Survey sur les techniques d'anonymisation : Théorie et Pratique

Benjamin NGUYEN

benjamin.nguyen@insa-cvl.fr

Laboratoire d'Informatique Fondamentale d'Orléans, INSA Centre Val de Loire

GDR Sécurité Informatique / GT Protection de la Vie Privée

Plan

1. *Qu'est ce que l'anonymat ?*
2. *La pseudonymisation*

Traités par Claire Levallois-Barth

3. Architecture d'anonymisation
4. Technique historique d'anonymisation
5. Evaluation du risque de réidentification
6. ~~Techniques classiques d'anonymisation~~
7. Méthodes statistiques classiques
8. Confidentialité différentielle (*Differential Privacy*)
9. ~~Quelques travaux de recherche personnels~~

GDPR et 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.* (…)

1. Définition de “l’identifiabilité”
2. Précision de la portée des techniques de réidentification : obligation de moyens

/!\ Moins contraignant que la loi “informatique et libertés” (1978)

Loi Informatique et Libertés

Article 8

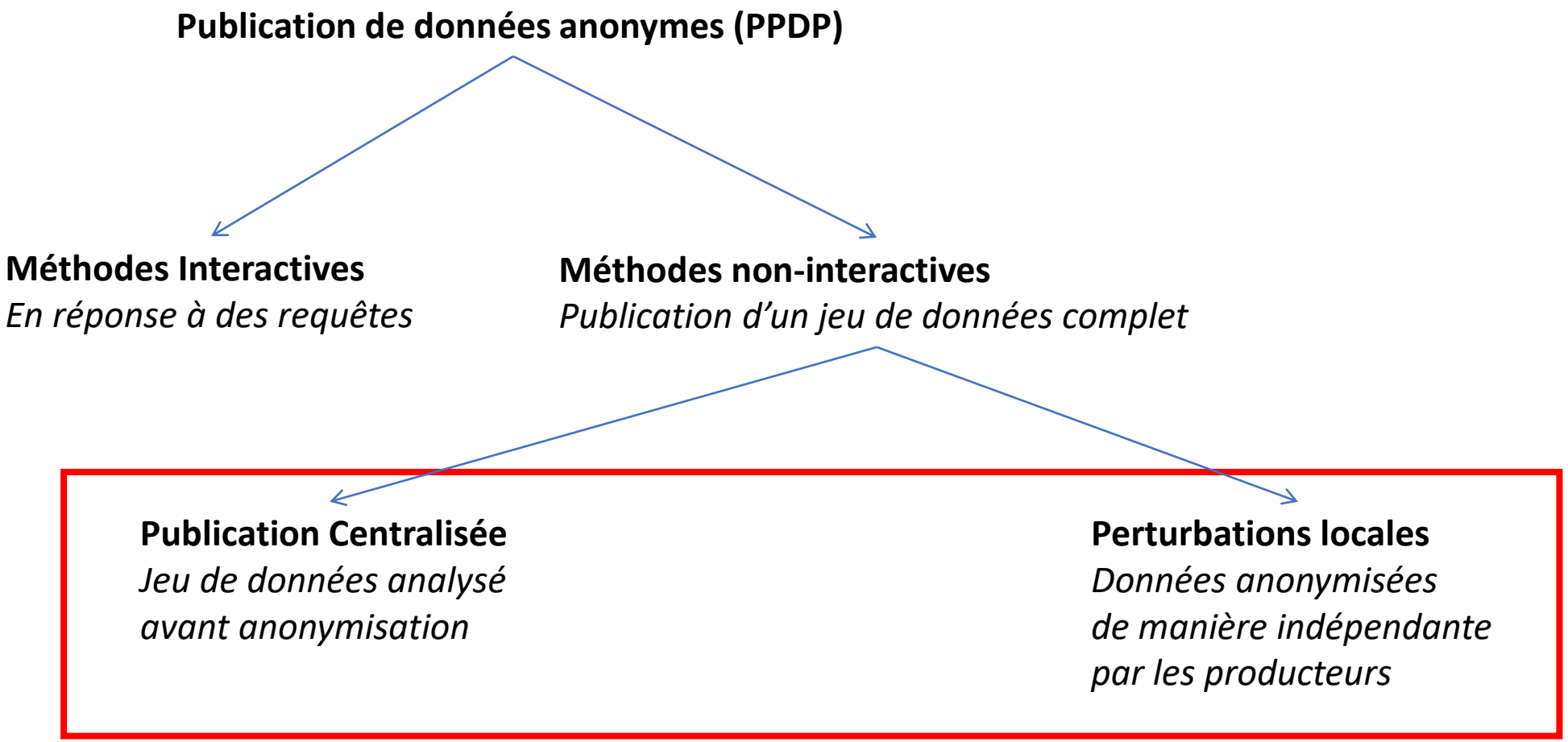
…

i) *Elle peut certifier ou homologuer et publier des référentiels ou des méthodologies générales aux fins de certification, par des tiers agréés ou accrédités selon les modalités mentionnées au h du présent 2°,* ***de la conformité à la présente loi de processus d'anonymisation des données à caractère personnel****, notamment en vue de la réutilisation d'informations publiques mises en ligne dans les conditions prévues au titre II du livre III du code des relations entre le public et l'administration.*

Architecture d’anonymisation

Comment anonymiser ?

