Évaluation de la vie privée lors de l'ouverture des données de trajets : l'étude de cas de Montréal.

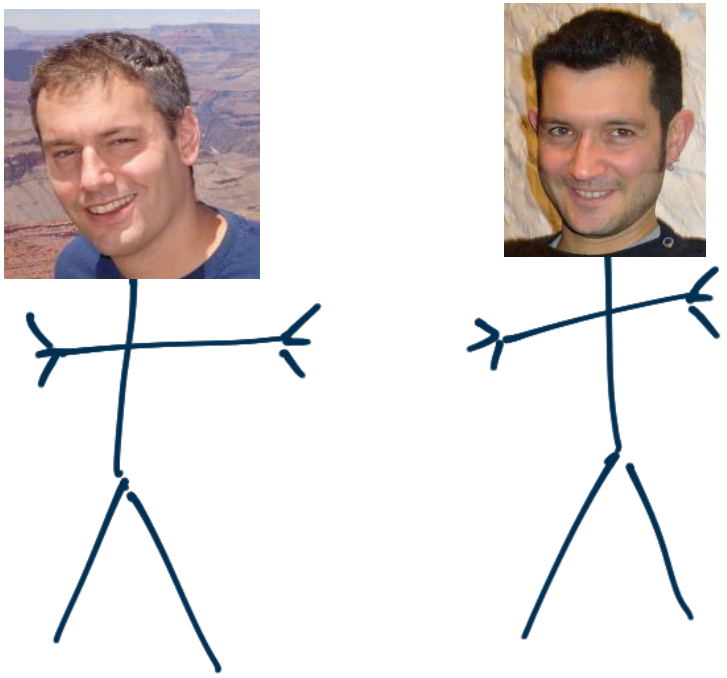
Sébastien Gambs, Marc-Olivier Killijian, Antoine Laurent

Mercredi 21 Octobre 2020

Séminaire Latece

Remerciements

Le travail de cette présentation est réalisé dans le cadre de ma maitrise, supervisée par :



Sébastien Gambs

Marc-Olivier Killijian

Connaissez-vous cet engin ?

Portails de données ouvertes

Montréal

CHICAGO DATA PORTAL

OPEN DATA
data.louisvilleky.gov

data.austintexas.gov
the official City of Austin open data portal

OPENDATA KC

Exemples de fuites d'information non désirées

Identification de stars grâce aux données de Taxi (tiré de Atockar 14)

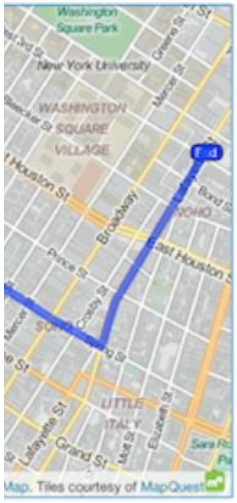
Exemples de fuites d'information non désirées

Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

- **Latest: Strava suggests military users 'opt out' of heatmap as row deepens**

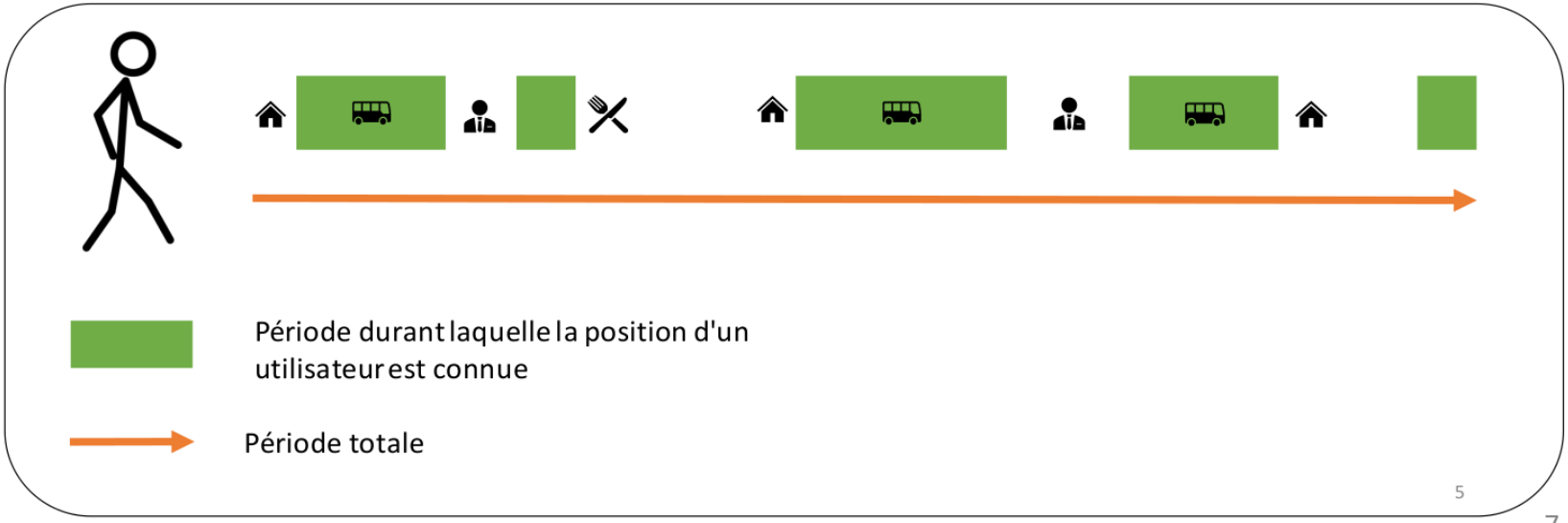
▲ A military base in Helmand Province, Afghanistan with route taken by joggers highlighted by Strava. Photograph: Strava Heatmap



