
FESTIVAL DES
LIBERTÉS
NUMÉRIQUES
Du 1er au 16 février 2020
3ème
édition

Atelier:
Mieux vaut prévenir que guérir: anonymisation de données de santé

Présenté par: Louis Philippe SONDECK

- Pourquoi lorsqu’on naît, on nous attribut un nom ?

- Le nom, est-ce vraiment utile pour identifier ?

- Le nom, est-ce vraiment utile pour identifier ?

Données stockées en local pour 3lift.com

tluid

Nom
tluid

Contenu
826947349496022380

Domaine
.3lift.com

Chemin d'accès
/

Envoyer pour
Toutes sortes de connexions

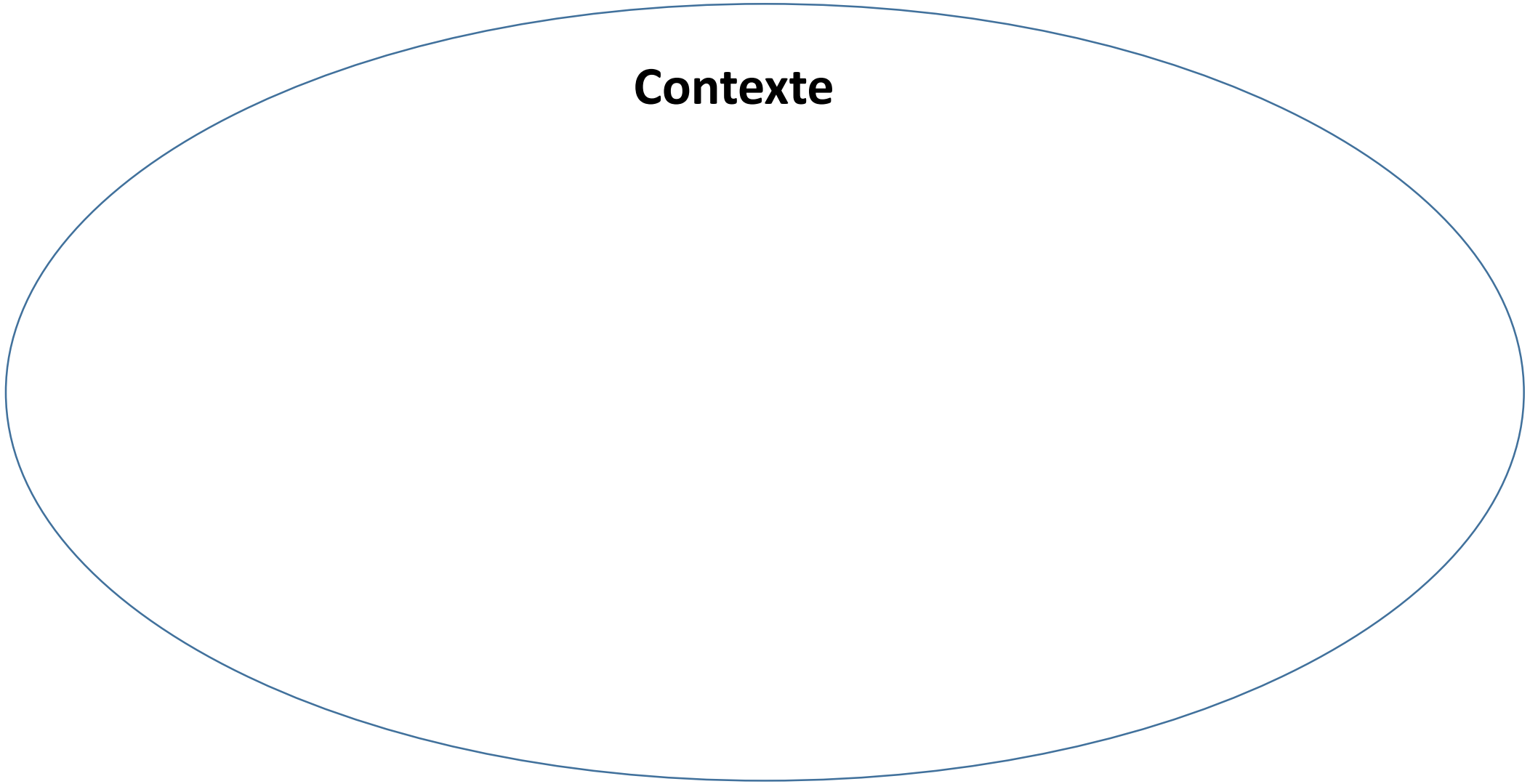
Accessible aux scripts
Oui

Date de création
jeudi 12 avril 2018 à 15:07:56

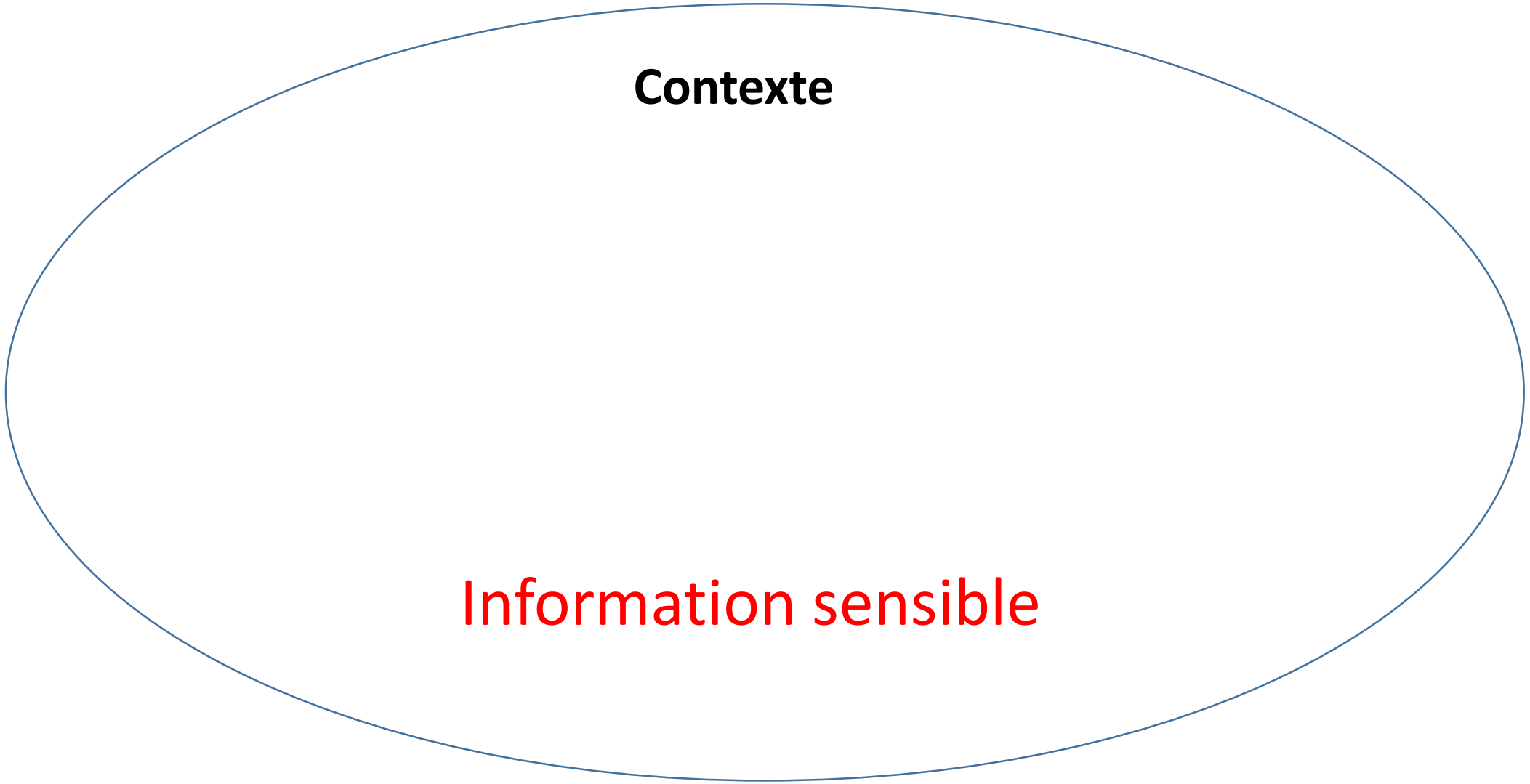
- Le nom, est-ce vraiment utile pour identifier ?

