

Techniques d'anonymisation

Benjamin NGUYEN

benjamin.nguyen@insa-cvl.fr

4AMRI

Plan

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

Qu'est ce que l'anonymat ?

Une réponse juridique : GDPR et données anonymes

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. 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. 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.

GDPR, données, et personne physique identifiable *i.e. pas anonyme*

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. (...)

1. Conditions d'application du règlement
2. Pseudonymisation

GDPR et Réidentification

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 l’ancienne loi “informatique et libertés” (1978)

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.

1. Droits de traitement des données anonymes : Le RGPD s'applique à des données personnelles. Par définition les données anonymes ne sont pas personnelles.

Note : l'anonymisation est un traitement, il doit donc être précisé dans les finalités du traitement lors de la collecte de données, et le consentement reçu.

Loi "Informatique et Libertés", art. 11-3

*(la CNIL) donne un avis sur la conformité aux dispositions de la présente loi des projets de règles professionnelles et des produits et procédures tendant à la protection des personnes à l'égard du traitement de données à caractère personnel, **ou à l'anonymisation de ces données**, qui lui sont soumis;*

Legitimité d'un processus d'anonymization : Avis du WP29 (2014)

Accordingly, the Working Party considers that anonymisation as an instance of further processing of personal data can be considered to be compatible with the original purposes of the processing but only on condition the anonymisation process is such as to reliably produce anonymised information in the sense described in this paper.

Note : Le groupe de travail dit “WP article 29” était un groupe de travail européen regroupant l'ensemble des CNILs européennes jusqu'en mai 2018. Il a été remplacé par le EDPB (European Data Protection Board)

Pourquoi anonymiser ?

- **Juridique** : Problématique de la gestion de données personnelles (cf. RGPD)

Des données anonymes ne tombent pas sous le coup du RGPD, le processus d'anonymisation, si

- **Vie privée / éthique** : Aucune raison de conserver des informations (identifiantes) si cette information n'est pas nécessaire
- **Sécurité** : Minimiser le risque lors de la conservation ou du traitement de données
- **Collaborer** : Fournir de véritables données exploitées par d'autres (e.g. modèles d'IA)

Linky



Ce qui est demandé par la loi

Intérêt de l'utilisateur

*/!\ Ethique et qualité des données ? (cf HDH)
Proportionnalité ?*

Intérêt du fournisseur de service

AEDP & EDPB “10 misunderstandings related to anonymization” (4/2021)

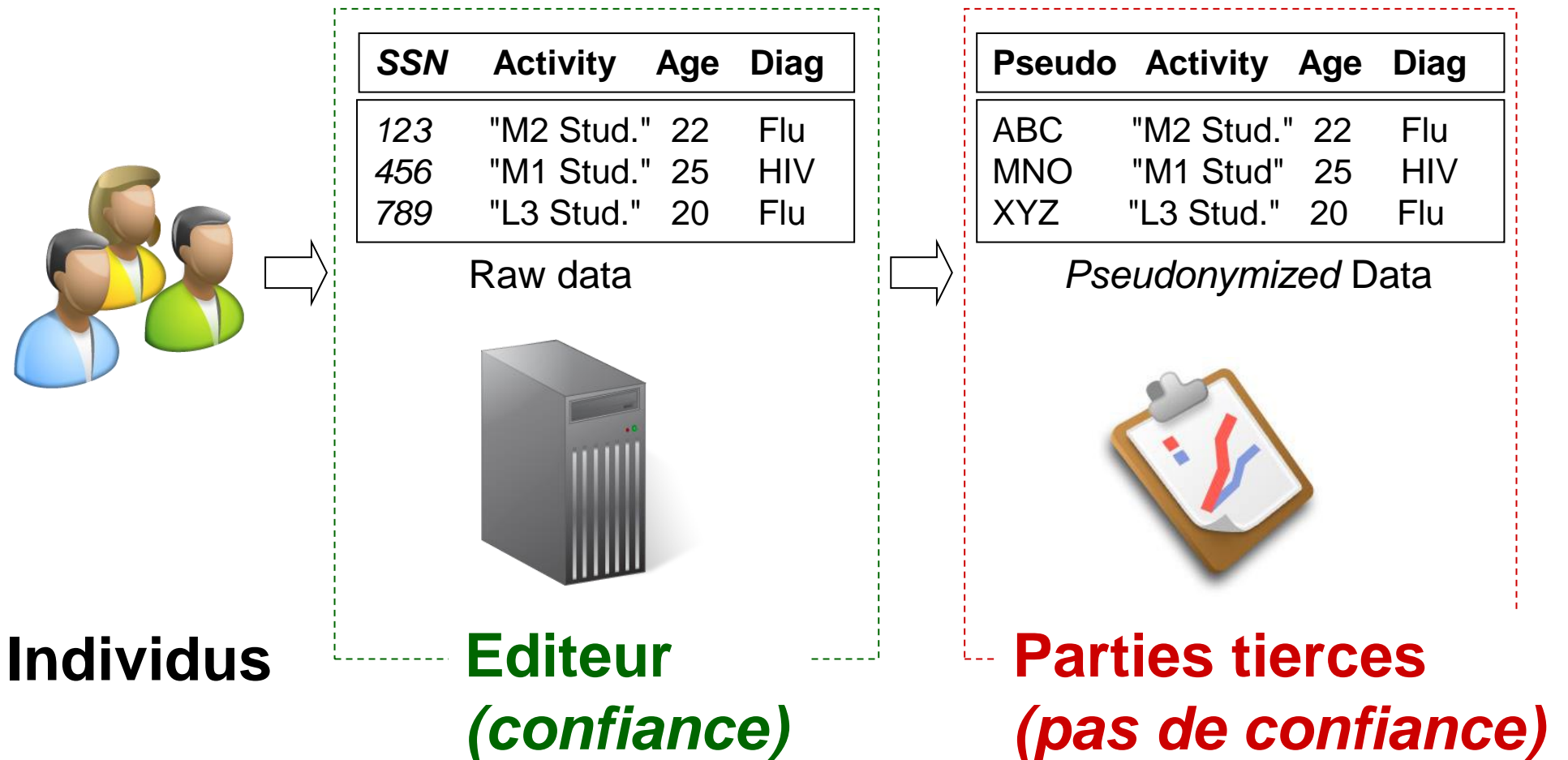
1. Pseudonymization is anonymization
2. Encryption is anonymization
3. Anonymization is always possible
4. Anonymization is forever
5. Anonymization always reduces the risk of reidentification to zero
6. Anonymization is a binary concept and cannot be measured
7. Anonymization can be fully automated
8. Anonymization makes the data useless
9. Following an anonymization process that others used successfully will lead our organisation to similar results
10. There is no risk and no interest of finding out who this data relates to

La Pseudonymisation

Ce que l'anonymisation n'est pas

La *pseudonymisation* :

Ce que l'anonymisation n'est pas



Attaque sur la pseudonymisation

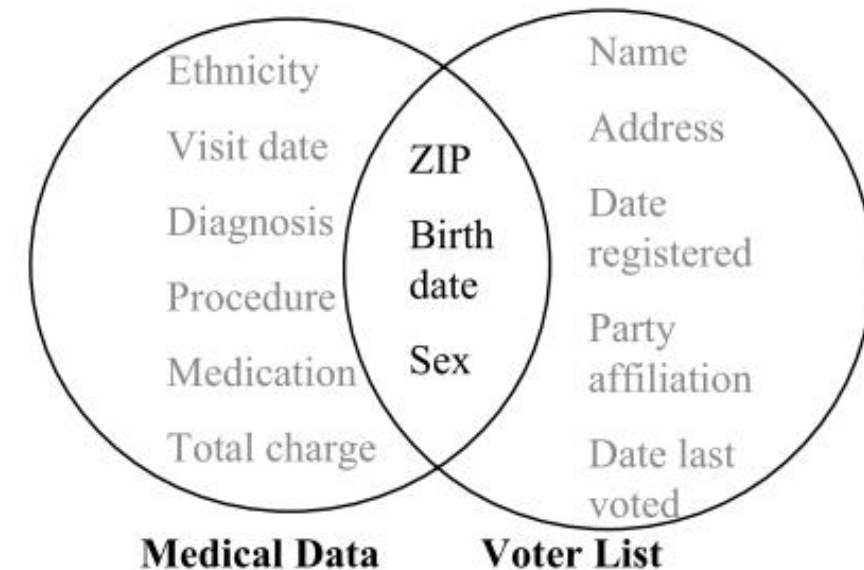
Sweeney 2002, *k*-anonymity: a model for protecting privacy (IJUFK-BS)

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} »