Identificatic)

Notions préliminaires

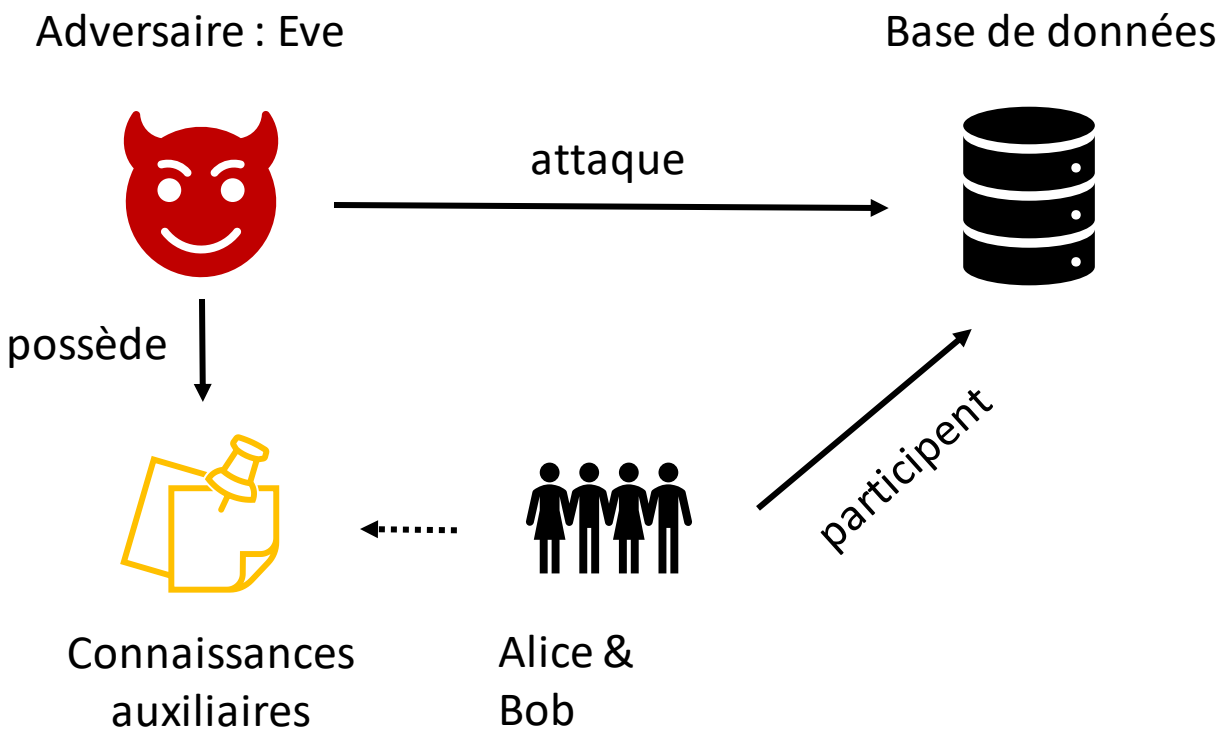
- Données Géolocalisées
- Points d'intérêts (POIS)
- Trajectoires
- Trajets



Attaques sur la mobilité

Outils d'évaluation de la vie privée

Attaques d'une base de données (de mobilité)



Objectif de l'attaque

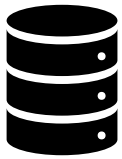
1. Réidentifier un individu dans la base de données attaquée
2. Prédire l'appartenance d'un individu
3. Chainer des morceaux d'informations de mobilité
4. Prédire les déplacements futurs
5. Identifier des points d'intérêts
6. Extraction de la sémantique des points
7. Apprendre les liens sociaux

Exemple de réidentification

Connaissance auxiliaire

Nom	Position 1	Position 3
Alice	**(3.68733, 43.40948)**	(3.73412, 43.40948)

Base de données attaquée



uuid	Position 1	Position 2	...	Position n
1ubsc438x	**(3.68733, 43.40948)**	(3.6, 43.40)	...	(3.734, 43.408)
39shHA912	(5.38, 49.98)	(5.4203, 49.6534)	...	(-73.7, 43.4)
0sas90buj	(3.79733, 43.51848)	(3.6, 43.40)	...	(3.8031, 43.50)
asad98wq7	(3.22523, 43.40)	Pas de valeurs	...	Pas de valeurs

