



NiPN

National Information
Platforms for Nutrition



Anonymisation des données

théorie et pratique

D.Blum - SOFRECO



NiPN

National Information
Platforms for Nutrition

Niamey - 28 au 31 mai 2019



Première journée

- Anonymisation, pseudonymisation
- Catégories de données d'une base de microdata
 - Identifiants nominatifs, données « sensibles », quasi-identifiants, autres données
- Ré-identification au sein d'une base de données pseudonymisée
- Mesures du risque de ré-identification au sein d'une base de données
 - estimation empirique
 - k-anonymat, l-diversité
 - autres méthodes



Quelques termes

- **Micro-données (microdata en anglais)**
 - base de données comportant des informations personnelles de granularité individuelle
- **Données agrégées**
 - tableau ou base de données comportant des informations relatives à des groupes d'au moins deux personnes
- **Variables floutées**
 - variables dont certaines modalités ont été constituées en plages de telle sorte que leur précision est réduite





NiPN
National Information
Platforms for Nutrition

**Anonymisation,
pseudonymisation**



Anonymisation, *stricto sensu*

- Faire disparaître les informations nominatives

Prénom	NOM	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
Mahamadou	ISSOUFOU	1952	Dandadji	Illéla	Tahoua	marié polygame	5	...
Brigi	RAFINI	07/04/1953	Iférouane	Arlit	Agadez	marié monogame	5	...
Guillaume	POIREL	...	...	...	France	marié monogame	3	...
...	...	...	...	...	...	...	...	...



Anonymisation, *stricto sensu*

- Solution n°1 : disparition pure et simple du prénom et du nom

date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
1952	Dandadji	Illéla	Tahoua	marié polygame	5	...
07/04/1953	Iférouane	Arlit	Agadez	marié monogame	5	...
...	...	...	France	marié monogame	3	...
...	...	...	...	...	...	...



Anonymisation, *stricto sensu*

- Solution n° 2 : pseudonymisation

« étiquette »	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
étiquette A	1952	Dandadji	Illéla	Tahoua	marié	5	...
étiquette B	07/04/1953	Iférouane	Arlit	Agadez	marié	5	...
étiquette C	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...



Pseudonymisation : principe

- Remplacer les informations nominatives par une « étiquette à la fois unique et non identifiante » appelée pseudonyme
 - unique : deux enregistrements distincts doivent comporter deux étiquettes distinctes
 - non identifiante : le contenu de l'étiquette ne doit pas fournir de renseignement directement ou indirectement nominatif
- Conserver **(ou non)** séparément la correspondance entre l'étiquette et les informations nominatives qu'elle remplace
 - à décider selon l'objectif (cf. diapos suivantes)
 - s'inspirer de l'organisation de l'ABS (Australian bureau of statistics)



Pseudonymes : table de correspondance

- Données d'origine non anonymes

Prénom	NOM	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
Mahamadou	ISSOUFOU	1952	Dandadji	Illéla	Tahoua	marié	5	...
Brigi	RAFINI	07/04/1953	Iférouane	Arlit	Agadez	marié	5	...
Guillaume	POIREL	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...



Pseudonymes : table de correspondance

- Données d'origine non anonymes

Prénom	NOM	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
Mahamadou	ISSOUFOU	1952	Dandadji	Illéla	Tahoua	marié	5	...
Brigi	RAFINI	07/04/1953	Iférouane	Arlit	Agadez	marié	5	...
Guillaume	POIREL	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...



Pseudonymes : table de correspondance

- Données d'origine non anonymes

Prénom	NOM	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
Mahamadou	ISSOUFOU	1952	Dandadji	Illéla	Tahoua	marié	5	...
Brigi	RAFINI	07/04/1953	Iférouane	Arlit	Agadez	marié	5	...
Guillaume	POIREL	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...



Pseudonymes : table de correspondance

- Données d'origine non anonymes

Prénom	NOM	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
Mahamadou	ISSOUFOU	1952	Dandadji	Illéla	Tahoua	marié polygame	5	...
Brigi	RAFINI	07/04/1953	Iférouane	Arlit	Agadez	marié monogame	5	...
Guillaume	POIREL	...	...	...	France	marié monogame	3	...
...	...	...	...	...	...	...	...	...

► Table de correspondance noms / pseudonymes + base de données pseudonymisées

Prénom	NOM	étiquette
Mahamadou	ISSOUFOU	étiquette A
Brigi	RAFINI	étiquette B
Guillaume	POIREL	étiquette C
...	...	...

étiquette	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
étiquette A	1952	Dandadji	Illéla	Tahoua	marié polygame	5	...
étiquette B	07/04/1953	Iférouane	Arlit	Agadez	marié monogame	5	...
étiquette C	...	...	...	France	marié monogame	3	...
...	...	...	...	...	...	...	...



Pseudonymisation : objectifs

- **soit pour permettre de « remonter » au recueil d'origine si besoin**
 - ⇒ conserver la correspondance « nom / pseudonyme » : **indispensable**
 - ⇒ usage **partagé** (producteur des données / service statistique)
- **soit pour faciliter la manipulation des enregistrements**
 - ⇒ conserver la correspondance : **totalelement inutile** (et même contre-productif)
 - ⇒ usage **interne** au service statistique



Pseudonymes à éviter, voire proscrire

- **Initiales du prénom et du nom**
 - MISSO-051, BRAFI-112, GPOIR-045, ...
- **Code identifiant la région, la campagne de collecte, etc.**
 - AGA2016-25087, MAR2017-11297, NIA2015-2564032, ...
- **Attribution de numéros séquentiels**

Ils peuvent révéler une information fondée sur un ordre naturel ou évident qui permet de « cibler » l'individu concerné :

 - tri alphabétique des noms
 - classement des villes et villages, départements, régions, etc.
 - déroulement chronologique de la collecte

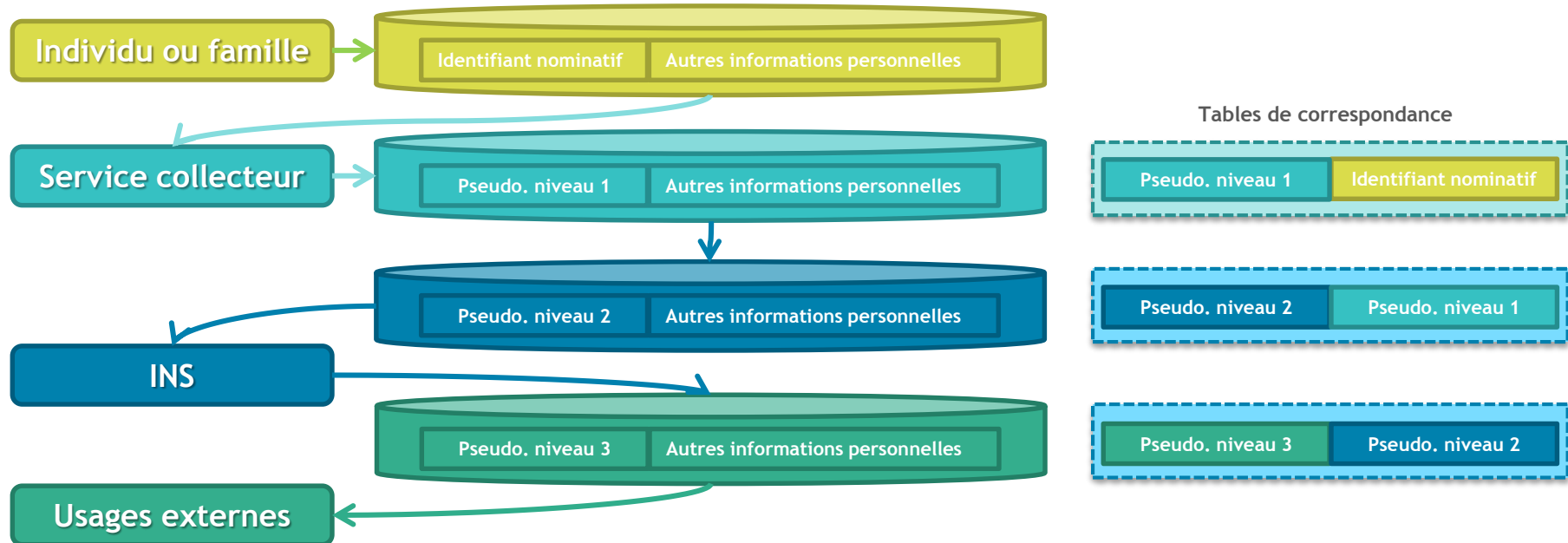


Comment pseudonymiser ?

