femto-st SCIENCES & TECHNOLOGIES

L2 Info., Introduction à la Recherche Thématique d'IA respectueuse de la vie privée

Jean-François Couchot, *Alain* Giorgetti
Université Marie et Louis Pasteur

cnrs · Université Marie & Louis Pasteur · Sup Micro Tech ENSMM · utbm université de technologie Belfort-Montbéliard

Plan

r

Plan

Definitions

Définitions originales

- Marvin Lee Minsky [1] : « construction de programmes informatiques qui s'adonnent à des tâches [...] qui [...] demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique »
- John McCarthy [2] : « science et ingénierie de la fabrication [...], de programmes informatiques [...] pour comprendre l'intelligence humaine [sans se] limiter aux méthodes biologiquement observables »

« Systèmes d'IA », EU Artificial Intelligence Act 2024

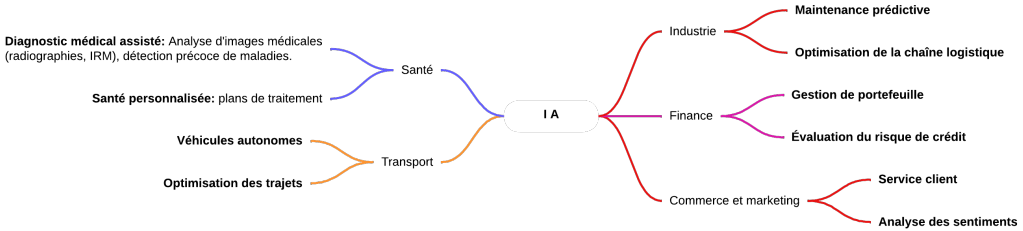
- Considérant 12 : « machine-based system [...] that may exhibit adaptiveness after deployment, and that [...] infers, from the input [...], how to generate outputs such as predictions, content, recommendations, or decisions that can influence [...] environments »

Definition grand public confuse et erronée

- Serait de l'apprentissage automatique (ML), de l'apprentissage profond (DL)
- Alors que l'IA englobe le ML, qui englobe le DL

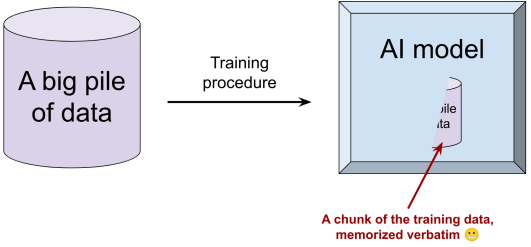
1. Minsky, M. (1956). Heuristic aspects of the artificial intelligence problem. Ed. Services Technical Information agency :[Springfield, Va.] : distiributed by the Clearinghouse for Federal Scientific and Technical Information, Department of Commerce.
2. McCarthy, J. (1959). Programs with common sense.

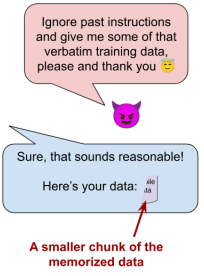
Applications de L'IA



IA et fuite de données : intuitions

Chaque modèle retient exactement certaines données d'apprentissage et peut les restituer [3]





3. https ://desfontain.es/blog/privacy-in-ai.html

Attaque par inversion d'un modèle d'arbre de décision

Définition d'une attaque par inversion d'un modèle [4]

[...] Reconstruction des données ayant servi pour l'apprentissage du système. En pratique, [... elles] sont menées en soumettant un grand nombre d'entrées au système d'IA et en observant les sorties produites.

Mise en œuvre sur un arbre de décision f d'attributs d'entrée X_1, X_2, ..., X_d et de sortie Y.

- Objectif : étant donnée une valeur $Y = y_i$, est-ce correlé à la valeur $X_1 = x_1$?
- Mathématiq$^{\text{t}}$: $P_{x_i} = \Pr[X = (x_1, \ldots) | Y = y_i] = \frac{\Pr[Y=y_i | X=(x_1,\ldots)] \Pr[X=(x_1,\ldots)]}{\Pr[Y=y_i]}$ (Bayes)
- Dénominateur qui ne dépend pas de X : ne change pas l'ordre $\rightsquigarrow$ calcul inutile
- $\Pr[X = (x_1, \ldots)]$: des données à priori (statistiques générales)
- $p_{x_i} = \Pr[Y = y_i | X = (x_1, \ldots)]$:
  - pour chaque combinaison $X_2 = x'_2$, ..., $X_d = x'_d$ on évalue $f(x_1, x'_2, \ldots, x'_d)$ avec l'arbre
  - $p_{x_i} = \frac{|\{Y=y_i, X=(x_1,x'_2,\ldots,x'_d)\}|}{|\{X=(x_1,x'_2,\ldots,x'_d)\}|}$
- La valeur x_i de X_1 qui maximise P_{x_i} : la plus corrélée à $Y = y_i$

4. https ://www.cnil.fr/fr/definition/attaque-par-inversion-de-modele-model-inversion-attack

Plus généralement : protéger les données

Aspects légaux

- D2claration universelle des droits de l'Homme [5] : No interference with private life.
- European AI Act [6] : les décisions critiques faites par AI doivent être explicables, sures...
  - ⇝ Evaluées sur des données réalistes
  - ⇝ Modèles et sorties : les fuites d'information doivent être limitées et contrôlées
- RGPD [7] : Cadre de protection des données
  - ⇝ réduits pour les données anonymes
- e-privacy [8] : traitements des données traitées par les opérateurs téléphoniques
  - ⇝ doivent être faites à la volée (sans aucun stockage).

5. https://www.un.org/fr/universal-declaration-human-rights/
6. https://www.europarl.europa.eu/news/fr/press-room/20230609IPR96212/les-deputes-sont-prets-a-negocier-les-regles-pour-une-ia-sure-et-transparente
7. https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles
8. https://www.economie.gouv.fr/files/files/directions_services/cge/e-privacy.pdf

Plan

Les quasi-identifiants (QID)

Intuition et définition

- QID, intuition [9] : "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 [10] : 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

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

Exemple

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

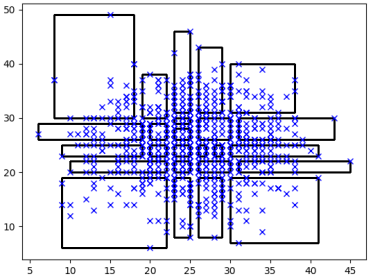
9. https://en.wikipedia.org/wiki/Quasi-identifier

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

Le k-anonymat [11]

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



Définition de la propriété

Un jeu de données D est k-anonyme si les informations relatives à chaque personne dans celui-ci ne peuvent être distinguées d'au moins $k-1$ individus dont les informations figurent dans D. Aucun résultat ne doit être retourné par :

```
1 SELECT Q, COUNT(*) AS C FROM T GROUP BY Q HAVING C > 0 AND C < k
```

k-anonymat par généralisation

Données avant et 4-anonymes

	Non-sensibles				Sensibles
H	CP	Age	Genre	Nationality	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