- Alice est l'utilisateur 1ubsc438 dans la base de données
- Alice est **unique** dans les données

Exemple d'attaque d'appartenance

- Eve veut seulement apprendre si Bob est dans les données
  - => la base de données doit être sensible (ex: personnes infectées par la Covid-19)

Nom	Point 1
Bob	**(3.6, 43.40)**

uuid	Position 1	Position 2	...	Position n
1ubsc438x	(3.68733, 43.40948)	**(3.6, 43.40)**	...	**(église)** (3.734, 43.408)
39shHA912	(5.38, 49.98)	(5.4203, 49.6534)	...	(-73.7, 43.4)
0sas90buj	(3.79733, 43.51848)	**(3.6, 43.40)**	...	**(église)** (3.8031, 43.50)
asad98wq7	(3.22523, 43.40)	Pas de valeurs	...	Pas de valeurs

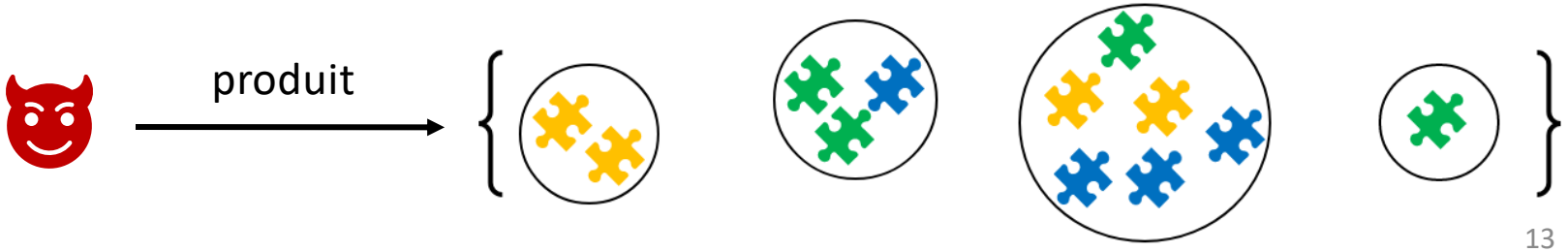
Bob est potentiellement dans la base de données

Attaque de chainage

- On peut voir la question de la chainabilité comme un problème de puzzles

- Chaque puzzle représente un individu dont les pièces sont ses informations de mobilité

- Si les puzzles sont mélangés, chainer les puzzles revient à proposer un regroupement proche des puzzles originaux