- **Trois méthodes distinctes**
 - Numérotation séquentielle
 - Numérotation aléatoire
 - « Hachage cryptographique »
- **Peuvent être mises en œuvre par le service collecteur des données**
- **Peuvent être combinées en cascade : étapes de pseudonymisation successives**
 - pseudonymisation initiale par le collecteur des données
 - sur-pseudonymisation lors de la centralisation auprès du service statistique
 - sur-pseudonymisation lors de la diffusion d'une extraction de la base de données



Pseudonymisation « en cascade »



Pseudonymisation : méthode n° 1

- **Numéro séquentiel**

- uniquement si les données ne sont pas fournies selon un tri préalable fondé sur un ordre naturel ou évident (ordre alphabétique des noms, classement habituel des villes et villages du département, des départements de la région, des régions du Niger, ordre chronologique de la collecte des données, etc.)
- peut être mis en œuvre par le service collecteur des données

Prénom	NOM	pseudonyme
Mahamadou	ISSOUFOU	00 000 001
Brigi	RAFINI	00 000 002
Guillaume	POIREL	00 000 003
...	...	...

pseudonyme	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
00 000 001	1952	Dandadji	Illéla	Tahoua	marié polygame	5	...
00 000 002	07/04/1953	Iférouane	Arlit	Agadez	marié monogame	5	...
00 000 003	...	...	...	France	marié monogame	3	...
...	...	...	...	...	...	...	...



Pseudonymisation : méthode n° 2

- **Numéro aléatoire**

- assez simple à développer et mettre en œuvre
- peut être mis à la disposition du service collecteur des données

Prénom	NOM	pseudonyme
Mahamadou	ISSOUFOU	870 145 032
Brigi	RAFINI	602 237 751
Guillaume	POIREL	537 501 223
...	...	...

pseudonyme	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
870 145 032	1952	Dandadji	Illéla	Tahoua	marié polygame	5	...
602 237 751	07/04/1953	Iférouane	Arlit	Agadez	marié monogame	5	...
537 501 223	...	...	...	France	marié monogame	3	...
...	...	...	...	...	...	...	...



Pseudonymisation : méthode n° 3 (1/2)

- « **Hachage cryptographique** »
 - indispensable uniquement dans le cas où le « chaînage » est nécessaire
 - chaînage chronologique : collectes itératives concernant les mêmes personnes
 - chaînage géographique : collectes concernant des personnes nomades
 - ne nécessite pas obligatoirement de conserver la table de correspondance !
 - en pratique on peut recourir aux fonctions de hachage publiques (MD5, SHA, etc.)
 - peut être mis à la disposition du service collecteur des données



Pseudonymisation : méthode n° 3 (2/2)

- « Message en entrée » constitué des informations nominatives
 - exemple **1 5 3 0 4 9 9 3 3 7 1 7 5 0 7 0 4 1 9 5 3 1**
- Application de la fonction de hachage sur le « message »
- « Clef de hachage » en sortie
 - exemple **9pZ82kPH165bnUj3h075MmAxx22yTM113Fdf4n63967mAaS3TTz8582qeZwSQ45**
 - longueur de la clef « suffisamment longue »
 - reproductibilité parfaite (le même message produit toujours la même clef)
 - taux de collision quasiment nul (deux messages distincts ne peuvent produire la même clef)
 - effet avalanche maximum (un changement infime du message produit une clef totalement différente)



Conserver la correspondance

Pourquoi conserver la table de correspondance ?

- pour « remonter » de l'enregistrement pseudonymisé au recueil d'origine
- pour « redescendre » du recueil d'origine à l'enregistrement pseudonymisé
 - sauf dans le cas du hachage cryptographique

Qui doit conserver la table de correspondance ?

- l'organisme qui est « en amont »

Comment conserver la table de correspondance ?

- cloisonnement





NiPN

National Information
Platforms for Nutrition

Catégories de données vis-à-vis de l'anonymisation

- Identifiants nominatifs directs ou indirects
- Données sensibles
- Quasi-identifiants
- Autres données



Identifiants (in)directement nominatifs

- **Directement nominatifs**
 - Prénom et NOM
- **Indirectement nominatifs, par utilisation de registres ou d'annuaires**
 - Adresse complète de résidence
 - Numéro de passeport, de carte nationale d'identité, de registre d'état-civil
 - Numéro de sécurité sociale, de permis de conduire
 - Numéro de téléphone
 - Adresse de courriel
- **Indirectement nominatifs, par croisement de bases et de registres**
 - Longitude + latitude du lieu de résidence => adresse complète de résidence
 - Adresse IP (internet protocol) => adresse complète de résidence



Données « sensibles »

- Données personnelles et confidentielles dont la révélation est une atteinte à la vie privée de la personne concernée
 - Source et montant des revenus
 - Descriptif et montant du patrimoine
 - Décisions administratives privées
 - Décisions judiciaires privées
 - État de santé (diagnostics, actes chirurgicaux, médicaments, ...)
 - ...
- La protection de la confidentialité des données sensibles est la seule justification de l'anonymisation



Quasi-identifiants (QID)

- **Tentative de définition**
 - Information plus ou moins précise qu'un « attaquant potentiel » peut détenir « assez couramment » au sujet d'un individu, mais ne permettant pas à elle seule de l'identifier au sein de la base de données
- **Risques liés aux QID**
 - Ré-identification par unicité des combinaisons de QID
 - Ré-identification par croisement avec d'autres sources



Exemples de QID non spécifiques

- **QID les plus courants (toutes bases de données)**
 - Date de naissance, âge
 - Sexe
 - Village, département, région de résidence
 - État marital, nombre d'enfants
 - Profession, catégorie socio-professionnelle
 - ...



Exemples de QID spécifiques

- **Base de données médico-administratives hospitalières (française)**
 - Date d'entrée, date de sortie, durée de séjour
 - Mode d'entrée, mode de sortie
 - Identification de l'établissement d'hospitalisation
- 2009 : 17 millions de séjours pseudonymisés
- 89% de ces séjours présentent des combinaisons de QID uniques
 - Accès aux données de santé de tous les individus hospitalisés en France



Exemples de QID spécifiques

- **Travaux de recherche américains**
 - Date de naissance complète
 - Sexe
 - Code postal de résidence complet
- 87% des résidents américains présentent des combinaisons uniques de ces trois QID





NiPN

National Information
Platforms for Nutrition



Ré-identification au sein d'une base de données pseudonymisée



Qu'entend-on par « ré-identification » ?

- Retrouver au sein d'une base de données les enregistrements relatifs à une personne dont on connaît déjà certains QID
- Retrouver des combinaisons de QID dont on dispose dans d'autres bases de données, qui fournissent l'identité des personnes concernées
 - cf. 1997 : Latanya Sweeney ré-identifie William Weld



Qu'entend-on par « ré-identification » ?

- **Principe**
 - fondé sur l'unicité (ou la rareté) de la combinaison des QID relatifs à la personne ciblée (ou au ménage ciblé)
- **Objectifs (rarement licites)**
 - lever la confidentialité sur la personne concernée
 - lever la confidentialité sur certains attributs relatifs à la personne concernée
 - les deux à la fois
- **Ne concerne pas nécessairement une célébrité**



Anonymisation *versus* ré-identification

- **Anonymisation (ou plutôt pseudonymisation)**
 - consiste à faire disparaître des informations inutiles pour le travail statistique proprement dit
 - processus technique relativement simple à maîtriser
- **Empêcher la ré-identification**
 - consiste à réduire, voire supprimer le risque lié à la combinaison des QID
 - sans supprimer les QID, indispensables au travail statistique proprement dit
 - processus technique un peu plus complexe à maîtriser



Y a-t-il un risque de ré-identification ?