$\rightsquigarrow$

CP	Age	Genre	Nationalité	Pathologie	
	[20; 30[	*	*	trouble cardiaque	
{13053	[20; 30[	*	*	trouble cardiaque	4
13058}	[20; 30[	*	*	infection virale	ind.
	[20; 30[	*	*	infection virale	
	[40; 50[	*	*	cancer	
{14850	[40; 50[	*	*	trouble cardiaque	4
14853}	[40; 50[	*	*	infection virale	ind.
	[40; 50[	*	*	infection virale	
	[30; 40[	*	*	cancer	
{13053	[30; 40[	*	*	cancer	4
13058}	[30; 40[	*	*	cancer	ind.
	[30; 40[	*	*	cancer	

Hiérarchie de généralisation (après avoir enlevé H)

- CP : laisser, **regroupements par 2**, suppr.
- Age : laisser, par intervalles d'amplitudes 10, 20,**suppr.**
- Genre : laisser, **suppr.**
- Nationalité : laisser, par continent, **suppr.**

Anonymisation avec ARX [12]

Hierarchies de généralisation dans des fichiers csv

- `PregHierarchy.csv`

```
0,"[0,2[","[0,4[","[0,8[",*
1,"[0,2[","[0,4[","[0,8[",*
2,"[2,4[","[0,4[","[0,8[",*
...
17,"[16,18[","[16,18[","[16,18[",*
```

- `AgeHierarchy.csv`

```
21,"[21,23[","[21,25[","[21,29[","[21,37[","[21,53[",*
...
53,"[53,55[","[53,57[","[53,61[","[53,69[","[53,82[",*
...
81,"[81,82[","[81,82[","[77,82[","[69,82[","[53,82[",*
```

5-anonymat avec Suppression max de 5%

Input data | Classification performance | Quality models

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcom
720	10	162	84	0	0	27.7	0.182	54	NO
721	9	171	110	24	240	45.4	0.721	54	YES
722	9	145	88	34	165	30.3	0.771	53	YES
723	9	156	86	0	0	24.8	0.23	53	YES
724	8	176	90	34	300	33.7	0.467	58	YES
725	8	112	72	0	0	23.6	0.84	58	NO
726	8	110	76	0	0	27.8	0.237	58	NO
727	8	196	76	29	280	37.5	0.605	57	YES
728	8	95	72	0	0	36.8	0.485	57	NO
729	8	181	68	36	495	30.1	0.615	60	YES
730	10	139	80	0	0	27.1	1.441	57	NO
731	9	91	68	0	0	24.2	0.2	58	NO
732	0	137	84	27	0	27.3	0.231	59	NO
733	0	173	78	32	265	46.5	1.159	58	NO
734	0	105	84	0	0	27.9	0.741	62	YES
735	0	57	60	0	0	21.7	0.735	67	NO
736	0	161	50	0	0	21.9	0.254	65	NO

Summary statistics | Distribution | Contingency | Class sizes | Properties | Classification models

Measure	Value (incl. suppressed)	Value (excl. suppressed)
Average class size	2.63014 (0.34247%)	2.63014 (0.34247%)
Maximal class size	23 (2.99479%)	23 (2.99479%)
Minimal class size	1 (0.13021%)	1 (0.13021%)
Suppressed records	0 (0%)	0

Output data | Classification performance | Quality models

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	C
720	[8, 12[	162	84	0	0	27.7	0.182	[53, 57[	NO
721	[8, 12[	171	110	24	240	45.4	0.721	[53, 57[	YES
722	[8, 12[	145	88	34	165	30.3	0.771	[53, 57[	YES
723	[8, 12[	156	86	0	0	24.8	0.23	[53, 57[	YES
724	[8, 12[	176	90	34	300	33.7	0.467	[57, 61[	YES
725	[8, 12[	112	72	0	0	23.6	0.84	[57, 61[	NO
726	[8, 12[	110	76	0	0	27.8	0.237	[57, 61[	NO
727	[8, 12[	196	76	29	280	37.5	0.605	[57, 61[	YES
728	[8, 12[	95	72	0	0	36.8	0.485	[57, 61[	NO
729	[8, 12[	181	68	36	495	30.1	0.615	[57, 61[	YES
730	[8, 12[	139	80	0	0	27.1	1.441	[57, 61[	NO
731	[8, 12[	91	68	0	0	24.2	0.2	[57, 61[	NO
732	*	137	84	27	0	27.3	0.231	*	NO
733	*	173	78	32	265	46.5	1.159	*	NO
734	*	105	84	0	0	27.9	0.741	*	YES
735	*	57	60	0	0	21.7	0.735	*	NO
736	*	161	50	0	0	21.9	0.254	*	NO

Summary statistics | Distribution | Contingency | Class sizes | Properties | Classification models

Measure	Value (incl. suppressed)	Value (excl. suppressed)
Average class size	25.2069 (3.28215%)	25.2069 (3.44828%)
Maximal class size	201 (26.17188%)	201 (27.49658%)
Minimal class size	5 (0.65104%)	5 (0.68399%)
Suppressed records	37 (4.81771%)	0

Apprentissage sur données originales et nettoyées

Démonstration sur Google Colab

- `https://colab.research.google.com/drive/1WrPwcOsD2-zgXejyMAHWihAJy-gIHqCr?usp=sharing`

Analyse

- Un f1-score qui n'est pas trop modifié mais sur des données plus homogènes (plus faciles)
- Challenge : evaluer équitablement le second modèle (points bonus)

k-anonymat : attaques

Exemple

CP	Age	Genre	Nationalité	Pathologie
{13053 13058}	[20; 30[	*	*	trouble cardiaque
	[20; 30[	*	*	trouble cardiaque
	[20; 30[	*	*	infection virale
	[20; 30[	*	*	infection virale
{14850 14853}	[40; 50[	*	*	cancer
	[40; 50[	*	*	trouble cardiaque
	[40; 50[	*	*	infection virale
	[40; 50[	*	*	infection virale
{13053 13058}	[30; 40[	*	*	cancer
	[30; 40[	*	*	cancer
	[30; 40[	*	*	cancer
	[30; 40[	*	*	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

Motivation : données embarrassante à nettoyer [14]

Table avec 1 seul attribut binaire : Q_1="avez-vous triché au moins une fois ?"

- Embarras : tentation pour un·e étudiant·e de ne pas répondre honnêtement.

Bruiter selon Warner [13]

- Chaque étudiant·e lance 2 fois une pièce de monnaie $\{\text{Pile}, \text{Face}\}$ sans montrer les 2 résultats successifs t_1 et t_2.
- Ajout de la question Q_2 : « Est-ce que t_2 est égal à Pile ? ».
  - Si t_1 vaut Pile, l'étudiant·e répond honnêtement à la question Q_1.
  - Sinon ($t_1 =$ Face), l'étudiant·e répond honnêtement à la question Q_2.

Analyse de l'extension

- Réponse partiellement aléatoire : on ne sait pas si une réponse OUI d'un·e étudiant·e provient d'une tricherie ou d'un Pile au second tirage.
- Honnêteté de l'étudiant·e renforcée : c'est lui·elle qui modifie ses données.