Limites de la pseudonymisation

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

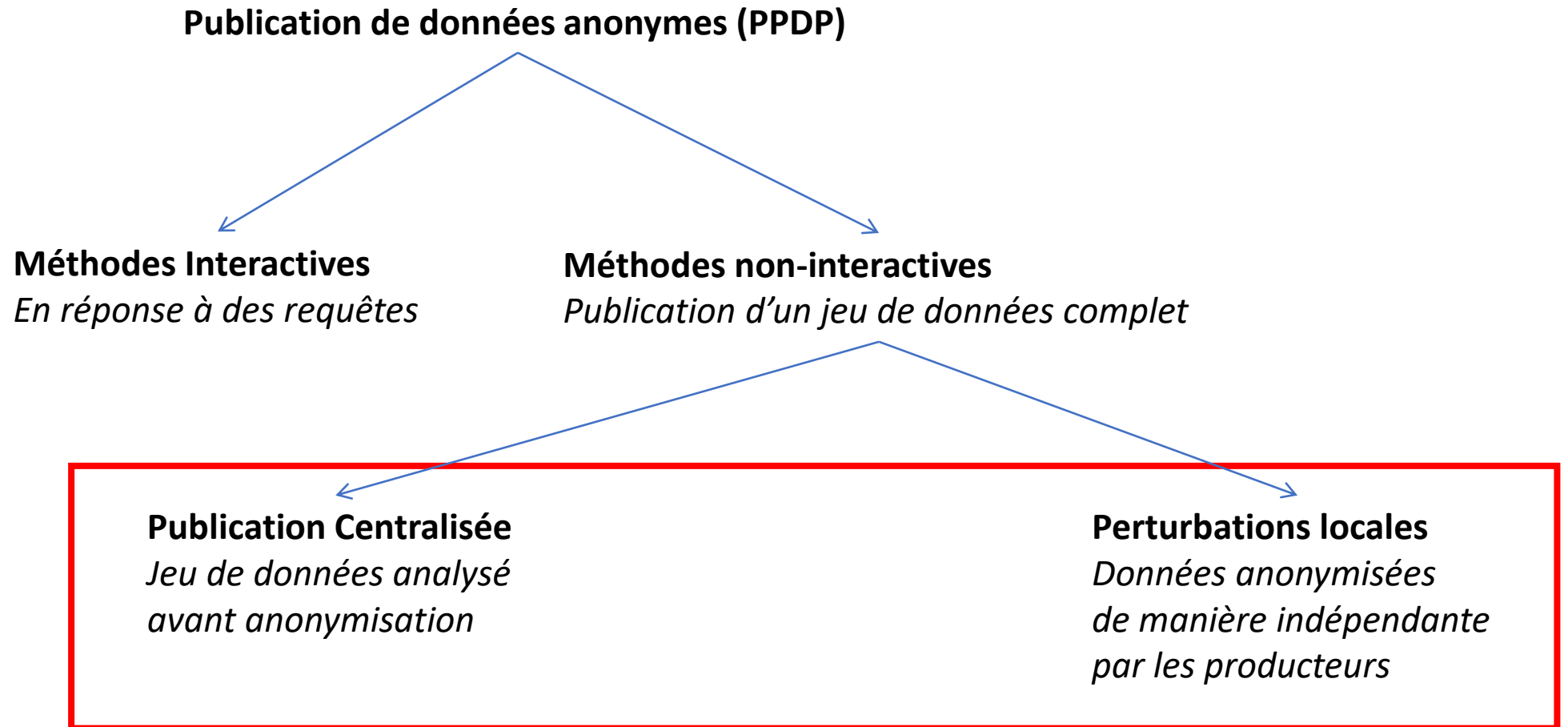
Faut-il pseudonymiser ?

- Oui, mais ça ne suffit pas !
- L'UE recommande néanmoins de pseudonymiser comme mesure préventive complémentaire à l'anonymisation.

Architecture d'anonymisation

Comment anonymiser ?

Classification des approches



Architecture classique d'anonymisation en *publication centralisée*

Contexte

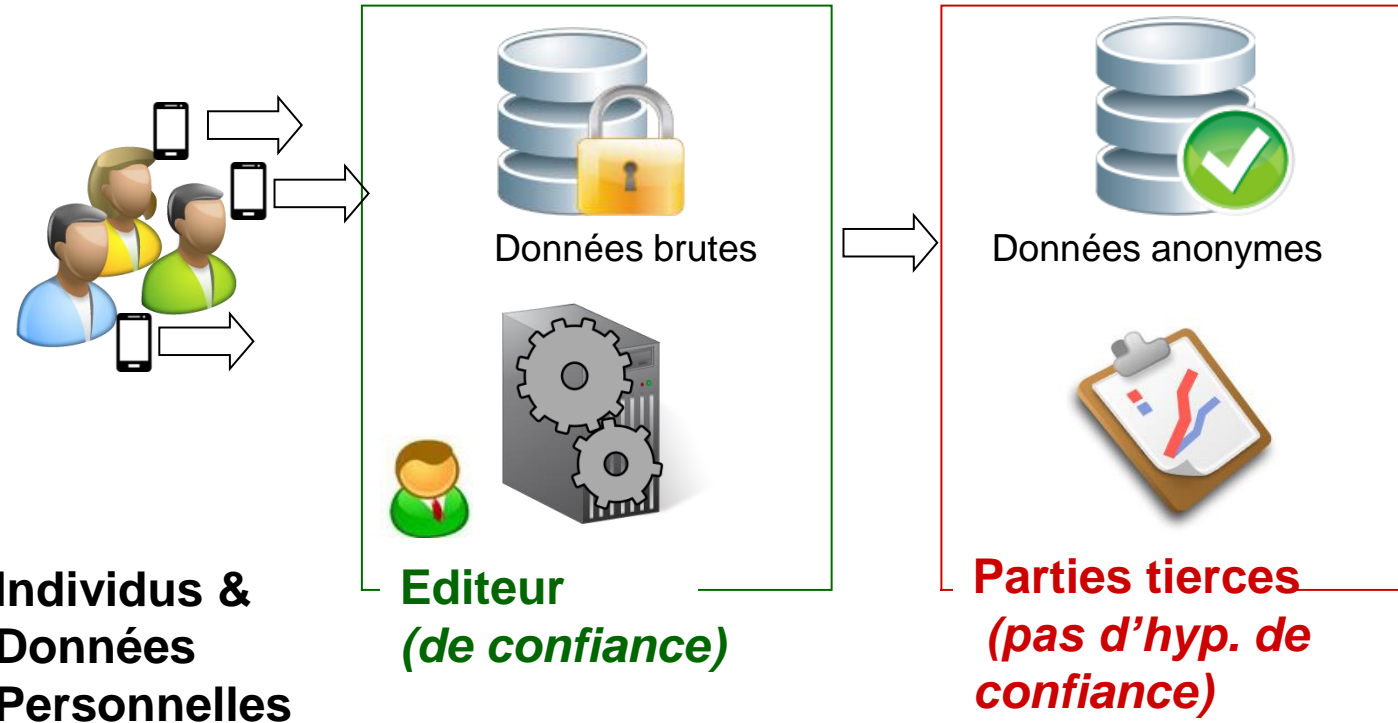
Des données personnelles, sensibles, issus de mesures, de capteurs, de questionnaires, etc

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

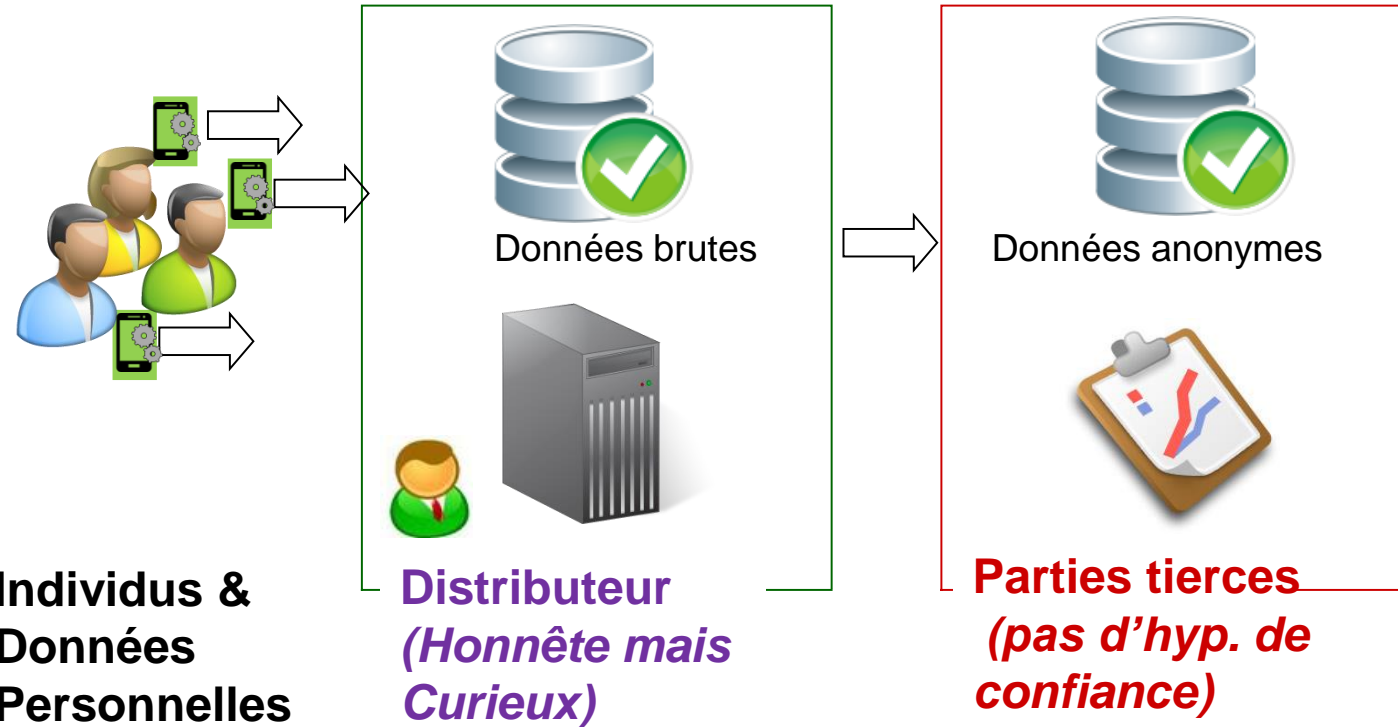
Des données personnelles, sensibles, issus de mesures, de capteurs, de questionnaires, etc

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 d'un processus d'anonymisation

- **Une définition et métrique de la “privacy” qui répondent à la question :**
Quelle protection proposer (et comment la mesurer) ?
- **Une définition et métrique d'utilité qui répondent à la question :**
Comment mesurer la perte d'utilité dans le processus utilisant des données anonymes au lieu des données exactes ?
- **Un algorithme d'anonymisation qui répond à la question :**
Comment protéger les données 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

Une première définition de données anonymes

Attaque sur la pseudonymisation

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

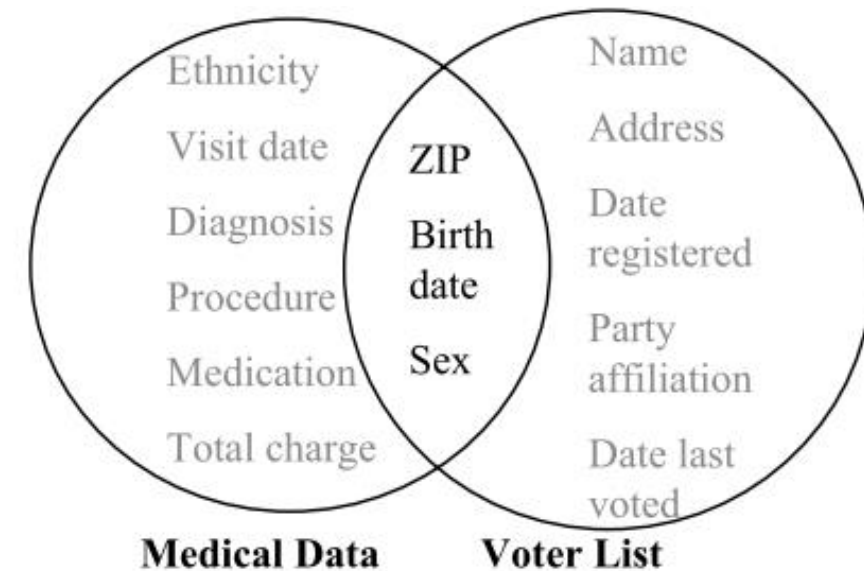
Sweeney a montré l'existence de quasi identifiants:

1- Des données médicales ont été “anonymisées” puis publiées

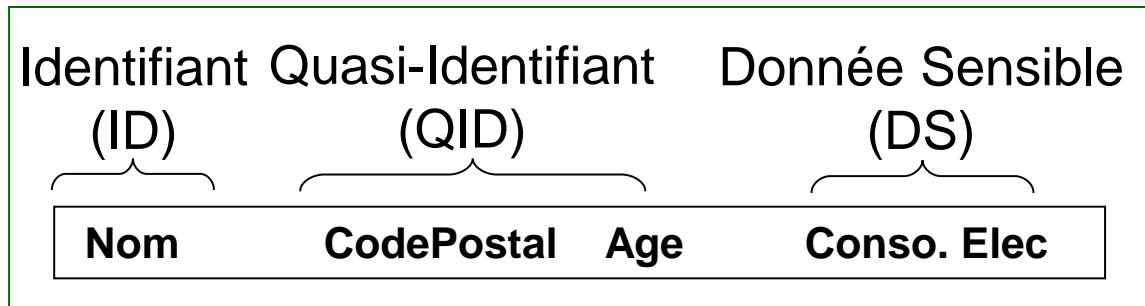
2- Une liste d'électeurs était disponible publiquement

→ 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} »

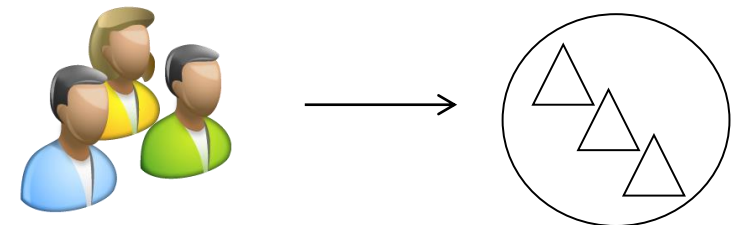


La naissance du k -anonymat



Pour chaque nuplet:

- Les identifiants doivent être retirés
- Le lien entre quasi-identifiant et données sensible doit être *obfusqué* mais doit rester globalement correct
- 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

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

Données brutes

k -anonymat par Bucketization [Xiao, Tao]

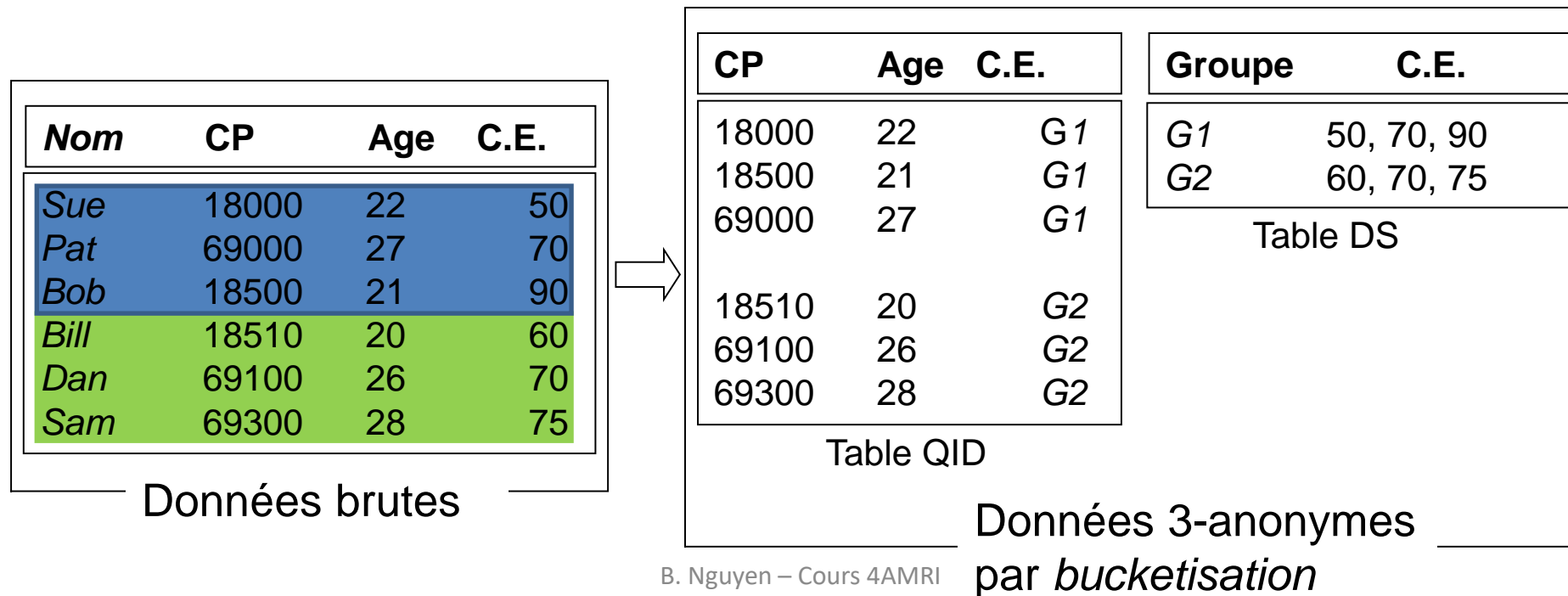
- **Idée** : construire des groupes de k nuplets

<i>Nom</i>	<i>CP</i>	<i>Age</i>	<i>C.E.</i>
<i>Sue</i>	18000	22	50
<i>Pat</i>	69000	27	70
<i>Bob</i>	18500	21	90
<i>Bill</i>	18510	20	60
<i>Dan</i>	69100	26	70
<i>Sam</i>	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.



k -anonymat par Bucketization [Xiao, Tao]

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

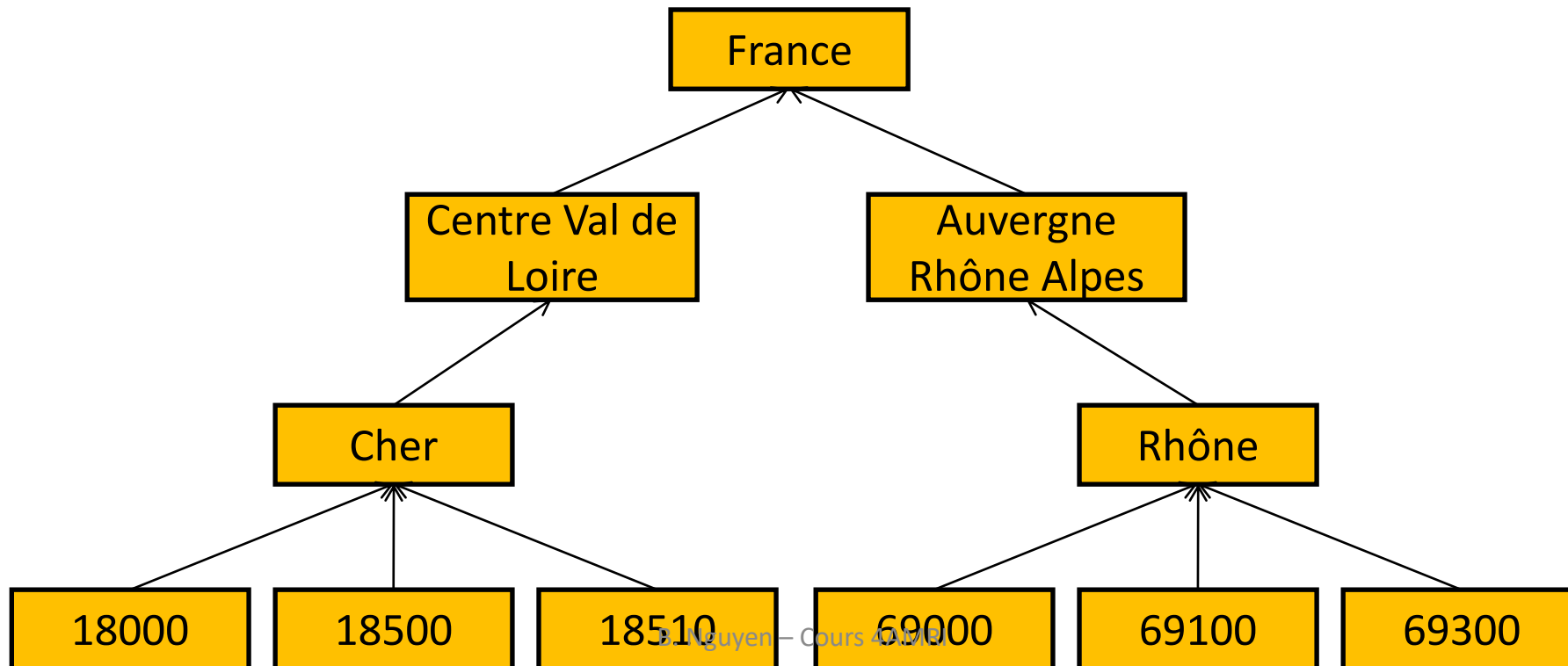
Ne pourrait-on pas *mieux* regrouper les données ?