Classification des approches



Architecture classique d'anonymisation en *publication centralisée*

Contexte

Des données personnelles, sensibles, issus de mesures, de capteurs, de questionnaires, etc

Objectif

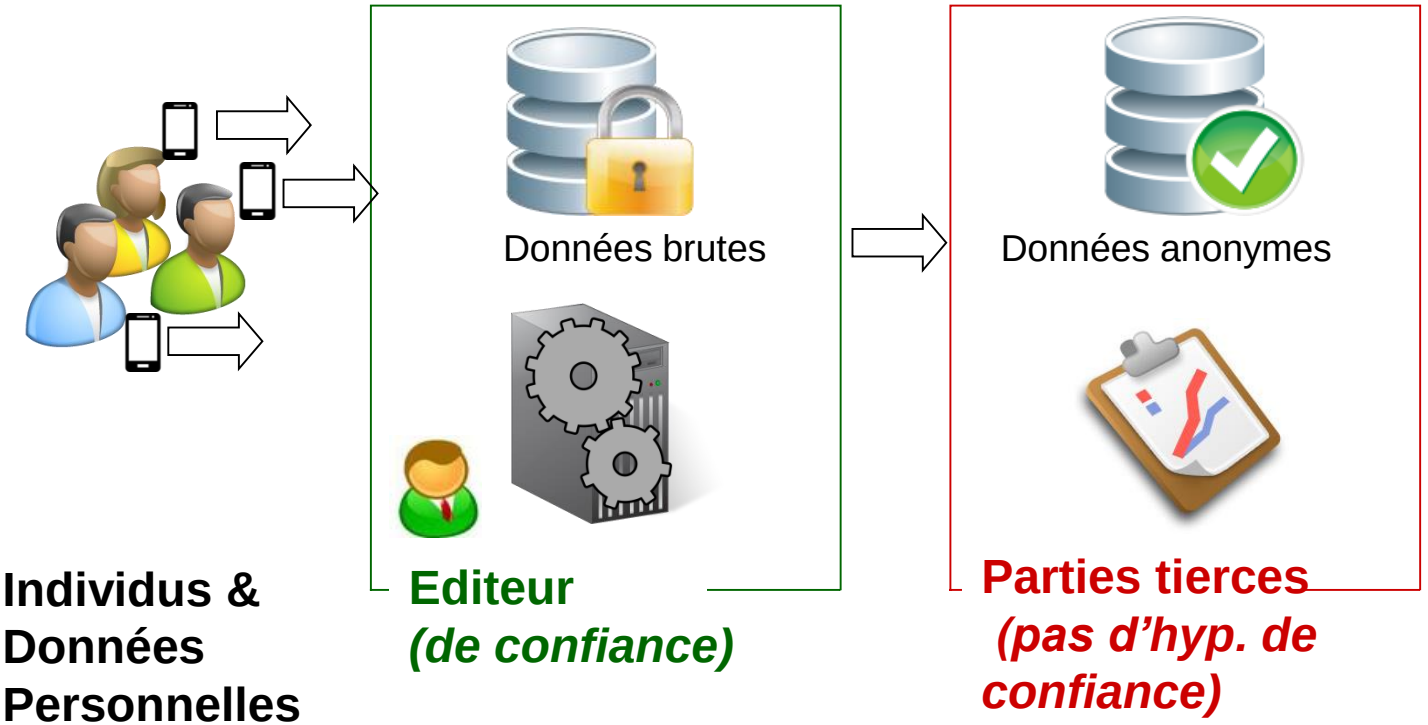
Poser des requêtes (agrégats, corrélations,…)

Contraintes

–Impossibilité d'utiliser un système spécifique interactif pour répondre aux requêtes

–Diffuser une fois le jeu de données, mais de telle sorte qu'il soit « inoffensif »

–Choisir un mécanisme d'anonymisation



http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Architecture classique d'anonymisation en *perturbation locale*

Contexte

Des données personnelles, sensibles, issus de mesures, de capteurs, de questionnaires, etc

Objectif

Poser des requêtes (agrégats, corrélations,…)