- **Dans « notre » base de données**
 - Établir la liste exhaustive des QID
 - Commencer par une estimation empirique
- **1^{er} cas : risque empirique faible**
 - ⇒ résoudre la question des cas extrêmes (*outliers*)
- **2^e cas : risque empirique moyen ou fort**
 - ⇒ calculer le risque réel
 - ⇒ se fixer des objectifs (seuils ou plafonds)
 - ⇒ résoudre une par une toutes les situations hors des limites



Cas particulier : l'échantillonnage

- **Moins la base de données est exhaustive, moins elle présente de risque de ré-identification**
 - il s'agit de l'exhaustivité par rapport à son propre périmètre
 - sous réserve de ne pas révéler certaines caractéristiques de la population dont l'échantillon est issu
 - sous réserve de ne pas pouvoir la recouper avec une autre base qui serait exhaustive et comporterait les mêmes QID
 - sous réserve de ne pas révéler qui fait partie de l'échantillon





NiPN

National Information
Platforms for Nutrition

Mesures du risque de ré- identification

au sein d'une base de données

méthode empirique
k-anonymat, *l*-diversité
autres méthodes



Estimation empirique : méthode

- **Pour chaque QID**
 1. classer les modalités du QID par « part relative » décroissante
 - soit sur la base du dénombrement réel
 - soit à dires d'expert (distributions connues pour être uniforme ou non, etc.)
 2. déterminer le nombre de modalités qui à elles seules totalisent au moins 80% de l'effectif total
- **Pour l'ensemble de la base**
 - calculer le produit de toutes les valeurs obtenues à l'étape 2 précédente
 - ce produit est l'estimation empirique du nombre de combinaisons de QID distinctes
 - diviser l'effectif total de la base par ce produit
 - au dessus de 30 le risque est faible
 - en dessous de 10 le risque est fort
 - entre 10 et 30 le risque est moyen



***k*-anonymat (*k*-anonymity)**

- **Dans une base de données**
 - on distingue toutes les combinaisons effectives des QID
 - on observe le nombre d'enregistrements relevant de chacune d'elles
 - la valeur la plus faible est appelé *k*-anonymat
- **Les « normes habituelles »**
 - *k*-anonymat inférieur à 5 : mauvais
 - *k*-anonymat compris entre 5 et 9 : médiocre
 - *k*-anonymat supérieur ou égal à 10 : satisfaisant
 - *k*-anonymat supérieur à 30 : très satisfaisant



Intérêt du k -anonymat

- **Que représente-t-il ?**
 - L'effectif de la combinaison de QID la plus rare
 - Un indicateur de la rareté des combinaisons de QID de toute la base
- **On peut le compléter par**
 - Nombre total de combinaisons de QID distinctes dans la base de données
 - Nombre de combinaisons de QID ayant pour effectif la valeur du k -anonymat
 - Effectif moyen des combinaisons de QID
- **A quoi sert-il ?**
 - C'est une métrique qui participe à la qualification du risque relatif à une base
 - Permet de « *monitorer* » l'évolution du risque de ré-identification obtenue par les mesures que l'on va mettre en œuvre pour réduire ce risque



Interprétation du k -anonymat > 10

Aucun individu ne se trouve dans un groupe de moins de 10 personnes



Interprétation du k -anonymat > 10

Aucun individu ne se trouve dans un groupe de moins de 10 personnes
donc

Même si un attaquant connaît des informations sur une personne, il ne pourra pas la distinguer dans la base parmi les 10 personnes similaires



Interprétation du k -anonymat > 10

Aucun individu ne se trouve dans un groupe de moins de 10 personnes

donc

Même si un attaquant connaît des informations sur une personne, il ne pourra pas la distinguer dans la base parmi les 10 personnes similaires

donc

Il est impossible pour un attaquant de tirer des conclusions relatives à un individu ou un ménage à partir des données sensibles de la base de données, en se fondant sur les QID dont il dispose relatifs à cet individu ou ce ménage



Interprétation du k -anonymat > 10

Aucun individu ne se trouve dans un groupe de moins de 10 personnes

donc

Même si un attaquant connaît des informations sur une personne, il ne pourra pas la distinguer dans la base parmi les 10 personnes similaires

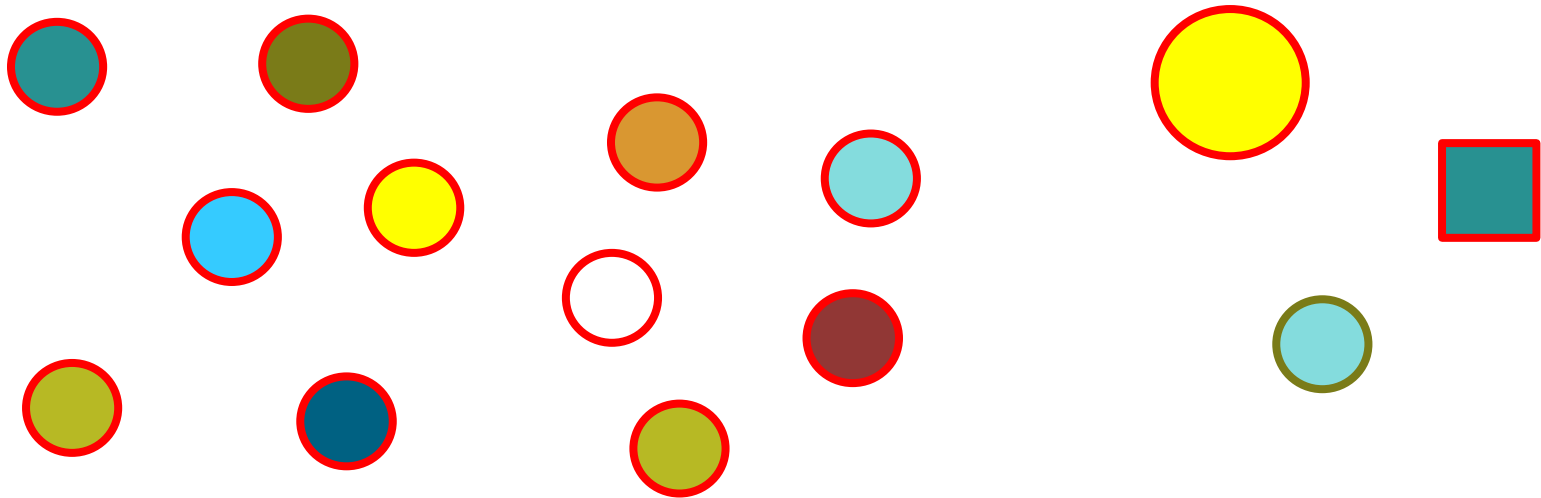
donc

Il est impossible pour un attaquant de tirer des conclusions relatives à un individu ou un ménage à partir des données sensibles de la base de données, en se fondant sur les QID dont il dispose relatifs à cet individu ou ce ménage

Sauf que ce n'est pas tout à fait suffisant !