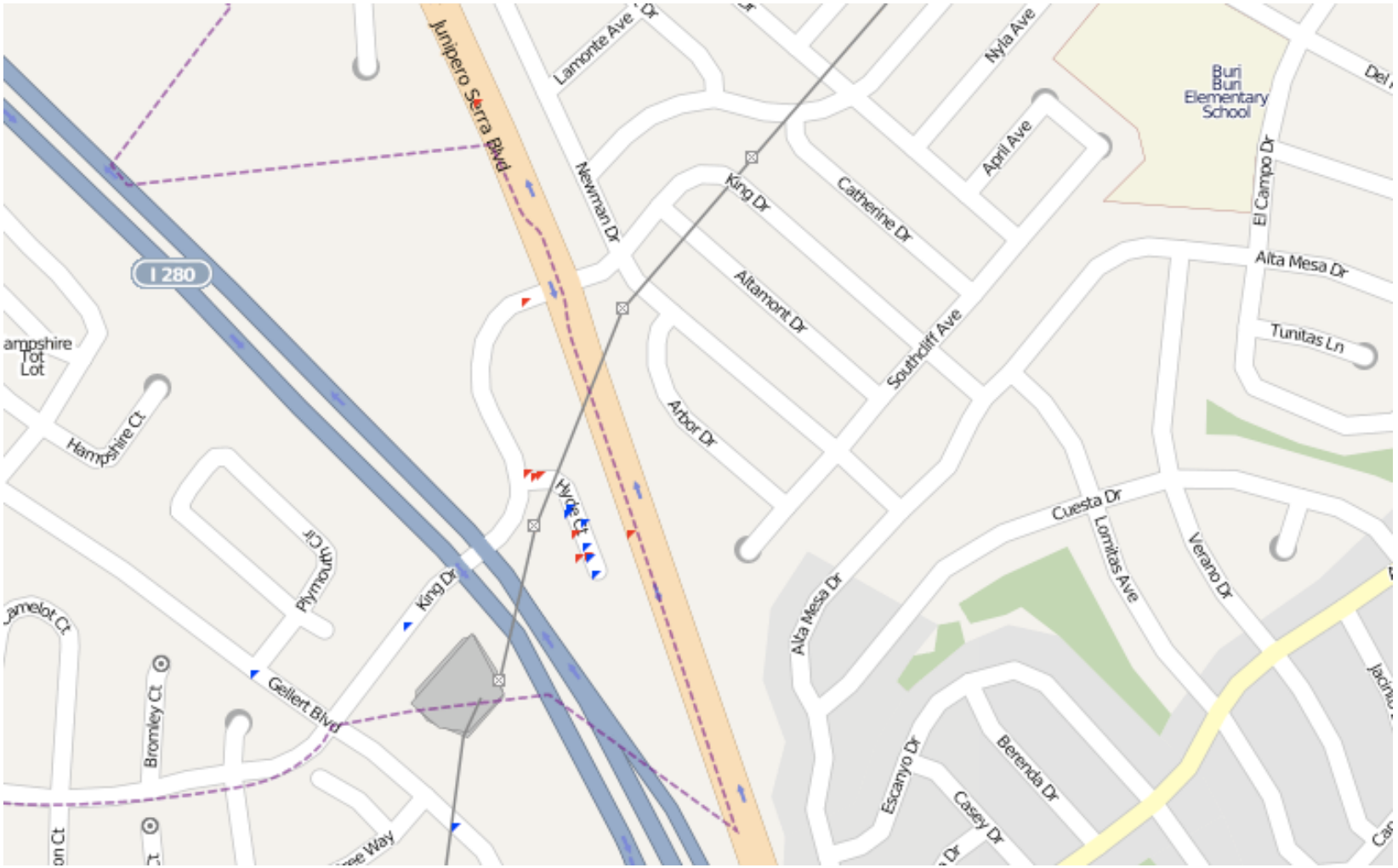


Identifier des points d'intérêts

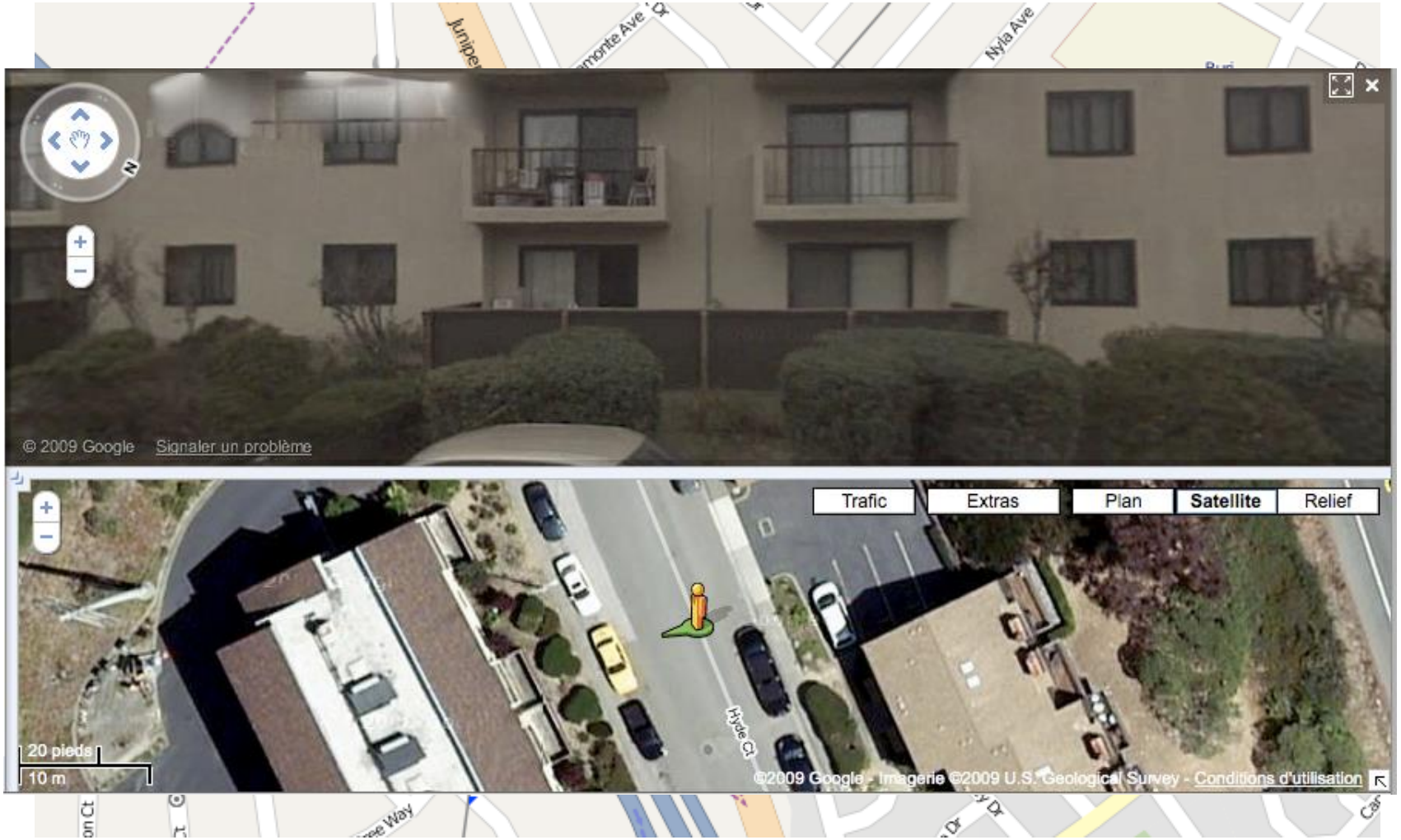
- D'abord :
  - Repérer les points où un utilisateur est immobile (ou presque)
  - Clustériser ces points