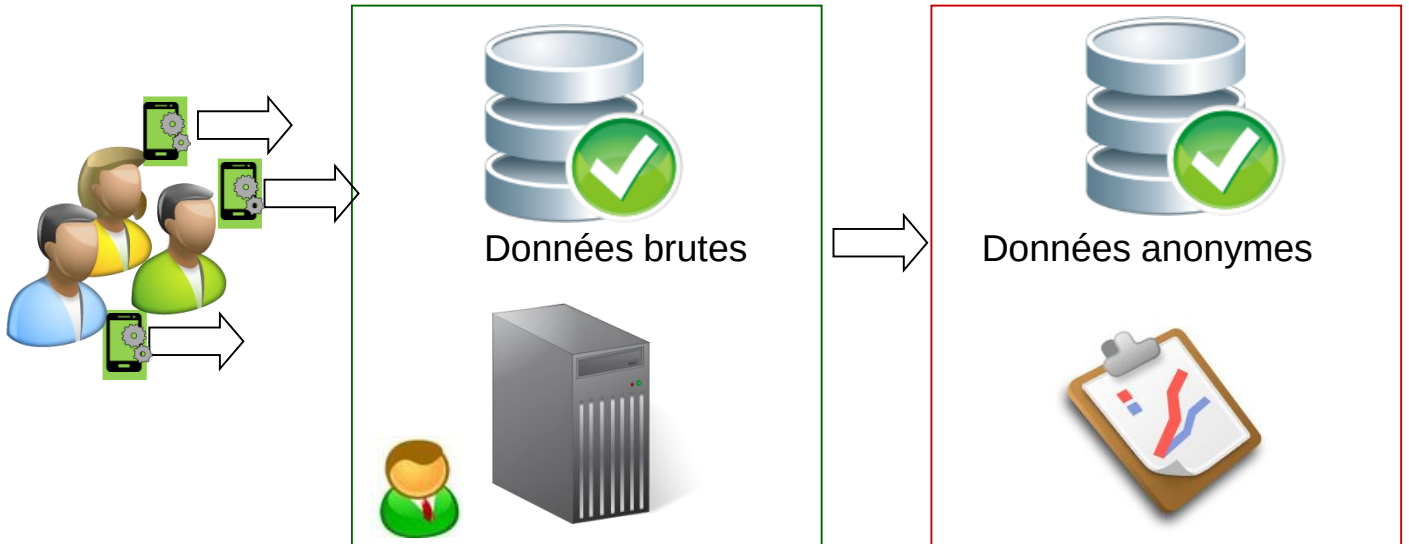
Contraintes

–Impossibilité d'utiliser un système spécifique interactif pour répondre aux requêtes

–Diffuser une fois le jeu de données, mais de telle sorte qu'il soit « inoffensif »

–Choisir un mécanisme d'anonymisation

–***S'assurer qu'il peut fonctionner en mode « perturbation locale »***



Composants de l'anonymisation

- **Une définition de la "privacy" qui répond à la question :** quelle protection proposer ?
- **Une métrique d'utilité qui répond à la question :** Comment mesurer la perte d'information dans le processus ?
- **Une algorithme d'anonymisation qui répond à la question :** Comment protéger les donnée tout en maximisant l'utilité des données ?
- **Un processus d'anonymisation :** qui permet de mettre en œuvre l'algorithme de manière sûre et sécurisée.

Technique historique d'anonymisation

Attaque sur la pseudonymisation

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

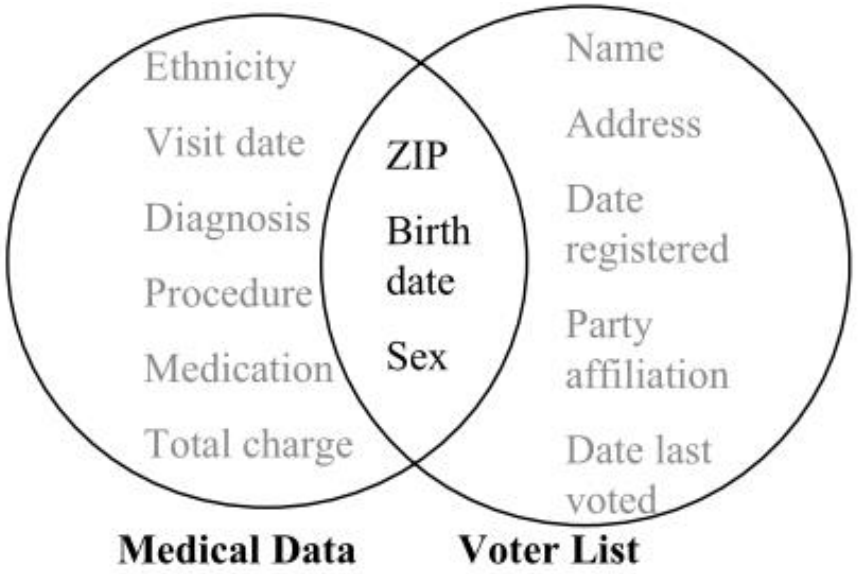
Sweeney a montré l'existence de quasi identifiants:

1- Des données médicales ont été "anonymisées" puis publiées

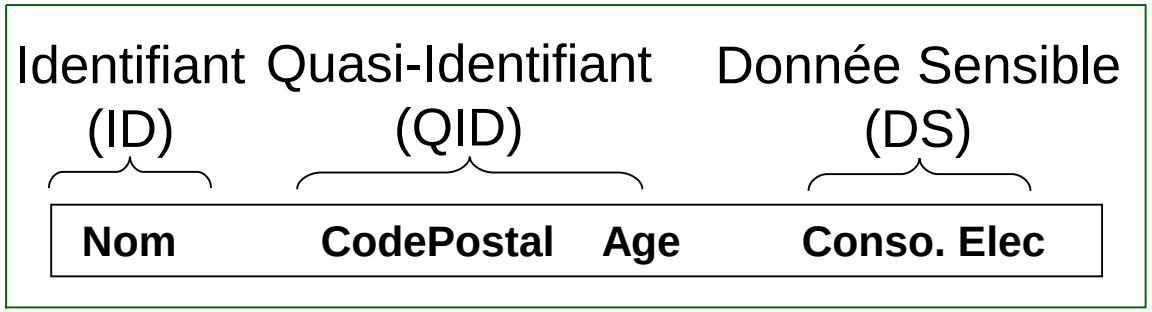
2- Une liste d'électeurs était disponible publiquement

→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} »



La naissance du *k*-anonymat

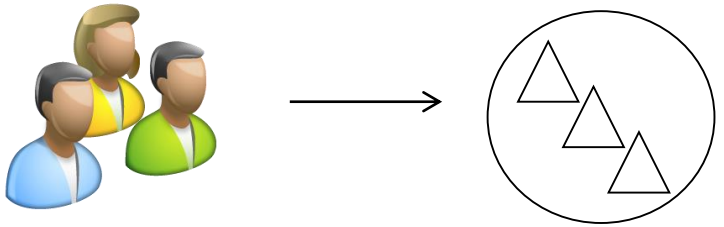


Pour chaque nuplet:

–Les identifiants doivent être retirés

–Le lien entre quasi-identifiant et données sensible doit être *obfusquée* mais doit rester globalement correcte

–Cette obfuscation est atteinte en permettant à chaque nuplet de correspondre à *k* DS différentes



