# Analyses respectueuses de la vie privée: initiation à la théorie et à la pratique
# Projet L2 CMI

Jean-François Couchot

## 1 Présentation

Lorsqu'on analyse des données pour faire de l'apprentissage par exemple, on se heurte tôt ou tard à la confidentialité de celles-ci. Le RGPD dit explicitement que sans consentement, ni stockage ni analyse donc ne peut être réalisé.

Des méthodes de nettoyage des données ont été mises en place pour permettre de mémoriser celles-ci tout en garantissant à leur propriétaires un certain niveau de confidentialité.

Parmi celles-ci on note les méthodes syntaxiques comme le $k$-anonymat [Swe02], la $l$-diversité [MKGV07]. D'autres méthodes probabilistes sont aujourd'hui mises en œuvre, notamment celles à base de confidentialité différentielle globale [DMNS06] ou locale [KLN$^+$11].

## 2 Objectifs

Ce projet se découpera en 5 étapes principales.

1. Mise en place d'algorithmes d'apprentissage simples (Naïve Bayes, $k$-moyenne, régression linaire) sur des jeux de données public. On utilisera la bibliothèque `scikit sklearn`[1]. On cherchera à comprendre le fonctionnement de ces algorithmes.
2. Etude de la théorie du $k$-anonymat [Swe02], de la $l$-diversité [MKGV07]. Appropriation ensuite de la confidentialité différentielle globale [DMNS06], puis locale [KLN$^+$11].
3. Utilisation d'un outil d'anonymisation de données comme `ARX`[2] pour nettoyer des jeux de données publiques Utilisation d'une bibliothèque de confidentialité différentielle `diffprivlib`[3].
4. Application des méthodes introduites en 1 sur les jeux de données nettoyés.
5. Exploitation d'algorithmes d'apprentissage implantant en interne des mécanismes de confidentialité différentielle et comparaison des résultats avec ceux obtenus en 4.

## Références

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[KLN$^+$11] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately ? *SIAM Journal on Computing*, 40(3) :793–826, 2011.

[MKGV07] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity : Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1) :3–es, 2007.

[Swe02] Latanya Sweeney. k-anonymity : A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05) :557–570, 2002.

1. `https://scikit-learn.org/stable/index.html`
2. `https://arx.deidentifier.org/`
3. `https://diffprivlib.readthedocs.io/en/latest/`