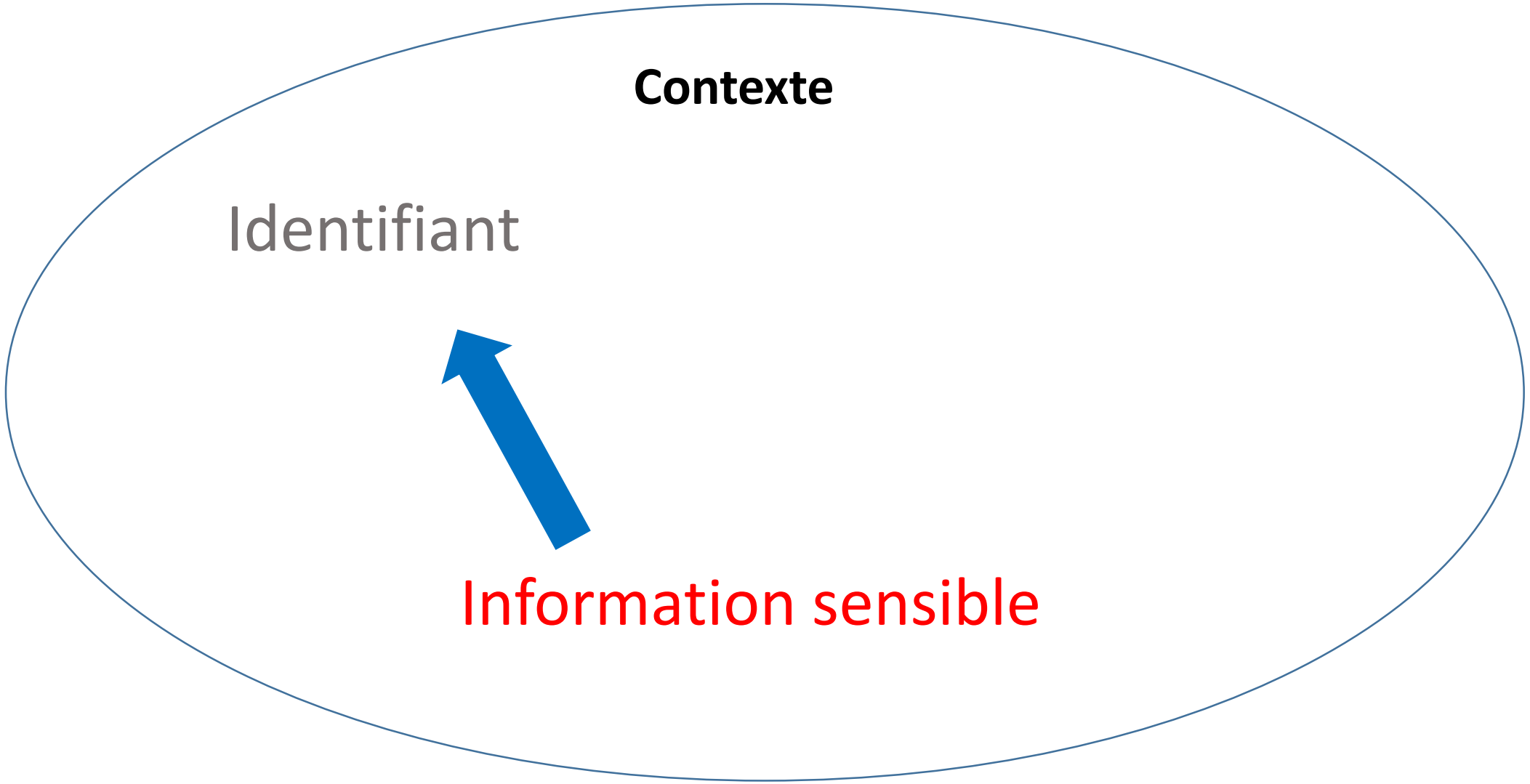
- Le nom, est-ce vraiment utile pour identifier ?

Identification

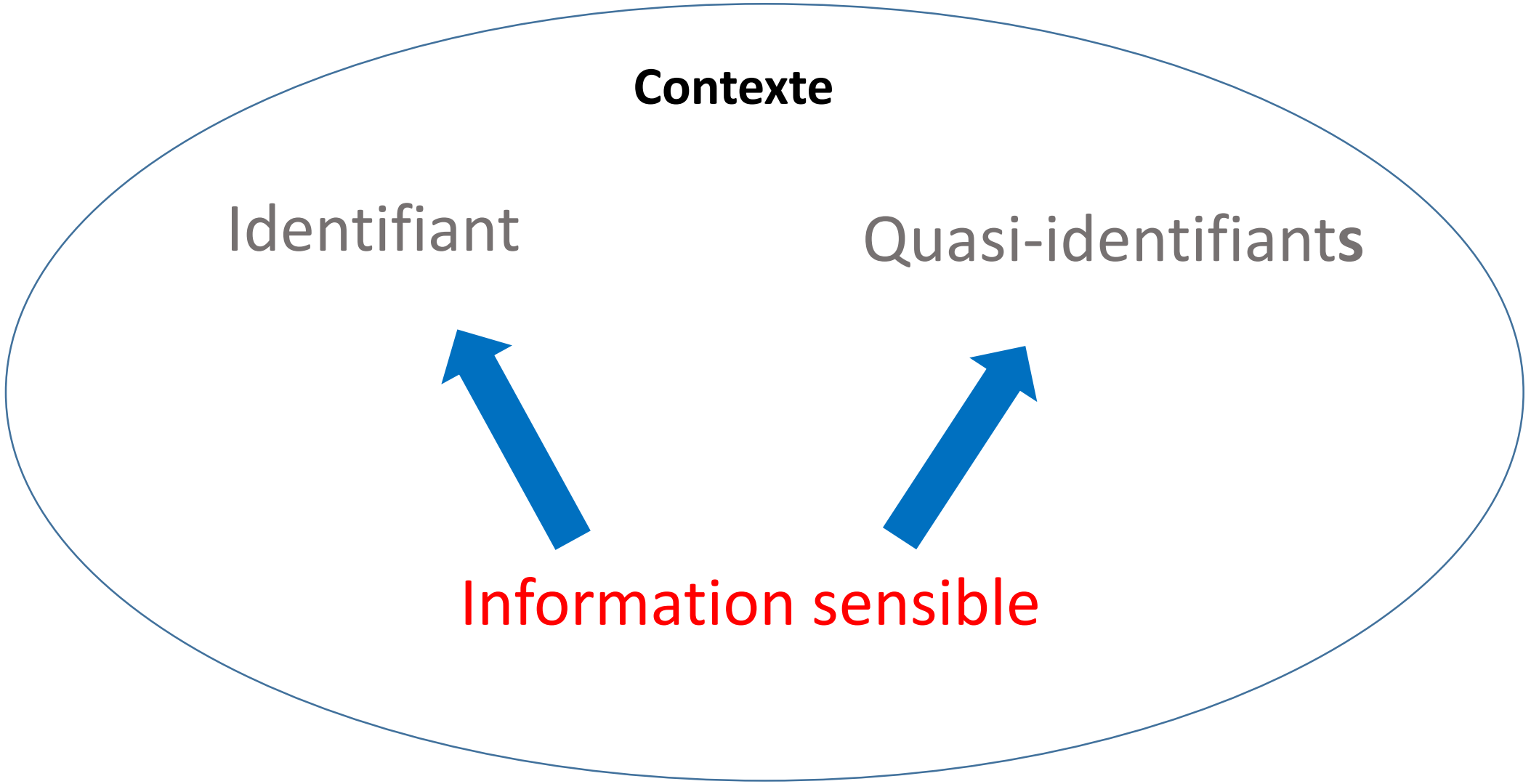


Identification


Contexte
Information sensible
Identification



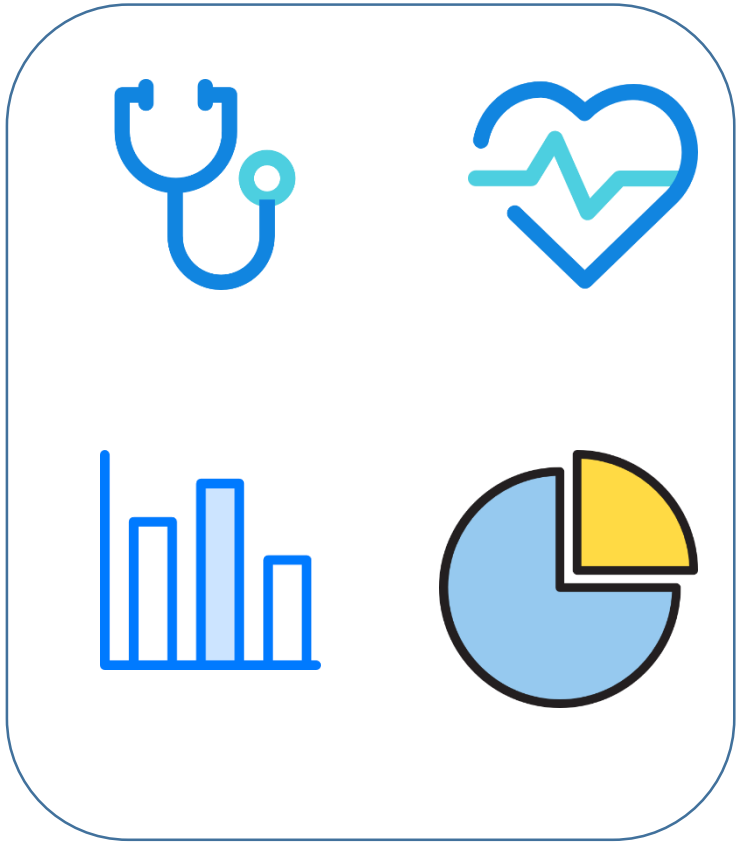
Identification



Identification

Plan

- Données personnelles de santé et intérêts pour la recherche
  - Qu'est ce qu'une donnée personnelle de santé
  - Quels intérêts pour la recherche
- La législation encadrant la recherche en santé
  - Le RGPD: éléments essentiels
  - Le cadre juridique applicable
  - Les risques en cas de non respect
- L'anonymisation des données
  - Cas réels de ré-identification
  - Principes et modèles
  - Evaluation de l'anonymisation
- Demo: anonymisation des données avec ARX



Données personnelles de santé et intérêts pour la recherche

Données personnelles et données personnelles de santé ?

Données personnelles

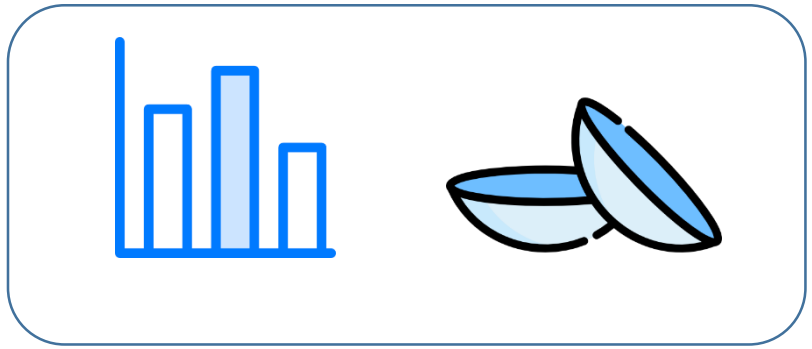
- Une **donnée personnelle** est toute information pouvant être rattachée **directement** ou **indirectement** à une **personne physique**

Données personnelles de santé

- Une **donnée personnelle de santé** est toute information permettant de tirer une **conclusion au regard de la santé** d'une personne. Elle doit être appréciée au **cas par cas**.