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

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

B. Nguyen – Cours 4AMRI

Données brutes

k -anonymat par généralisation

[Sweeney]

Idée :

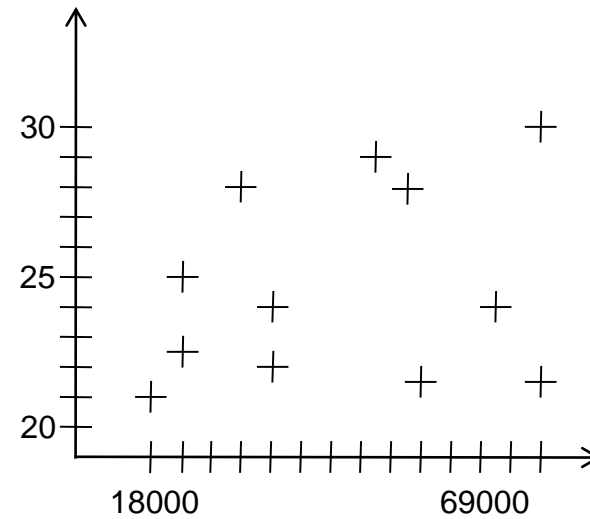
1. se doter pour chaque attribut du QID d'un arbre de généralisation
2. Généraliser la valeur de certains attributs jusqu'à ce que tous les nuplets soient identiques à au moins $k-1$ autres

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

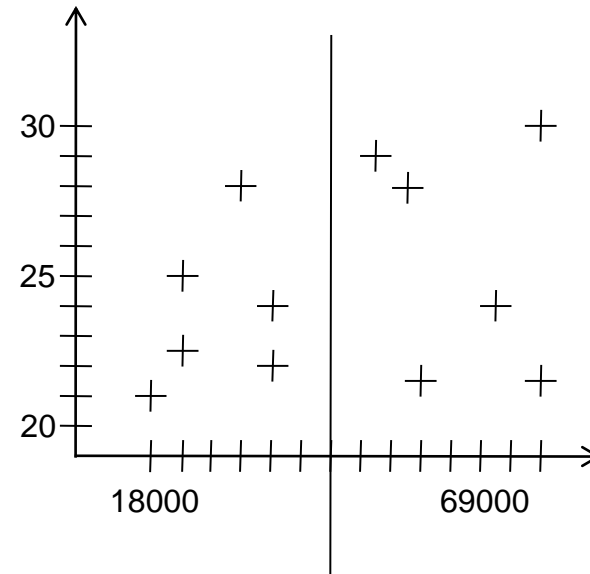
B. Nguyen – Cours 4AMRI

Données 3-anonymes

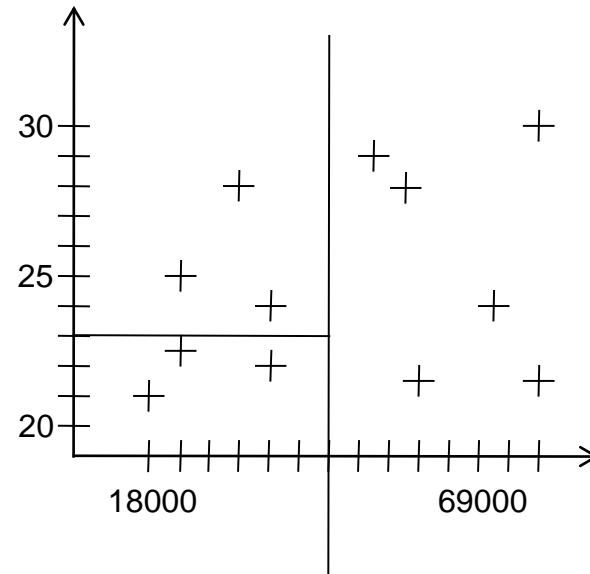
Implémentation : Algorithme de Mondrian [LeFevre *et al.*]



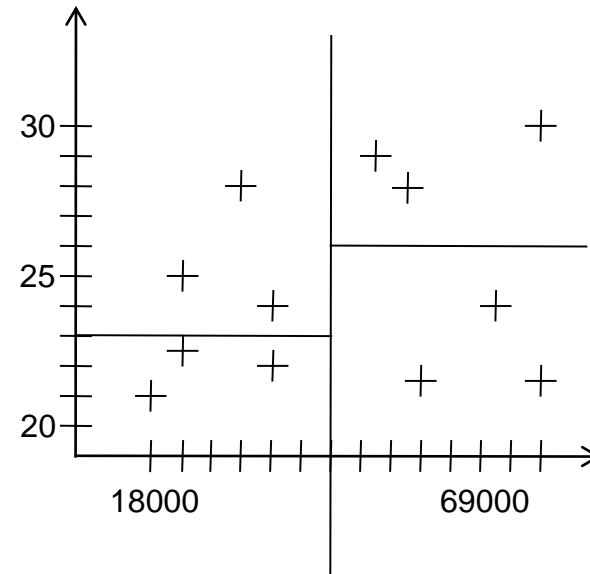
Implémentation : Algorithme de Mondrian [LeFevre *et al.*]



Implémentation : Algorithme de Mondrian [LeFevre *et al.*]

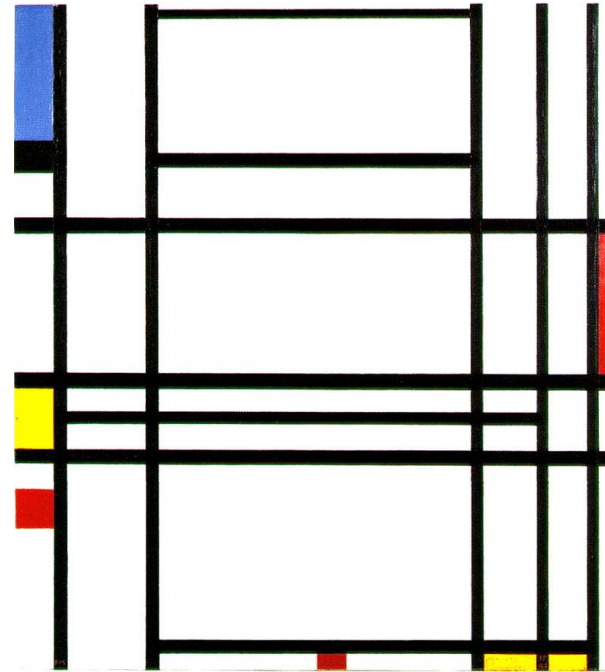


Implémentation : Algorithme de Mondrian [LeFevre *et al.*]



Implémentation : Algorithme de Mondrian

[LeFevre *et al.*]