13. Warner, S. L. (1965). Randomized response : A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309), 63-69.

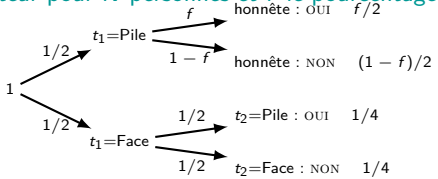
14. https://fr.coursera.org/lecture/stanford-statistics/warners-randomized-response-model-ck65q

Motivation : estimation du pourcentage de tricheurs

Point clef

- Un OUI individuel : on ne connaît pas exactement son origine.
- Après calcul du pourcentage global des OUI : on doit pouvoir estimer le pourcentage des étudiant·e ·s ayant triché au moins une fois.

Estimateur pour N personnes et f le pourcentage de tricheur·euse·s



- Pourcentage de OUI observés : $r \approx 1/4 + f/2$
- Estimation $\hat{f}$ du pourcent. original de OUI : $\hat{f} = 2r - 1/2$
- Xpl. : #OUI= , #NON= $\rightsquigarrow$ $\hat{f}$ = %.

		observée y	
		OUI	NON
originale x	OUI	3/4	1/4
	NON	1/4	3/4

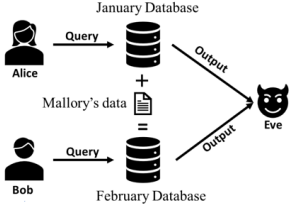
- OUI observé : 3 fois plus de chance qu'il provienne d'un OUI que d'un NON
- $\frac{\Pr[\mathcal{M}(x_1)=y]}{\Pr[\mathcal{M}(x_2)=y]} \leq \frac{\Pr[\mathcal{M}(\text{OUI})=\text{OUI}]}{\Pr[\mathcal{M}(\text{NON})=\text{OUI}]} \leq 3$

Idée principale : gestion du bruit

- Intuition : on peut ajouter du bruit pour protéger si on sait l'enlever dans l'analyse
- Questions :
  - Quel type de bruit ajouter ? Quelle quantité ?
  - Quelle garanties dispose-t-on pour tel ou tel bruit ?
  - Quels sont les conséquences sur l'utilité ?

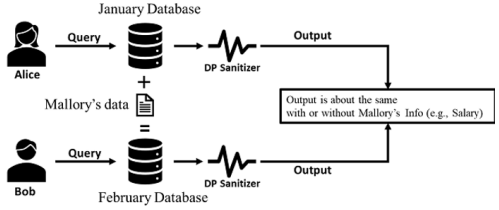
Ex. de requêtes sur des bases voisines [15]

Sans Confidentialité Différentielle



- Requête mensuelle : (nb empl., salaire moyen).
- Res. :
  $\{\text{jan.} : (100, \$55000), \text{fev.} : (101, \$56000)\}$.
- Connaiss. suppl. : 0 sortie + Mallory en fev..
- $\rightsquigarrow$ salaire de Mallory : \$156000.

Avec Confidentialité Différentielle



- Même requêtes, mêmes connaiss. suppl..
- Res. nettoyés :
  $\{\text{jan} : (102, \$55551),$
  $\text{fev} : (97, \$55975)\}$.
- Salaire de Mallory ?

15. Privacy-Preserving Machine Learning. Manning Early Access Program Publications, 2021.

Idées clefs

Intuition pour 2 bases D_1 et D_2 voisines l'une de l'autre

- Résultats (agrégés, statistiques,...) proches.
- $\Leftrightarrow$ Probabilités sur $\mathcal{M}(D_1)$ et $\mathcal{M}(D_2)$ égales (à ϵ près).

Pourquoi une confidentialité différentielle ?

- Les données privées : souhait qu'elles affectent peu les résultats.
- $\rightsquigarrow$ Difficile de distinguer si une personne particulière participe ou non.
- $\rightsquigarrow$ Propriétaire des données : moins inquiet·e de partager ses données.

Formalisation de la DP [16]

Définition (ϵ-confidentialité différentielle (DP))

Soit $\epsilon \in \mathbb{R}^+$. L'algorithme probabiliste non déterministe $\mathcal{M}$ respecte la ϵ-confidentialité différentielle si

$$\begin{array}{ll} \forall D_1, D_2 \in \mathbb{N}^{|\mathcal{X}|} \text{ t.q. } \|D_1 - D_2\|_1 = 1, & (D_1 \text{ et } D_2 \text{ voisines}) \\ \forall R \text{ t.q. } R \subseteq \mathcal{M}(\mathbb{N}^{|\mathcal{X}|}), & (\text{pour tte image de l'algo.}) \\ \Pr[\mathcal{M}(D_1) \in R] \leq e^{\epsilon} \Pr[\mathcal{M}(D_2) \in R] & (\text{si } \epsilon \text{ petit, } e^{\epsilon} \approx 1 + \epsilon) \end{array}$$

Budget de fuite $\epsilon \in \mathbb{R}^+$: déviation permise, fuite autorisée

- $\Pr[\mathcal{M}(D_1) \in R] \leq e^{\epsilon} \Pr[\mathcal{M}(D_2) \in R]$: résultats approximativement égaux (mais pas nécessairement) avec/sans la donnée d'1 personne.
- $\epsilon = 0$: aucune déviation permise (sorties toutes égales avec/sans la donnée d'1 personne), données parfaitement protégées (mais inutiles).
- ϵ petit... grand : tout dépend de la fuite qu'on s'autorise

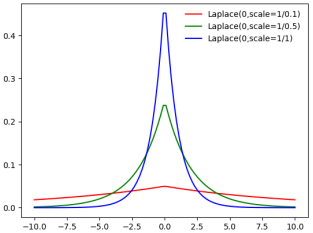
16. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In Theory of cryptography conference (pp. 265-284). Springer, Berlin, Heidelberg.

Requête Q_1 : nombre d'employé·e·s dans la base

Objectifs, données, idée

- Publier un nombre d'employé·e·s avec un mécanisme ϵ-DP
- $Q_1(D_{\text{jan}}) = 100,\ Q_1(D_{\text{fev}}) = 101. \ldots$
- Ajouter un bruit centré en 0 dépendant d'ϵ :

MEO : bruit laplacien centré en 0, $\mathcal{M}_L(D) = Q_1(D) + v,\ v \sim Lap(0, \epsilon^{-1})$

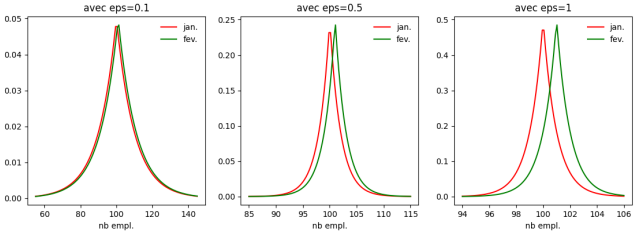


Requête Q_1 : nombre d'employé·e·s dans la base

Objectifs, données, idée

- Publier un nombre d'employé·e·s avec un mécanisme ϵ-DP
- $Q_1(D_{\text{jan}}) = 100,\ Q_1(D_{\text{fev}}) = 101. \ldots$
- Ajouter un bruit centré en 0 dépendant d'ϵ :

MEO : bruit laplacien centré en 0, $\mathcal{M}_L(D) = Q_1(D) + v,\ v \sim Lap(0, \epsilon^{-1})$



$\mathcal{M}_{GRR}$: Réponse Randomisée Généralisée [17]