Les garanties du *k*-anonymat

→ Probabilité de « Record linkage » = 1/k

(retrouver exactement quel n-uplet est lié à une valeur sensible de la base)

k-anonymat par Bucketization [Xiao, Tao]

- **Idée :** construire des groupes de *k* nuplets

Nom	CP	Age	C.E.
Sue	18000	22	50
Pat	69000	27	70
Bob	18500	21	90
Bill	18510	20	60
Dan	69100	26	70
Sam	69300	28	75

Données brutes

k-anonymat par Bucketization [Xiao, Tao]

- **Idée :** construire des groupes de *k* nuplets

Nom	CP	Age	C.E.
Sue	18000	22	50
Pat	69000	27	70
Bob	18500	21	90
Bill	18510	20	60
Dan	69100	26	70
Sam	69300	28	75

Données brutes

k-anonymat par Bucketization [Xiao, Tao]

- **Idée :** construire des groupes de *k* nuplets puis diviser ces informations en deux tables QID et DS.

Nom	CP	Age	C.E.
Sue	18000	22	50
Pat	69000	27	70
Bob	18500	21	90
Bill	18510	20	60
Dan	69100	26	70
Sam	69300	28	75

Données brutes

CP	Age	C.E.
18000	22	*G1*
18500	21	*G1*
69000	27	*G1*
18510	20	*G2*
69100	26	*G2*
69300	28	*G2*

Table QID

Groupe	C.E.
G1	50, 70, 90
G2	60, 70, 75

Table DS

Données 3-anonymes par *bucketisation*

k-anonymat par Bucketization [Xiao, Tao]

- **Avantage :** facile à mettre en œuvre et à implémenter
- **Désavantage :** l'utilité des données n'est pas claire

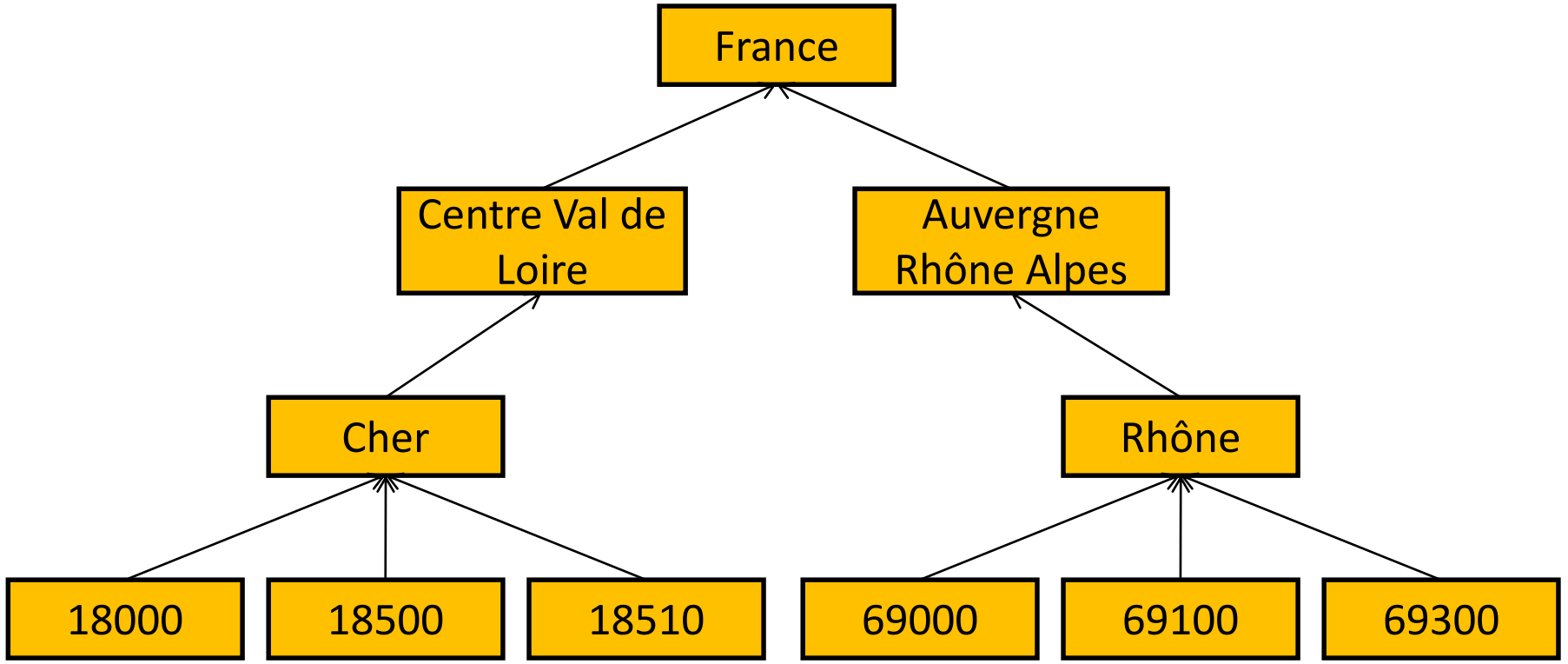
Ne pourrait-on pas *mieux* regrouper les données ?

k-anonymat par généralisation

[Sweeney]

Idée :

1.se doter pour chaque attribut du QID d'un arbre de généralisation



k-anonymat par généralisation

[Sweeney]

Idée :

1. se doter pour chaque attribut du QID d'un arbre de généralisation
2. Généraliser la valeur de certains attributs jusqu'à ce que tous les nuplets soient identiques à au moins *k-1* autres

