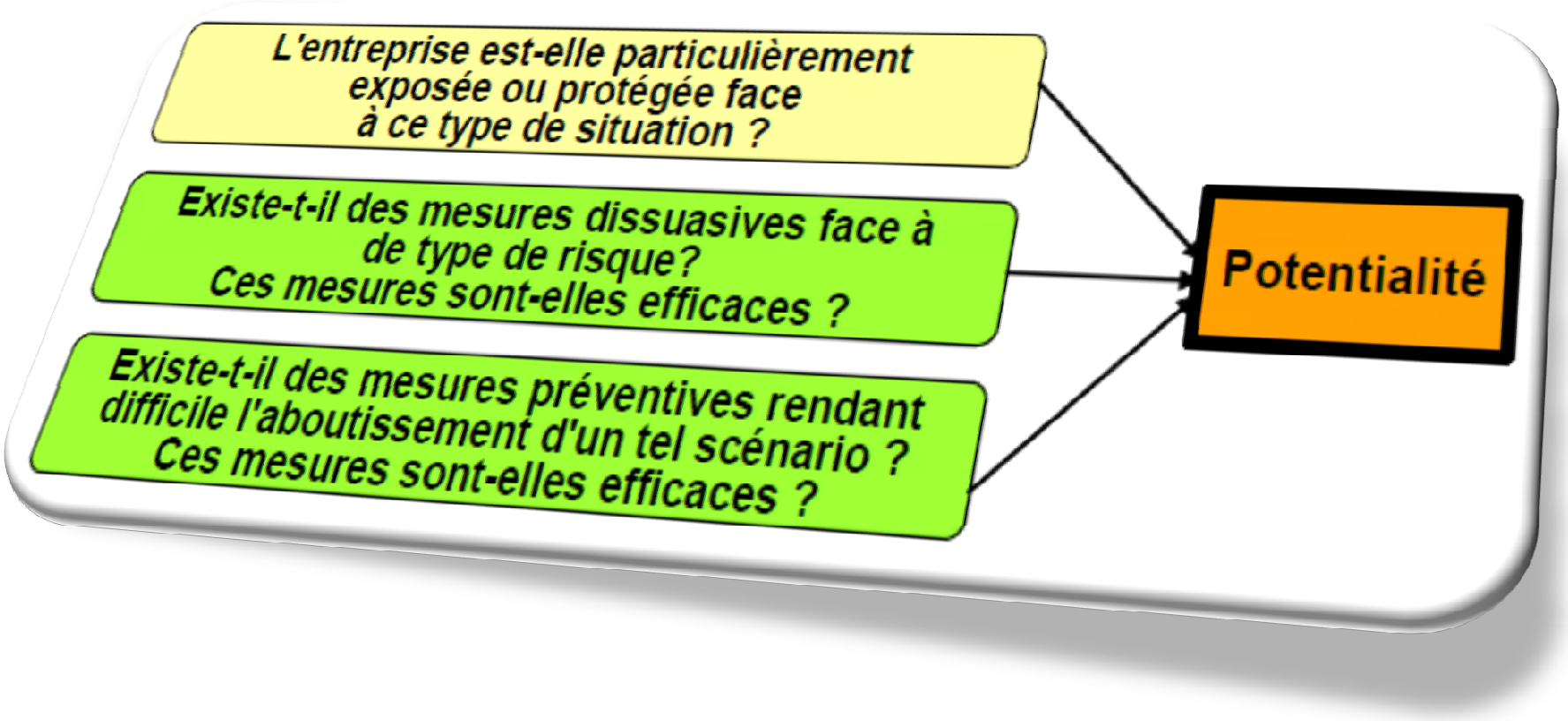
Analyse d'une situation de risque et quantification du risque

- **Potentialité du risque**
  - Probabilité d'occurrence
  - Fonction des mesures de sécurité mises en place