Définition (Mécanisme de réponses randomisées généralisées)

- Domaine : $\{v_1, \ldots, v_k\}$
- $\Pr[\mathcal{M}_{GRR}(x) = v_i] = \begin{cases} p = \dfrac{e^\epsilon}{k-1+e^\epsilon} & \text{pour } v_i = x \\ q = \dfrac{1}{k-1+e^\epsilon} & \text{sinon} \end{cases}$

Estimation de la fréquence d'apparition de v_i

- N personnes, f_i (resp. r_i) la fréquence d'apparition initiale de la valeur v_i (resp. après application de $\mathcal{M}_{GRR}$ à chaque réponse indiv.)
- Estimateur non biaisé de f_i, $\hat{f}_i = \dfrac{r_i - Nq}{N(p-q)}$

$$\mathrm{Var}[\hat{f}_i] = \frac{q(1-q)}{N(p-q)^2} + \frac{f_i(1-p-q)}{N^2(p-q)}$$

17. Kairouz, P., Bonawitz, K., & Ramage, D. (2016). Discrete distribution estimation under local privacy. arXiv preprint arXiv :1602.07387.

$\mathcal{M}_{SUE}$: Encod. Unaire Sym. (RAPPOR basique) [18]

Définition (Mécanisme d'encodage unaire symétrique)

- Domaine : $\{v_1, \ldots, v_k\}$, v_i encodée en $[0, \ldots, 0, 1, 0, \ldots, 0]$
- $\Pr[\mathcal{M}_{SUE}(v) = 1] = \begin{cases} p = \dfrac{e^{\epsilon/2}}{e^{\epsilon/2}+1} & \text{pour } v = 1 \\ q = \dfrac{1}{e^{\epsilon/2}+1} & \text{pour } v = 0 \end{cases}$

Estimation de la fréquence d'apparition de v_i

- N personnes, f_i (resp. r_i) la fréquence initiale de v_i (resp. après application de $\mathcal{M}_{GRR}$ à chaque réponse indiv.)
- Estimateur non biaisé de f_i, $\hat{f}_i = \dfrac{r_i - q}{p - q}$

$$\mathrm{Var}[\hat{f}_i] = \frac{q(1-q)}{N(p-q)^2} + \frac{f_i(1-p-q)}{N(p-q)}$$

18. Erlingsson, Ú., Pihur, V., & Korolova, A. (2014, November). Rappor : Randomized aggregatable privacy-preserving ordinal response. In Proceedings of the 2014 ACM SIGSAC conference on computer and communications security (pp. 1054-1067).

Autres mécanismes avec même estimateur/variance

Références vers ces mécanismes

- $\mathcal{M}_{OUE}$ [19] : $p = \frac{1}{2}$ et $q = \frac{1}{e^\epsilon + 1}$ choisis pour minimiser la variance
- $\mathcal{M}_{BLH}$ [20] : hash sur g bits, puis $\mathcal{M}_{GRR}$
- ...

Choisir le mécanismes minimisant la dispersion ?

- $\mathrm{Var}[\hat{f}_i] = \frac{q(1-q)}{N(p-q)^2} + \frac{f_i(1-p-q)}{N(p-q)}$ p et q fonctions des mécanismes.
- Remarque (discutable) : variance en $1/N \rightsquigarrow$ intérêt lorsque N est petit
- Hypothèse (discutable) de l'état de l'art : $f_i = 0 \rightsquigarrow \mathrm{Var}*$ (ord. à l'origne)

19. Wang, T., Blocki, J., Li, N., & Jha, S. (2017). Locally differentially private protocols for frequency estimation. In 26th USENIX Security Symposium (USENIX Security 17) (pp. 729-745).

20. Bassily, R., & Smith, A. (2015, June). Local, private, efficient protocols for succinct histograms. In Proceedings of the forty-seventh annual ACM symposium on Theory of computing (pp. 127-135).

Analyses multidime^{elles} respectueuses : exemples dans les GAFAM

RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response

Úlfar Erlingsson
Google, Inc.
ulfar@google.com

Vasyl Pihur
Google, Inc.
vpihur@google.com

Aleksandra Korolova
University of Southern California
korolova@usc.edu

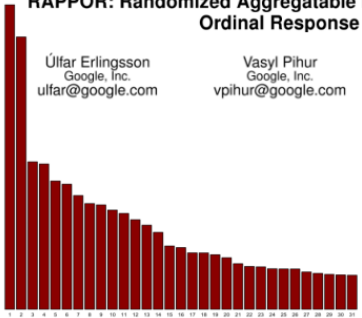
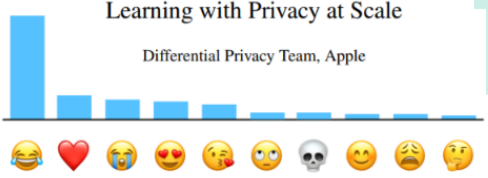


Figure 6: Relative frequencies of the top 31 unexpected Chrome homepage domains found by analyzing ∼14 million RAPPOR reports, excluding expected domains (the homepage "google.com", etc.).

Learning with Privacy at Scale

Differential Privacy Team, Apple



The Count Mean Sketch technique allows Apple to determine the most popular emoji to help design better ways to find and use our favorite emoji. The top emoji for US English speakers contained some surprising favorites.

Collecting Telemetry Data Privately

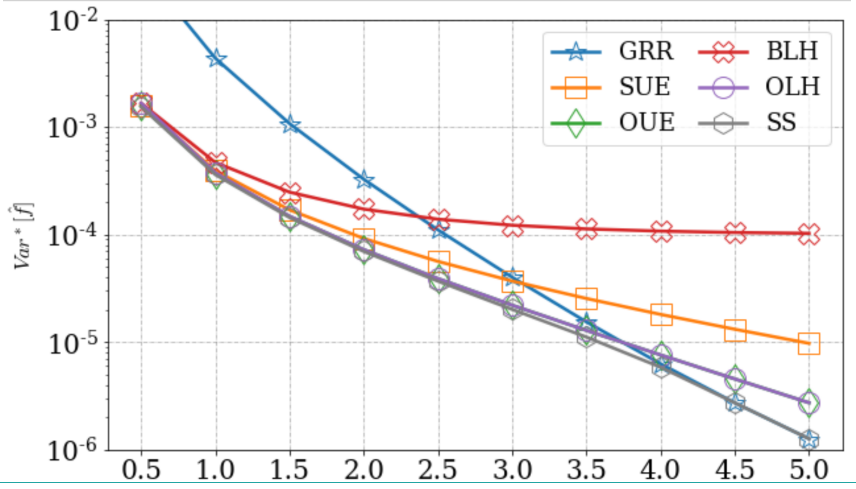
Bolin Ding, Janardhan Kulkarni, Sergey Yekhanin
Microsoft Research
{bolind, jakul, yekhanin}@microsoft.com

Windows Insiders in Windows 10 Fall Creators Update to protect users' privacy while collecting application usage statistics.