- Ensuite ?
  - Les premiers points et derniers points de la journée -> domicile
  - Le point le plus fréquenté la journée -> lieu de travail
  - Un point entre le domicile et le travail -> école/loisir/sport ?

Identifier des points d'intérêts



Identifier des points d'intérêts



Les mécanismes de défense

Problème de l'ouverture de données : publier des données de mobilité *<<privacy-safe>>* sans connaitre l'utilisation qui va en être faite.

Solution : Assainir des données avant la publication

Ingrédients de l'assainissement

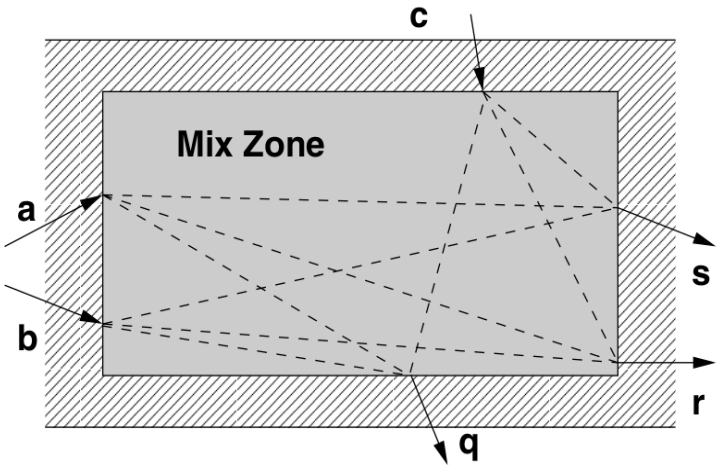
1. Modèle de respect de la vie privée : qu'est-ce que cela signifie pour des données publiées d'être respectueuses de la vie privée ?
2. Méthodes d'assainissement : comment modifier les données pour atteindre la propriété définie par le modèle de respect de la vie privée ?
3. Mesure d'utilité : comment mesurer l'utilité des données résultantes ?

Modèles ad-hoc d'assainissement

- Les données sont perturbées sans garanties formelles de la vie privée.
- Souvent utilisé lorsque :
  - Les autres modèles de vie privée dégradent trop les données.
  - Ils sont plus directs à appliquer sur les données.

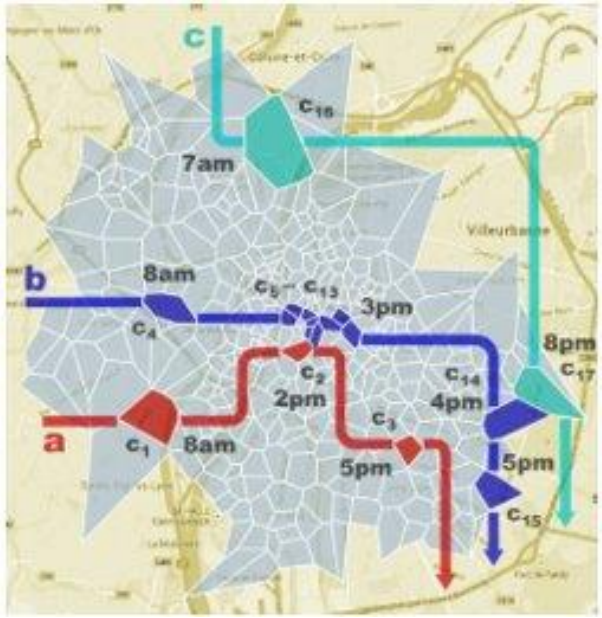
Zones de mixage : ad-hoc (Beresford et Stajano, 2003)