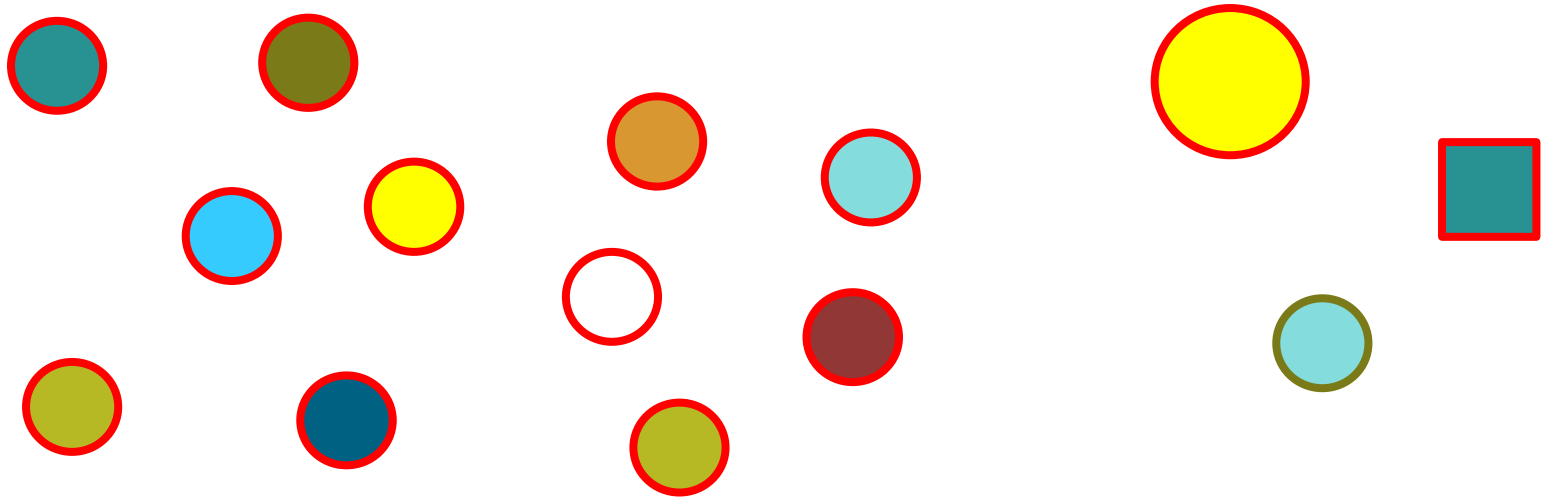


QID = forme, taille, contour
donnée sensible = couleur

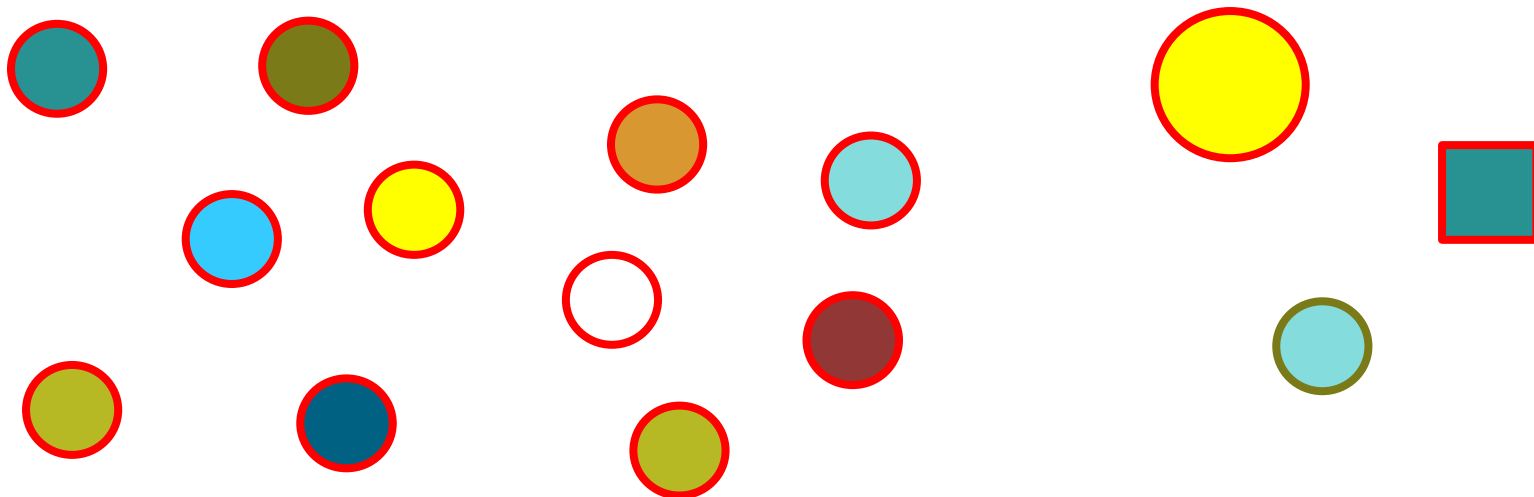


QID = forme, taille, contour
donnée = couleur

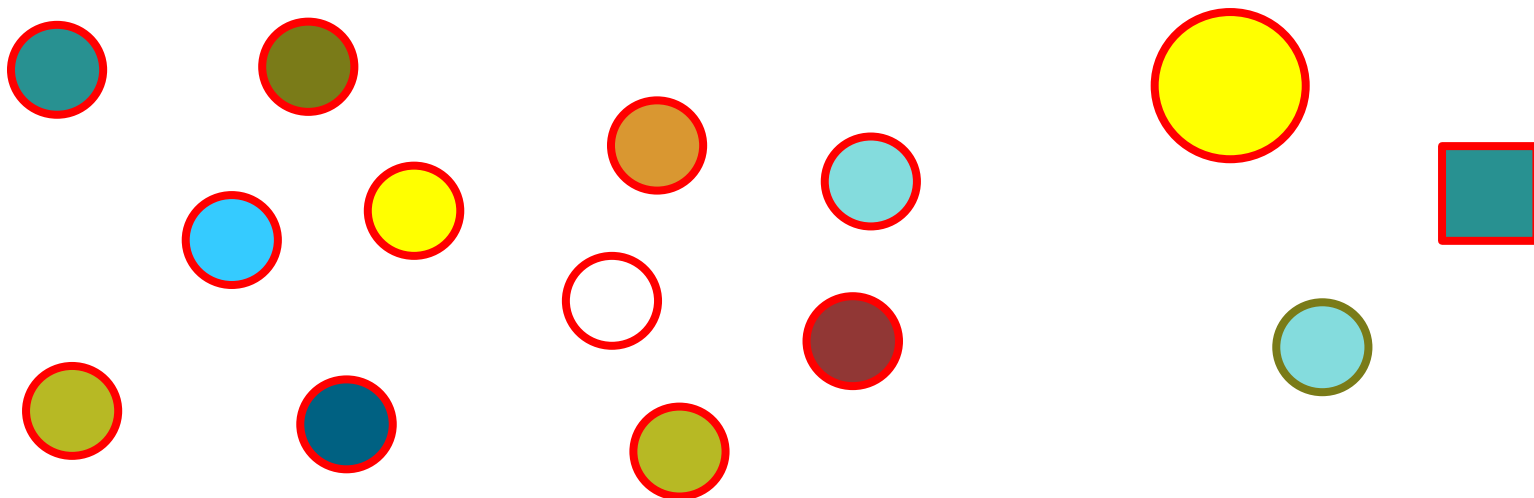
enfant
adulte



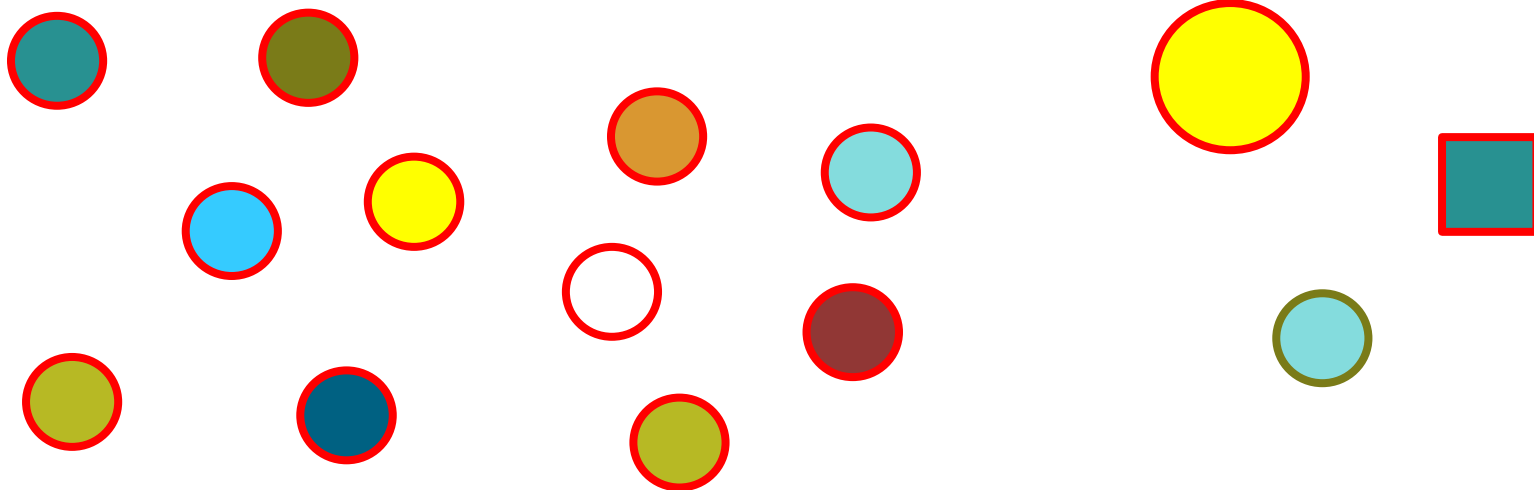
QID = forme taille contour
donnée couleur



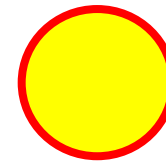
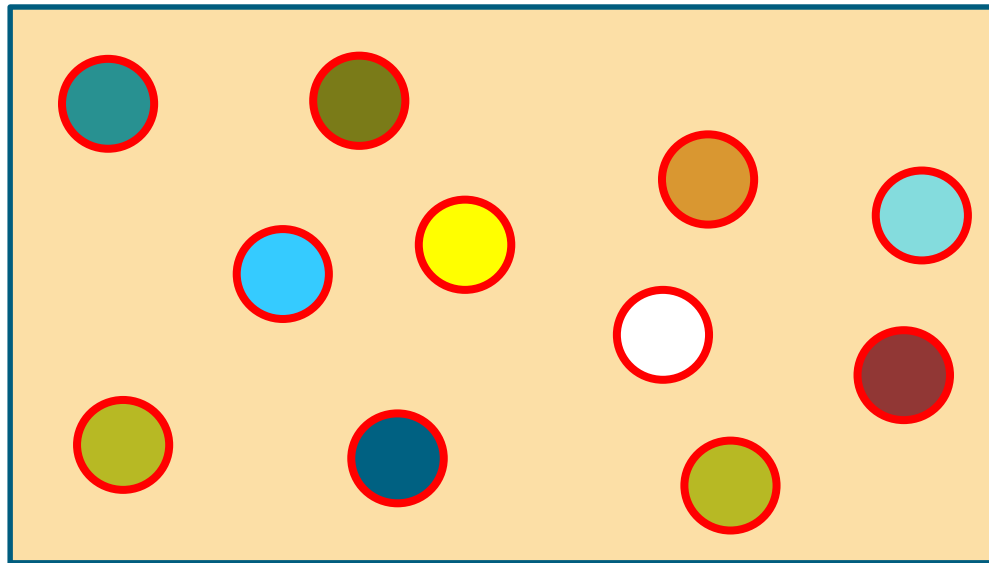
QID = forme **taille** contour
donné **enfant** **adulte** **taille** **sexe**



