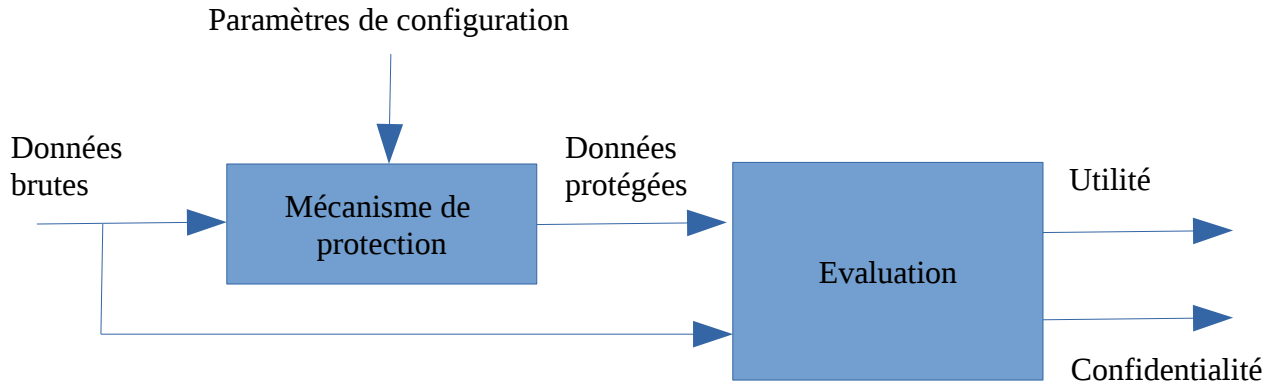
Challenge d'annonymisation de données
Protection de traces de mobilité

antoine.boutet@insa-lyon.fr, mathieu.cunche@insa-lyon.fr, benjamin.nguyen@insa-cvl.fr

Le but de ce challenge est de vous familiariser avec la protection d'un ensemble de traces de mobilité. Plus précisément, vous allez recevoir les traces de mobilité d'une centaine d'utilisateurs. Il vous faudra mettre en place des mécanismes de protection et les évaluer. L'évaluation de mécanisme de protection est réalisé en comparant les données protégées aux données brutes comme illustré ci-dessous et en mesurant la qualité (ou la détérioration) des données protégées et leur niveau de confidentialité.



Métrique de confidentialité

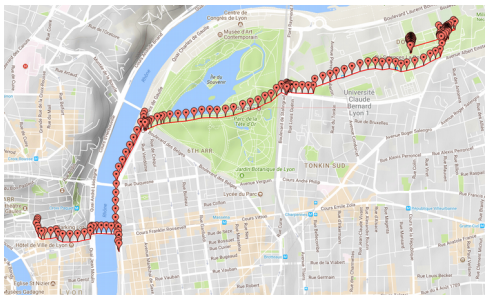
Dans le cadre de ce challenge, la métrique de confidentialité est une mesure de re-identification des utilisateurs. Plus précisément, on mesure le ratio des utilisateurs correctement re-identifiés par fenêtre de 1 semaines.

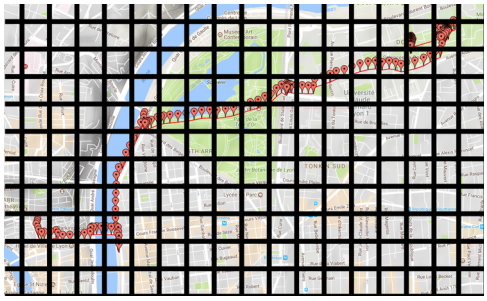
$$\Re - identification = \frac{1}{NF} \sum_{i=1}^{N} \sum_{j=1}^{F} CorrectedMatchings(ij)$$

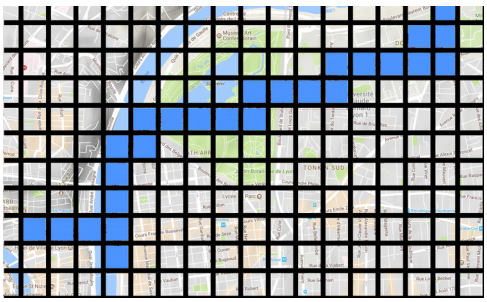
ou N est le nombre d'utilisateur, F le nombre de fenêtre, et la fonction CorrectMatching(ij) qui renvoie 1 si l'utilisateur i est correctement re-identifié sur la fenêtre j.

Métrique d'utilité

Plusieurs métriques d'utilité peuvent être considérées. Pour faciliter le calcul de certaines métriques, nous allons segmenter l'espace avec une grille de taille variable. On pourra ainsi manipuler la liste des cellules où l'utilisateur à séjourné au lieu de chaque point de collecte associé à ses coordonnées. Les images ci-dessous illustrent cette représentation.







Voici quelques exemples de métrique d'utilité qui considèrent chaque utilisateur individuellement :

- **Déplacements effectués :** calcul de la différence de zone de couverture, mesurée par le nombre de cellules différentes ou l'utilisateur à séjourné. Cette métrique évolue en fonction de la taille de cellule considérée. On peut aussi considérer uniquement les cellules où l'utilisateur a séjourné au-delà d'un certain temps, ou ne considérer uniquement les jours en semaine ou le week-end.

- **Distorsion :** calcul de la distorsion de différentes manières. On peut considérer une distorsion purement spatiale en calculant la différence en terme de cellule (en faisant varier la taille de la cellule), et on peut considérer une distorsion spatio-temporelle en calculant la différence de temps passée dans chaque cellule qui est identique à la cellule originale.

- **Extraction des Points d'Intérêts** : un point d'intéré (POI) est définit comme une cellule où l'utilisateur reste un certain temps. Ces POIs peuvent représenter par exemple le lieu d'habitation (détection d'un séjour quotidien la nuit), un lieu de travail (détection d'un séjour quotidien la semaine en journée), ou d'une activité (détection d'un séjour ponctuel). Cette métrique peut comparer les x POIs les plus visités entre les données brutes et les données protégées.

Une métrique d'utilité peut également considérer l'ensemble du jeu de données :

- **Croisements :** identification et comparaison des cellules où le plus grand nombre d'utilisateur circule.