

Benchmarking RGPD

PAR : Abdellatif EL Mahdaoui

ENCADRANTE : Mme Besma Zeddini

# **Benchmarking RGPD**

Techniques cité par notre encadrante (Obligatiore):

* **Masquage (Masking)** :  
   Remplacement d'une donnée par un caractère ou une valeur fictive (ex: numéro de téléphone remplacé par XXXX).
* **Pseudonymisation :**  
   Remplacement des identifiants directs par un pseudonyme ou identifiant fictif réversible par une clé de correspondance séparée.
* **Généralisation :**  
   Remplacement des données précises par des valeurs plus générales (ex: date de naissance précise par une tranche d’âge, ville précise remplacée par région ou pays).
* **Perturbation (Distorsion) :** Modification légère des données originales en ajoutant un « bruit » statistique pour rendre difficile l’identification directe.
* **Agrégation (Tranchage) :**  
   Regroupement des données individuelles en ensembles agrégés (ex: moyenne d'âge au lieu d'âges individuels).

Autres (Si on a le temps):

* **Suppression (Suppression de champs ou d'attributs) :**  
   Suppression complète des attributs sensibles ou identifiants directs (ex : suppression du nom, prénom, adresse).
* **Permutation :**  
   Mélange des données au sein d'un même attribut afin d’éviter les rapprochements directs avec les individus.
* **K-anonymat (K-anonymity) :** Assure qu'un individu ne peut être distingué parmi au moins "k" autres individus partageant les mêmes caractéristiques.
* **L-diversité (L-diversity) :**  
   Renforce le K-anonymat en assurant qu'un attribut sensible possède au moins "l" valeurs différentes dans chaque groupe anonymisé.
* **T-closeness :**  
   Technique améliorée assurant que la distribution d'un attribut sensible au sein d’un groupe anonymisé reste proche de la distribution globale de ce même attribut dans l'ensemble des données.
* **Tokenisation :**  
   Remplacement d'informations sensibles par des "jetons" ou tokens sans lien direct, souvent utilisé pour anonymiser des transactions financières ou médicales.

Comment tester?

1. Utiliser <https://fakerjs.dev/> pour générer une dataset
2. On applique les techniques une par une (Python)
3. On calcule (Durant le script python) ces critères:

|  |  |  |  |
| --- | --- | --- | --- |
| **Critique / Critère** | **Description** | **Peut être testé en Python ?** | **Comment le tester ?** |
| ⏱️ **Temps d'exécution** | Temps pris pour appliquer la méthode d’anonymisation | ✅ Oui | time.time() ou time.perf\_counter() avant/après |
| 💾 **Mémoire utilisée** | Quantité de RAM consommée pendant l’opération | ✅ Oui | psutil ou memory\_profiler |
| 📉 **Perte d'information** | Écart entre les données originales et anonymisées | ✅ Oui | distance Levenshtein, RMSE, ou simple == pour taux de modification |
| 🔁 **Réversibilité** | Peut-on retrouver les données originales ? | ✅ Oui (simulé) | implémenter une "attaque de ré-identification" simple |
| 🔐 **Risque de ré-identification** | Nombre ou pourcentage de lignes pouvant être croisées avec données externes | ✅ Oui (approximativement) | comparer clés directes (nom, email, tel) entre deux datasets |
| 🔢 **Entropie / Diversité** | Mesure du désordre ou de diversité d’un champ après anonymisation | ✅ Oui | scipy.stats.entropy() ou simple len(df[col].unique()) |
| 🎯 **Précision des données** | Est-ce que les données gardent leur valeur informative | ✅ Oui | test sur corrélation ou distributions statistiques |
| 🧪 **Robustesse à l’échelle** | Est-ce que la méthode fonctionne bien sur gros datasets (scalabilité) | ✅ Oui | tester sur jeux de données 10k, 100k, 1M lignes |

Le test:

1. Installation des bibliothèques: 

**pandas:** manipulation de dataset, application des methodes sur les colonnes (Comme si on a un csv)

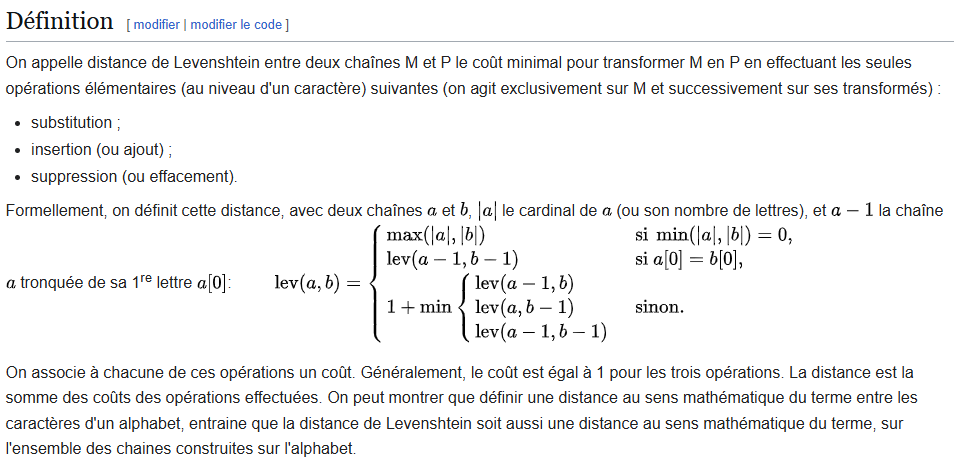
**faker:** géneration des datasets simulés avec des identifiants sensibles