QID = forme **taille** contour
donnée **enfant** **adulte** **sexe** **revenus**



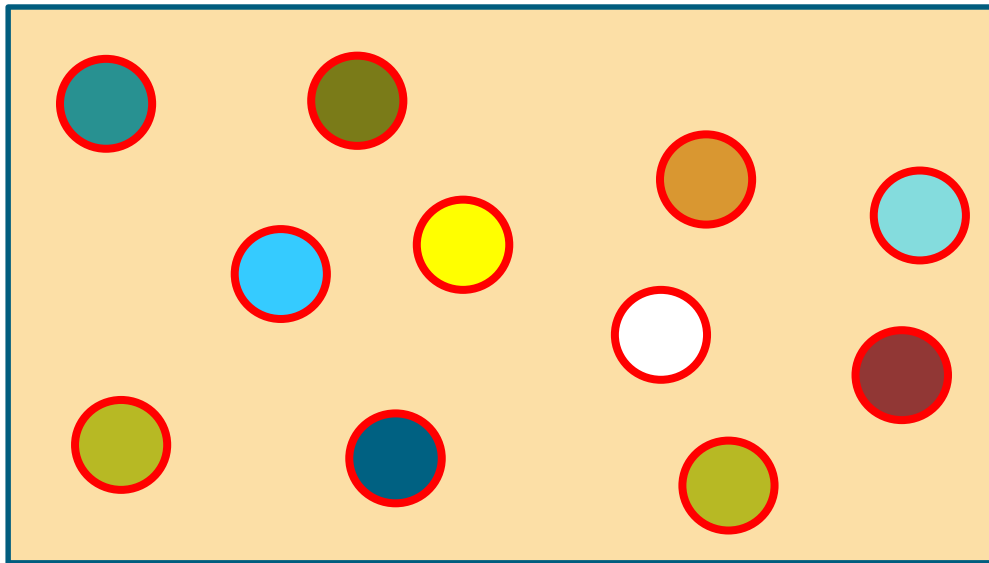
QID = forme, taille, contour
donnée sensible = couleur



rond, petit, contour rouge



QID = forme, taille, contour
donnée sensible = couleur



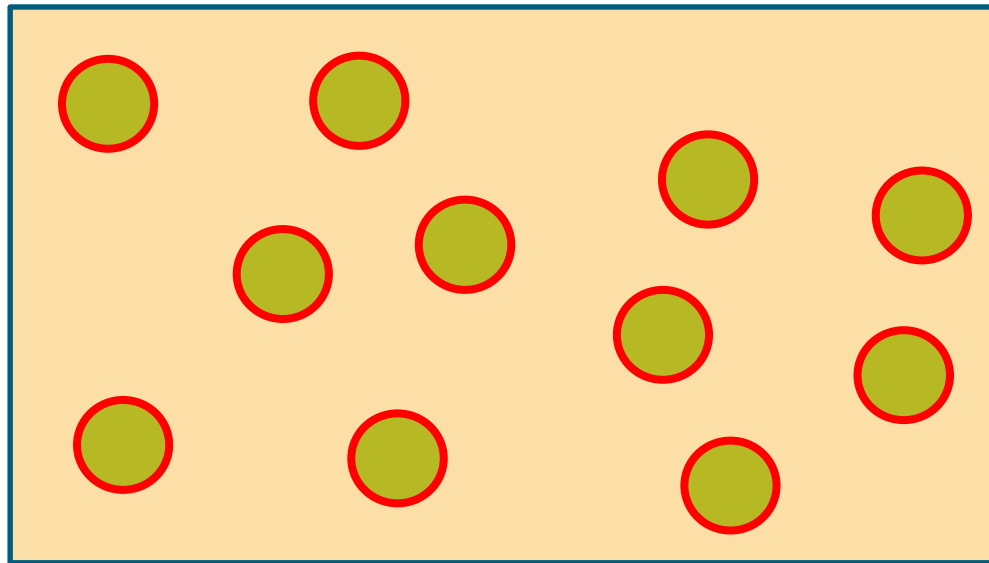
Il est impossible de lever la confidentialité des données sensibles relatives à tel ou tel membre du groupe.

L'hétérogénéité des couleurs le prouve.

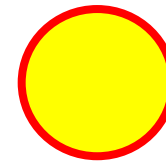
rond, petit, contour rouge



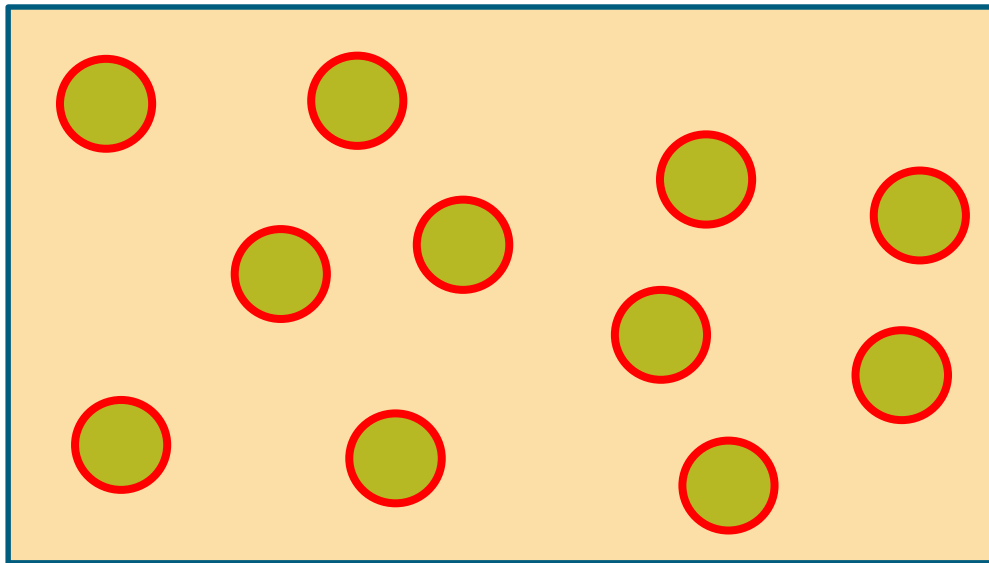
QID = forme, taille, contour
donnée sensible = couleur



rond, petit, contour rouge



QID = forme, taille, contour
donnée sensible = couleur



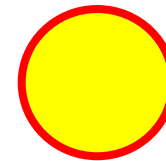
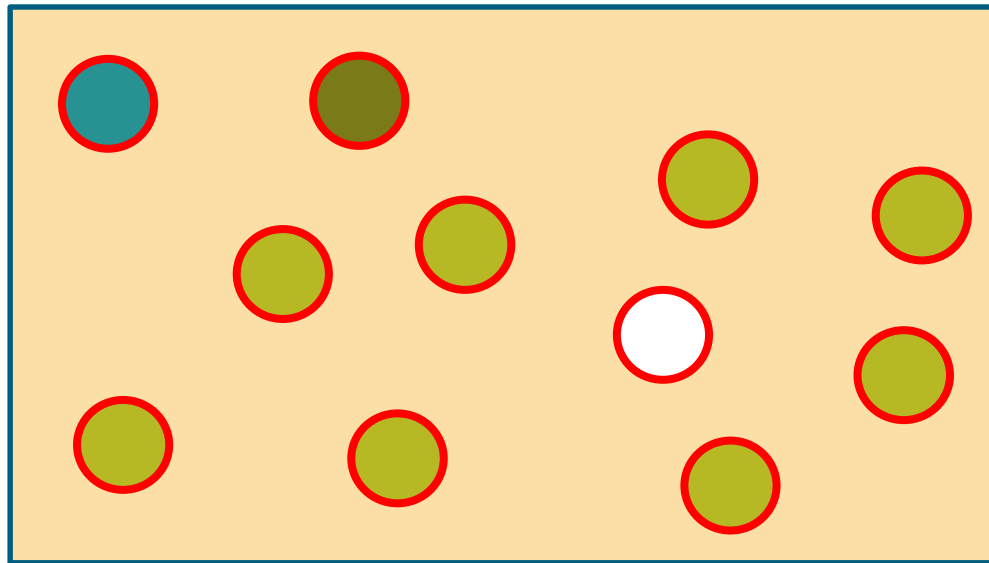
Dans ce cas limite, peu importe de savoir quel enregistrement correspond à quel individu : il est évident que tous les individus du groupe présentent la même information sensible.