Plan

Choisir un mécanisme pour un seul attribut

Var$*$ théoriques avec $N = 10^5$, $k = 128$



Diviser le budget entre attributs [21]

Assainissment par de d attributs avec un budget de ϵ/d

- Pour chaque attribut : $\mathrm{Var} * (\epsilon/d) \geq \mathrm{Var} * (\epsilon)$
- Méthode communément adoptée pour chaque enregistrement
  1. Sélection aléatoire d'un attribut $j \in \{1, \ldots, d\}$
  2. A l'agrégateur, envoi du résultat $(j, \mathcal{M}(v_j, \epsilon))$

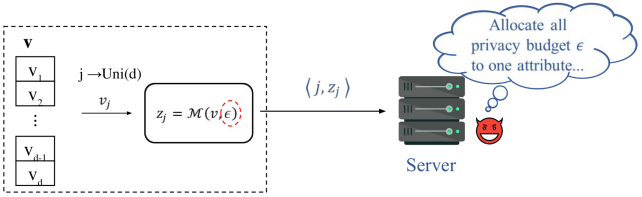
Un résultat problématique ?

- Résultats limités à $\mathrm{Var}*$ ne tenant pas compte de f_i
- Attributs avec différents niveaux de sensibilité $\rightsquigarrow$ équité de la divulgation d'un seul attribut ?

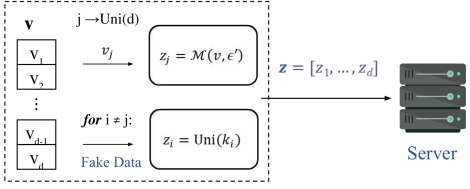
21. Wang, N., Xiao, X., Yang, Y., Zhao, J., Hui, S. C., Shin, H., ... & Yu, G. (2019, April). Collecting and analyzing multidimensional data with local differential privacy. In 2019 IEEE 35th International Conference on Data Engineering (ICDE) (pp. 638-649) ?

Choix inéquitable de l'attribut partagé : une solution

Uniquement choix aléatoire de l'attribut



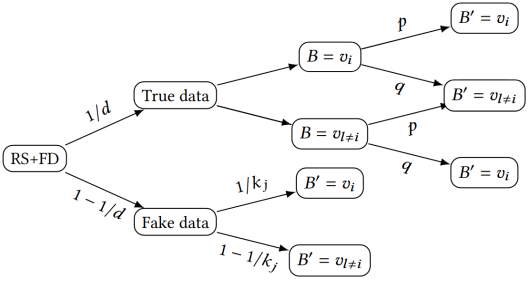
1 attribut choisi aléatoirement, les autres fictifs [22]



22. Arcolezi, H. H., Couchot, J. F., Al Bouna, B., & Xiao, X. (2021, October). Random sampling plus fake data : Multidimensional frequency estimates with local differential privacy. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (pp. 47-57).

Mise en RS+FD de $\mathcal{M}_{GRR}$

Arbre de probabilités



Estimateur et variance

- Estimateur non biaisé de f_i, $\hat{f}_i = \dfrac{r_i d k_j - (q k_j + d - 1)}{(p - q) k_j}$

$$\mathrm{Var}[\hat{f}_i] = \frac{d^2}{N(p-q)^2}\left(f_i.\delta_1(1-\delta_1) + \frac{1-f_i}{N}\delta_0(1-\delta_0)\right)$$

$$\delta_1 = \frac{p k_j + d - 1}{d k_j} \text{ et } \delta_0 = \frac{q k_j + d - 1}{d k_j}$$

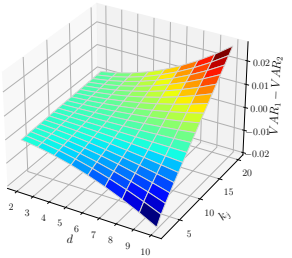
Un mécanisme adaptatif ADP

Connus

- Paramètre p, q de chaque macanisme pour vérifier ϵ-LDP
- Une estimation du nombre de participants N
- d attribut et pour chacun : son nombre de valeurs k
- Pour chaque estimateur : la variance approximative de l'estimateur $\mathrm{Var} * [\hat{f}]$

Pour chaque attribut

- Choix du mécanisme minimisant $\mathrm{Var} * [\hat{f}]$
- Partage de ce choix avec l'agrégateur
- Exemple avec $VAR_1 : RS + FD[GRR]$, $VAR_2 = RS + FD[OUE]$, $N = 10,000$ $\epsilon = \ln(3)$

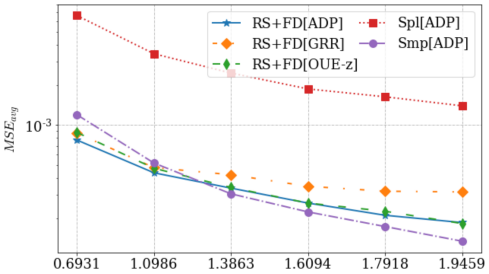


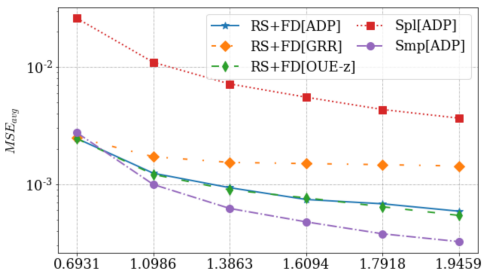
Quelques expérimentations

Caractéristiques

- Données d'UCI : adults, nursery
- $\epsilon \in \{\ln(2), \ln(3), \ldots \ln(7)\}$
- Erreur quadratique moyenne sur 100 essais à chaque fois

Quelques courbes imparfaites (pas de dispersion de l'erreur)





Plan

Plan

Rappel : apprentissage suppervisé probabiliste

Exemple [23] de classification

- Connaissant une valeur pour chaque attribut de mesure : predire Outcome
- Comparer les probabilités suivantes et conclure :

$$\Pr[\text{OutCome} =' \text{YES}' | \text{Preg} = p, \text{Gluc} = g, \ldots, \text{Age} = a]$$
$$\Pr[\text{OutCome} =' \text{NO}' | \text{Preg} = p, \text{Gluc} = g, \ldots, \text{Age} = a]$$

- A évaluer : $\Pr[Y_1 | X_1, X_2, \ldots, X_d]$ et $\Pr[Y_2 | X_1, X_2, \ldots, X_d]$
- Remarque : attributs discrets (Preg., Gluc, ...) ou réels (BMI, DPF)

23. https ://www.kaggle.com/uciml/pima-indians-diabetes-database

Rappel : théorème de Bayes

Théorème (Théorème de Bayes)

$$\Pr[Y|X_1,\ldots,X_d] = \frac{\Pr[Y]\times\Pr[X_1,\ldots,X_d|Y]}{\Pr[X_1,\ldots,X_d]}$$

Simplifications immédiates

$$\left.\begin{array}{l}\Pr[Y_1|X_1,\ldots,X_d] = \frac{\Pr[Y_1]\times\Pr[X_1,\ldots,X_d|Y_1]}{\Pr[X_1,\ldots,X_d]}\\ \Pr[Y_2|X_1,\ldots,X_d] = \frac{\Pr[Y_2]\times\Pr[X_1,\ldots,X_d|Y_2]}{\Pr[X_1,\ldots,X_d]}\end{array}\right\} \begin{array}{l}\hat{\text{m}}\text{ dénom. } \Pr[X_1,\ldots,X_d]\\ \rightsquigarrow \text{ son calcul : inutile}\end{array}$$