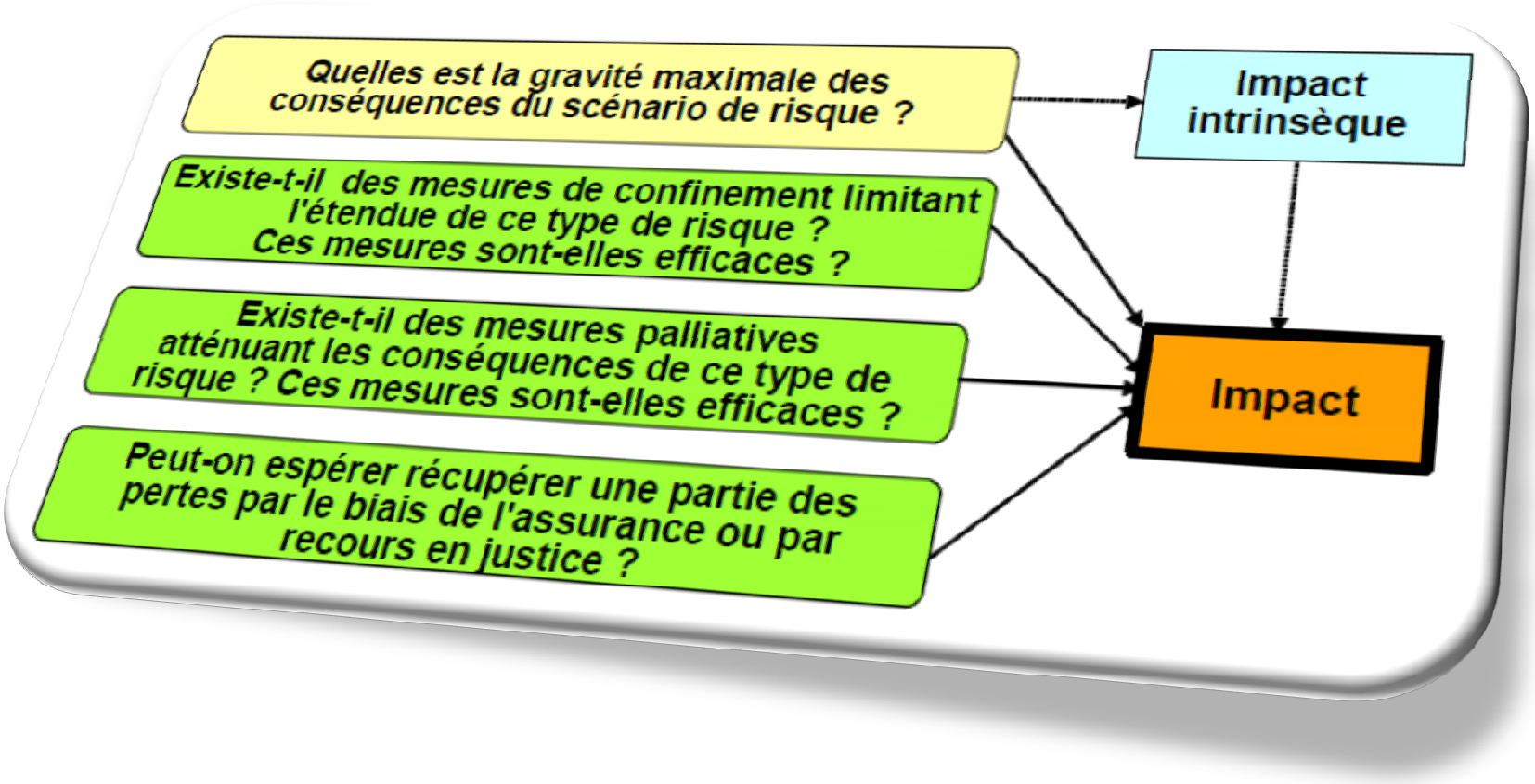
- **Impact du risque**
  - Gravité des conséquences directes et indirectes qui découleraient de l'occurrence du risque
  - Fonction de l'impact maximum ou intrinsèque, défini lors de l'analyse des enjeux, et des mesures de sécurité adaptées

