ISIFC3-esanté, Sécurité Appliquée

Données de santé : introduction à la protection de la vie privée

Jean-François Couchot

Université de Franche-Comté, UFR-ST

Plan

Big data et vie privée

Données de santé : dé-identifier les documents ?

Anonymisation par pseudonymisation : attaquable

PVP : k-anonymat et extensions

Plan

Plan

PVP : Exacerbé par le Big Data

Big Data & Data mining

- L'exploration de données (Data Mining) : inférence de connaissances intéressantes à partir de grandes quantités de données (Big Data)
- Tendances générale : exploration des données en croissance // Big Data
- Analyse/extraction de connaissance : techniquement réalisable aujourd'hui

Big Data : application et volume

- Domaines : veille économique, découverte scientifique, santé, profilage
- Marché du Big Data en santé : $\approx$ 67,82 G$ en 2025 (Globe News Wire)

- 90% de toutes les données : créées au cours des deux dernières années (IBM)
- 97,2% des organisations : investissent en Big Data et IA. (New Vantage)
- Offres d'emploi dans le domaine : $\approx$ 2,7M en 2020. (Forbes)

Plan

Problématique

Contexte médical

- Apprentissage automatique :
  - Performant même sur des données non structurées (textuelles)
  - Problèmes accessibles : dossiers similaires, flux aux urgences, codes CIM10,. . .
  - Réalisable informatiquement par des spécialistes en apprentissage
  - Nécessité de "partager $\rightsquigarrow$ dé-identifier" les données
- Dilemme : partager ou dégrader les données

Questions

- Qu'est-ce qui régit le traitement des données de santé (textuelles) ?
- Techniques pour dé-identifier les documents ? Robustesse ?
- Approches d'apprentissage respectueux ?

Plan

Données de santé ? [1]

Europe : RGPD, cons. 35

- Les données à caractère personnel concernant la santé **devraient comprendre** les données [...] qui révèlent des informations sur l'état de santé physique ou mentale [...] d'une personne.
- Ensemble contenant au moins les informations :
  - collectées lors d'une inscription à un services de santé (NSS)
  - obtenues lors de tests/examens sur le corps : résultats d'analyses biologiques, génétiques ;
  - concernant une maladie, un handicap, des antécédents, un traitement clinique ou l'état physiologique ou biomédical d'une personne.
  - qui deviennent des données de santé, par croisement avec d'autres données : nombre de pas quotidien + mesure de poids ?

Question

- Pourquoi laisser tant de liberté à l'interprétation ? Avantages, inconvénients ?

1. https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante

Un régime juridique particulier [2]

Aperçu des différentes législations susceptibles de s'appliquer

- loi Informatique et Libertés (art. 8 et chapitre IX) ;
- dispositions sur le secret (art. L. 1110-4 du CSP) ;
- dispositions relatives aux référentiels de sécurité et d'interopérabilité des données de santé (art. L. 1110-4-1 du CSP) ;
- dispositions sur l'hébergement des données de santé (art. L. 1111-8 et R. 1111-8-8 et s. du CSP) ;
- dispositions sur la mise à disposition des données de santé (art. L. 1460-1 et s. du CSP) ;
- interdiction de procéder à une cession ou à une exploitation commerciale des données de santé (art. L. 1111-8 du CSP, art. L 4113-7 du CSP)...

Et ailleurs ?

Aux USA (HIPAA ?), au Canada ?

2. https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante

RGPD : Considérant 26

Applicabilité du RGPD aux données

- Applicable : sur des données à caractère personnel, susceptibles de permettre d'identifier (directement ou non) la personne.
- Non applicable : aux données initialement anonymes, ou rendues au moyen d'une démarche

« **Il n'y a dès 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**. »

RGPD : Considérant 26

Identification d'une personne physique

« 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. 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. »

Conséquences

- Nécessite d'évaluer régulièrement la qualité des fichiers rendus anonymes par le passé qui ont été diffusés (dont les acteurs sont responsables), de s'appuyer sur des méthodes éprouvé, publiées, dont les faiblesses sont connues pour réaliser cet anonymat, enfin de documenter l'anonymisation et les analyses de celle-ci.
- Repose sur la robustesse de la méthode d'anonymisation mise en place