Nom	**CP**	**Age**	**Conso Elec**
Sue	18000	22	50
Pat	69000	27	70
Bob	18500	21	90
Bill	18510	20	60
Dan	69100	26	70
Sam	69300	28	75

Données brutes

k-anonymat par généralisation

[Sweeney]

Idée :

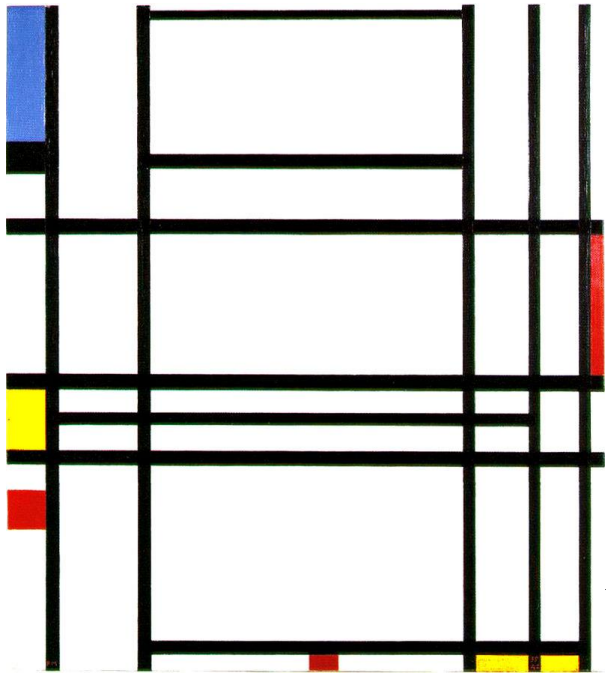
1. se doter pour chaque attribut du QID d'un arbre de généralisation
2. Généraliser la valeur de certains attributs jusqu'à ce que tous les nuplets soient identiques à au moins *k-1* autres

CP	Age	Conso Elec
Cher	[20-24]	50
Rhône	[25-29]	70
Cher	[20-24]	90
Cher	[20-24]	60
Rhône	[25-29]	70
Rhône	[25-29]	75

Données 3-anonymes

Implémentation : Algorithme de Mondrian

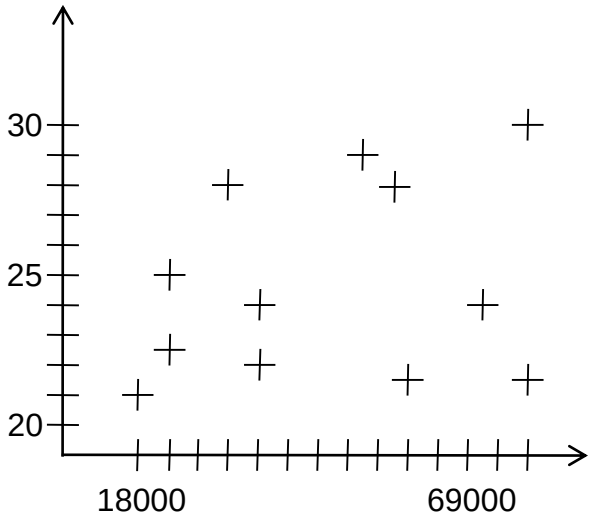
[LeFevre *et al.*]



Composition nr 10
Piet Mondrian

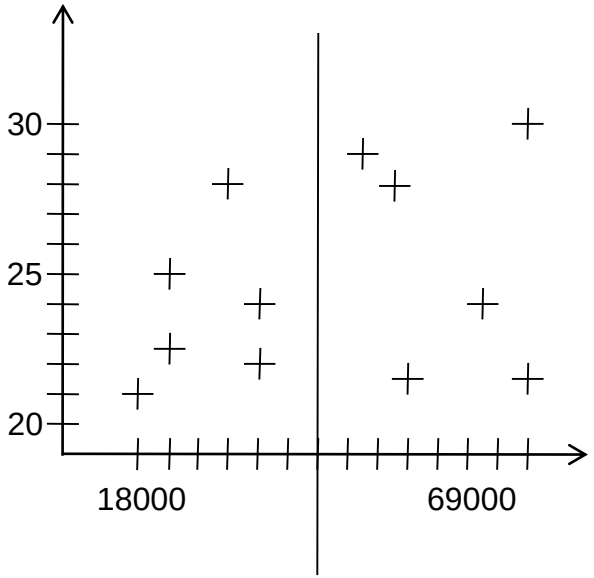
Implémentation : Algorithme de Mondrian

[LeFevre *et al.*]



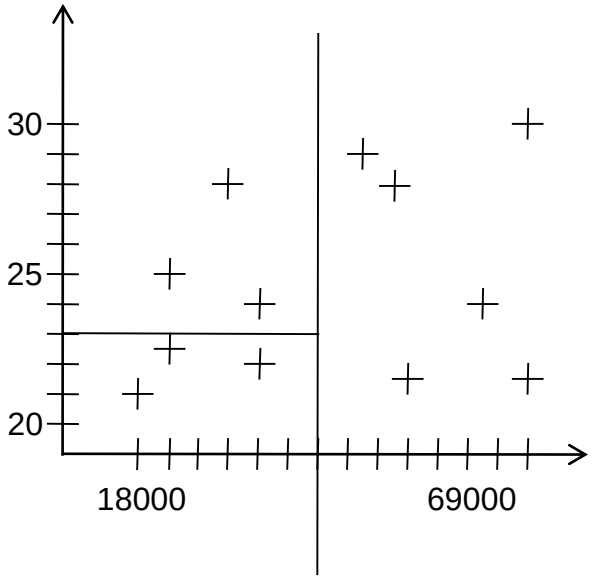
Implémentation : Algorithme de Mondrian

[LeFevre *et al.*]