Le risque de voir levée la confidentialité des données sensibles relatives à tel ou tel membre du groupe est élevé.

rond, petit, contour rouge



QID = forme, taille, contour
donnée sensible = couleur



rond, petit, contour rouge

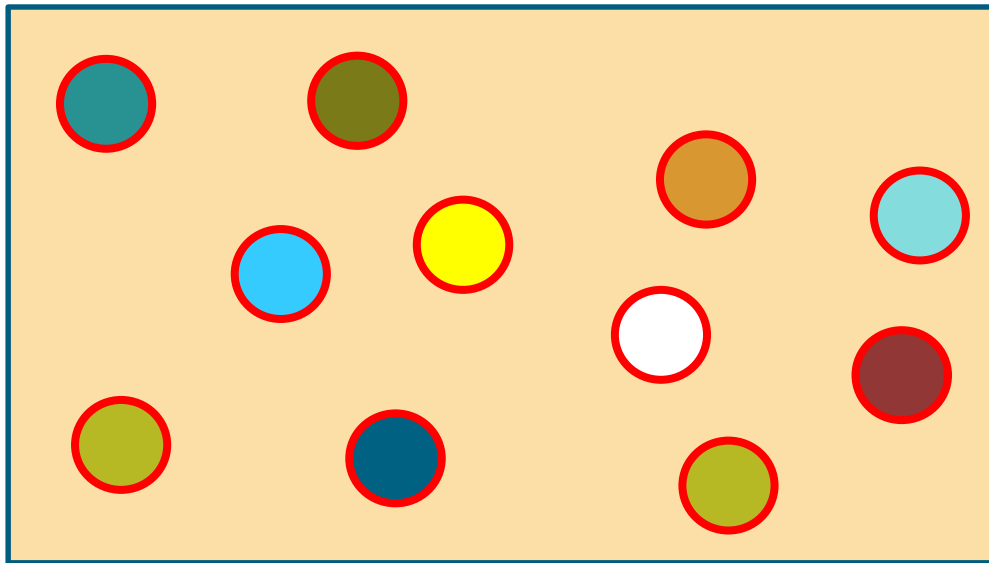


l-diversité (*l*-diversity)

- **Dans une base de données**
 - on distingue toutes les combinaisons effectives des QID
 - on observe le nombre de modalités distinctes des données sensibles que présentent chacune d'elles, qu'on appelle sa diversité
 - la valeur la plus faible est appelé *l*-diversité
- **Les « normes habituelles »**
 - *l*-diversité inférieure à 3 : mauvais
 - *l*-diversité supérieure ou égale à 3 : satisfaisant
 - *l*-diversité supérieure ou égale à 5 : très satisfaisant



QID = forme, taille, contour
donnée sensible = couleur



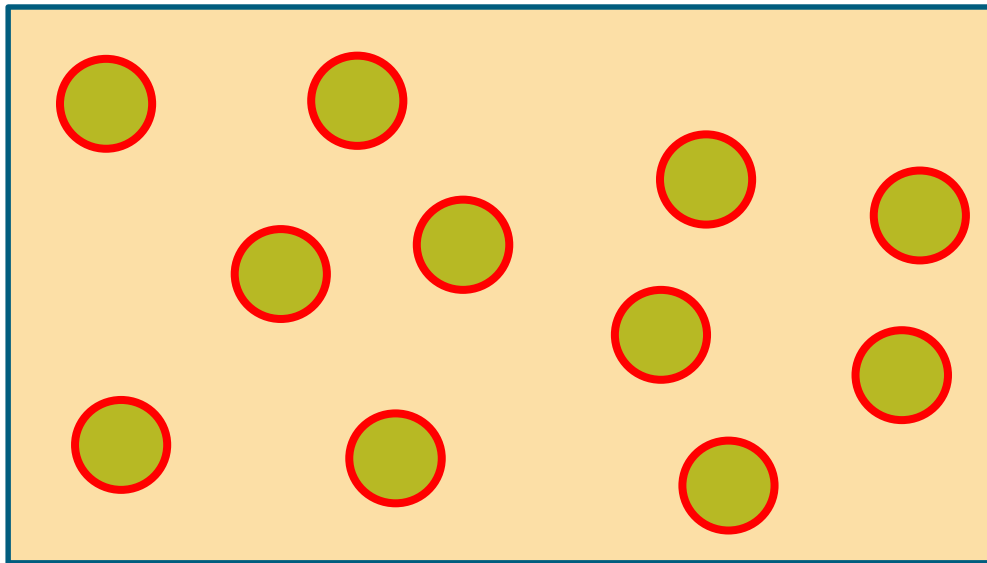
Nombre de modalités distinctes
des données sensibles (couleur)

10

rond, petit, contour rouge



QID = forme, taille, contour
donnée sensible = couleur



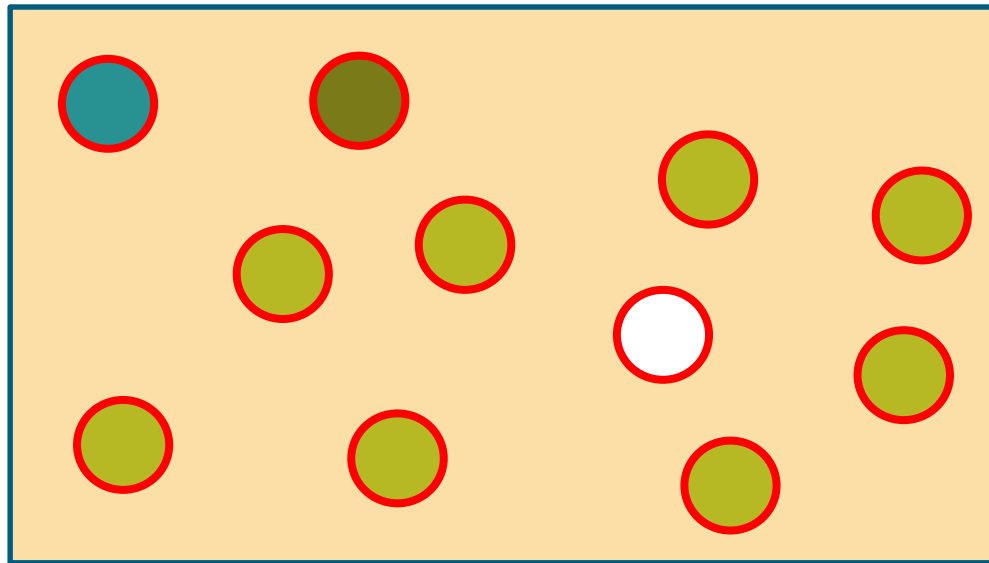
Nombre de modalités distinctes
des données sensibles (couleur)

1

rond, petit, contour rouge



QID = forme, taille, contour
donnée sensible = couleur



Nombre de modalités distinctes
des données sensibles (couleur)

4

rond, petit, contour rouge

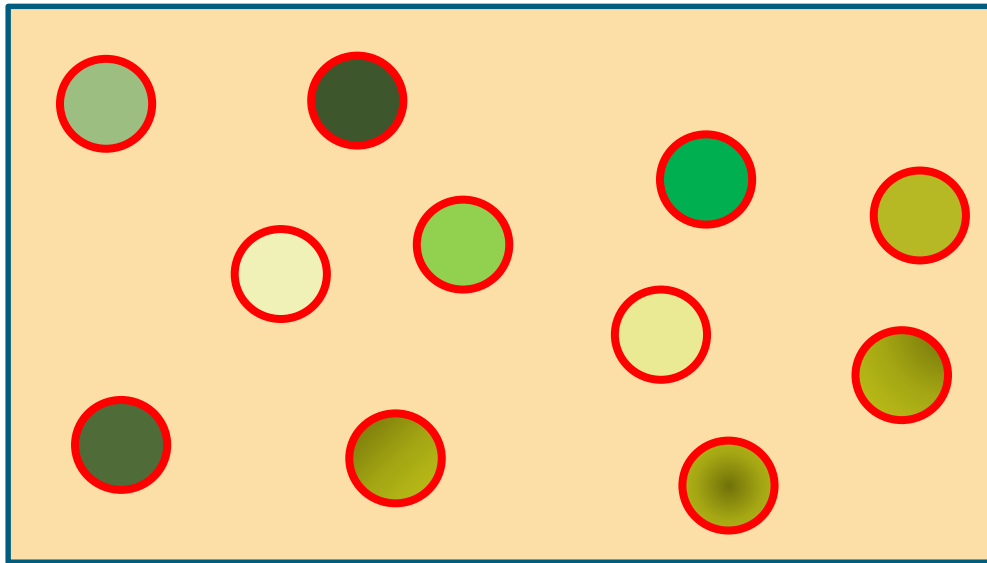


Intérêt de la *I*-diversité

- **Que représente-t-elle ?**
 - L'hétérogénéité des données sensibles de la combinaison de QID la plus uniforme
 - Un indicateur de l'hétérogénéité au sein de toutes les combinaisons de QID
- **On doit la compléter par**
 - Proportion de combinaisons ayant une diversité égale à 1
 - Proportion de combinaisons ayant une diversité égale à 2
 - Proportion de combinaisons ayant une diversité de 3 ou plus
- **A quoi sert-elle ?**
 - C'est une métrique qui participe à la qualification du risque relatif à une base
 - Permet de « *monitorer* » l'évolution du risque de ré-identification obtenue par les mesures que l'on va mettre en œuvre pour réduire ce risque