- On peut énumérer **3 catégories** de données de santé:
  - **Par nature** (ex: prestations de soins, résultats d'examens…)
  - **Par croisement** (ex: poids avec d'autres données)
  - **En raison de leur destination** (nombre de pas…)

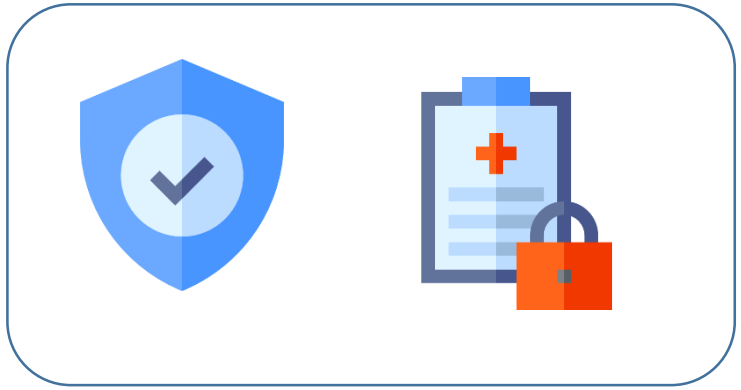
Quels intérêts pour la recherche



Usages de données personnelles en santé

Les données de santé peuvent être utilisées à différentes fins dans le domaine de la recherche:

- **la prévention le diagnostique et les actes thérapeutiques**
- **l'amélioration de la santé au quotidien** (par exemple via l**es objets connectés)** :
  - **Lentilles connectés (**mesure la teneur **en glucose** dans les larmes…),
  - **montre connectée** (mesure de la **qualité de sommeil** le nombre de **calories** brulées…),
  - **chaussures/chaussettes connectées** (conseils pour bien courir, mesure du **rythme cardiaque**…)



Le cadre juridique applicable en matière de recherche en santé

Le cadre juridique applicable à la recherche utilisant des données de santé

Textes de référence

- **Loi 78-17 modifiée - Informatique, Fichiers et Libertés**
- **Règlement européen 2016-678 - RGPD**
- **Le code de la santé publique**

Procédures

- **Engagement à respecter une MR + Inscription du traitement au registre des activités de traitement**

OU

- **Demande d'autorisation (INDS, CEREES, CNIL)**

Eléments d'appréciation

- **Le périmètre des données considérées:**
  - **Recherche *interne* ou *externe***
  - **Données personnelles de santé**: Recherche Impliquant la Personne Humaine **(RIPH)**
  - **Données non personnelles:** Recherche N'Impliquant pas la Personne Humaine **(RNIPH)**
- **La finalité du traitement:**
  - Recherche **interventionnelle à risques et contraintes importants (RIPH1)**
  - Recherche interventionnelle **à risques et contraintes minimes (RIPH2)**
  - Recherche observationnelle **(RIPH3)**
- **La méthodologie utilisée**: les méthodologies de références **(MR-00x)**

Le RGPD: éléments essentiels

1. Le respect des principes juridiques

+

2. La gestion des risques de sécurité des données

=

3. Conformité

Principes juridiques

1. **La finalité (art. 5.1(b) du RGPD)** : explicite, légitime, déterminée
2. **La base juridique (art. 6 du RGPD)**
3. **La minimisation de données (art. 5.1 (c) du RGPD):** adéquates, pertinentes et non excessives
4. **L'information (art. 12, 13 et 14 du RGPD):** claire et complète
5. **La durée de conservation (art. 5.1 (e) du RGPD)**: limitée
6. **Les droits (art. 15, 16, 17, 18, 20 et 21 du RGPD)**

Risques de sécurité

- **Confidentialité**: accès non autorisé
- **Intégrité**: modification non autorisée
- **Disponibilité**: disparition des données

sanctions

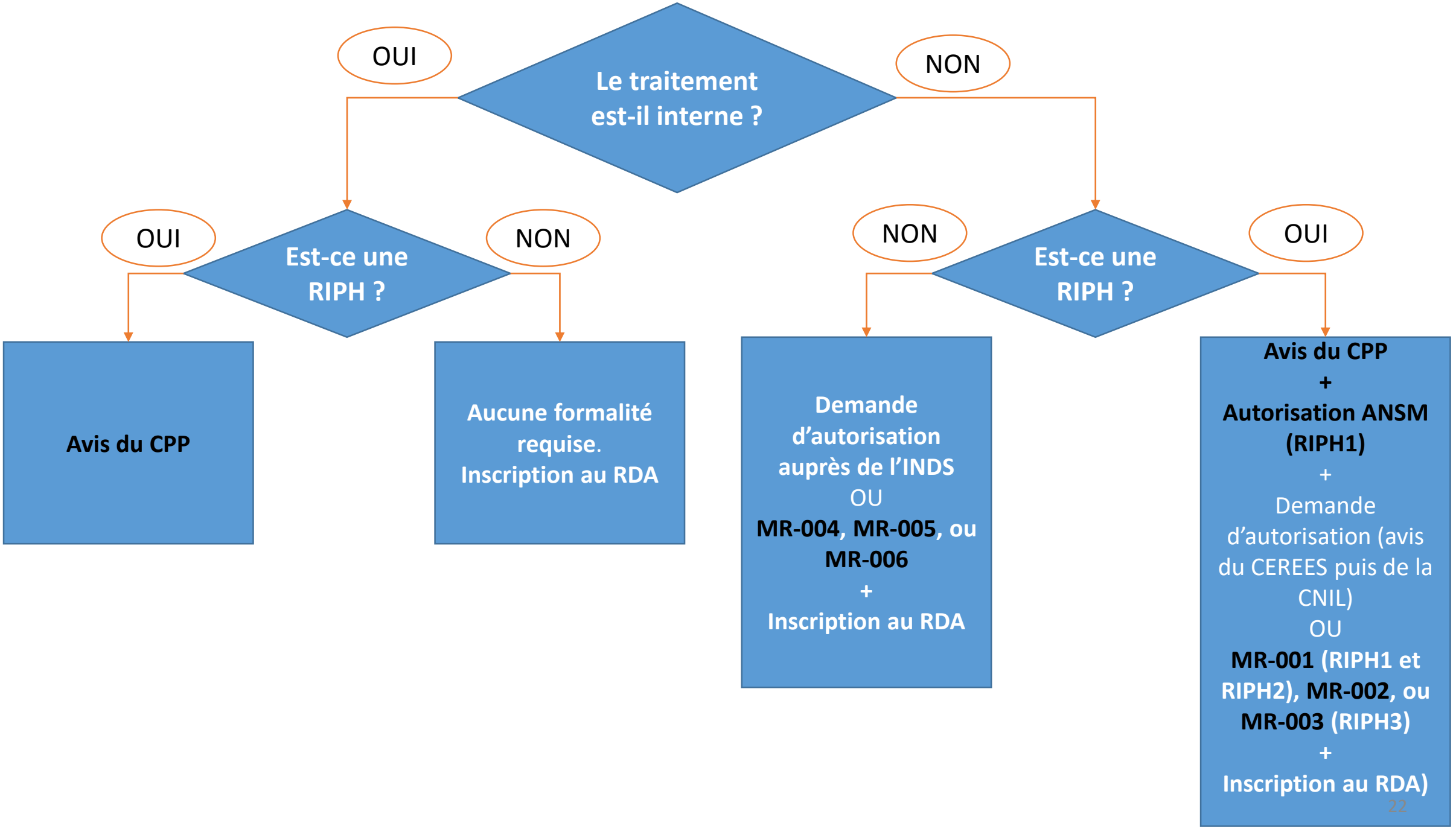
- 4% du CA mondial ou 20 millions d'euros

- **Recherche « interne »:** données **thérapeutiques/médicales** recueillies **par** le personnel médical et **exclusivement pour** un usage thérapeutique/médical

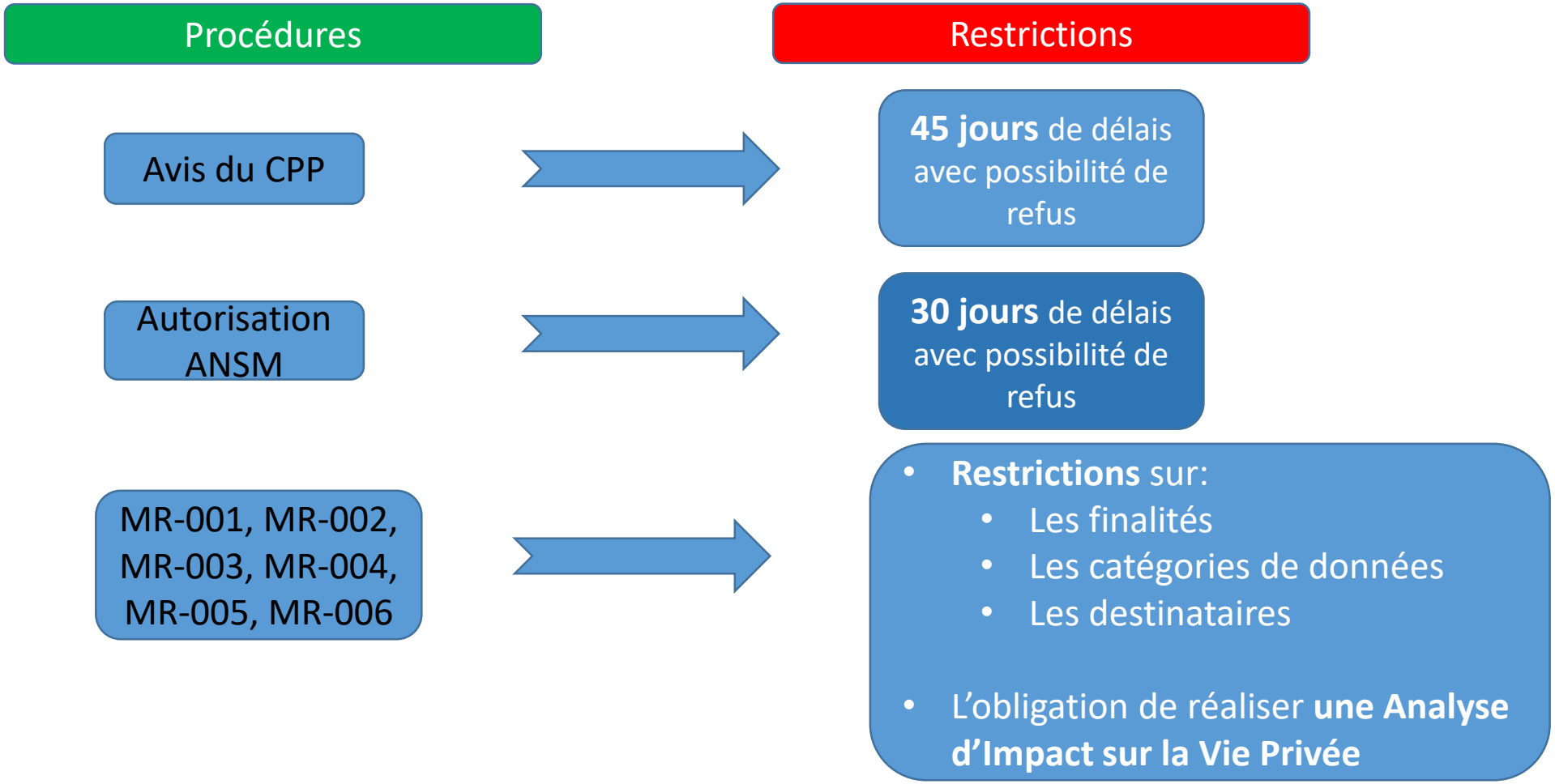
RECHERCHE IMPLIQUANT LA PERSONNE HUMAINE (AVIS FAVORABLE DU CPP REQUIS)			RECHERCHE N'IMPLIQUANT PAS LA PERSONNE HUMAINE
Recherche interventionnelle **(RIPH1)**	Recherche interventionnelle à risques et contraintes minimes **(RIPH2)**	Recherche non interventionnelle **(RIPH3)**	**(RNIPH)**
Sans objet Ces recherches nécessitent de recueillir des données spécifiques et supplémentaires par rapport à la prise en charge habituelle du patient. Ces recherches doivent donc être réalisées dans le cadre du chapitre IX de la loi Informatique et Libertés.			**Aucune formalité** Inscription au registre des activités de traitement