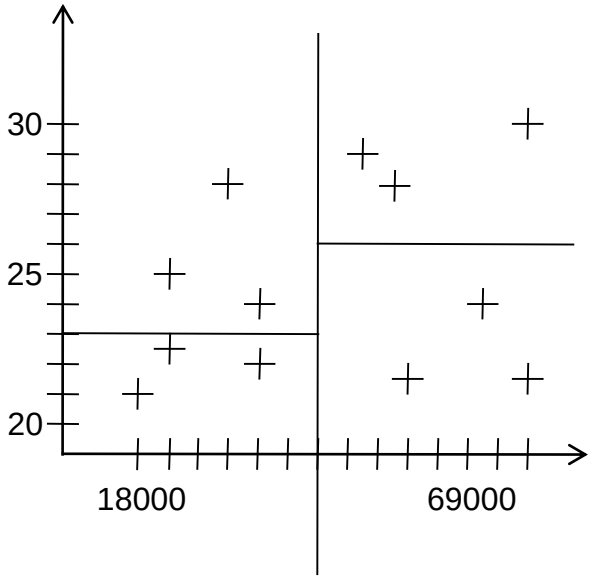
Implémentation : Algorithme de Mondrian

[LeFevre *et al.*]



Implémentation : Algorithme de Mondrian

[LeFevre *et al.*]



k-anonymat par généralisation

[Sweeney]

Cette technique permet de poser des requêtes SQL de type agrégat.

SELECT CP, AVG(Conso)

FROM T

GROUP BY CP

k-anonymat par généralisation

[Sweeney]

Cette technique permet de poser des requêtes SQL de type agrégat.

SELECT CP, AVG(Conso)

FROM T

GROUP BY CP

CP	C.E.
18000	50
69000	70
18500	90
18510	60
69100	70
69300	75

Données brutes

k-anonymat par généralisation

[Sweeney]

Cette technique permet de poser des requêtes SQL de type agrégat.

```
SELECT CP, AVG(Conso)
FROM T
GROUP BY CP
```

CP	C.E.
Cher	66.67
Rhône	71.67

Données anonymisées

Compromis “confidentialité” / utilité

/!\ Comment mesurer l’utilité ? /!\

Mise en application du *k-anonymat* avec ARX

ARX Data Anonymisation Tool

- Disponible en opensource et gratuit sur : https://arx.deidentifier.org/anonymization-tool/
- Projet porté par l'Université Technique de München

ARX

Data Anonymization Tool

ARX is a comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analyzing the usefulness of output data.

The software has been used in a variety of contexts, including commercial big data analytics platforms, research projects, clinical trial data sharing and for training purposes.

ARX is able to handle large datasets on commodity hardware and it features an intuitive cross-platform graphical user interface. You can find further information here, or directly proceed to our downloads section.

Evaluation du risque de réidentification

Attaque par le biais des QID

Que connait l'adversaire ?

- Métriques pour évaluer l'impact d'un attribut sur l'identification : Prosecutor Risk / Journalist Risk / Marketeer Risk : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2528029/

- Unicité de la population

Métriques d'évaluation du risque de réidentification

- D'après : Khaled El Emam, et Fida Kamal Dankar, *Protecting Privacy Using k-Anonymity*, in J Am Med Inform Assoc. 2008 Sep-Oct; 15(5): 627–637
- *Prosecutor risk :*
- *Re-identify a specific individual (known as the prosecutor re-identification scenario). The intruder (e.g., a prosecutor) would know that a particular individual (e.g., a defendant) exists in an anonymized database and wishes to find out which record belongs to that individual.*

Métriques d'évaluation du risque de réidentification

- D'après : Khaled El Emam, et Fida Kamal Dankar, *Protecting Privacy Using k-Anonymity*, in J Am Med Inform Assoc. 2008 Sep-Oct; 15(5): 627–637
- *Journalist risk :*
- *Re-identify an arbitrary individual (known as the journalist re-identification scenario). The intruder does not care which individual is being re-identified, but is only interested in being able to claim that it can be done. In this case the intruder wishes to re-identify a single individual to discredit the organization disclosing the data.*

Métriques d'évaluation du risque de réidentification

- D'après : Khaled El Emam, et Fida Kamal Dankar, *A Method for Evaluating Marketer Re-identification Risk* , in PAIS'10 (voir : https://www.researchgate.net/profile/Fida_Dankar/publication/220774036_A_method_for_evaluating_marketer_re-identification_risk/links/55e6827b08aec74dbe74ea64.pdf
- *Marketeer risk :*
- *An intruder wishes to re-identify as many records as possible in the disclosed database. We assume that the intruder lacks any additional information apart from the matching quasi-identifiers.*

Modele

- Base de données privée : U avec |U|=n
- Base de données connue de l'attaquant : D et |D|=N
- X l'ensemble de toutes les classes d'équivalence possibles
- Z = $\{z_i\}$ une classe d'équivalence
- J le nombre de classes d'équivalences total possible, ~J le nombre de vraies classes d'équivalence
- f_j le nombre d'enregistrements de la classe d'équivalence j dans U
- F_j le nombre d'enregistrements de la classe d'équivalence j dans D

Calcul du risque

Theorem 1. The expected proportion of U records that can be disclosed in a random mapping from U to D is.

$$\lambda = \sum_{j=1}^{\tilde{J}} \frac{f_j / F_j}{n} \qquad \text{.....(1)}$$

Note that if $n = N$ then $\lambda = \frac{\tilde{J}}{N}$.