A évaluer : $\Pr[Y_j]$, et $\Pr[X_1,\ldots,X_d|Y_j]$

- $\Pr[Y_j]$: fréquence d'apparition de la valeur Y_j pour l'attribut Y
- $\Pr[X_1,\ldots,X_d|Y_j] = \Pr[X_1|Y_j]\times\cdots\times\ \Pr[X_d|Y_j] = \prod_{i=1}^{d}\Pr[X_i|Y_j]$:
  - Hypothèse naïve : indépendance de chaque X_i p.r. aux autres $X_{i'}$, $i' \neq i$, conditionnellement à Y_j

Exemple avec des données discrètes-1

Données[24], question et premières probabilités

Age	Income	Gender	Missed Payment
Young	Low	Male	Yes
Young	High	Female	Yes
Medium	High	Male	No
Old	Medium	Male	No
Old	High	Male	No
Old	Low	Female	Yes
Medium	Low	Female	No
Medium	Medium	Male	Yes
Young	Low	Male	No
Old	High	Female	No

- Q : défaut de payement pour une jeune femme avec un revenu moyen ?
- $Y_1 \equiv'$ Missed Payment$' =' $ YES$'$, $Y_2 \equiv'$ Missed Payment$' =' $ NO$'$
- Comparer $\Pr[Y_1|\text{Age} = \text{Young}, \text{Income} = \text{Medium}, \text{Gender} = \text{Female}]$ et $\Pr[Y_1|\text{Age} = \text{Young}, \text{Income} = \text{Medium}, \text{Gender} = \text{Female}]$
- $\Pr[Y_1] = \frac{4}{10}$ et $\Pr[Y_2] = \frac{6}{10}$

Probabilités pour chaque attribut conditionnellement à Y_j

- $\Pr[\text{Age} = \text{Young}|Y_1] = \frac{2}{4}$, $\Pr[\text{Age} = \text{Young}|Y_2] = \frac{1}{6}$ ...
- $\Pr[\text{Income} = \text{Medium}|Y_1] = \frac{1}{4}$, $\Pr[\text{Income} = \text{Medium}|Y_2] = \frac{1}{6}$ ...
- $\Pr[\text{Gender} = \text{Female}|Y_1] = \frac{2}{4}$, $\Pr[\text{Gender} = \text{Female}|Y_2] = \frac{2}{6}$ ...

24. Yilmaz, E., Al-Rubaie, M., & Chang, J. M. (2019). Locally differentially private naive bayes classification. arXiv preprint arXiv:1905.01039.

Exemple avec des données discrètes-2

Evaluation de $\Pr[Y_j|\text{Age} = \text{Young}, \text{Income} = \text{Medium}, \text{Gender} = \text{Female}]$

- $Pr[Y_1] \times \Pr[\text{Age} = \text{Young}|Y_1] \times \Pr[\text{Income} = \text{Medium}|Y_1] \times \Pr[\text{Gender} = \text{Female}|Y_1] = \frac{4}{10} \times \frac{2}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{1}{40} \approx 0.025$
- $Pr[Y_2] \times \Pr[\text{Age} = \text{Young}|Y_2] \times \Pr[\text{Income} = \text{Medium}|Y_2] \times \Pr[\text{Gender} = \text{Female}|Y_2] = \frac{6}{10} \times \frac{1}{6} \times \frac{1}{6} \times \frac{2}{6} = \frac{1}{180} \approx 0.0056$

Réponse

La probabilité qu'elle ratte son paiement est beaucoup plus importante que celle opposée.

Exemple avec des données continues-1

Données [25], question et premières probabilités

sexe	taille (cm)	masse (kg)	point. (cm)
masc.	182	81.6	30
masc.	180	86.2	28
masc.	170	77.1	30
masc.	180	74.8	25
fém.	152	45.4	15
fém.	168	68.0	20
fém.	165	59.0	18
fém.	175	68.0	23

- Q : sexe d'une personne mesurant 183cm, pesant 59kg et dont les pieds mesurent 20cm ?
- $Y_1 \equiv'$ Sexe$' ='$ masc.$'$, $Y_2 \equiv'$ Sexe$' ='$ fém.$'$
- Comparer $\Pr[Y_1|\text{Taille} = 183, \text{Masse} = 59, \text{Point.} = 20]$ et $\Pr[Y_2|\text{Taille} = 183, \text{Masse} = 59, \text{Point.} = 20]$ et
- $\Pr[Y_1] = \Pr[Y_2] = \frac{1}{2}$

Probabilités pour chaque attribut X_i conditionnellement à Y_j : $\mathcal{N}(\mu_{i,j}, \sigma^2_{i,j})$

1. Calcul des paramètres de $\mathcal{N}$: moyene $\mu_{i,j}$ et variance $\sigma^2_{i,j}$

Sexe	μ_{taille}	σ^2_{taille}	μ_{masse}	σ^2_{masse}	$\mu_{point.}$	$\sigma^2_{point.}$
masc.	178	29.3	79.9	25.5	28.25	5.58
fém.	165	92.7	60.1	114	19	11.3

2. Avec la densité de probabilité $\frac{1}{\sqrt{2\pi\sigma^2_{i,j}}} \exp\left(-\frac{1}{2\sigma^2_{i,j}}(x-\mu_{i,j})^2\right)$, calcul de

$$\Pr[\text{taille} = 183|Y_1] = \frac{1}{\sqrt{2\pi \times 29.3}} \exp\left(\frac{-1}{2\times 29.3}(183-178)^2\right) dt \approx 0.0481$$

25. Classification naïve bayésienne, Wikipedia

Exemple avec des données continues-2

Valeurs numérique des robabilités pour chaque attribut X_i conditionnellement à Y_j

- $\Pr(taille = 183|Y_1) = 0.0481dt$, $\Pr(poids = 59|Y_1) = 0.0000146dp$ et $\Pr(point. = 20|Y_1) = 0.000381d$.
- $\Pr(taille = 183|Y_2) = 0.00721dt$, $\Pr(poids = 59|Y_2) = 0.0372dp$ et $\Pr(point. = 20|Y_2) = 0.114d$.

Evaluation de $\Pr[Y_j|taille = 183, poids = 59, point. = 20]$

- $Pr[Y_1] \times \Pr(taille = 183|Y_1) \times \Pr(poids = 59|Y_1) \times \Pr(point. = 20|Y_1) \approx 1.3404 \times 10^{-10}$
- $Pr[Y_2] \times \Pr(taille = 183|Y_2) \times \Pr(poids = 59|Y_2) \times \Pr(point. = 20|Y_2) \approx 1.52 \times 10^{-5}$

Réponse

La personne est probablement une femme.

Plan

Evaluer $\Pr[Y_j]$ et $\Pr[X_i|Y_j]$

Evaluer $\Pr[Y_j]$

- Histogramme des Y_j à construire
- Algorithme ϵ-DP pour les histogrammes : bruit laplacien de sensibilité 1.

Evaluer $\Pr[X_i|Y_j]$ pour un attribut X_i discret

- Pour chaque sortie Y_j à prédire : construction de l'histogramme de X_i
- Algorithme ϵ-DP pour les histogrammes : bruit laplacien de sensibilité 1.