- **Risque = Potentialité $\times$ Impact**

Potentialité du risque



Impact du risque



Grille d'aversion du risque pour la quantification

- **Utilisation d'une échelle de valeur de dysfonctionnement**

I = 4	G = 2	G = 3	G = 4	G = 4
I = 3	G = 2	G = 3	G = 3	G = 4
I = 2	G = 1	G = 2	G = 2	G = 3
I = 1	G = 1	G = 1	G = 1	G = 2
	P = 1	P = 2	P = 3	P = 4

Démarche

- **Transformation par application des techniques de floutage**
  - K-anonymity
  - L-diversity

- **5 étapes**
  - Classification des attributs
  - Définition des nomenclatures
  - Choix des valeurs de K et L
  - Génération de l'Open Data
  - Traitement des cas rares

Technique de protection par floutage
Classification des attributs

- **Identifiant**
  - Attribut (ou ensemble d'attributs) qui identifie un enregistrement de façon unique
  - Correspond à la clé
  - On suppose que les identifiants ont été anonymisés (chiffrement)
- **Quasi identifiant**
  - Attribut (ou ensemble d'attributs) qui peut être utilisé pour identifier un enregistrement avec une forte probabilité
  - Attributs pour lesquels un attaquant peut facilement obtenir l'accès
  - Exemple: <Age, Sexe, Adresse>
- **Sensible**
  - Attributs que l'on veut conserver secret
- **Autre**
  - Les autres attributs
  - Ils ne sont pas sensibles mais l'on suppose que l'attaquant ne peut pas facilement y avoir accès

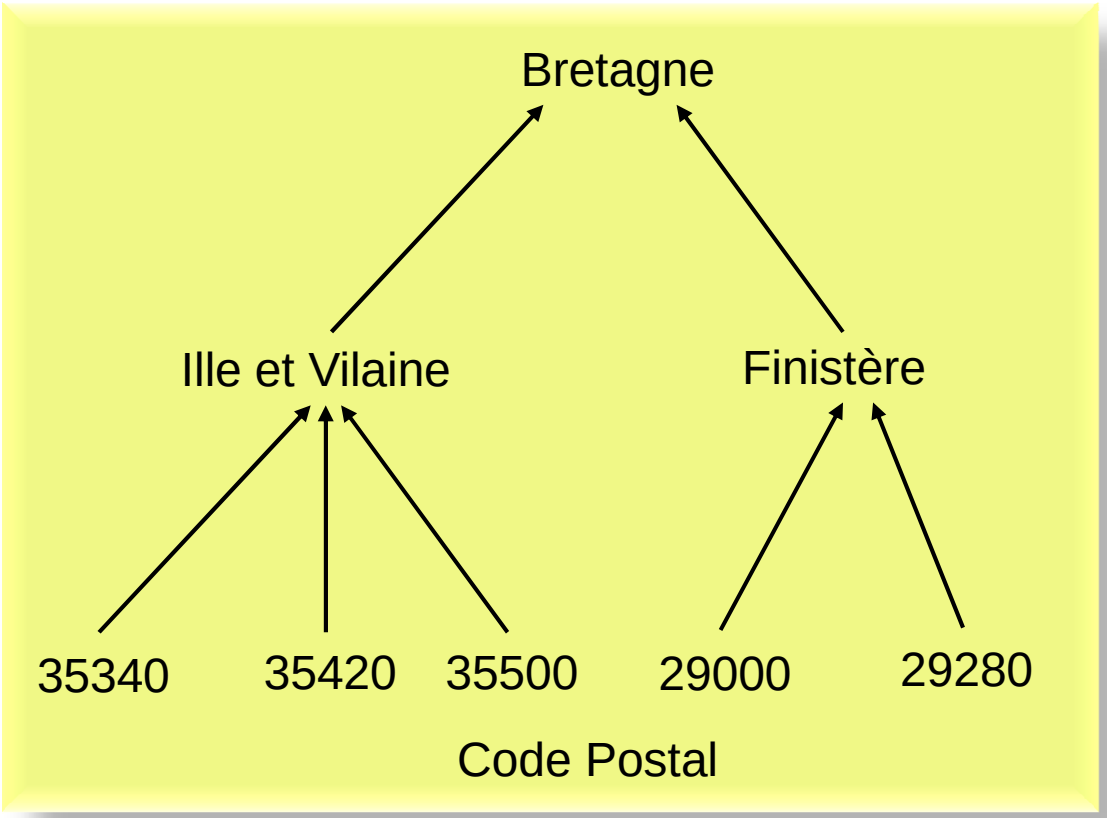
Technique de protection par floutage : La K-anonymity