Calcul du risque

$$R_p = 1 \Big/ \min_j \left(f_j \right)$$

$$R_J = 1 \Big/ \min_j \left(F_j \right)$$

Mise en application de la qualité de l'anonymisation avec ARX

Méthodes statistiques

Quelques techniques compatibles avec la méthode de perturbation locale

Méthode de réponse aléatoire

Cadre : réponse Oui/Non

- Donner une probabilité *p* de dire la vérité à chaque individu et de mentir avec une probabilité *(1-p)*
- *En général : p=0.5 + ε*
- Estimateur:
  - Soir π proportion de la populaiton pour laquelle la vraie réponse est « Oui »
  - La proportion attention de « Oui » est :
    $P(Oui) = (\pi * p) + (1 - \pi)*(1 - p)$
    → $\pi = [P(Oui) - (1 - p)] / (2p - 1)$
  - Si m/n individus ont répondu « oui », π_{est} estime π vaut :
    $\pi_{est} = [m/n - (1 - p)] / (2p - 1)$

Differential Privacy

L’approche “à la mode”

Differential privacy

Dwork 2006, *Differential Privacy* (ICALP)

- Le problème principal du *k*-anonymat est que la sécurité depend des connaissances de l'attaquant.
- Un *framework* a été proposé en 2006 par Dwork. Il permet de quantifier le risque de participation dans une base de données par rapport à un algorithme d'anonymisation

On dit qu'un algorithme (aléatoire) satisfait la contrainte (ε,δ)-differential privacy si :

-Pour toute paire de bases de données D_1 et D_2 (dites adjacentes) qui ne diffèrent que par la présence ou non d'un individu

-Pour tout résultat Ω de l'algorithme,

Il existe ε tel que :

$$Pr[A(D_1) = \Omega] \leq e^{\varepsilon} Pr[A(D_2)=\Omega]+\delta$$

Mécanisme de Laplace

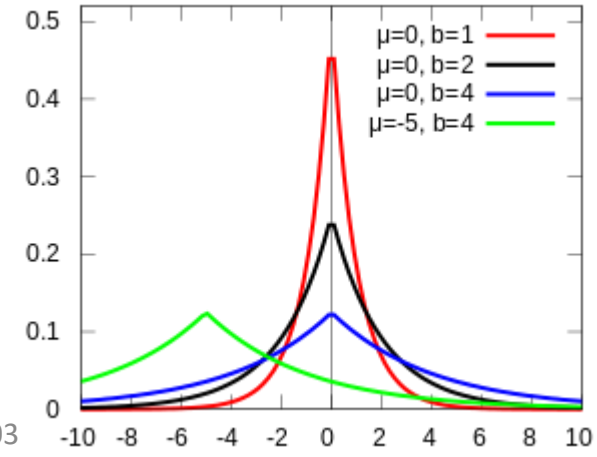
Dwork a définit le "*mécanisme de Laplace*" qui dit que si on bruite une fonction avec une valeur tirée d'une distribution de Laplace, centrée sur 0 et d'échelle $\Delta f/\varepsilon$ alors ce mécanisme respecte la contrainte de differential privacy

Definition 4 (ℓ_1-sensitivity). *The ℓ_1-sensitivity of a function $f : \mathbb{N}^{|\chi|} \rightarrow \mathbb{R}^k$ is :*

$$\Delta f = \max_{\substack{x,y \in \mathbb{N}^{|\chi|} \\ ||x-y||_1 = 1}} ||f(x) - f(y)||_1$$

Definition 7 (Laplace Distribution). *The Laplace Distribution with scale b is the distribution with probability density function*

$$Lap(x|b) = \frac{1}{2b} \exp(-\frac{|x|}{b})$$



Mécanisme d'Exponentiation

Ce mécanisme classe tous les outputs potentiels $r \in R$ pour un jeu de données initial D selon une fonction de score (ie à valeurs réelles).

Puis elle choisit aléatoirement l'un de ces outputs selon une distribution qui donne une plus grande probabilité aux outputs ayant le meilleur score.

Definition 3 (Exponential mechanism [44]). For any function $s : (\mathcal{D}_m \times \mathcal{R}) \to \mathbb{R}$, the *exponential mechanism* $\mathcal{E}_s^\epsilon(D, \mathcal{R})$ chooses and outputs an element $r \in \mathcal{R}$ with probability proportional to $\exp\left(\frac{s(D,r)\epsilon}{2\Delta s}\right)$, where the *sensitivity* Δs of the function s is defined as

$$\Delta s := \max_{r \in \mathcal{R}} \max_{D_1, D_2 \in \mathcal{D}_m : |D_1 \oplus D_2| = 1} |s(D_1, r) - s(D_2, r)|.$$

It can be seen that it is important to use score functions which assign higher scores to outputs with higher quality while having a low sensitivity. The privacy guarantees provided are as follows:

Theorem 2. *For any function* $s : (\mathcal{D}_m \times \mathcal{R}) \to \mathbb{R}$, $\mathcal{E}_s^\epsilon(D, \mathcal{R})$ *satisfies* ϵ*-differential privacy [44].*

Séries temporelles :
Théorème de composition [Dwork 06]

Theorem 3.14. Let $\mathcal{M}_1 : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathcal{R}_1$ be an ε_1-differentially private algorithm, and let $\mathcal{M}_2 : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathcal{R}_2$ be an ε_2-differentially private algorithm. Then their combination, defined to be $\mathcal{M}_{1,2} : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathcal{R}_1 \times \mathcal{R}_2$ by the mapping: $\mathcal{M}_{1,2}(x) = (\mathcal{M}_1(x), \mathcal{M}_2(x))$ is $\varepsilon_1 + \varepsilon_2$-differentially private.