- **Recherche multicentrique et/ou accès aux données en dehors de l'équipe de soins**

RECHERCHE IMPLIQUANT LA PERSONNE HUMAINE (RIPH) (AVIS FAVORABLE DU CPP REQUIS)			RECHERCHE N'IMPLIQUANT PAS LA PERSONNE HUMAINE (RNIPH)
Recherche interventionnelle **(RIPH1)**	Recherche interventionnelle à risques et contraintes minimes **(RIPH2)**	Recherche non interventionnelle **(RIPH3)**	**(RNIPH)**
Demande d'**autorisation** de recherche **ou** **Engagement** de conformité **MR-001** + Inscription au **registre des activités**	Demande d'**autorisation** de recherche **ou** **Engagement** de conformité **MR-001** + Inscription au **registre des activités**	Demande d'**autorisation** de recherche **ou** **Engagement** de conformité **MR-002** ou **MR-003** + Inscription au **registre des activités**	Dépôt auprès de l'**INDS** de la demande d'**autorisation** de recherche (avis **CEREES** puis avis **CNIL**) **ou** **Engagement** de conformité **MR-002** ou **MR-003** + Inscription au **registre des activités**


Le traitement est-il interne ?
OUI
NON
Est-ce une RIPH ?
OUI
NON
Avis du CPP
Aucune formalité requise.
Inscription au RDA
Est-ce une RIPH ?
NON
OUI
Demande d'autorisation auprès de l'INDS
OU
MR-004, MR-005, ou MR-006
+
Inscription au RDA
Avis du CPP
+
Autorisation ANSM (RIPH1)
+
Demande d'autorisation (avis du CEREES puis de la CNIL)
OU
MR-001 (RIPH1 et RIPH2), MR-002, ou MR-003 (RIPH3)
+
Inscription au RDA)

Contraintes liées à l'utilisation des données personnelles de santé (RIPH & RNIPH)

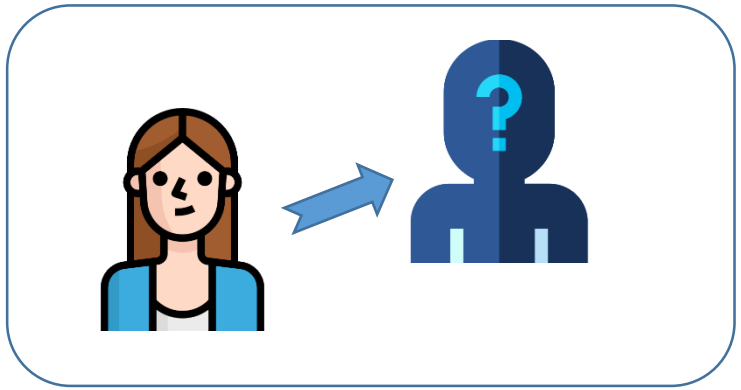


Intérêt de l'anonymisation

RGPD
(MR, CPP…)

Données personnelles

Données anonymisées = Données non personnelles



Anonymisation des données: principes et modèles

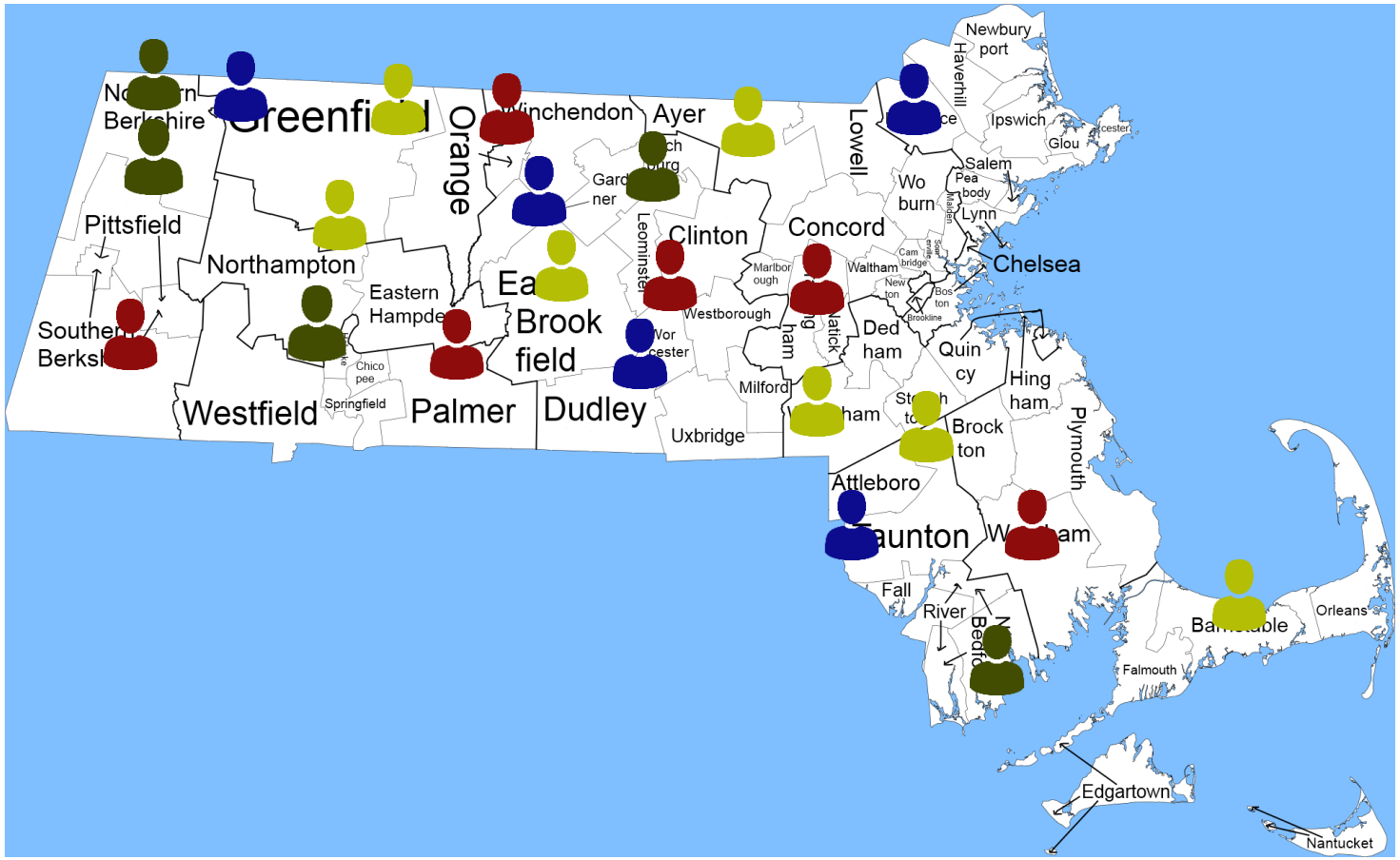
Cas réel de ré-identification

- **En 2001, une chercheuse du nom de L. *Sweeney*** a pu **ré-identifier** des personnes en croisant des données « anonymisées » publiées par une société d'assurance, avec des données d'un fichier électorale en clair, ceci, en se basant uniquement sur 3 attributs: ***le code postal, la date de naissance, le genre***

- Elle a ainsi déduit des ***informations médicales*** du **gouverneur de l'Etat du Massachussets** (William Weld).