Composition nr 10
Piet Mondrian

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
```

Compromis “confidentialité” / utilité
/!\ Comment mesurer l’utilité ? /!\

CP	C.E.
Cher	66.67
Rhône	71.67

Données anonymisées

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

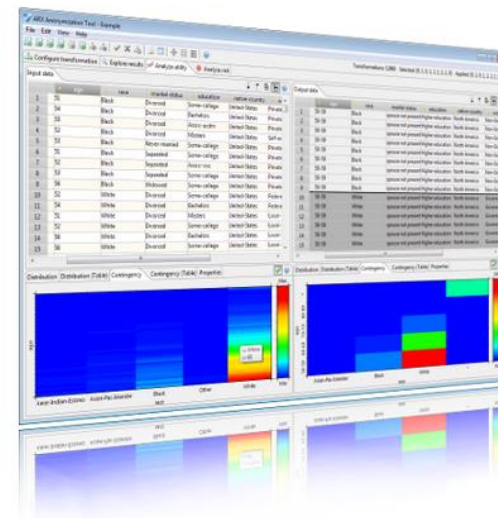
ARX

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 connaît 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.*



Former Governor
Bill Weld
Of Massachusetts

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.*



Thelma Arnold
Aka user #4417749
AoL Search

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, $\sim J$ 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.

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

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

Calcul du risque

$$R_p = \frac{1}{\min_j (f_j)}$$

$$R_j = \frac{1}{\min_j (F_j)}$$

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

Techniques classiques d'anonymisation

Des multiples modèles d'anonymat : L-diversité, T-closeness, PRAM, Echantillonnage, Differential Privacy ...

Principal problème du k -anonymat

Et si les valeurs sensibles sont toutes les mêmes ?

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

Données brutes

CP	Age	C.E.
Cher	[20-24]	50
Rhône	[25-29]	70
Cher	[20-24]	90
Cher	[20-24]	60
Rhône	[25-29]	70
Rhône	[25-29]	70

Données anonymes

→ La consommation électrique d'un habitant du Rhône est 70kWh / mois !

L-diversité

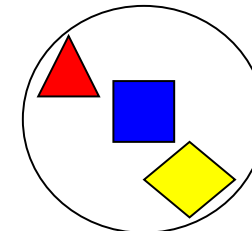
[Machanavajjhala *et al.* 06]

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

Données brutes

CP	Age	C.E.
France	[20-29]	50
France	[20-29]	70
France	[20-29]	90
France	[20-29]	60
France	[20-29]	70
France	[20-29]	70

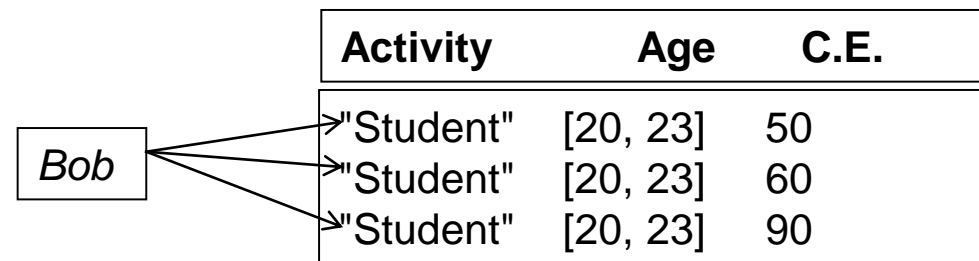
Données anonymes
et diverses



*Obtenu en plaçant des contraintes
sur les classes d'équivalence.*

Les garanties de la l -diversity

- Un individu dont le QID appartient à une classe et qui a participé à la release peut être associé à n'importe laquelle des L valeurs sensibles avec une probabilité donnée
 - Par exemple, Bob peut être associé à n'importe laquelle des valeurs {Flu, HIV, Cancer} avec ici une probabilité identique
- → probabilité d'Attribute linkage = $1/L$



Intuition

- Il faut s'assurer que chaque groupe k -anonyme est également assez « divers » (varié)
 - Chaque classe doit être associée à au moins L valeurs sensibles “bien représentées”
 - “bien représentée” a une définition variable
- **Conséquences :**
 - perte de précision 😞
 - gain d'anonymat 😊

t -closeness ou comment aller trop loin ?

- Intuition: Dans chaque classe, la distribution des données sensibles doit être proche de la distribution générale, avec au plus une variation d'un facteur t
- → La connaissance a posteriori de l'adversaire est la même que sa connaissance a priori (qui contient la connaissance globale de la distribution)
- Example:

CP	Age	C.E.
Cher	[20-24]	40
Cher	[20-24]	70
Cher	[20-24]	70
Rhône	[25-29]	40
Rhône	[25-29]	70
Rhône	[25-29]	70

Données anonymes

Mise en application de la *L-diversité* avec ARX

Méthodes statistiques

Quelques techniques compatibles avec la méthode de perturbation locale

Méthode de réponse aléatoire Compatible « Local Differential Privacy »

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* : $p=0.5 + \varepsilon_{RR}$
- Estimateur:
 - Soit π proportion de la population pour laquelle la vraie réponse est « Oui »
 - La proportion attendue de « Oui » est :
$$P(\text{Oui}) = (\pi * p) + (1 - \pi) * (1 - p)$$
$$\rightarrow \pi = [P(\text{Oui}) - (1 - p)] / (2p - 1)$$
 - Si m/n individus ont répondu « oui », π_{est} estime π vaut :
$$\pi_{\text{est}} = [m/n - (1 - p)] / (2p - 1)$$

Post-Randomization Matrix (PRAM)

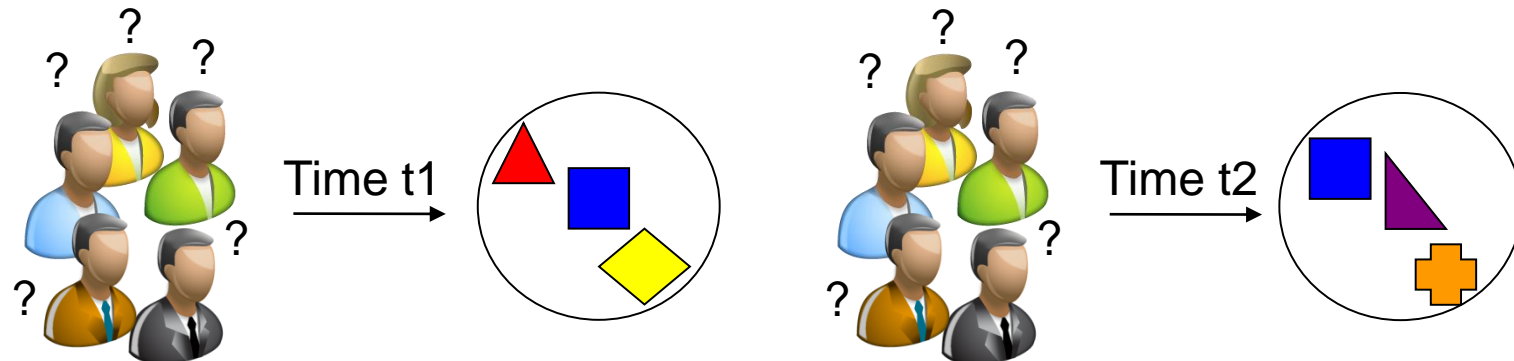
Peut être vu comme de la local differential privacy

- Utilisée dans le logiciel Mu-Argus (Eurostat)
- On définit une matrice de probabilité de transition d'une valeur vers une autre, puis on applique ces probas.