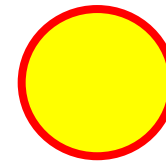


l-diversité : une difficulté pratique



11 nuances de vert, mais vert quand même

rond, petit, contour rouge



***l*-diversité : une difficulté pratique**

- **Les modalités précises des données sensibles peuvent être distinctes mais très proches**
 - un attaquant pourrait finalement tirer des conclusions « assez précises » sur tous les individus qui composent le groupe
 - la confidentialité des informations relatives à tous les individus du groupe est compromise, même si l'on ne peut déterminer qui est précisément chaque individu du groupe
 - ⇒ pour le calcul de la diversité, il serait justifié d'utiliser une variable moins précise, de sorte que toutes ces modalités n'en fassent plus qu'une seule
 - ⇒ pour améliorer la protection des données sensibles, il faudrait renforcer la diversité du groupe, par exemple en augmentant sa taille (fusion avec un autre groupe, dont les QID seraient proches)



I-diversité : autre difficulté pratique

- Il y a souvent plusieurs données sensibles : de laquelle doit-on mesurer la *I*-diversité ?
 - on retient la *I*-diversité de la variable sensible qui est la moins diverse
 - soit elle est bien connue des gens du métier ou des experts et alors on ne calcule la *I*-diversité que pour cette variable
 - soit on hésite entre plusieurs variables sensibles, et dans ce cas on calcule la *I*-diversité pour chacune de ces variables pour ne retenir finalement que la plus faible



Autre méthode de calcul du risque de ré-identification

- pour une combinaison donnée des QIDs
 - nombre d'individus présentant cette combinaison dans la population : $\mathcal{K}$
 - nombre d'individus présentant cette combinaison dans l'échantillon : k
- risque de ré-identification pour cette combinaison
 - si l'on sait que l'individu recherché se trouve dans l'échantillon : $1 / k$
 - si on l'ignore : $k / \mathcal{K}$
- remarques
 - calculer un risque pour chacune des combinaisons
 - la plupart du temps on connaît k mais pas $\mathcal{K}$
 - il faut estimer $\mathcal{K}$ à partir de k
 - mais en général, si le facteur d'échantillonnage est $\mathcal{E}$, le risque est dans l'intervalle $[\mathcal{E} / 2 , \mathcal{E} * 2,5]$
 - exemple, pour un échantillonnage au 100^{ème} le risque est compris entre 1/200 et 1/40
 - ne dispense pas de se préoccuper de la k -diversité !



NiPN

National Information
Platforms for Nutrition



Anonymisation des données

théorie et pratique

D.Blum - SOFRECO



NiPN

National Information
Platforms for Nutrition

Niamey - 28 au 31 mai 2019



Deuxième journée

- Travaux pratiques



Résumé de l'épisode précédent

1. suppression du risque d'identification immédiat

- suppression des identifiants individuels directs et/ou indirects
- pseudonymisation (le plus souvent)

2. contrôle du risque de « ré-identification »

- évaluation du risque
 - estimation empirique
 - métrique habituelle : ℓ -anonymat et ℓ -diversité
- déterminer quelle stratégie de réduction du risque mettre en œuvre

Hachage cryptographique avec Excel (1/2)

```
Function ToBase64String(rabyt)
  With CreateObject("MSXML2.DOMDocument")
    .LoadXML "<root />"
    .DocumentElement.DataType = "bin.base64"
    .DocumentElement.nodeTypedValue = rabyt
    ToBase64String = Replace(.DocumentElement.text, vbLf, "")
  End With
End Function
```

```
Function ComputeHash(fld)
  Dim text As Object
  Dim SHA512 As Object
  Set text = CreateObject("System.Text.UTF8Encoding")
  Set SHA512 = CreateObject("System.Security.Cryptography.SHA512Managed")
  ComputeHash = ToBase64String(SHA512.ComputeHash_2((text.GetBytes_4(fld))))
End Function
```



Hachage cryptographique avec Excel (2/2)

Hachage cryptographique.xlsm - Microsoft Excel

Accueil Insertion Mise en page Formules Données Révision Affichage PDF Architect 5 Creator

E6 $\text{=ComputeHash(CONCATENER("grain de sel";A6;"/";B6;"/";C6;"/";D6))$

	A	B	C	D	E
1	Prénom	NOM	sexe	date de naissance	Clef de hachage
2	Tryphon	TOURNESOL	masculin	04/12/1947	boDNYN0Q0hwoMF7ZCVfHv/YtdQNVeUV2FpcRC018vhsK0x592LNcGkUF17snVuTM8c6+2MsVETiXbZvGrmnIXw==
3	Capitaine	HADDOCK	masculin	12/04/1945	cCCOzqxCSKc3lRA/iyasKzbXaliYIzMEY/o9/ZHvRKnnctNmd8g3VQgp6iHcfLqFb79RAzYvXDgjTFFBEKbSg==
4	Tintin	MILOU	masculin	22/05/1907	eJpbZQhMC43SZLnFtakVRSeF2vWlkV67okDn3fOg3daHJeirFT+WruLzXilFlWELdccQhgwBaKhXB5luXqd2Zw==
5	Alfred	HITCHCOCK	masculin	13/08/1899	HTedVqhicdghsUdtTtuJcE4fQ00FowJc9e6VX6P9BEebXHwdnp9JCqG9PW3gmwf6hr0DR1sW3UhPO+MGT14lCA==
6	Bianca	CASTAFIORE	féminin	20/02/1948	rJnL1B/1842/PSPwr6UaRKQHy18kZBxnRNh817+2Q3hgQn0jOy8DYf1Ie7MfW7OzZMKCKKyK9c51kRL8SdmatQ==

Feuil1

Prêt

165 %

$\text{= ComputeHash(CONCATENER("grain de sel";A6;"/";B6;"/";C6;"/";D6))}$



Calcul du *k*-anonymat avec SQL (1/2)

	région	sexe	âge	...	situation matrimoniale	dépenses nourriture	dépenses de santé	...	ressources
Pseudonyme	QID n°1	QID n°2	QID n°3	...	QID n°x	donnée sensible n°1	donnée sensible n°2	...	donnée sensible n°y
	q1	q2	q3	...	qx	v1	v2	...	vy
1	1	1	41	...	1	...	...	...	...
2	3	1	39	...	2	...	...	...	...
3	1	2	42	...	2	...	...	...	...
4	2	1	41	...	2	...	...	...	...
5	2	2	38	...	1	...	...	...	...
6	5	2	42	...	2	...	...	...	...
7	8	2	15	...	2	...	...	...	...
8	6	1	50	...	3	...	...	...	...
9	1	1	45	...	2	...	...	...	...
10	4	2	40	...	3	...	...	...	...
11	4	1	38	...	1	...	...	...	...
...	...	...	...	...	...	...	...	...	...