Cas réel de ré-identification

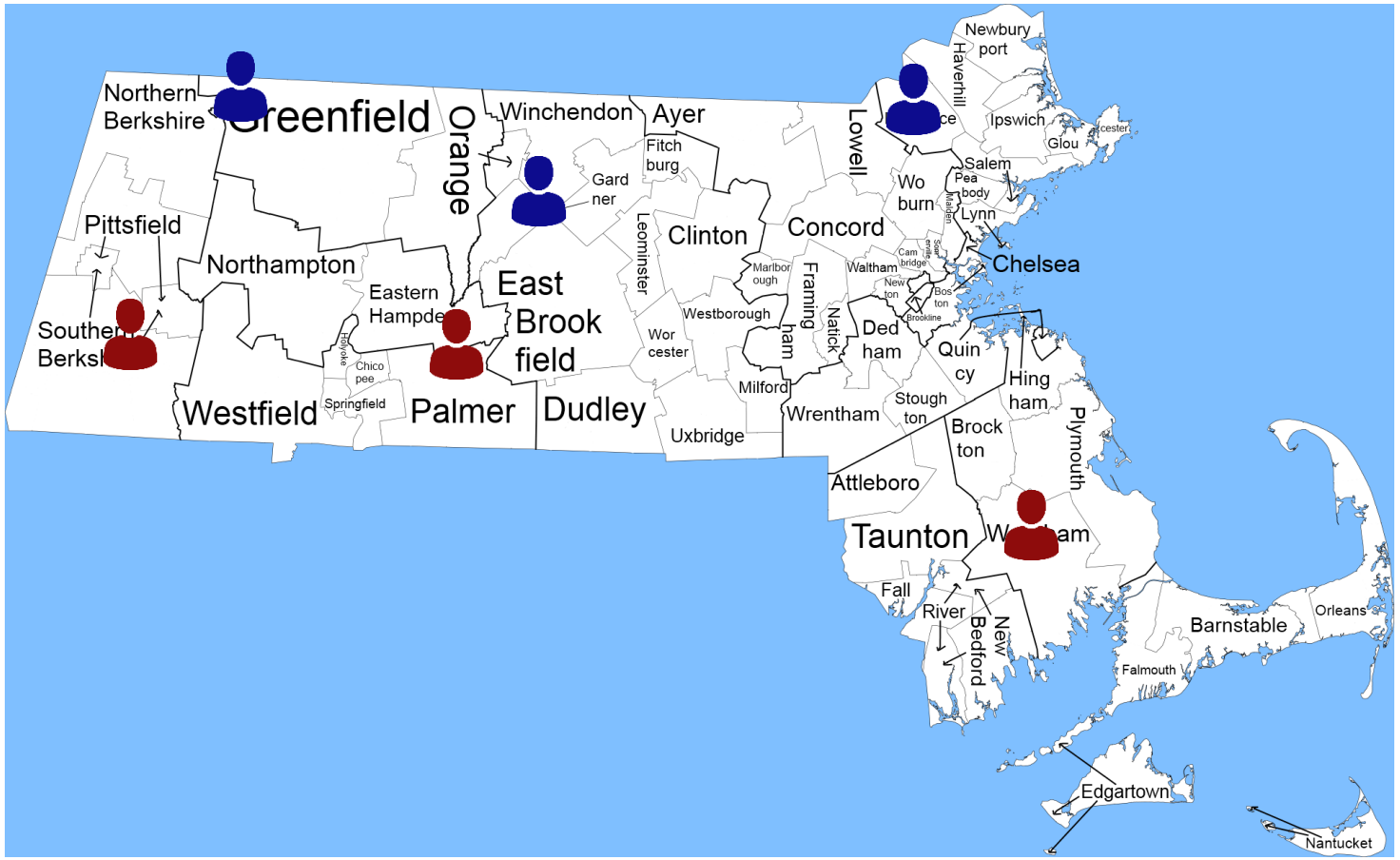
Données de santé
« anonymisées » publiées par le **GIC (Group Insurance Company)**, l'organisme chargé de l'achat des contrats d'assurance des employés de l'Etat



Cas réel de ré-identification

Nombre de personnes ayant pour information:

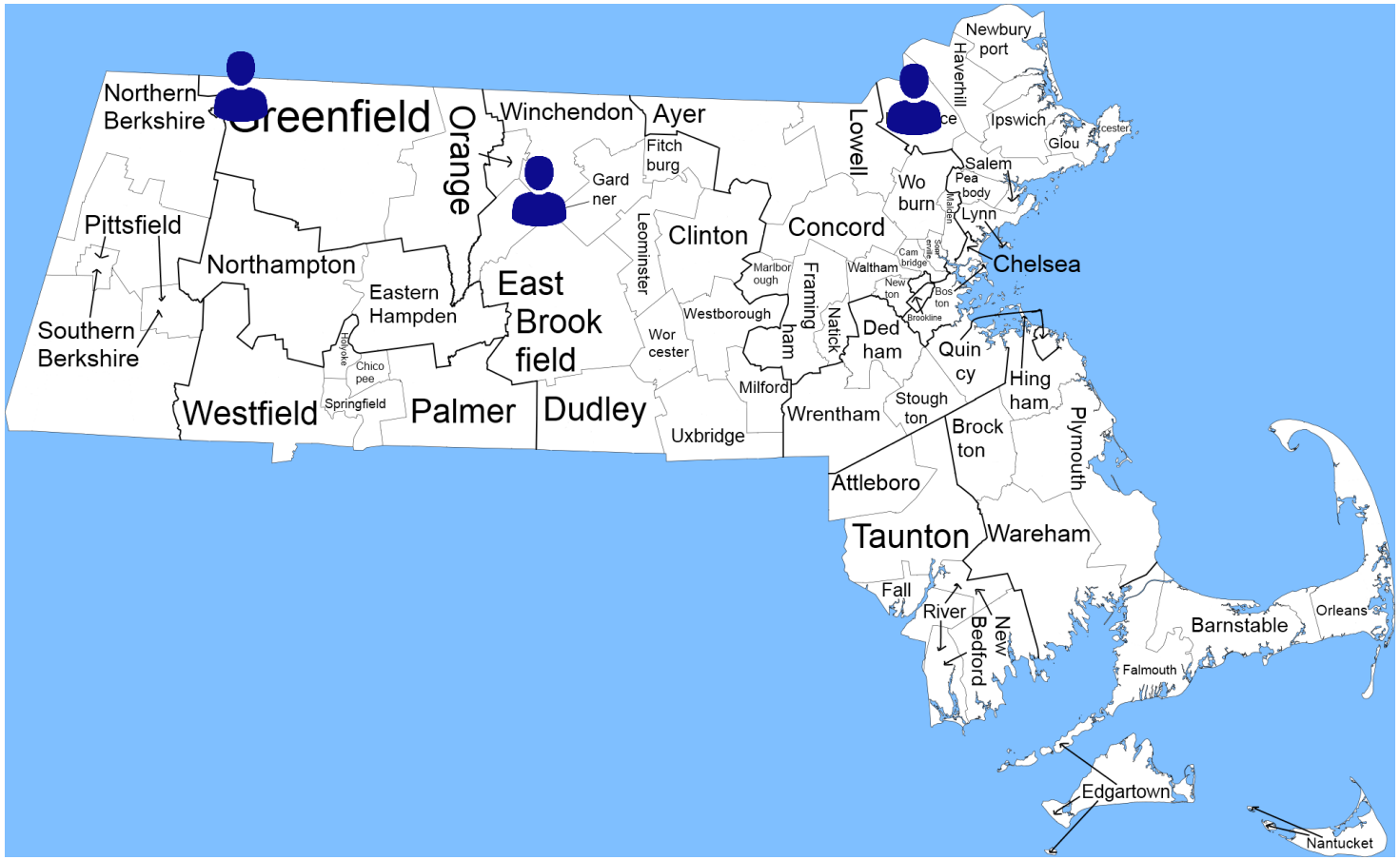
1. Date de naissance: 31/07/1945



Cas réel de ré-identification

Nombre de personnes ayant pour information:

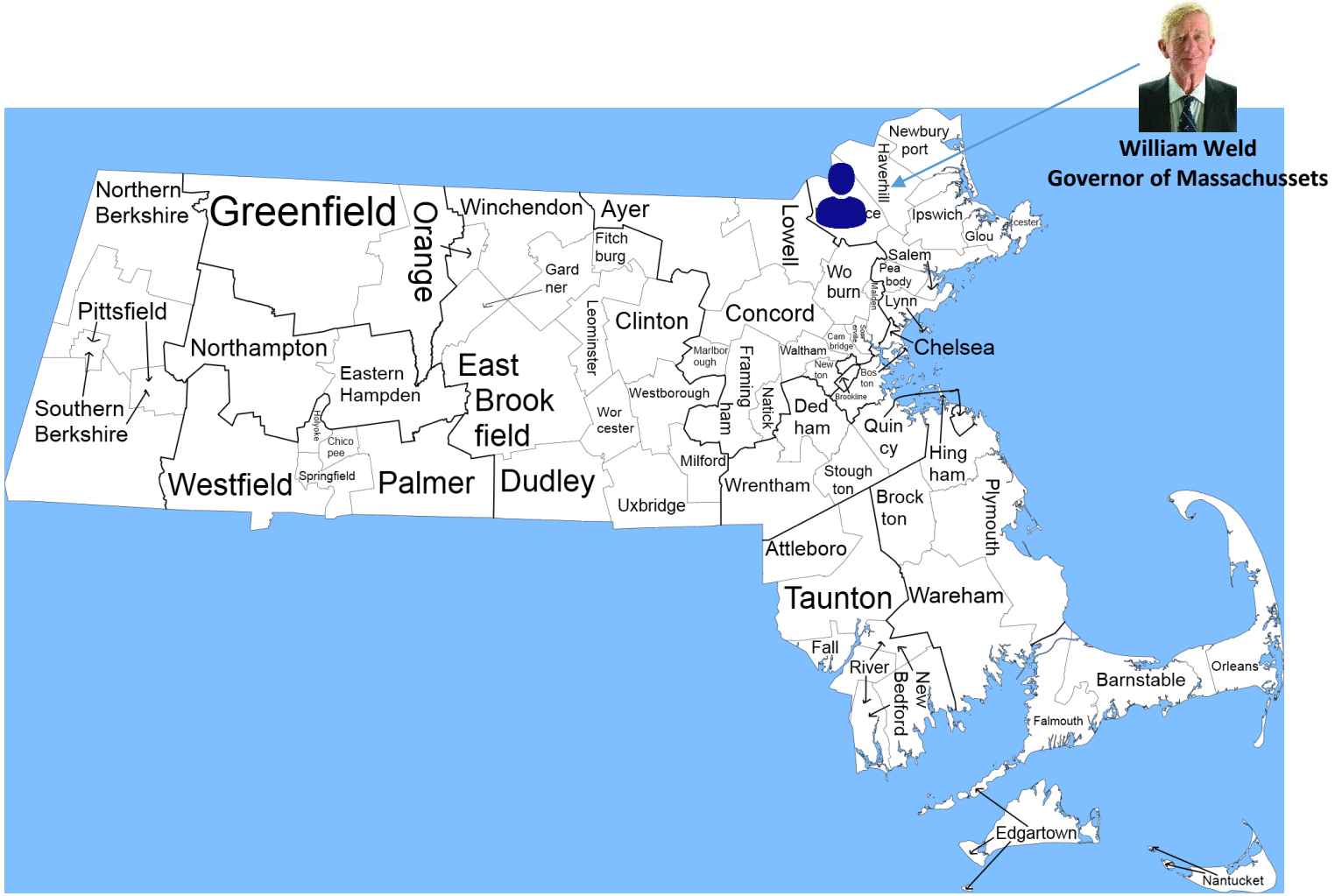
1. Date de naissance: 31/07/1945
2. Genre: Masculin