	Grippe	Covid
Grippe	0.75	0.25
Covid	0.1	0.9

Méthode d'échantillonnage

- Donner une probabilité de participation $P_{participation}$ à chaque individu et dans chaque version.
 - Avantages :
 - On ne peut pas être sûr de savoir qui a participé dans la version → sécurité ?
 - Inconvénients :
 - Difficulté de faire des études transversales
 - Une donnée publiée est forcément vraie → risque de réidentification



Differential Privacy

Garanties formelles

Local Differential Privacy

Jordan & Wainwright [2013]

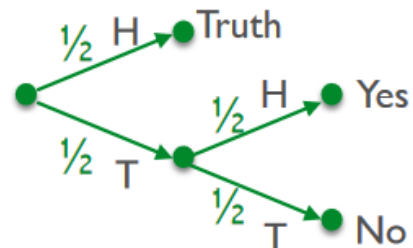
Definition Let \mathcal{X} be a set of possible values and \mathcal{Y} the set of noisy values. A mechanism \mathcal{K} is ϵ -locally differentially private (ϵ -LDP) if for all $x_1, x_2 \in \mathcal{X}$ and for all $y \in \mathcal{Y}$

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

or equivalently, using the conditional probability notation:

$$p(y | x) \leq e^\epsilon p(y | x')$$

For instance, the Randomized Response protocol is $(\log 3)$ -LDP



		y	
		yes	no
x	yes	3/4	1/4
	no	1/4	3/4

**Algorithme « Randomized Response »
avec $\epsilon_{RR} = 0.25$**

Differential privacy

Dwork 2006, *Differential Privacy* (ICALP)

- Le problème principal du k -anonymat est que la sécurité dépend 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 :

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

Mécanisme de Laplace

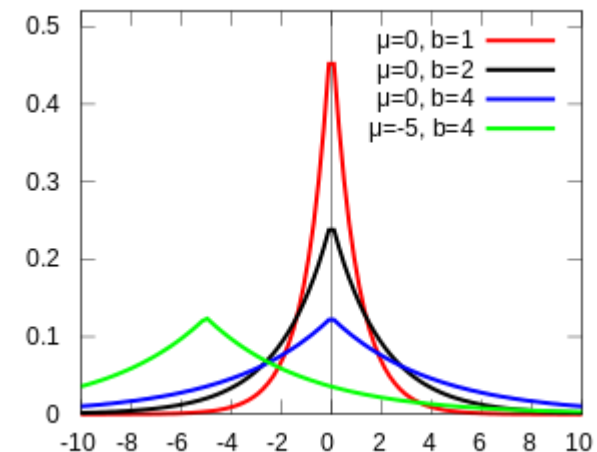
Dwork a défini le “*mécanisme de Laplace*” qui dit que si on bruite une fonction avec une valeur tirée d’une distribution de Laplace, centrée sur 0 et d’échelle $\Delta f/\epsilon$ alors ce mécanisme respecte la contrainte de differential privacy

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

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

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

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



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 $s : (\mathcal{D}_m \times \mathcal{R}) \rightarrow \mathbb{R}$, 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 $s : (\mathcal{D}_m \times \mathcal{R}) \rightarrow \mathbb{R}$, $\mathcal{E}_s^\epsilon(D, \mathcal{R})$ 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 $\mathcal{M}_{[k]}(x) = (\mathcal{M}_1(x), \dots, \mathcal{M}_k(x))$, then $\mathcal{M}_{[k]}$ is $(\sum_{i=1}^k \varepsilon_i, 0)$ -differentially private.

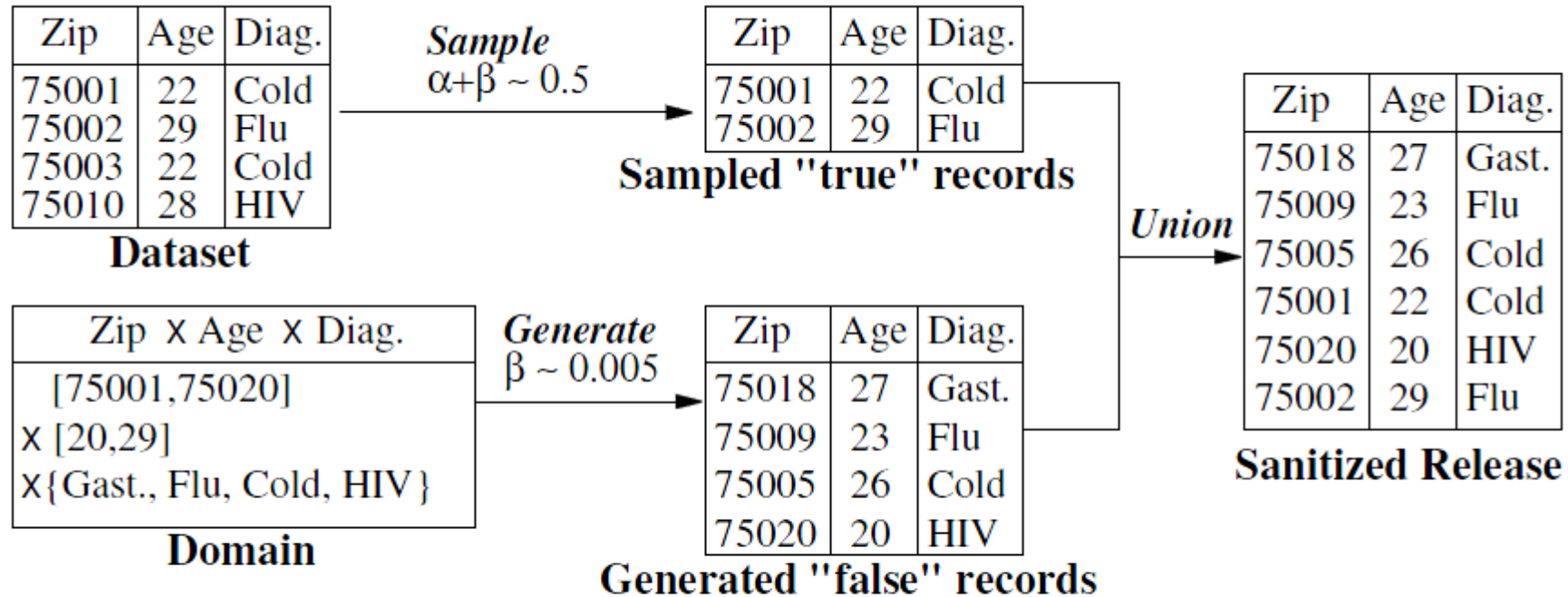
Modèle très général

Il suffit de définir l'adjacence !

Graphes :

- *Edge differential privacy* : deux graphes sont voisins s'ils ne diffèrent que par une arête
- *Node differential privacy* : deux graphes sont voisins s'ils ne diffèrent que par un noeud (et toutes les arêtes incidentes à ce noeud)

α, β – algorithm [Rastogi *et al.*]



On peut calculer des valeurs d'agrégats de type COUNTs avec un estimateur :

$$Q_{\text{Cold}} = (n_{\text{sanitized}} - \underbrace{\beta \cdot n_{\text{Domain}}}_{=2}) / \alpha = 0.5$$

Lier les approches k -anonymat et DP

J.Domingo-Ferrer [2015]

Dernièrement, des travaux ont eu lieu pour essayer de lier les approches classiques et la differential privacy.

Possibilité de réaliser une anonymisation de type t -closeness tout en respectant des garanties de type differential privacy.

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

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

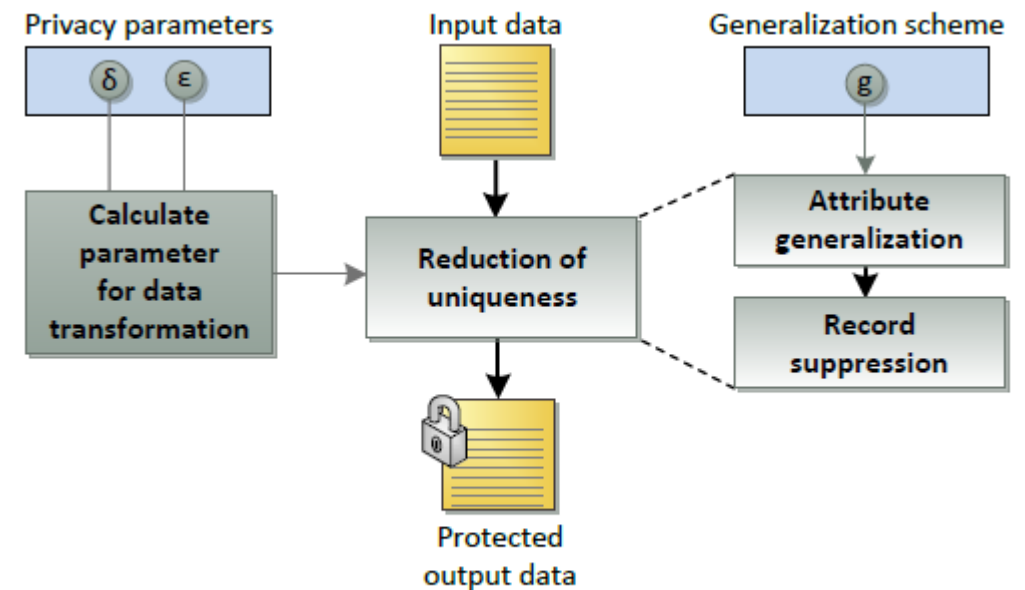


Fig. 5. Overview of the anonymization operator.

Idée : Choix aléatoire de paramètres de k -anonymisation

Autres travaux de recherche

Personnalisation de l'anonymat

Sécurisation du processus d'anonymisation

Anonymisation de processus complexes

Plateforme d'anonymisation et réidentification

Personnalisation du k -anonymat

Michel, Nguyen, Pucheral [DATA'17]

Quasi-identifïer		Sensitive
ZIP	Age	Condition
13***	40	Cancer
13***	45	Heart disease
13***	34	Heart disease
13***	38	Viral Infection
13***	[24,32]	Viral Infection