**numpy:** géneration des données numérique, calculs rapides sur les colonnes

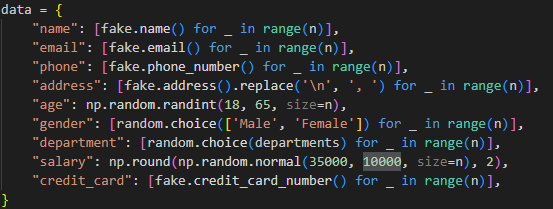
**memory-profiler:** mesure de mémoire

**psutil:** suivie de ram cpu (complément à memory-profiler)

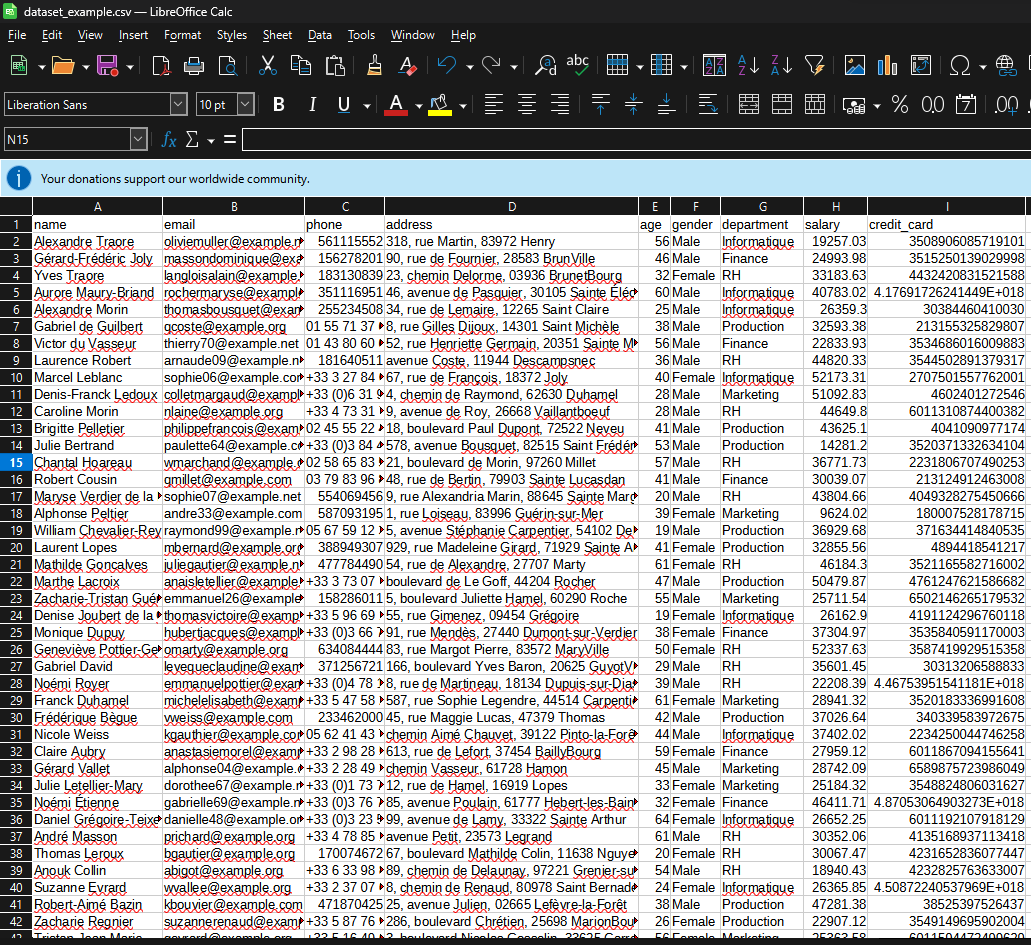
**python-Levenshtein:** calcul de la distance de Levenshtein entre les valeurs originales et anonymisées (perte d’information textuelle)



**scipy:** calculs statistiques avancés (entropie, tests de Kolmogorov-Smirnov, corrélation, etc.)

1. Génération d’une dataset fake avec 10000 lignes, et des champs comme: 





[*https://github.com/faker-ruby/faker/tree/main/doc/default*](https://github.com/faker-ruby/faker/tree/main/doc/default)