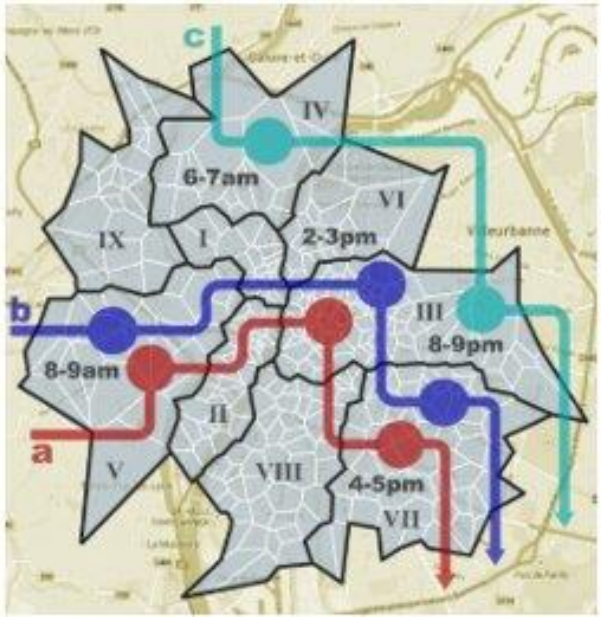
- Protège contre le chainage des trajets qui rentrent dans la mix-zone et des trajets qui en sortent :
  - Lorsqu'un utilisateur rentre dans la zone de mixage -> pas d'enregistrement de points.
  - Les pseudonymes sont changés lors de la sortie de la zone.
- Problèmes :
  - Si l'entropie est faible, les trajets sont chainables !
  - L'emplacement des mix-zones est très important.

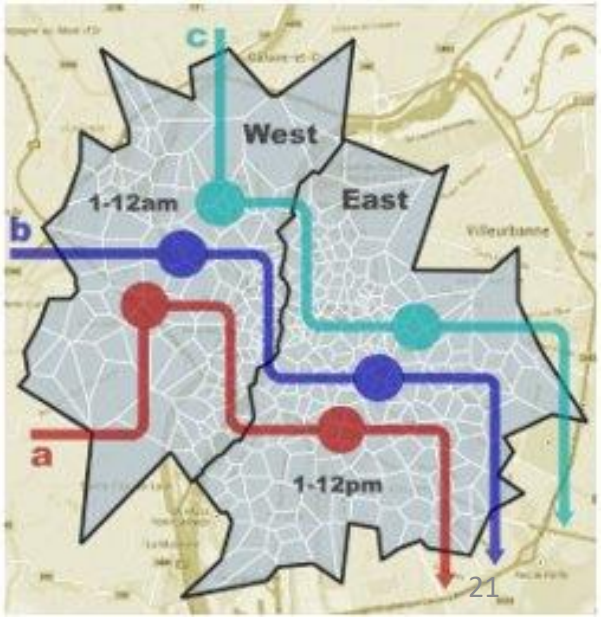


K-anonymat et trajectoires

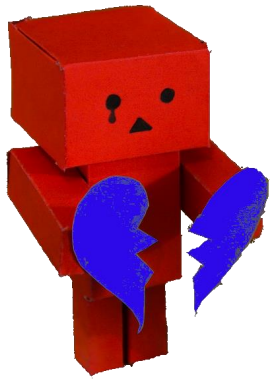
- Quasi-identifiants : tous les points spatio-temporels.
- Attributs sensibles : identité d'un utilisateur, ses POIs, maladies, etc. Il n'y a pas de définition formelle.
- Indistinguabilité : Chaque trajectoire doit être indistinguable de k-1 autres.







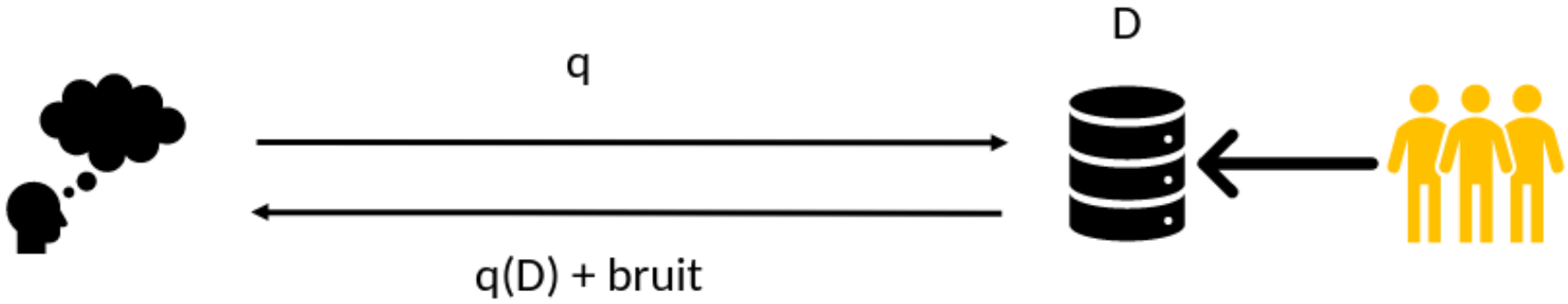
Limites du k-anonymat



- Garantie de vie privée dépend fortement du paramètre k :
  - Un <<gros>> k requiert de grosses perturbations des données.
- Garantie seulement contre l'unicité des individus.
  - Pas d'anonymat si toutes les personnes de votre groupe visitent un hôpital.
- Si deux jeux de données k-anonymisés sont publiés (ex: deux données d'années différentes) la k-anonymité n'est plus garantie.
- Ne protège pas des attaques d'appartenance, co-localisation ou divulgation de lieux visités.