Cas réel de ré-identification

Nombre de personnes ayant pour information:

1. Date de naissance: 31/07/1945
2. Genre: Masculin
3. Code postal: 02152



Cas réel de ré-identification

Id	Age	Sexe	Code Postal	Maladie
#1234	56	M	75013	VIH
#2345	56	M	02152	Cancer
#3456	27	F	02152	Malaria

Fichier Assurance (« anonymisé »)

Noms	Age	Sexe	Code Postal	Nbr Electeurs
J.Sparraw	26	M	75013	200
W.Weld	56	M	02152	450
J.Doe	27	F	75013	300

Fichier Electoral

- La ré-identification s'est faite en croisant 2 jeux de données

Cas réel de ré-identification

Id	Age	Sexe	Code Postal	Maladie
#1234	56	M	75013	VIH
#2345	56	M	02152	Cancer
#3456	27	F	02152	Malaria

Fichier Assurance (« anonymisé »)

Noms	Age	Sexe	Code Postal	Nbr Electeurs
J.Sparraw	26	M	75013	200
W.Weld	56	M	02152	450
J.Doe	27	F	75013	300

Fichier Electoral

- La ré-identification s'est appuyée sur une combinaison bien précise d'informations (**Age = 56, Sexe = M** et **Code postal = 02152**)

Cas réel de ré-identification

Id	Age	Sexe	Code Postal	Maladie
#1234	56	M	75013	VIH
#2345	56	M	02152	Cancer
#3456	27	F	02152	Malaria

Fichier Assurance (« anonymisé »)

Noms	Age	Sexe	Code Postal	Nbr Electeurs
J.Sparraw	26	M	75013	200
W.Weld	56	M	02152	450
J.Doe	27	F	75013	300

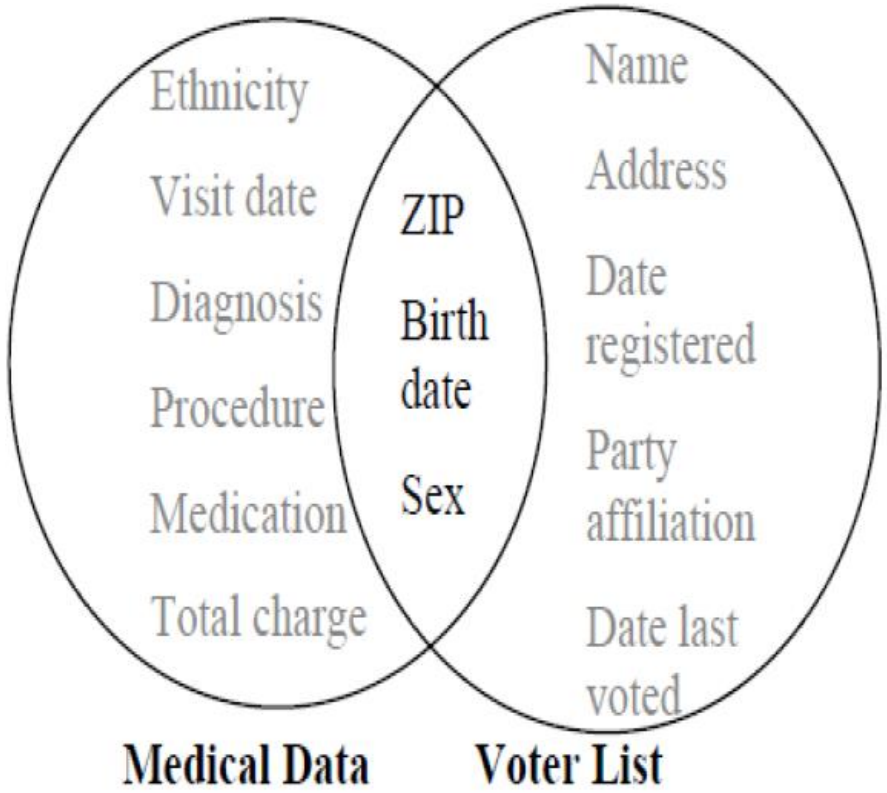
Fichier Electoral

- **Pour la ré-identification**:
  - Contexte: 2 jeux de données (d'assurance et de vote)
  - Information sensible: « Cancer »
  - Quasi-identifiants: **Age = 56, Sexe = M** et **Code postal = 02152**
  - Identifiant: **Nom = W.Weld**

- Dans ce contexte le nom **W.Weld** et l'ensemble: **{Age = 56, Sexe = M, Code postal = 02152}** sont équivalent

Cas réel de ré-identification

- 87% des américains peuvent être identifiés à partir du triplet: *date de naissance, sexe, code postal*



Définitions

- **Anonymisation:** « *Processus par lequel des informations personnellement identifiables sont* ***altérées de façon irréversible*** *de sorte que la personne à laquelle se rapporte l'information ne peut plus être* ***identifiée directement ou indirectement*** »

- L'anonymisation est le seul procédé permettant de transformer des données personnelles en données qui ne sont plus personnelles et donc qui sortent du champs d'application du RGPD

Terminologie

Identifiant: attribut permettant de ***caractériser de manière unique*** un individu dans un ***groupe d'individus.*** On distingue 2 types d'identifiants:

- **Identifiant direct** **nom**, **prénom**
- **Identifiant indirect**: adresse, **numéro de téléphone**, **NIR**, **empreinte digitale**…

Quasi-identifiant: attribut pouvant être ***combiner avec d'autres attributs*** pour identifier un individu (ex: **code postal**, **vêtements**, **genre**…)

Attribut sensible: attribut pouvant révéler des ***informations compromettantes*** sur un individu (ex: **maladie**, **salaire, données de géolocalisation**…)




Terminologie

Identifiant	QI	QI	QI	Att Sensible
Noms	**Age**	**Sexe**	**Code Postal**	**Maladie**
J.Sparraw	26	M	75013	VIH
W.Weld	26	M	75002	Cancer
J.Doe	27	F	75013	Corona Virus

Identifiant indirect

Les objectifs de l'anonymisation

- L'anonymisation vise à résoudre **le compromis** entre:
  - **Confidentialité**: empêcher la ré-identification des personnes
  - **Viabilité des données**: assurer la cohérence des données anonymisées par rapport au besoin cible

Confidentialité des données

Le G29 (groupe des CNILs Européens) publie un avis sur les techniques d'anonymisation qui défini 3 risques pour **la confidentialité** des données:

1. **L'individualisation:** Est-il toujours possible **d'isoler** une partie ou la totalité des enregistrements identifiant d'un individu ?

2. **La corrélation:** Est-il possible de **relier entre eux des enregistrements** relatifs à **un individu** ou à **un groupe d'individu** ?

3. **L'inférence:** Est-il possible de **déduire** la valeur d'un attribut à partir des valeurs d'un ensemble d'autres attributs ?

Le risque de ré-identification s'évalue selon ces 3 critères

Le risque d'individualisation

- **Le risque d'individualisation** évalue la capacité à isoler une personne à partir d'une valeur ou une combinaison de valeurs

Données auxiliaires

Noms	Age	Sexe	Code Postal	Nbr Electeurs
J.Doe	27	F	75013	300
W.Weld	26	M	75002	450

Données pseudonymisées

Noms	Age	Sexe	Code Postal	Maladie
12489	26	M	75013	Cancer
13245	26	M	75002	Cancer
12345	27	F	75013	Malaria

Le risque d'inférence

- **Le risque d'inférence** évalue la possibilité de déduire les valeurs d'un attribute à partir d'un autre. Par exemple, être âgé de **26 ans** et être **masculin** permet de déduire la maladie **Cancer**

Données auxiliaires

Noms	Age	Sexe	Nbr Electeurs
W.Weld	26	M	450

Données pseudonymisées

Noms	Age	Sexe	Code Postal	Maladie
12489	26	M	75013	Cancer
13245	26	M	75002	Cancer
12345	27	F	75013	Malaria

Le risque de corrélation

- **Le risque de corrélation** concerne la capacité à corréler deux jeux de données pour enrichir la connaissance sur les personnes concernées. Par exemple, être agé de **26 ans** et être **masculin** implique que l'on reside soit dans le **75013** soit dans le **75002**.

Données auxiliaires

Noms	Age	Sexe	Code Postal
J.Sparraw	26	M	75013

Données pseudonymisées

Noms	Age	Sexe	Code Postal
12489	26	M	75013
13245	26	M	75002
12345	27	F	75013

Viabilité des données

Hierarchies de valeurs de l'attribut Age

Age	Age*	Age**
22	2*	2*
22	2*	2*
22	2*	2*
45	>= 40	>= 30
63	>= 40	>= 30
40	>= 40	>= 30
35	3*	>= 30
35	3*	>= 30
32	3*	>= 30

- Est-il possible d'utiliser les données anonymisées pour répondre à des besoins identifiés ?
- Le besoin doit être caractériser au préalable, et pris en compte pendant le processus d'anonymisation
- La viabilité dépend de la manière dont les valeurs des attributs sont hiérarchisées

Les modèles d’anonymisation

Les modèles d'anonymisation

- **La généralisation:** dilution des attributs en modifiant leur échelle ou leur ordre de grandeur. Par exemple:
  - **le k-anonymat** qui assure qu'une ***information sensible*** ne puisse être associée à moins de ***k personnes*** dans un jeu de données.