Corollary 3.15. Let $\mathcal{M}_i : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathcal{R}_i$ be an $(\varepsilon_i, 0)$-differentially private algorithm for $i \in [k]$. Then if $\mathcal{M}_{[k]} : \mathbb{N}^{|\mathcal{X}|} \rightarrow \prod_{i=1}^{k} \mathcal{R}_i$ is defined to be $\mathcal{M}_{[k]}(x) = (\mathcal{M}_1(x), \ldots, \mathcal{M}_k(x))$, then $\mathcal{M}_{[k]}$ is $(\sum_{i=1}^{k} \varepsilon_i, 0)$-differentially private.

Local Differential Privacy

Jordan & Wainwright [2013]

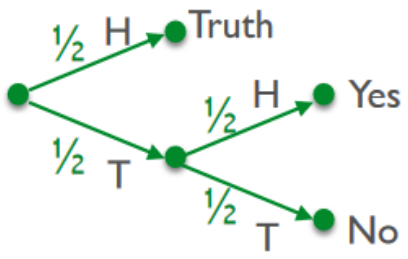
Definition Let $\mathcal{X}$ be a set of possible values and $\mathcal{Y}$ the set of noisy values. A mechanism $\mathcal{K}$ is ε-locally differentially private (ε-LDP) if for all $x_1, x_2 \in \mathcal{X}$ and for all $y \in \mathcal{Y}$

$$P[\mathcal{K}(x) = y] \leq e^{\varepsilon} \, P[\mathcal{K}(x') = y]$$

or equivalently, using the conditional probability notation:

$$p(y \mid x) \leq e^{\varepsilon} \, p(y \mid x')$$

For instance, the Randomized Response protocol is (log 3)-LPD



x \ y	yes	no
yes	¾	¼
no	¼	¾

Algorithme « *Randomized Response* » avec ε = 0.25

Source : C. Palamidessi

Algorithme SafePub (ARX)

Bild, Kuhn, Passer [2018]

Input: Dataset D, Parameters ϵ_{anon}, ϵ_{search}, δ, *steps*
Output: Dataset S

1: Draw a random sample D_s from D ▷ (ϵ_{anon})
2: Initialize set of transformations G
3: **for** (Int $i \leftarrow 1, ..., steps$) **do**
4: Update G
5: **for** $(g \in G)$ **do**
6: Anonymize D_s using g ▷ $(\epsilon_{anon}, \delta)$
7: Assess quality of resulting data
8: **end for**
9: Probabilistically select solution $g \in G$ ▷ (ϵ_{search})
10: **end for**
11: **return** Dataset D_s anonymized using the best solution selected in Line 9 ▷ $(\epsilon_{anon}, \delta)$

Fig. 4. High-level design of the SafePub mechanism. The search strategy is implemented by the loop in lines 3 to 10.

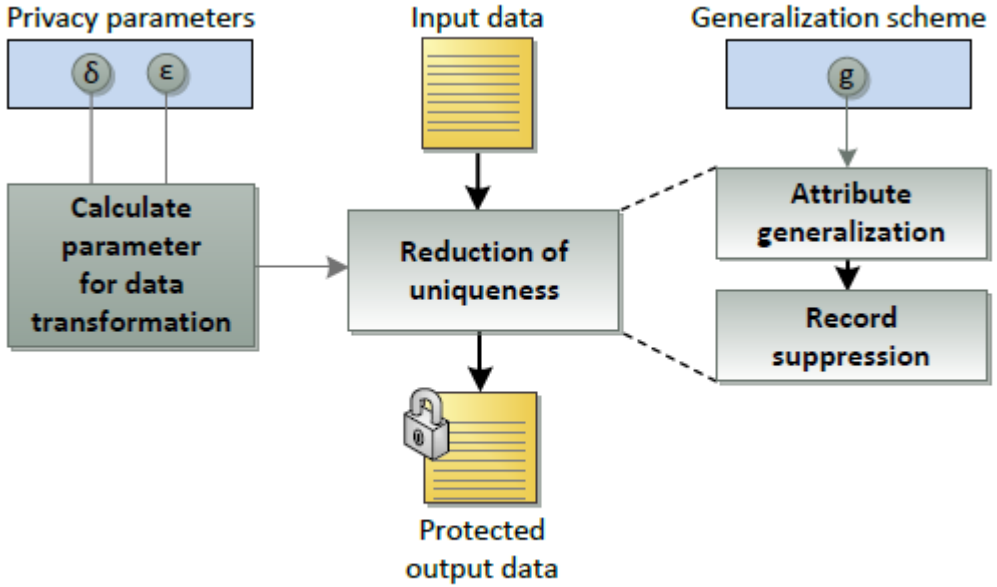


Fig. 5. Overview of the anonymization operator.

Idée : Choix aléatoire de paramètres de *k-anonymisation*

Conclusion

L’anonymisation est un compromis

- Il est capital
  - D’être capable d’évaluer le risque
  - D’être capable d’évaluer l’utilité des données une fois anonymisées
- Quel modèle utiliser ?
  - La technique de *differential privacy* assez à la mode ces dernières années dans la communauté de recherche en informatique
  - Les techniques « syntaxiques » restent une approche souvent acceptable de réduction de risque (tout comme la pseudonymisation)
- Il n’est pas possible de donner de garanties absolues !
  - Le RGPD demande une obligation de moyens