La Differential Privacy (DP) sauvera-t-elle l'anonymisation ?

- Context original : requêtes d'agrégation (comptage, somme). . .
  - Exemple : q = SELECT COUNT(*) FROM MTL-TRJ.TRIPS WHERE QUARTIER == NDG
- Objectif principal : cacher l'impact de la participation d'un individu sur le résultat agrégé.



Impact d'epsilon sur la qualité des données

	NOISE BARRIER	Epsilon	Age
 1 person age 22		100	22
 1 person age 22		1.0	24
 1 person age 22		0.1	-115

Tirée de la présentation : PETS Keynote Address - Deploying Differential Privacy for the 2020 Census of Population and Housing

Impact d'epsilon sur la qualité des données

	NOISE BARRIER	5,000 (50%) runs	9,500 (95%) runs
 1 person age 22		Median(age): 9 → 73	Median(age): 0→ 104
 10 people, all age 22		Median(age): 17 → 61	Median(age): 0→ 103
 100 people, all age 22		Median(age): 21 → 22	Median(age): 21→ 22

Tirée de la présentation : PETS Keynote Address - Deploying Differential Privacy for the 2020 Census of Population and Housing

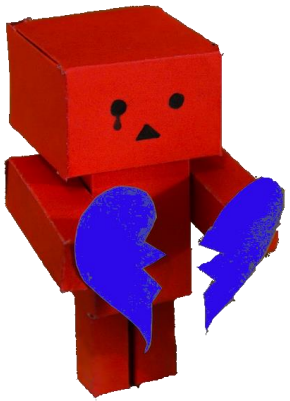
Géo-Indistingabilité : DP pour les localisations

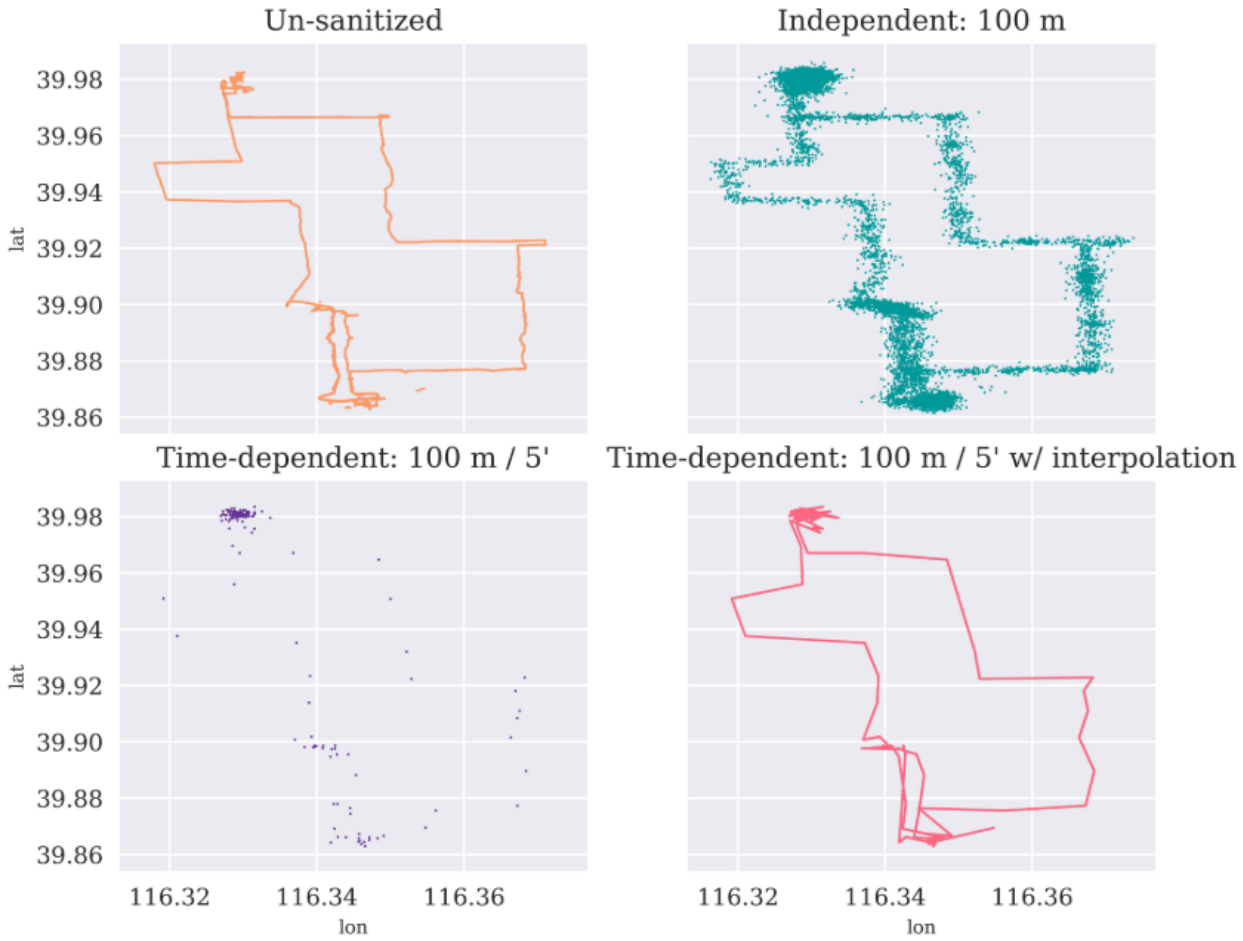
- Idée : un adversaire ne peut pas déduire la position d'un utilisateur dans un rayon r.