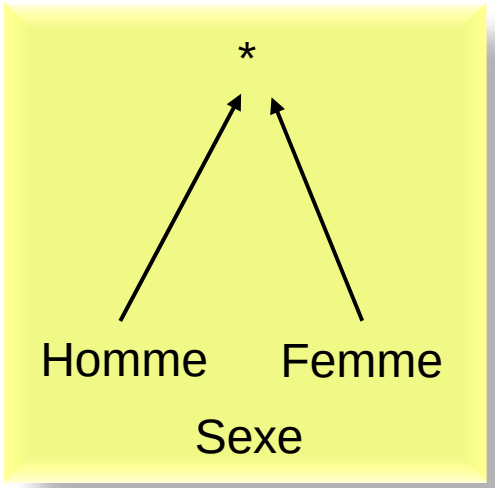
- **Une base est k-anonymisée si l'information concernant chaque individu contenu dans la base ne peut pas être distinguée d'au moins k-1 autres individus qui apparaissent également dans la base**

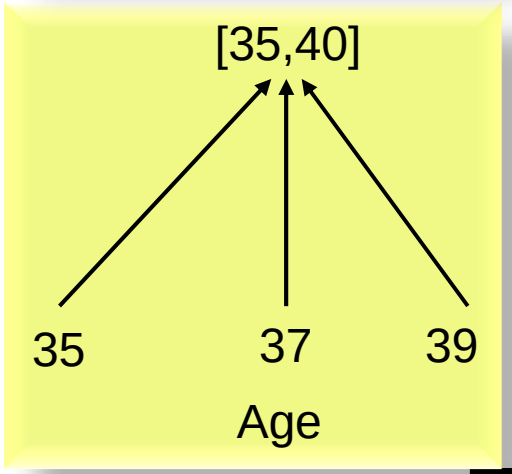
- **Chaque quasi-identifiant doit apparaître dans au moins k enregistrements**
  - Exemple : < Age, Sexe, Code postal >

- **Principe**
  - Algorithme de généralisation
  - Remplacer chaque quasi-identifiant par des valeurs moins spécifiques jusqu'à obtenir un groupe de k valeurs identiques
  - Plusieurs algorithmes ont été définis

Technique de protection par floutage
Définition des nomenclatures







Technique de protection par floutage
La K-anonymity

Pseudonyme	Identifiants indirects			Variable sensible non identifiante
	Age	*Sexe*	*Code postal*	*Maladie*
290388	76	Femme	42300	Cirrhose
276209	86	Femme	73270	Bronchite
251057	68	Femme	73270	Hépatite C
186704	111	Femme	73270	Hépatite C
219687	17	Homme	75014	Insuffisance cardiaque
223818	31	Homme	75014	bronchite
182604	38	Homme	93120	Grippe
183501	42	Homme	75012	Diabète
175545	55	Homme	75016	Diabète
205972	47	Homme	91000	Diabète

K = 3

Pseudonyme	*âge*	*Sexe*	*Région*	*Maladie*
290388	> 60	Femme	Rhône-Alpes	Cirrhose
276209	> 60	F*emme*	Rhône-Alpes	Bronchite
251057	> 60	Femme	Rhône-Alpes	Hépatite C
186704	> 60	Femme	Rhône-Alpes	Hépatite C
219687	< 40	Homme	Ile-de-France	Insuffisance cardiaque
223818	< 40	Homme	Ile-de-France	bronchite
182604	< 40	Homme	Ile-de-France	Grippe
183501	[40,60]	Homme	Ile-de-France	Diabète
175545	[40,60]	Homme	Ile-de-France	Diabète
205972	[40,60]	Homme	Ile-de-France	Diabète