Plan

Pseudonymisation

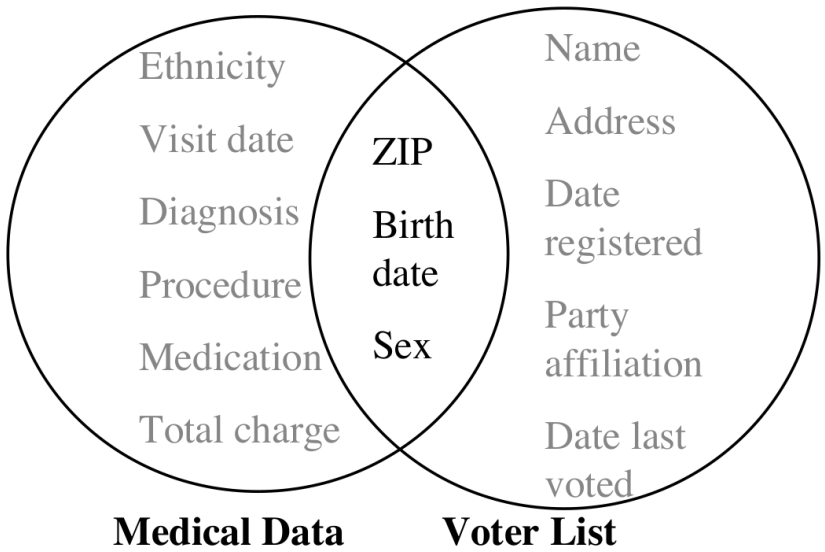
- Champs identifiants : supprimés et remplacés par un id (H(NSS)).

	Non-sensibles				Sensibles
H	CP	Age	Genre	Nationalité	Pathologie
1	13053	28	H	russe	trouble cardiaque
2	13068	29	H	américaine	trouble cardiaque
3	13068	21	F	japonaise	infection virale
4	13053	23	H	américaine	infection virale
5	14853	49	H	indienne	cancer
6	14853	48	F	russe	trouble cardiaque
7	14850	47	H	américaine	infection virale
8	14850	49	F	américaine	infection virale
9	13053	31	H	américaine	cancer
10	13053	37	H	indienne	cancer
11	13068	36	F	japonaise	cancer
12	13068	35	F	américaine	cancer

- Avantage : calculs identiques à ceux sur la base de données initiale (Age moyen/cancer=37,8)

Pseudonymisation : attaque par quasi-identifiants [3]

- Base de donnée médicale pseudonymisée et publique



- Liste d'électeurs publique, recensement USA, 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"
- CP, genre, date de naissance : quasi-identifiants
- Identification de données médicales du gouverneur Weld

3. Sweeney, L. (2002). k-anonymity : A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

Pseudonymisation : attaque par intersection [4]

2006, diffusion par AOL de 20M requêtes, 658K utilisateur sans nom

AnonID	Query	QueryTime
1326	“holiday mansion houseboat”	2006-03-29
1326	“back to the future”	2006-04-01
591476	“english spanish translator”	2006-03-20
591476	“panama vacations”	2006-03-20
591476	“breast reduction”	2006-03-23
591476	“volunteer work at hospitals in brooklyn”	2006-05-24
591476	...	...
591476	“how to secretly poison your ex”	2006-03-12

Thelma Arnold, 62 ans, veuve vivant à Lilburn, Ga., réidentifiée en 3 j.

AnonID	Query
4417749	“people with last name ’Arnold”’
4417749	“landscapers in Lilburn, Ga”
4417749	“60 single men”
4417749	“dog that urinates on everything”
4417749	dog-related queries

⇝ Suppression hâtive des données sur le site d’AOL.

4. BARBARO, Michael, ZELLER, Tom, et HANSELL, Saul. A face is exposed for AOL searcher no. 4417749. New York Times, 2006, vol. 9, no 2008, p. 8.