13***	[24,32]	Cancer
13***	[24,32]	Cancer
13***	[24,32]	Heart Disease

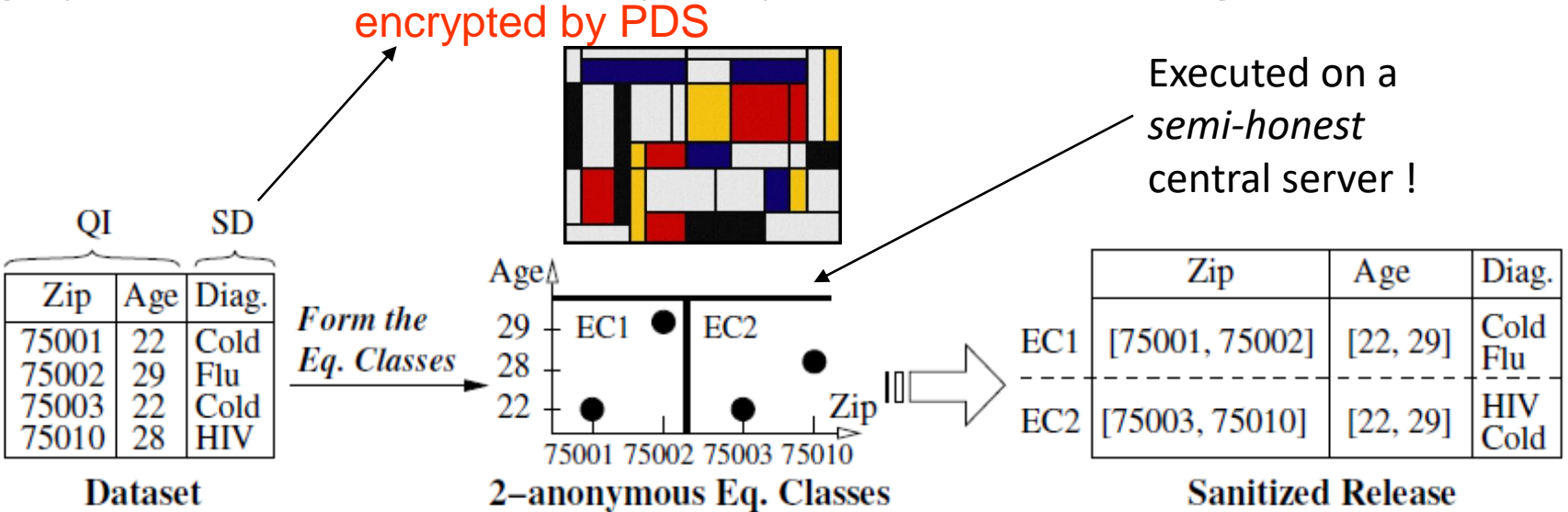
Table: k -anonymity [Swe02]

Quasi-identifïer		Sensitive	k_i
ZIP	Age	Condition	k_i
13***	40	Cancer	4
13***	45	Heart disease	4
13***	34	Heart disease	3
13***	38	Viral Infection	3
131**	[31,32]	Viral Infection	2
131**	[31,32]	Cancer	2
130**	[24,27]	Cancer	2
130**	[24,27]	Heart Disease	2

Table: k_i -anonymity

S curisation du processus d'anonymisation

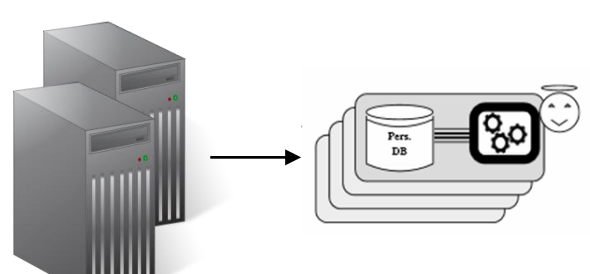
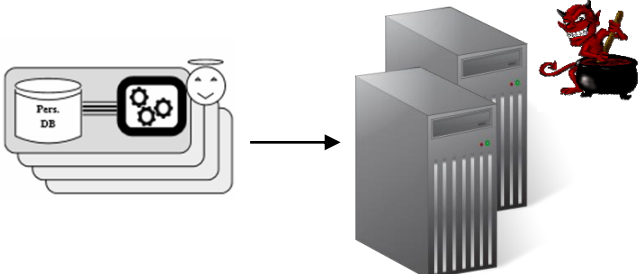
Allard, Nguyen, Pucheral *K-anonymity & Mondrian Algorithm (PST'11 award)*



1) Collection Phase

2) Construction Phase

3) Sanitization Phase

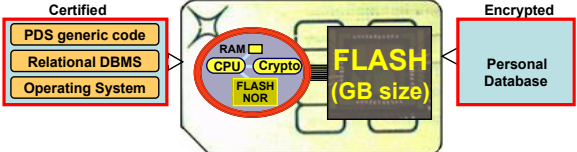


$$t_i = (QI_i, E(SD_i))$$

$$QI_i \rightarrow EC_j$$

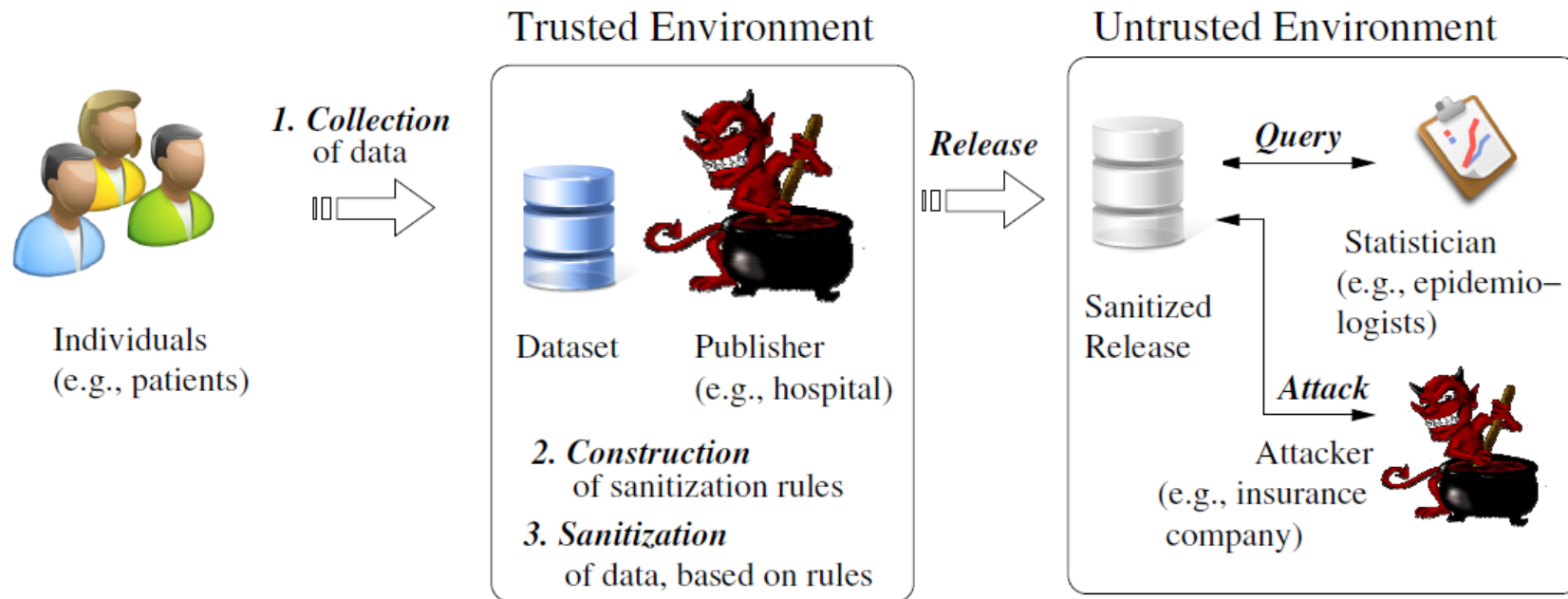
$$t_i = (EC_j, E(SD_i))$$

$$t_i = (EC_j, SD_i)$$



Sécurisation du processus d'anonymisation

Allard, Nguyen, Pucheral [DAPD'13]



Individuals :
Private Data

Publisher
(UNTRUSTED)
→ **Secure Computation**
Needed

Recipients
(no trust assumption)
→ **Privacy Models**
K-anon, L-div, Dif. Priv.

K -cores et differential privacy

Eichler, Rossi, Nguyen, Berthomé, Vazirgiannis [APVP17]

Théorème :

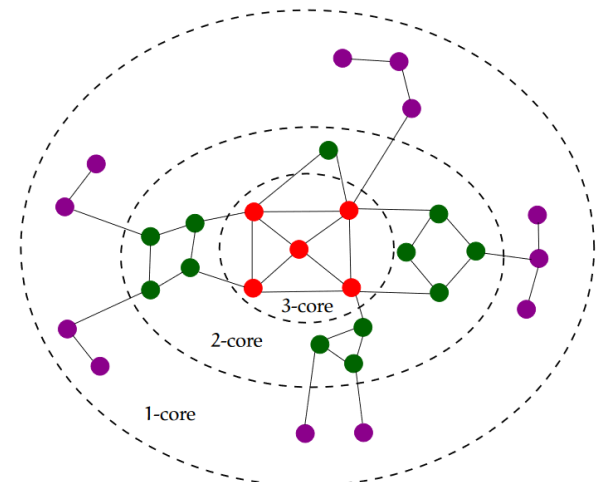
Il est possible de publier les k -cores calculés de manière distribuée en respectant le critère de differential privacy.

Algorithm 2: Naïve ϵ -DP Montresor Algorithm

While not converged do Montresor

When converged, publish $k + Y$ where Y is a IID random variables drawn from $Lap(1/\epsilon)$.

On peut donc sécuriser le résultat, mais peut-on sécuriser le processus ? En effet, au cours de l'exécution de l'algorithme de Montresor, les degrés sont échangés, or les degrés ne sont pas differential private !



● Core number $c = 1$ ● Core number $c = 2$ ● Core number $c = 3$

Enjeu : adversaire semi-honnête (honnête-mais-curieux)

Les informations échangées pour calculer les k -cores sont sensibles ! En particulier, la première étape est d'envoyer le degré.

→ Il faut donc sécuriser l'exécution de l'algorithme lui même !

- Idée pour l'étape 1 (publication des degrés) : anonymisation DP du graphe (intuition : borner le degré max du graphe).
- Idée pour l'étape 2 (échange sécurisé des données) : envoyer une valeur DP anonyme.
- **Problème** : l'algorithme de Montresor n'est pas prévu pour, car le comportement de la valeur du k -core local n'est plus monotone !

Algorithme

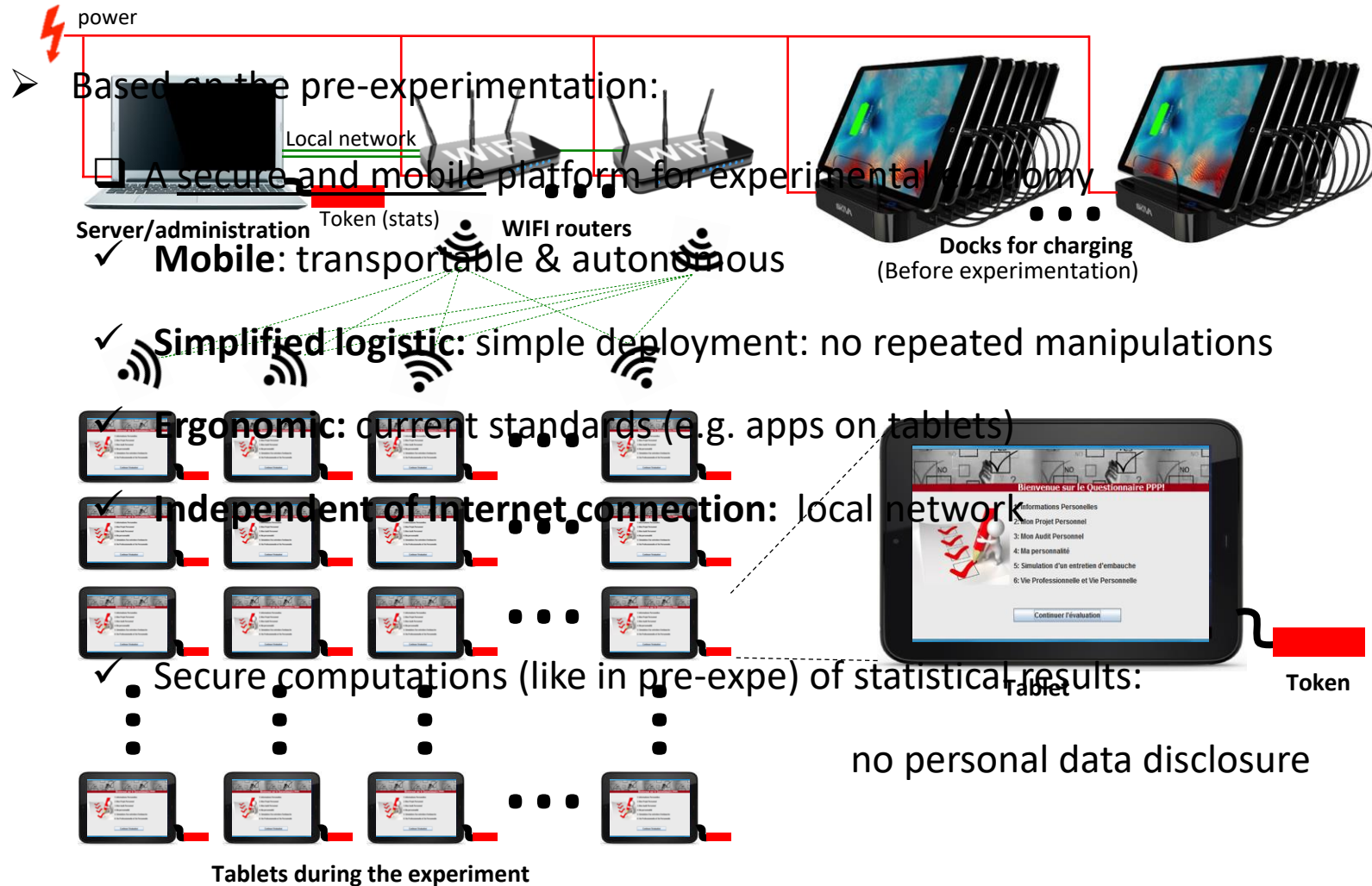
Algorithm 3: ϵ -DP Montresor Algorithm

While not converged do Montresor but publish $k^{(i)} + Y^{(i)}$ as estimated coreness, where $Y^{(i)}$ is a IID random variable drawn from $Lap(1/\epsilon')$ with $\epsilon' = \epsilon/Z$ where Z is an estimate of the number of steps the Montresor algorithm will take to converge.

Lorsque la valeur déclarée d'un voisin augmente, il suffit de ne rien faire !

Laboratoire d'expérimentation sociale mobile

Katsouraki, Bouganim, Nguyen (EDBT, 2016)



DARC/ INSAnonym

Plateforme d'anonymisation et de réidentification (APVP20, APVP21, APVP22)

Problème : Comment montrer qu'une technique d'anonymisation est suffisamment robuste ?

Plusieurs manières existent :

- “Feeling” (k-anonymat/L-diversité)
- Preuves “formelles” dûes au modèle (Differential Privacy)
- Critères G29 (cf controverse Diffix)

Ces approches sont souvent dépendantes du type de données !

→ DARC : Un banc d'essai/compétition et confrontation des algorithmes de ré-identification pour une sorte de jeu de données spécifique ~ analyse de risque ?

On retrouve ces mêmes approches d'une manière générale en sécurité informatique et IA

Problème supplémentaire : UTILITE DES DONNEES !

DARC : Une approche d'évaluation de type "attaque/défense"

- Principe :
 - L'anonymisation est évaluée **expérimentalement** au travers d'une compétition publique
- Résultat et exploitation :
 - Développement d'algorithmes d'anonymisation