- **La randomisation:** altération de la véracité des données afin d'affaiblir le lien entre les données et l'individu. Par exemple:
  - **La confidentialité différentielle:** qui protège les données en y ajoutant du bruit

Généralisation vs Randomisation

- **Généralisation (k-anonymat):**
  - **Avantage**: permet un meilleur contrôle de la viabilité des données
  - **Inconvénient**: confidentialité difficile à maitriser lorsque la volumétrie des données est importante

- **Randomisation (Confidentialité différentielle):**
  - **Avantage**: fournit des garanties de confidentialité formelles
  - **Inconvénient**: viabilité des données difficile à maîtriser

Les techniques de la généralisation

- La généralisation se décline en 3 principales techniques:
  - le **k-anonymat**,
  - la **l-diversité** (qui améliore la k-anonymat)
  - la **t-proximité** (qui « améliore » la l-diversité)

- Les améliorations font référence à la capacité de résister aux attaques sur la confidentialité (*individualisation, corrélation, inférence*)

La k-anonymisation réduit le risque d'individualisation

• Le **k-anonymat** assure qu'une ***information sensible*** ne puisse être associée à moins de ***k personnes*** dans un jeu de données

Original

	Age	Disease
1	22	lung cancer
2	22	lung cancer
3	22	lung cancer
4	45	stomach cancer
5	63	diabetes
6	40	aids
7	35	aids
8	35	flu
9	32	diabetes

Generalization Table

	Age	Age*
1	22	2*
2	22	2*
3	22	2*
4	45	≥ 40
5	63	≥ 40
6	40	≥ 40
7	35	3*
8	35	3*
9	32	3*

3-anonymous

	Age*	Disease
1	2*	lung cancer
2	2*	lung cancer
3	2*	lung cancer
4	≥ 40	stomach cancer
5	≥ 40	diabetes
6	≥ 40	flu
7	3*	aids
8	3*	aids
9	3*	diabetes

L’attaque d’inférence

• Le **k-anonymat** est vulnérable à l’attaque d’inference. Par exemple, on déduit dans que toutes les personnes âgées d’une vingtaine d’année ont le cancer du poumon

Table 3-anonymisée

	Age*	Disease
1	2*	lung cancer
2	2*	lung cancer
3	2*	lung cancer
4	≥ 40	stomach cancer
5	≥ 40	diabetes
6	≥ 40	flu
7	3*	aids
8	3*	aids
9	3*	diabetes

L’attaque d’inférence

• Le **k-anonymat** est vulnérable à l’attaque d’inference. On déduit avec une probabilité de 2/3 que toutes les personnes âgées d’une trentaine d’années ont le SIDA

Table 3-anonymisée

	Age*	Disease
1	2*	lung cancer
2	2*	lung cancer
3	2*	lung cancer
4	≥ 40	stomach cancer
5	≥ 40	diabetes
6	≥ 40	flu
7	3*	aids
8	3*	aids
9	3*	diabetes

La l-diversité réduit le risque d'inférence (Homogeneity Attack)

- La l-diversité **améliore** la k-anonymisation: diversification des valeurs de l'attribut sensible; diminue le risque d'inférence

Table 3-anonymisée

	Age*	Disease
1	2*	lung cancer
2	2*	lung cancer
3	2*	lung cancer
4	≥ 40	stomach cancer
5	≥ 40	diabetes
6	≥ 40	flu
7	3*	aids
8	3*	aids
9	3*	diabetes

Table 3-diversifiée

	Age**	Disease
1	< 40	lung cancer
2	< 40	lung cancer
3	< 40	lung cancer
4	≥ 40	stomach cancer
5	≥ 40	diabetes
6	≥ 40	flu
7	< 40	aids
8	< 40	aids
9	< 40	diabetes

L’attaque d’inférence sémantique (similarity attack)

- La l-diversité est vulnérable à l’attaque d’inférence sémantique

Table 3-diversifiée

	ZIP Code*	Age*	Salary	Disease
1	355**	2*	4K	colon cancer
2	355**	2*	5K	stomach cancer
3	355**	2*	6K	lung cancer
4	3581*	≥ 40	7K	stomach cancer
5	3581*	≥ 40	12K	diabetes
6	3581*	≥ 40	9K	aids
7	355**	3*	8K	aids
8	355**	3*	10K	flu
9	355**	3*	11K	lung cancer

L'attaque d'inférence sémantique (similarity attack)

Considérations sémantiques:

- **low =>** {4K, 5K, 6K},
- **medium** => {7K, 8K, 9K},
- **high** => {10K, 11K, 12K}

Table 3-diversifiée

	ZIP Code*	Age*	Salary	Disease
1	355**	2*	4K	colon cancer
2	355**	2*	5K	stomach cancer
3	355**	2*	6K	lung cancer
4	3581*	$\geq$ 40	7K	stomach cancer
5	3581*	$\geq$ 40	12K	diabetes
6	3581*	$\geq$ 40	9K	aids
7	355**	3*	8K	aids
8	355**	3*	10K	flu
9	355**	3*	11K	lung cancer

La t-closeness réduit le risque d'inférence sémantique

- La *t-proximité* **améliore** la ***l-diversité*:** diversification des valeurs sensibles en prenant en compte la sémantique des données

Table 3-diversifiée

	ZIP Code*	Age*	Salary	Disease
1	355**	2*	4K	colon cancer
2	355**	2*	5K	stomach cancer
3	355**	2*	6K	lung cancer
4	3581*	≥ 40	7K	stomach cancer
5	3581*	≥ 40	12K	diabetes
6	3581*	≥ 40	9K	aids
7	355**	3*	8K	aids
8	355**	3*	10K	flu
9	355**	3*	11K	lung cancer

0.167-proximité par rapport à Salary

	ZIP Code*	Age**	Salary	Disease
1	3556*	≤ 40	4K	colon cancer
3	3556*	≤ 40	6K	lung cancer
8	3556*	≤ 40	10K	flu
4	3581*	≥ 40	7K	stomach cancer
5	3581*	≥ 40	12K	diabetes
6	3581*	≥ 40	9K	aids
2	3550*	≤ 40	5K	stomach cancer
7	3550*	≤ 40	8K	aids
9	3550*	≤ 40	11K	lung cancer

Comparaison des techniques en fonction des risques

Méthodes	Risque d'individualisation	Risque de corrélation	Risque d'inférence
Pseudonymisation	OUI	OUI	OUI
K-anonymisation	NON	OUI	Peut-être pas
L-diversité	NON	OUI	Peut-être pas
T-closeness	NON	OUI	Peut-être pas

Evaluation de l’anonymisation

Evaluation de la viabilité des données

Evaluation de la viabilité

- Il existe plusieurs manières d'évaluer la viabilité:
  - **De manière générique**:
    - **Evaluation de la perte de données**: certaines techniques engendrent des pertes de données l'une de méthodes d'évaluation de la viabilité consiste à évaluer cette perte de données
    - **Evaluation de la quantité d'information:** celle perdue lors de l'anonymisation. Se base sur l'entropie, qui est proportionnelle à la viabilité des données
  - **De manière spécifique**: en exécutant le test directement sur les données anonymisées

Evaluation générique de la viabilité des données

- **Evaluation de la perte de données**: les suppressions peuvent être globales (toute la ligne) ou locales (certaines colonnes).

- **Evaluation de l'incertitude:** basée sur le calcul d e l'entropie

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log(p(x))$$

Calculation of entropy to assess data viability

Hiérarchies des valeurs de l'attribut Age

Age	Age*	Age**
22	2*	2*
22	2*	2*
22	2*	2*
45	>= 40	>= 30
63	>= 40	>= 30
40	>= 40	>= 30
35	3*	>= 30
35	3*	>= 30
32	3*	>= 30