Plan

Plan

Le k-anonymat[3] et les quasi-identifiants -1

Quasi-identifiants : QID

- QID, intuition[5] : "des éléments qui ne sont pas en eux-mêmes des identificateurs uniques, mais qui sont suffisamment bien corrélés avec une entité pour pouvoir être combinés avec d'autres quasi-identifiants afin de créer un identificateur unique"
- QID, définition[6] : Les attributs de $Q \subseteq \{A_1, \ldots, A_M\}$ sont quasi-identifiants de la relation T si la requête suivante retourne au moins un résultat

```
SELECT Q FROM T
GROUP BY Q
HAVING COUNT(*)=1
```

- (CP, genre, date de naissance) : triplets uniques dans 87% des cas $\rightsquigarrow$ quasi-identifiants

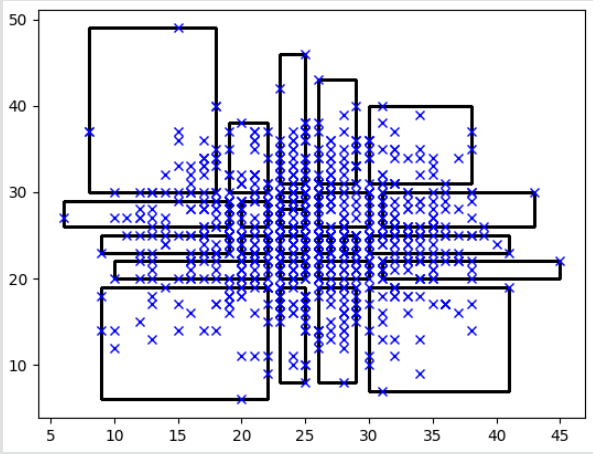
5. https://en.wikipedia.org/wiki/Quasi-identifier

6. Nguyen, B., & Castelluccia, C. (2020). Techniques d'anonymisation tabulaire : concepts et mise en oeuvre. arXiv preprint arXiv :2001.02650.

Le k-anonymat^3 et les quasi-identifiants -2

Intuition : regrouper les QID pour casser l'unicité

- Niveau de détail des valeurs des QID : à réduire pour qu'il y ait au moins k individus différents dont les QIDs sont égaux
- Individus avec mêmes QIDs : font partie de la même classe d'équivalence



k-anonymat par généralisation

Réduction des niveaux de détail

Sur l'exemple :

- CP : laisser, **regroupements par 2**, suppr.
- Age : laisser, par intervalles d'amplitudes 10, 20,**suppr.**
- Genre : laisser, **suppr.**
- Nationalité : laisser, par continent, **suppr.**
- $\rightsquigarrow 3 \times 4 \times 2 \times 3 = 72$ combinaisons de généralisation !

Regroupement par classes d'équivalence de card. $\geq$ à 4

CP	Age	Genre	Nationalité	Pathologie
	[21; 31[	*	*	trouble cardiaque
{13053	[21; 31[	*	*	trouble cardiaque
13058}	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
	[41; 50[	*	*	cancer
{14850	[41; 50[	*	*	trouble cardiaque
14853}	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
	[31; 41[	*	*	cancer
{13053	[31; 41[	*	*	cancer
13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

4 individus

4 individus

4 individus

Plan

C_{AVG} : nbre. moy. d'élements par classe normalisé

Définition pour $|T|$ enregistrements et propriétés

$$C_{AVG} = \frac{|T|}{|EQs|} \times \frac{1}{k}$$

- et $|EQs|$: nbre. de classes d'équiv.
- $\frac{|T|}{|EQs|}$ nbre. moy. d'éléments par classe $(\geq k) \rightsquigarrow C_{AVG} \geq 1$
- Utilité décroissante à mesure que C_{AVG} croît.

Exemple avec 4-anonymat

CP	Age	Genre	Nationalité	Pathologie
	[21; 31[	*	*	trouble cardiaque
{13053	[21; 31[	*	*	trouble cardiaque
13058}	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
	[41; 50[	*	*	cancer
{14850	[41; 50[	*	*	trouble cardiaque
14853}	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
	[31; 41[	*	*	cancer
{13053	[31; 41[	*	*	cancer
13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

- $C_{AVG} = \frac{12}{3} \times \frac{1}{4} = 1$
- Optimal pour cette métrique

Loss : Perte dûe à la généralisation

Définition pour $|T|$ enregistrements, n QID et propriétés

$$Loss = \frac{1}{n|T|} \sum_{j=1}^{|T|} \sum_{i=1}^{n} \frac{R_{i_j}}{R_i}$$

- R_{i_j}/R_i : rapport acquis/acquérables
  - discret : $(|\text{généralisation } i_j| - 1)/(|\text{Qid } i| - 1)$
  - continu : $(U_{i_j} - L_{i_j})/(U_i - L_i)$
- Moyenne des acquis/acquérables, utilité décroissante // *Loss* croît