- **La k-anonymity n'est pas suffisante pour assurer la confidentialité si un attribut dans un groupe n'est pas correctement diversifié**

Technique de protection par floutage
La L-Diversity

- **Principe**
  - Garantir que les données sensibles dans chaque groupe de quasi-identifieur sont diversifiées
- **Plusieurs variantes**
  - Distinct L-Diversity
    - La plus simple
    - Mais possibilité d'attaques fréquentielles
  - Probabilistic L-Diversity
  - Entropy L-Diversity
  - T-Closeness
    - La distribution des attributs sensibles dans chaque groupe doit être proche de la distribution dans la base de données globale
- **Remarque**
  - La diversification est seulement possible si l'on dispose de suffisamment d'enregistrements dans la base !

Technique de protection par floutage
La L-Diversity

K = 3, L = 1

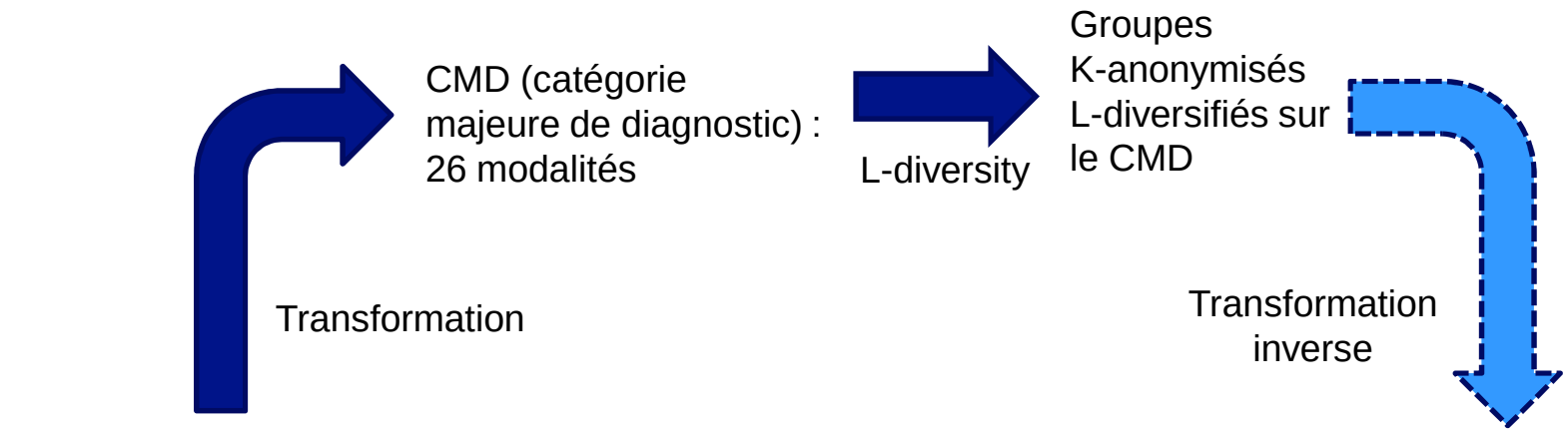
Pseudonyme	Tranche d'âge	Sexe	Région	Maladie
290388	> 60	Femme	Rhône-Alpes	Cirrhose
276209	> 60	Femme	Rhône-Alpes	Bronchite
251057	> 60	Femme	Rhône-Alpes	Hépatite C
186704	> 60	Femme	Rhône-Alpes	Hépatite C
219687	< 40	Homme	Ile-de-France	Insuffisance cardiaque
223818	< 40	Homme	Ile-de-France	bronchite
182604	< 40	Homme	Ile-de-France	Grippe
183501	[40,60]	Homme	Ile-de-France	Diabète
175545	[40,60]	Homme	Ile-de-France	Diabète
205972	[40,60]	Homme	Ile-de-France	Diabète