- **L'entropie est proportionnelle à la viabilité des données**

Calcul de l'entropie des hiérarchies des valeurs de l'attribut Age

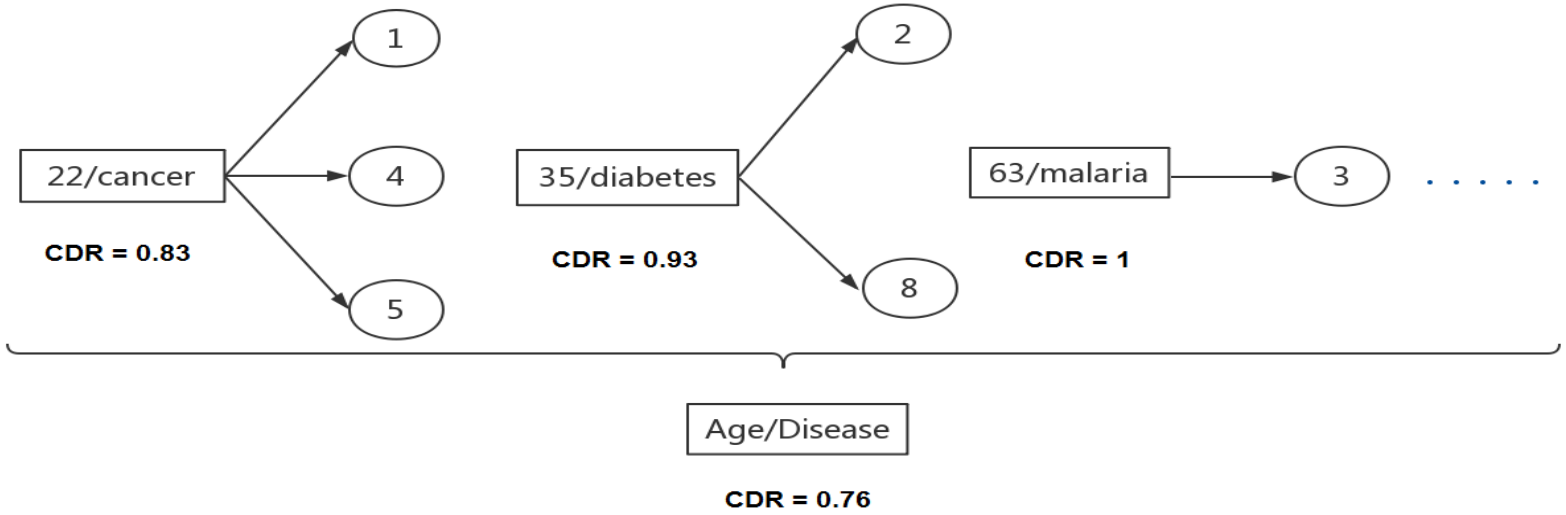
QI	H()
Age	2,41
Age*	1,58
Age**	0,91

Evaluation de la confidentialité des données

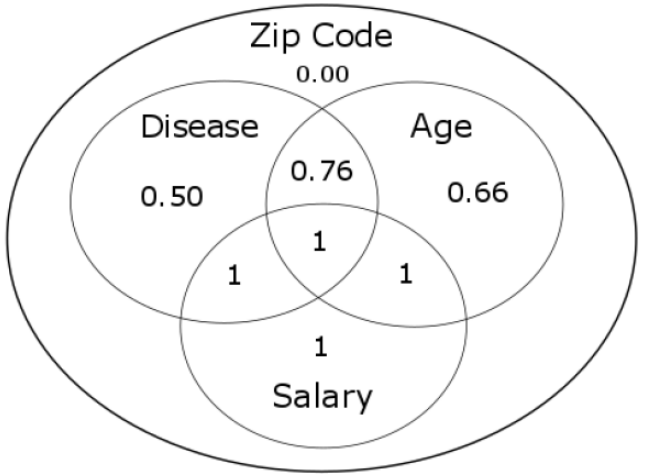
Calcul d'un CDR

Example Table

	ZIP Code	Age	Salary	Disease
1	35000	22	4K	cancer
2	35000	35	5K	diabetes
3	35000	63	3K	malaria
4	35000	22	13K	cancer
5	35000	22	8K	cancer
6	35000	35	15K	malaria
7	35000	45	9K	malaria
8	35000	35	7K	diabetes
9	35000	40	11K	diabetes

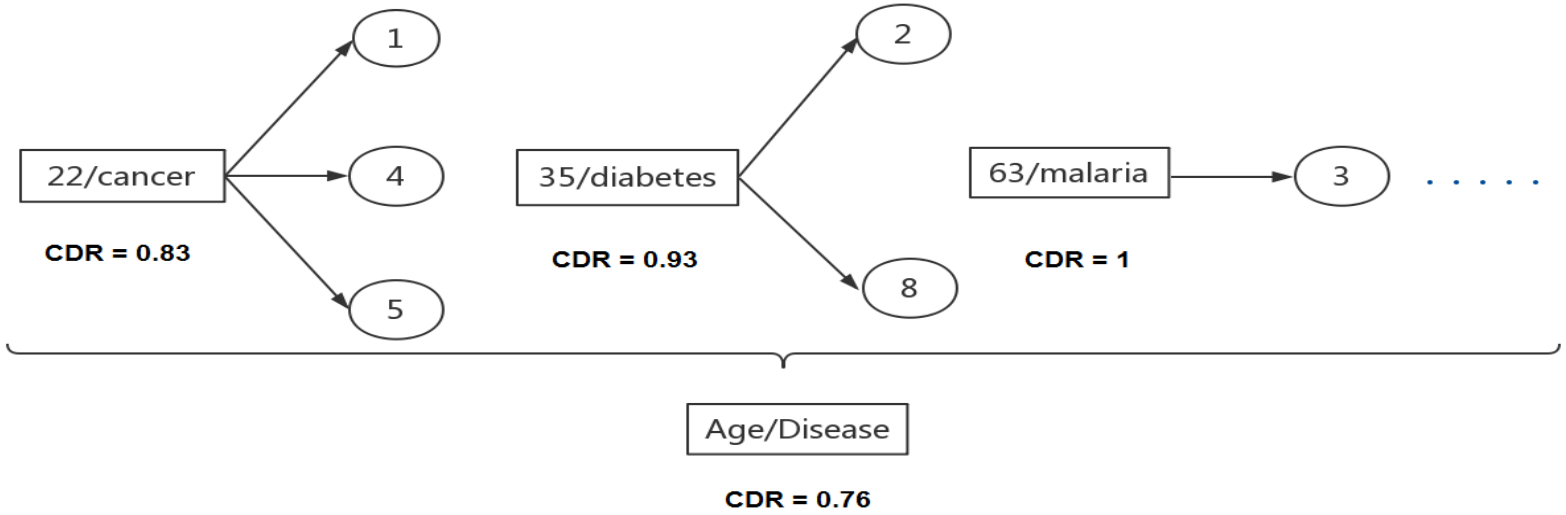


Calcul d'un CDR



Example Table

	ZIP Code	Age	Salary	Disease
1	35000	22	4K	cancer
2	35000	35	5K	diabetes
3	35000	63	3K	malaria
4	35000	22	13K	cancer
5	35000	22	8K	cancer
6	35000	35	15K	malaria
7	35000	45	9K	malaria
8	35000	35	7K	diabetes
9	35000	40	11K	diabetes



Evaluation de l'attaque d'individualisation

Original

	Age	Disease
1	22	lung cancer
2	22	lung cancer
3	22	lung cancer
4	45	stomach cancer
5	63	diabetes
6	40	aids
7	35	aids
8	35	flu
9	32	diabetes

Generalization Table

	Age	Age*
1	22	2*
2	22	2*
3	22	2*
4	45	≥ 40
5	63	≥ 40
6	40	≥ 40
7	35	3*
8	35	3*
9	32	3*

3-anonymous

	Age*	Disease
1	2*	lung cancer
2	2*	lung cancer
3	2*	lung cancer
4	≥ 40	stomach cancer
5	≥ 40	diabetes
6	≥ 40	flu
7	3*	aids
8	3*	aids
9	3*	diabetes

Evaluation de l'attaque d'individualisation

X	Y	$DR_X(Y)$
Age	2*	1
Age	≥ 40	0.78
Age	3*	0.87
Age	Age*	0.66

Evaluation de l'attaque d'inférence

3-anonymous Table

	Age*	Disease
1	2*	lung cancer
2	2*	lung cancer
3	2*	lung cancer
4	≥ 40	stomach cancer
5	≥ 40	diabetes
6	≥ 40	flu
7	3*	aids
8	3*	aids
9	3*	diabetes

3-diverse Table

	Age*	Disease
1	< 40	lung cancer
2	< 40	lung cancer
3	< 40	lung cancer
4	≥ 40	stomach cancer
5	≥ 40	diabetes
6	≥ 40	flu
7	< 40	aids
8	< 40	aids
9	< 40	diabetes

Evaluation de l'attaque d'inférence

X	Y	$DR_X(Y)$
3-anonymous Table		
Disease	2*	1
Disease	≥ 40	0.70
Disease	3*	0.83
Disease	Age*	0.52
3-diverse Table		
Disease	< 40	0.40
Disease	≥ 40	0.78
Disease	Age*	0.18

Evaluation de l'attaque d'inférence sémantique

0,167-closeness Table w.r.t. Salary

	ZIP Code*	Age**	Salary	Disease
1	3556*	≤ 40	4K	colon cancer
3	3556*	≤ 40	6K	lung cancer
8	3556*	≤ 40	10K	flu
4	3581*	≥ 40	7K	stomach cancer
5	3581*	≥ 40	12K	diabetes
6	3581*	≥ 40	9K	aids
2	3550*	≤ 40	5K	stomach cancer
7	3550*	≤ 40	8K	aids
9	3550*	≤ 40	11K	lung cancer

3-diverse Table

	ZIP Code*	Age*	Salary	Disease
1	355**	2*	4K	colon cancer
2	355**	2*	5K	stomach cancer
3	355**	2*	6K	lung cancer
4	3581*	≥ 40	7K	stomach cancer
5	3581*	≥ 40	12K	diabetes
6	3581*	≥ 40	9K	aids
7	355**	3*	8K	aids
8	355**	3*	10K	flu
9	355**	3*	11K	lung cancer

- **SP3 = {{4K, 6K, 10K}, {7K, 12K, 9K}, {5K, 8K, 11K}}**

Evaluation de l'attaque d'inference sémantique

X	Y	$sDR_X(Y)$
3-diverse Table		
SP_3	2*	0.81
SP_3	≥ 40	1
SP_3	3*	0.81
SP_3	Age*	0.61
SP_3	355**	0.58
SP_3	3581*	1
SP_3	**ZIP Code***	**0.58**
t-closeness Table		
SP_3	≤ 40	0.58
SP_3	≥ 40	1
SP_3	Age**	0.58
SP_3	3550*	1
SP_3	3581*	1
SP_3	3556*	1
SP_3	**ZIP Code***	**1**

Démo: anonymisation des données avec ARX

Louis Philippe SONDECK
louissondeck.com
Consultant expert RGPD et sécurité des donnés
Tel: 06 45 07 54 56

Annexes

Exemples de données longitudinales

Table originale

Subjects	ZIP Code	Age	Visit Date	Disease
1	75018	22	21/01/2019	cancer
2	75002	35	13/03/2019	diabetes
3	75001	63	02/02/2019	flu
1	75018	22	27/01/2019	aids
1	75018	22	13/01/2019	cancer
2	75002	35	01/04/2019	malaria
3	75001	63	01/01/2019	malaria
4	75007	35	24/01/2019	diabetes
5	75015	40	12/07/2019	malaria

Table données stables

Subjects	ZIP Code	Age
1	75018	22
2	75002	35
3	75001	63
4	75007	35
5	75015	40

Table données variables

Subjects	Visit Date	Disease
1	21/01/2019	cancer
2	13/03/2019	diabetes
3	02/02/2019	flu
1	27/01/2019	aids
1	13/01/2019	cancer
2	01/04/2019	malaria
3	01/01/2019	malaria
4	24/01/2019	diabetes
5	12/07/2019	malaria