Exemple avec 4-anonymat

CP	Age	Genre	Nationalité	Pathologie
{13053 13058}	[21; 31[	*	*	trouble cardiaque
	[21; 31[	*	*	trouble cardiaque
	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
{14850 14853}	[41; 50[	*	*	cancer
	[41; 50[	*	*	trouble cardiaque
	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
{13053 13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

- Nat. $= \{\text{russe}, \text{am.}, \text{jap.}, \text{ind.}\} \rightsquigarrow$ $R_{\text{Nat}} = 3$, $R_{\text{CP}} = 3$, $R_{\text{Genre}} = 1$,
- $R_{\text{Age}} = 50 - 21 = 29$
- $Loss = \frac{1}{4\times 12}(8 \times (\frac{1}{3} + \frac{10}{29} + \frac{1}{1} + \frac{3}{3}) + 4 \times (\frac{1}{3} + \frac{9}{29} + \frac{1}{1} + \frac{3}{3})) = \frac{2}{3}$

Plan

k-anonymat : attaques

Exemple

CP	Age	Genre	Nationalité	Pathologie
	[21; 31[	*	*	trouble cardiaque
{13053	[21; 31[	*	*	trouble cardiaque
13058}	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
	[41; 50[	*	*	cancer
{14850	[41; 50[	*	*	trouble cardiaque
14853}	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
	[31; 41[	*	*	cancer
{13053	[31; 41[	*	*	cancer
13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

} 4 individus

} 4 individus

} 4 individus

Attaques

- Homogénéité :
  - $\oplus$ Patient de 35 ans connu $\rightsquigarrow$ cancer.
  - $\ominus$ Patient de 29 ans connu $\rightsquigarrow$ ~~cancer~~.
- Connaissance supplémentaire : un japonais de 21 ans, P(trouble cardiaque|japonais)=faible $\rightsquigarrow$ infection virale.

Plan

l-diversité pour contrer l'homogénéité

Définition et remarque

- Classe d'équivalence : l-diverse si contient au moins l valeurs "représentatives" par donnée sensible
- Base de données : l-diverse si toutes ses classes d'équivalence le sont
- Représentatives : valeurs distinctes $\subset$ répartition sensée...

4-anonymat et 3-diversité par valeurs distinctes

Quasi-IDentifiants				Sensibles
CP	Age	Genre	Nationalité	Pathologie
130**	[21; 41[	*	*	trouble cardiaque
130**	[21; 41[	*	*	trouble cardiaque
130**	[21; 41[	*	*	infection virale
130**	[21; 41[	*	*	infection virale
130**	[21; 41[	*	*	cancer
130**	[21; 41[	*	*	cancer
130**	[21; 41[	*	*	cancer
130**	[31; 41[	*	*	cancer
148**	[41; 50[	*	*	cancer
148**	[41; 50[	*	*	trouble cardiaque
148**	[41; 50[	*	*	infection virale
148**	[41; 50[	*	*	infection virale

} 3 val. sensibles $\neq$

} 3 val. sensibles $\neq$

Avec des éléments $\neq$ mais sémantiqu^t proches ?

Exemple agrégé avec salaire et maladies sensibles

CP	Age	Salaire	Pathologie
476**	2*	3K	ulcère gastrique
476**	2*	4K	gastrite
476**	2*	5K	cancer de l'estomac
4790*	$\geq$ 40	6K	gastrite
4790*	$\geq$ 40	11K	grippe
4790*	$\geq$ 40	8K	bronchite
476**	3*	7K	bronchite
476**	3*	9K	pneumonie
476**	3*	10K	cancer de l'estomac

- Satisfait la 3-diversité par valeurs distinctes.

Fuite due à la non prise en compte de la proximité de valeurs.

Déductions possibles de la connaissance que Bob est dans la classe 1 :

- son salaire ([3K-5K]) est relativement bas
- souffre de l'estomac (toutes les pathologies y sont liées)