Calcul du *k*-anonymat avec SQL (2/2)

	région	sexe	âge	...	situation matrimoniale	dépenses nourriture	dépenses de santé	...	ressources
Pseudonyme	QID n°1	QID n°2	QID n°3	...	QID n°x	donnée sensible n°1	donnée sensible n°2	...	donnée sensible n°y
	q1	q2	q3	...	qx	v1	v2	...	vy
1	1	1	41	...	1	...	...	...	...
...	...	...	...	...	...	...	...	...	...

```
SQL> select q1, q2, q3, [...], qx, count(*) as k
      from [nom_de_la_table]
      group by q1, q2, q3, [...], qx
```





NiPN
National Information
Platforms for Nutrition



Réduction des risques de ré-identification et de divulgation

Actions sur les QID
Actions sur les données sensibles



Troisième journée

- **Méthodes pour limiter le risque de ré-identification**
 1. intervenir sur les QIDs
 - suppression des quasi-identifiants (QIDs) inutiles
 - « floutage » des QIDs utiles (sans réduire leur « utilisabilité »)
 2. intervenir sur les données sensibles
 - il ne s'agit plus à proprement parler de techniques d'anonymisation
 - trois types de méthodes



Autrement dit

- **Base nominative de données brutes**
 - suppression des informations directement ou indirectement nominatives
 - introduction de pseudonymes
- **Base pseudonymisée de données brutes**
 - évaluer empiriquement le risque de ré-identification par les combinaisons de QID
 - calculs exacts du *k*-anonymat et de la *l*-diversité
- **Et si $k\text{-anonymat} < 10$ ou $l\text{-diversité} < 3$**
 - réduire le risque en réduisant le nombre de combinaisons (jouer sur les QIDs)
 - réduire le risque de divulgation en réduisant la précision des données sensibles



Réduction du risque de ré-identification

Intervenir sur les QIDs

Les méthodes

- méthodes non perturbatrices
 - « floutage » des QIDs (= recodage en plages de valeurs)
 - suppression locale de certaines valeurs des QIDs
- méthodes perturbatrices
 - bruitage des QIDs
 - post-randomisation des QIDs
 - permutation des valeurs de QIDs
- création de jeux de données virtuelles

Action sur les QIDs : « floutage » (1/2)

- **Terminologie officielle : « généralisation »**
 - réduire la précision des QID en regroupant plusieurs modalités
 - QID quantitatifs : plages de valeurs (âge, ressources, dépenses, etc.)
 - QID catégoriels ou qualitatifs : catégories plus larges
 - commencer par flouter les QID ayant le plus grand nombre de modalités
 - localisations géographiques (directes ou indirectes)
 - dates
 - référentiels détaillés (professions, etc.)
 - âge



Action sur les QIDs : « floutage »

(2/2)

- **Garder à l'esprit qu'en général**
 - réduire les modalités d'un facteur de 10 réduit le nombre de groupes d'un facteur de 2, voire 3 mais pas plus
 - réduire les modalités des QID
 - agit très peu sur l'effectif de la combinaison de QID la plus rare (le k -anonymat)
 - mais agit nettement sur les effectifs des combinaisons les plus denses
- **Conséquence**
 - plutôt qu'une généralisation globale, mettre en œuvre une généralisation locale (ou « différenciée »)
 - avec un inconvénient : les modalités sont inhomogènes



Réduction du risque de ré-identification

Intervenir sur les QIDs

Les méthodes

- méthodes non perturbatrices
 - « floutage » des QIDs (= recodage en plages de valeurs)
 - suppression locale de certaines valeurs des QIDs
- méthodes perturbatrices
 - bruitage des QIDs
 - post-randomisation des QIDs
 - permutation des valeurs de QIDs
- création de jeux de données virtuelles

Perturbation des QIDs

- **Création de bruit**
 - ajout de valeurs aléatoires (généralement faibles) aux QIDs quantitatifs
 - sous réserve de préserver les moyennes, variances, corrélations...
- **Permuter les valeurs de QIDs de deux enregistrements**
 - sous réserve de réciprocité pour ne pas biaiser l'ensemble
 - sous réserve de ne pas créer des absurdités détectables



Réduction du risque de divulgation de données sensibles

(mais ce n'est plus de l'anonymisation)

Les méthodes (= masquage des données)

- méthodes non perturbatrices
 - « floutage » des données sensibles (= recodage en plages de valeurs)
 - suppression locale de certaines valeurs des données sensibles
- méthodes perturbatrices
 - bruitage des données sensibles
 - micro-agrégation des données sensibles
- création de jeux de données virtuelles

Dans tous les cas, il faut ensuite évaluer la perte d'information

« Masquage » des données

- **Principe**
 - modifier les données sensibles de certains enregistrements
 - tout en conservant « certaines » de leurs caractéristiques statistiques
 - moyennes, corrélations, etc.



Perturbation de l'information

- **Création de bruit**
 - ajout de valeurs aléatoires (généralement faibles) aux valeurs recueillies
 - sous réserve de préserver les moyennes, variances, corrélations...
- **Permuter certaines données sensibles de deux enregistrements**
 - sous réserve de réciprocité pour ne pas biaiser l'ensemble
 - sous réserve de ne pas créer des absurdités détectables
- **Micro-agrégation**
 - on remplace les données par une « moyenne locale »



« détails » à prendre en considération

- **Échantillonnage**
 - attention au croisement avec une base de données externe
- **Suppression d'enregistrements**
 - sous réserve d'en ajuster l'impact
- **Traitement des *outliers* (valeurs extrêmes)**
 - limite inférieure
 - limite supérieure
- **Suppression d'une donnée trop discriminante**
 - transformée en valeur manquante
 - sous réserve de ne pas communiquer sur ce « cas exceptionnel »





NiPN

National Information
Platforms for Nutrition



Anonymisation des données

théorie et pratique

D.Blum - SOFRECO



NiPN
National Information
Platforms for Nutrition

Niamey - 28 au 31 mai 2019



Résumé de l'épisode précédent

1. interventions sur les QIDs

- sélection itérative du « meilleur » QID à re-traiter
 - suppression de QIDs inutiles
 - « floutage » de certaines modalités de QIDs
- calcul itératif du k-anonymat sous Excel

2. interventions sur les données sensibles

- « floutage » de certaines modalités de données sensibles
- calcul itératif de la l-diversité sous Excel

Quatrième journée

- **Essais manuels de méthodes perturbatrices**
 1. sur les QIDs
 - bruitage
 - permutations
 - calculs itératifs du k-anonymat et de la l-diversité
 2. sur les données sensibles
 - bruitage
 - micro-agrégation
 - calculs itératifs de la l-diversité



Recommandations

1. Ne pas réduire le sujet à la question de la pseudonymisation
2. Ne pas espérer le logiciel-miracle
3. Fixer les objectifs en fonction du type de base de données et du type de diffusion
4. L'échantillonnage peut rassurer à tort
5. Huit à dix QID au maximum
6. Attention aux pseudo-QID qui sont en fait des vrais identifiants indirectement nominatifs
7. Maîtriser l'approche empirique pour le calcul du risque de ré-identification
8. Se concentrer sur les combinaisons de QID de faible effectif
9. Privilégier la « généralisation locale » des QID
10. Recourir à la « généralisation locale » des données sensibles en cas de besoin



Encore des recommandations

- ▶ Garder à l'esprit : la diffusion en *open data* est en général « *one shot* »
- ▶ Maîtriser le processus de A à Z
- ▶ Prévoir trois mois à temps plein pour une base de 150 à 200 variables dont 5 à 10 QID
- ▶ Questions ciblées par courriel : dblum@le-pmsi.fr



Vos questions et commentaires



**Merci pour votre attention,
et bonnes pseudonymisations
non ré-identifiantes !**





NiPN

National Information
Platforms for Nutrition



Anonymisation des données

théorie et pratique

D.Blum - SOFRECO



NiPN
National Information
Platforms for Nutrition

Niamey - 28 au 31 mai 2019



Démographie du Niger

	1977	1988	2001	2012	2013	2014	2015	2016	2017	2018
Niger	5 102 990	7 251 626	11 060 291	16 993 563	17 679 760	18 389 164	19 124 883	19 865 067	20 651 070	21 466 863
Agadez	124 985	208 828	321 639	478 833	495 211	512 148	529 658	547 755	566 447	585 737
Diffa	167 389	189 091	346 595	586 648	607 732	627 529	648 049	669 307	691 356	714 242
Dosso	693 207	1 018 895	1 505 864	2 034 324	2 113 733	2 195 788	2 280 703	2 368 651	2 459 812	2 554 379
Maradi	949 747	1 389 433	2 235 748	3 365 969	3 511 327	3 663 102	3 821 593	3 987 165	4 160 231	4 340 983
Tahoua	993 615	1 308 598	1 972 729	3 304 193	3 432 320	3 564 239	3 699 907	3 839 457	3 983 172	4 131 384
Tillabéri	928 849	1 328 283	1 889 515	2 701 408	2 814 086	2 930 976	3 052 368	3 155 731	3 280 333	3 409 676
Zinder	1 002 225	1 411 061	2 080 250	3 506 758	3 653 746	3 806 825	3 966 348	4 132 321	4 305 953	4 487 009
Niamey	242 973	397 437	707 951	1 015 430	1 051 605	1 088 557	1 126 257	1 164 680	1 203 766	1 243 453





NiPN

National Information
Platforms for Nutrition



Anonymisation des données

théorie et pratique

D.Blum - SOFRECO



NiPN

National Information
Platforms for Nutrition

Niamey - 28 au 31 mai 2019