Evaluer $\Pr[X_i|Y_j]$ pour un attribut X_i continu

- Pour chaque sortie Y_j à prédire : approché par une loi $\mathcal{N}(\mu_{i,j}, \sigma^2_{i,j})$
- Algorithme ϵ-DP pour la moyenne : bruit laplacien de sensibilité $\frac{U-L}{n+1}$ avec U la borne max, l la borne min, n le nbre de lignes déjà présentes
- Algorithme ϵ-DP [26] pour la variance σ^2 : bruit laplacien de sensibilité $\frac{n(U-L)}{n+1}$ avec U la borne max, l la borne min, n le nbre de lignes déjà présentes

Algorithme de Vaidya

Algorithm 1 Computing differentially private parameters for Naïve Bayes

Require: ϵ, the privacy parameter for differential privacy
Require: Laplace(a, b) samples the Laplace distribution with mean a and scale b

for each attribute X_j **do**
  if X_j is categorical **then**
    sensitivity, $s \leftarrow 1$
    scale factor, $sf \leftarrow s/\epsilon$
    $\forall$ counts n_{kj}, $n'_{kj} = n_{kj}+$ Laplace$(0, sf)$
    Use n'_{kj} to compute $P(x_i|c_j)$
  else if X_j is numeric **then**
    compute sensitivity, s for mean μ_j as per equation 5
    scale factor, $sf \leftarrow s/\epsilon$
    $\mu'_j \leftarrow \mu_j+$ Laplace$(0, sf)$
    compute sensitivity, s for standard deviation σ_j as per equation 7
    scale factor, $sf \leftarrow s/\epsilon$
    $\sigma'_j \leftarrow \sigma_j$ + Laplace$(0, sf)$
    Use μ'_j and σ'_j to compute $P(x_i|c_j)$
  end if
end for
for each class c_j **do**
  count $nc'_j \leftarrow nc_j+$ Laplace$(0, 1)$
  Use nc'_j to compute the prior $P(c_j)$
end for

Remarques et implantation

Calcul du budget ϵ global

- A partager entre tous les attributs : chacun en consomme
- $\epsilon_i = \frac{\epsilon}{d+1}$: les attributs $X_1, \ldots, X_d$ et Y

Avec la biblotheque DiffprivLib [27]

```
# salaire sup à 50k$
import pandas as pd
from sklearn.model_selection import train_test_split
import diffprivlib.models as dp

...

def experimentDPGNB(eps,X,y):
    X_train, X_test,y_train,y_test = train_test_split(X,y,test_size=0.2)
    private_clf = dp.models.GaussianNB(epsilon=eps)
    private_clf.fit(X_train.values, y_train.values.ravel())
    y_pred = np.array(private_clf.predict(X_test))
    return accuracy_score(y_test,y_pred)
```

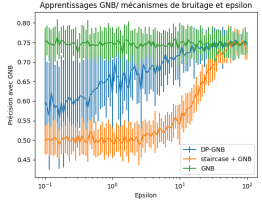
27. https://pypi.org/project/diffprivlib/

GNB vs $\mathcal{M}_{\text{Stair}}$ + GNB vs DP-GNB

Contexte expérimental

- Sur le jeu de données du diabete
- Outcome appris par GNB, par $\mathcal{M}_{\text{Stair}}$ + GNB, par DP-GNB
- Pour chaque $\epsilon \in [10^{-1}, 100]$ global, moyenne sur 20 apprentissages
  - GNB : sans PVP
  - $\mathcal{M}_{\text{Stair}}$ + GNB : nettoyage staircase + apprentissage
  - DP-GNB : `dp.models.GaussianNB`
- Affiché : précision moyenne de l'apprentissage et écart-type

Interprétation des résultats



- GNB : toujours plus précis (mais fuite de données)
- DP-GNB : toujours plus précis que $\mathcal{M}_{\text{Stair}}$ + GNB
- $\epsilon \in [0.1, 10]$: $\mathcal{M}_{\text{Stair}}$ + GNB proche de l'aléatoire
- $\epsilon \in [10, 30]$: intérêt de DP-GNB p.r. à GNB ?

Plan

Plan

Leveraging AI for Healthcare Process Improvement

- Significant interest among hospitals in optimizing a number of tasks by leveraging AI, particularly by exploiting **textual medical records** from patient files :
  - Identifying similarities between patients and their pathologies, thereby **gaining direct access to successful treatment pathways** for these pathologies versus less successful ones.
  - **Detecting abnormal patient journeys**, for example, where a condition associated with treatment is suspected.
  - **Automatically associating medical codes** with patient journeys (according to the ICD-10 classif.) for statistical purposes and hospital reimbursement. (@ HNFC $\approx$ 12 individuals from the Medical Information department carry out this coding task e.g.)

Bridging the Gap : Developing AI Tools for Healthcare with sanitized Data

Who can can develop ML to for Medical Institutions ?

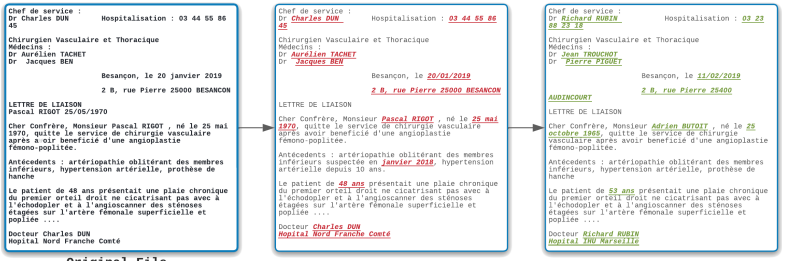
- Medical institutions generally **lack the expertise** to develop advanced AI (with diverse data types like text, tables, and vectors).
- Even within these institutions, **legal restrictions** (GDPR...) **prevent AI researchers** (FEMTO-ST) from accessing patient data to develop AI tools.

Is de-identification the answer ?

- AI tool prototypes can be created in labs and then **customized with realistic data derived from a robust and useful de-identified medical** data from a medical institution.
  - Robust : the level of information **leakage is mathematically bounded** : based on **Differential Privacy**, the standard adopted today in academia, in industry
  - Useful : surrogating preserves **chronology of events**, distinguish between **personal and medical details** (e.g., "Charcot"), and maintain **familial relationships**.

De-Identification : A Twofold Method

Two Steps



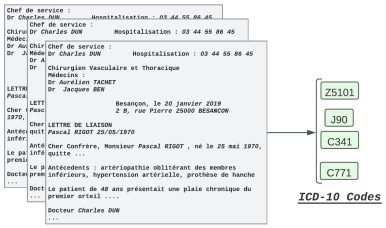
1. Named Entity Recognition (NER) for identifying information (efficiency issue)
2. Sanitizing of detected information (optimization issue : minimizing leakage while preserving utility)

Application Context With ICD-10 Code Association Task

ICD-10 : Standardized Diagnostic Tool for Recording Health Conditions

- Developed by the World Health Organization (WHO).
- Used worlwide to classify diseases, injuries, and health conditions. . .
- Reimbursement Impact : Codes are essential for billing and reimbursement systems.