K = 4, L = 3

Pseudonyme	Tranche d'âge	Sexe	Région	Maladie
290388	> 60	Femme	Rhône-Alpes	Cirrhose
276209	> 60	Femme	Rhône-Alpes	Bronchite
251057	> 60	Femme	Rhône-Alpes	Hépatite C
186704	> 60	Femme	Rhône-Alpes	Hépatite C
219687	< 60	Homme	Ile-de-France	Insuffisance cardiaque
223818	< 60	Homme	Ile-de-France	bronchite
182604	< 60	Homme	Ile-de-France	Grippe
183501	< 60	Homme	Ile-de-France	Diabète
175545	< 60	Homme	Ile-de-France	Diabète
205972	< 60	Homme	Ile-de-France	Diabète

Technique de protection par floutage
La L-Diversity

- **Application au PMSI**



Analyse des cas rares

- **Valeurs rares de quasi identifiants**
  - Peu fréquent dans les grosses bases de données
  - Exemple : âge > 110
  - A prendre en compte dans la nomenclature

- **Associations rares de valeurs de quasi identifiants**
  - Correspondent aux cas les plus fréquents
  - Exemple : patient qui se fait soigner dans un établissement éloigné de son domicile

Gestion des cas rares

- **Dans la pratique, un faible pourcentage de cas rares suffit à remettre en cause le niveau d'anonymisation**
  - Dégradation très significative de la fonction d'utilité

Nom de la variable	Nature de la variable	*Niveau*	*Nombre de Modalités*
Sexe	Quasi-identifiant	Niveau 0	2
Âge	Quasi-identifiant	Niveau 1	19
Lieu de résidence	Quasi-identifiant	Niveau 1	99
Numéro Finess	Quasi-identifiant	Niveau 4	1
Durée d'hospitalisation	Quasi-identifiant	Niveau 4	1
Nombre de clés d'indentification		**3762**	
CMD, catégorie majeure de diagnostic	Donnée sensible	En clair	26

Anonymisation de l'ensemble de la base PMSI (K = 10, L = 3)

ARX

Anonymisation de la base PMSI après élimination de 3% des cas les plus rares (K = 10, L = 3)

ARX

Nom de la variable	Nature de la variable	*Niveau*	*Nombre de Modalités*
Sexe	Quasi-identifiant	Niveau 0	2
Âge	Quasi-identifiant	Niveau 1	19
Lieu de résidence	Quasi-identifiant	Niveau 1	99
Numéro Finess	Quasi-identifiant	Niveau 2	23 (22 régions+ les DOM regroupés)
Durée d'hospitalisation	Quasi-identifiant	Niveau 1	12
Nombre de clés d'indentification		**1038312**	
CMD, catégorie majeure de diagnostic	Donnée sensible	En clair	26

Gestion des cas rares

Solutions envisageables

- Suppression
  - Enregistrement, tout ou partie des occurrences d'une valeur donnée dans la table, cellule
  - Résultats biaisés dans certains types de traitement
- Bruitage
  - Permutation
    - Dissociation des quasi-identifiants des attributs sensibles
    - Données non modifiées
    - Résultats plus précis en comparaison à la généralisation
  - Perturbation
    - Ajout de bruit aléatoire
    - Calculs de moyennes et de corrélations sont préservés
    - Protection faible lorsque la corrélation entre les attributs est forte
- Insertion de données synthétiques
- Floutage avec plusieurs niveaux d'anonymisation

Conclusion

- **Faut-il une réglementation des paramètres d'anonymisation ?**
  - Valeur de K, valeur de L
- **Le consentement – les préférences des individus concernés par les données, est-ce une solution pour fix ces paramètres ?**
  - Problème de coût et d'efficacité
- **Passage à l'échelle**
  - Absence d'implémentation d'algorithmes d'anonymisation pour les trop gros volumes de données
  - Besoin d'adapter les algorithmes pour gérer du big data
- **Validation des données en sortie d'un processus d'anonymisation**
  - Niveau de protection des données personnelles, niveau d'utilisabilité des données obtenues
  - Vers des Centres d'Anonymisation des Données