 - Développement d'attaques génériques, qui peuvent être testées sur les données anonymisées et sur de nouveaux algorithmes
- Besoins :
 - Avoir une connaissance de certaines contraintes sur les données, ou savoir quel type d'algorithme de fouille de données soit utilisé par la suite, pour calibrer l'utilité
- Une approche un peu similaire est proposée / en cours de développement (P. Francis MPI / General Data Anonymity GDA-Score)

Composants d'un processus d'anonymisation

- **Une définition et métrique (modèle) de la “privacy”** qui répondent à la question : *Quelle protection proposer (et comment la mesurer) ?* → *Évalué expérimentalement par la qualité des attaques (et non posé comme paramètre du problème)*
- **Une définition et métrique d'utilité** qui répondent à la question : *Comment mesurer la perte d'utilité dans le processus utilisant des données anonymes au lieu des données exactes ?* → *Proposé par les concepteurs du challenge plateforme*
- **Un algorithme d'anonymisation** qui répond à la question : *Comment protéger les données tout en maximisant l'utilité des données ?* → *Évalué par la plateforme, peut être basé sur les techniques classiques de sécurisation (e.g. Differential Privacy, k-anonymat, etc.)*
- **Un processus d'anonymisation** : *Qui permet de mettre en œuvre l'algorithme de manière sûre et sécurisée.* → *Orthogonal à notre approche, non traité ici*

Modèle d'attaque : Attaquant à connaissance totale

ATTAQUANT A CONNAISSANCE TOTALE

- L'attaquant connaît la base de données initiale
 - L'attaquant voit la base de données anonymisée (et mélangée)
 - L'attaquant doit dire quel ligne anonymisée correspond à quelle ligne dans la base de données initiale
-
- 2 compétitions en une
 - Meilleure défense = meilleure utilité des données, pondéré par le nombre de lignes de la base réidentifiées → utile si le fournisseur de données ne donne pas de méthode d'anonymisation
 - Meilleure attaque = avoir réidentifié un maximum de lignes dans toutes les soumissions des concurrents → utile si le fournisseur de données veut tester ses méthodes d'anonymisation

Méthodes d'anonymisation

- Au libre choix des participants
 - K-anonymat (i.e. rendre similaire plusieurs lignes en effaçant ou généralisant)
 - Differential Privacy (i.e. perturber les données en utilisant une distribution particulière)
 - Ou tout autre méthode développée par les équipes
 - Toutes les méthodes doivent utiliser un pseudonyme (généré aléatoirement)

/!\ Contraintes sur la possibilité de changer de pseudonyme (ici : le pseudonyme doit être constant sur une semaine ou un mois)

Architecture Générale

Fichier donné par l'organisateur

Développé par DARC, exécuté sous le contrôle de l'organisateur

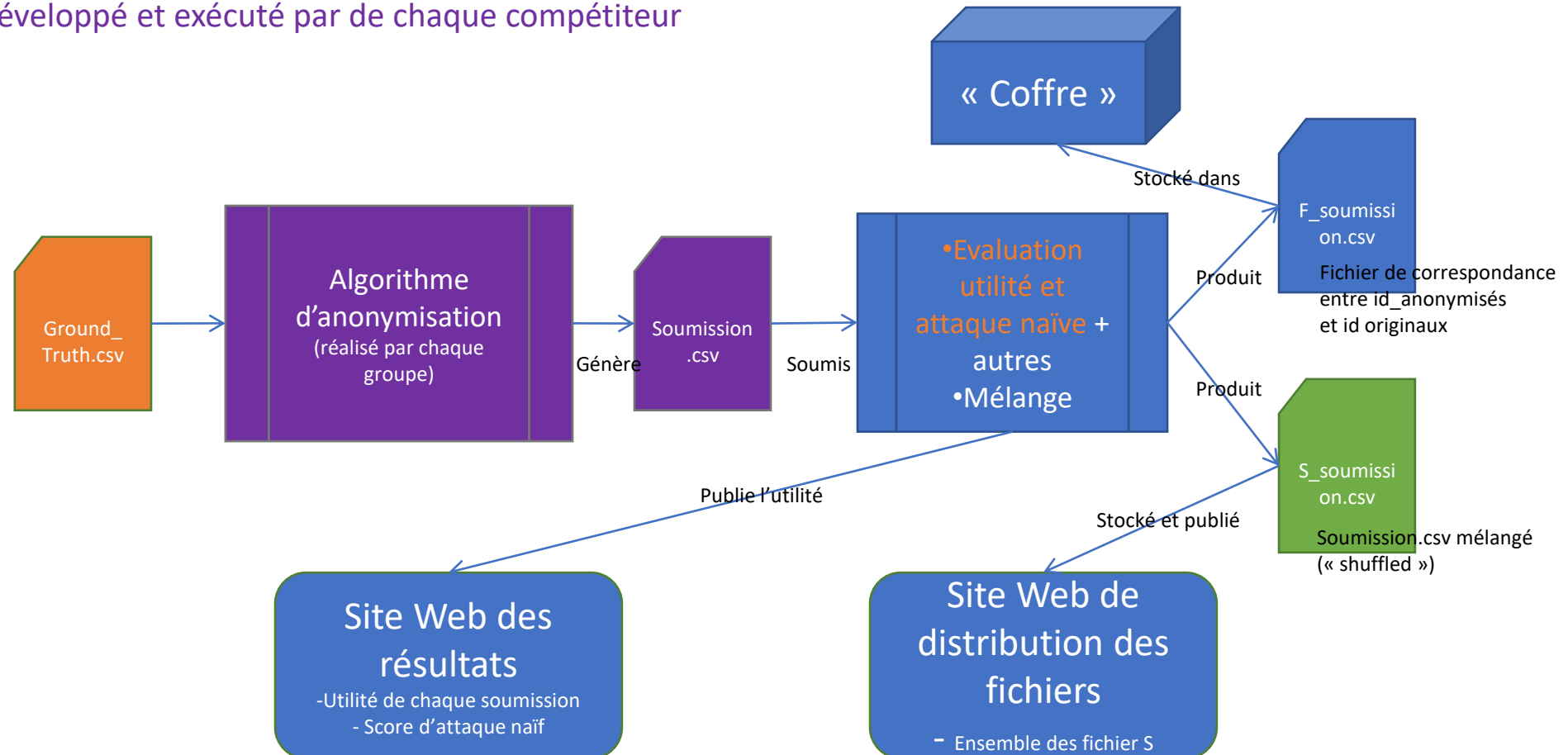
Algorithme développé par DARC, fichier généré par l'organisateur et mis à disposition

Algorithme développé et exécuté par de chaque compétiteur

Modèle d'attaque : Plateforme (choix orga)

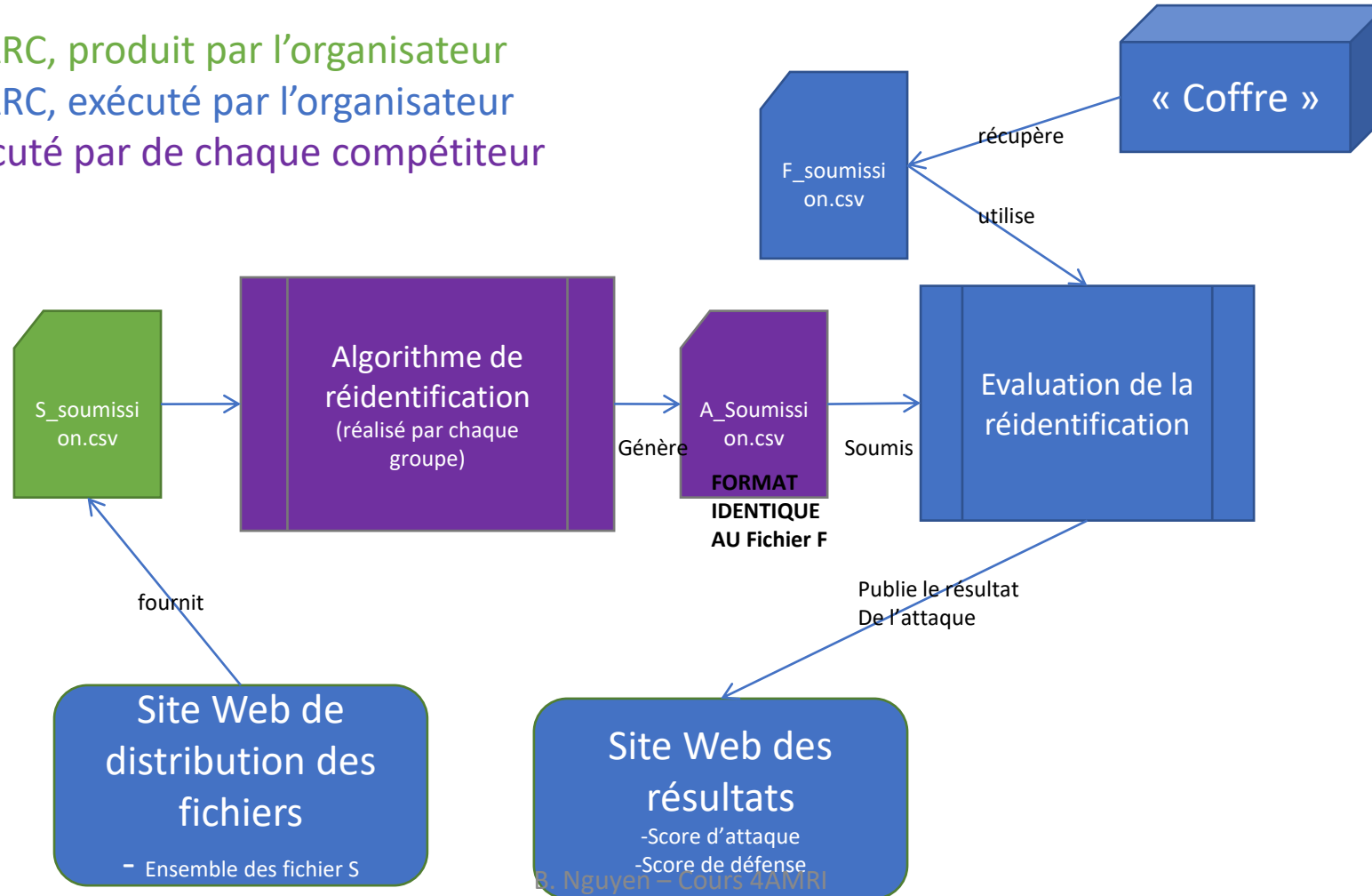
Données et métriques : Organisateur

Attaques et défenses : Participants



Architecture de réidentification

Réalisé par DARC, produit par l'organisateur
Réalisé par DARC, exécuté par l'organisateur
Réalisé et exécuté par de chaque compétiteur



DARC : Une approche d'évaluation de type "attaque/défense"

- Principe :
 - L'anonymisation est évaluée **expérimentalement** au travers d'une compétition publique
- Résultat et exploitation :
 - Développement d'algorithmes d'anonymisation
 - Développement d'attaques génériques, qui peuvent être testées sur les données anonymisées et sur de nouveaux algorithmes
- Besoins :
 - Avoir une connaissance de certaines contraintes sur les données, ou savoir quel type d'algorithme de fouille de données soit utilisé par la suite, pour calibrer l'utilité
- Une approche un peu similaire est proposée / en cours de développement (P. Francis MPI / General Data Anonymity GDA-Score)

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 pragmatique 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