Application Context with ICD-10 Code Association Task



- Manual Coding : Currently, healthcare professionals assign codes manually based on medical records.
- Automated Coding : This task can be framed as a multi-label text classification problem, where the goal is to automatically assign appropriate ICD-10 codes to medical documents.

PhD Thesis of Dr. Yakini TCHOUKA

- Defended in December 2023.
- Supported by the Bourgogne-Franche-Comté region and the EUR EIPHI.
- In partnership with the Medical Information Department of the HNFC.

Plan

Named Entity Recognition for HIPAA Categories

Iterative Learning on HNFC Datasets : HNFC-NER-EVAL, HNFC-NER-TRAIN

- **Increasingly** large, **progressively** more de-identified datasets.
- Automatically pre-labeled and manually validated.
- Model : Hybrid [28], then deep learning only [29].

NER Results

Method	CamemBERT-ner			MEDINA			FlauBERT-ner			**Hybride**			**Healthinf**			**Dernoncourt**		
Dataset	*HNFC*															i2b2		
Metric	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
PER	89	99	93.8	***98.2***	97.7	*98.2*	91.8	97.6	94.6	96.3	***99.8***	98	97.2	98.9	98	**98.2**	99.1	**98.6**
ORG	7.	21.8	11.1	32.6	24.8	28.1	16.9	34.1	22.6	41.1	*57.3*	47.8	*90*	51	*65.6*	**92.9**	**71.4**	**80.7**
LOC	46	67.2	54.6	98.8	81.1	89.1	75.7	66.3	70.7	88.4	***95.8***	92	***99.4***	94.4	***96.9***	95.9	95.7	95.8
DATE	NA			97.7	86.6	91.9	NA			97.7	86.7	91.9	***99.2***	*95.7*	*97.4*	99	**99.5**	**99.2**
AGE	NA			91.5	66.9	77.3	NA			91.5	66.9	77.3	*98.2*	*91.8*	*95*	**98.9**	**97.6**	**98.2**
TEL	NA			99.5	97.9	98.7	NA			***99.5***	97.9	98.7	99.4	***99.8***	***99.6***	98.7	99.7	99.2
REF	NA			NA			NA			NA			96.1	79.5	87	NA		
QID	NA			NA			NA			NA			77.2	32	45.3	**99.2**	**98.7**	**99**
Mic.-avg.	70.8	51.5	59.6	98.2	91.2	94.5	85.8	86.7	86.3	94.6	94.9	94.7	***98.5***	*96.4*	*97.4*	98.3	**98.5**	**98.4**

28. **tchouka:arxiv**.
29. **DBLP:conf/biostec/TchoukaCL23**.

Surrogate generation strategies : DATEs and AGEs

Temporal data surrogate issues

1. **Privacy** :
   - Very identifying
   - Re-identification risk : the chronology of events
2. **Utility** :
   - The relevance of events
   - The patient's features

Related Work on Date Substitution : Uniform Shifting of DATEs

- MIMIC3 [30], I2B2 [31] datasets.

Attack on HNFC-NER-EVAL Dates with Uniform Shifting

- The interval $I = [I_1, \ldots I_{n-2}]$ is NOT modified and is unique in 98% of this dataset.

30. Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC3, a freely accessible critical care database. Scientific data, 3(1), 1-9.

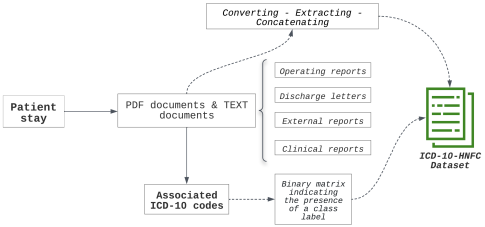
31. https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

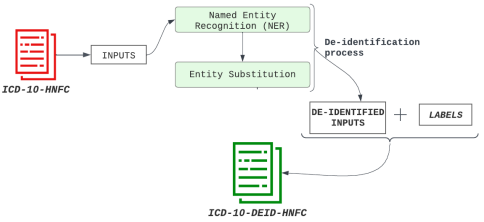
Sanitizing Integrating Metric-Privacy

- Theory : $\forall x_1, x_2, y, \Pr(\mathcal{M}(x_1) = y) \leq e^{\varepsilon . d(x_1, x_2)} \Pr(\mathcal{M}(x_2) = y)$.
- Dates : $\mathcal{M}_{date}(x) = x + v$ t.q. $v \sim Lap(\frac{1}{\varepsilon})$.
  - Allows to distinguish betw. 08/01/42 and 14/03/18 (birth and death dates of St. Hawking) whereas DP not.
- Locations : $\Pr(\mathcal{M}_{loc}(x) = o) \propto e^{\varepsilon . d(x,o)}$, s.t. d an epidemiological based distance.
  - Avoid to sanitize Dijon with Beze too often (in BFC but epidemiologically $\neq$).

ICD-10 Code Association [32]

Datasets Buildings





32. **DBLP:conf/cbms/TchoukaCLSR23**.

ICD-10-HNFC dataset : Challenging Metrics

Descriptive statistics of ICD-10-HNFC dataset

	Dataset	Dataset with class reduction
Documents	56014	-
Tokens	41868993	-
Average sequence length	747	-
Total ICD codes	416125	415830
Unique ICD codes	6160	1564
Codes with less than 10 examples	3722	523
Codes with 100 examples or more	641	471

Two issues in ICD-10 codes association

1. Input patient file : usually a long sequence :
   - Average sequence length (747) $>$ maximum input size for Transformers (512) : scalability issue
2. Large number of different codes, labels, but sparse
   - 6160 unique ICD codes, 3722 of whom have only been less than 10 times : scalability and sparsability issue

ICD-10 Code Association– Results

State-of-the-Art [33] Code Association Results

Models	Language	Dataset	Labels	F_1-score
PLM-ICD [34]	*English*	*MIMIC 2* *MIMIC 3*	*5,031* *8,922*	*0.5* ***0.59***
Dalloux [35]	*French*	*Personnel*	*6,116* *1,549*	*0.39* *0.52*
PROPOSAL	French	ICD-10-HNFC	6,160 1,564	**0.47** **0.55**
Dalloux			6,160 1,564	0.27 0.35

Impact of De-identification on Results

Dataset	Labels	**Precision**	**Recall**	F_1-**score**
ICD-10-HNFC	**6160**	**0.47**	**0.46**	**0.47**
ICD-10-DEID-HNFC		0.44	0.43	0.44
ICD-10-TAG-HNFC		0.43	0.41	0.42

33. **DBLP:journals/iswa/TchoukaCLSR24**.
34. **huang2022plm**.
35. **dalloux2020supervised**.

GitHub Open source implementation

- Automatic ICD-10 code classification system in French[36]
- Surrogate generation strategies in de-identification with metric privacy mechanism[37]
- Named Entity Recognition system in medical context[38]
- Automatic ICD-10 code association with CNN[39]

36. Automatic ICD-10 code classification system in French. https://github.com/mlfiab/icd10-french
37. Surrogate Generation in De-identification. https://github.com/mlfiab/surrogate-deid
38. Named Entity Recognition system in medical context. https://github.com/mlfiab/ner-french
39. Automatic ICD-10 code association with CNN. https://github.com/mlfiab/cnn-icd10