Exemple : On ne veut pas révéler si l'utilisateur est sur la rue St-Laurent ou St-Urbain, en revanche ce n'est pas un problème qu'on apprenne qu'il est à Montréal et non à New York.

- Marche bien pour cacher des points (ex: POIS).

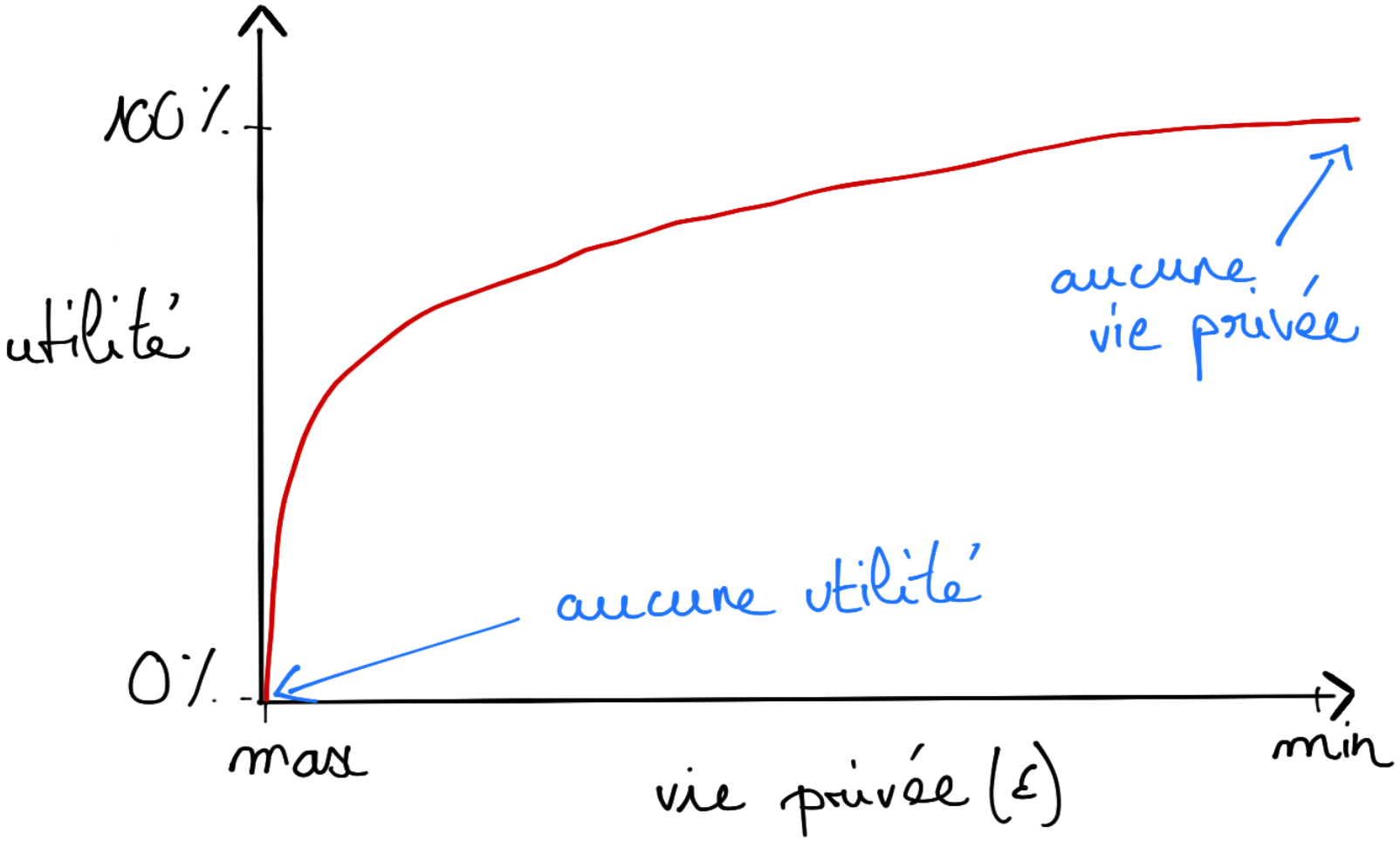
Geo-Ind et trajectoires





Tirée de (Di Luzio et al., 2019)

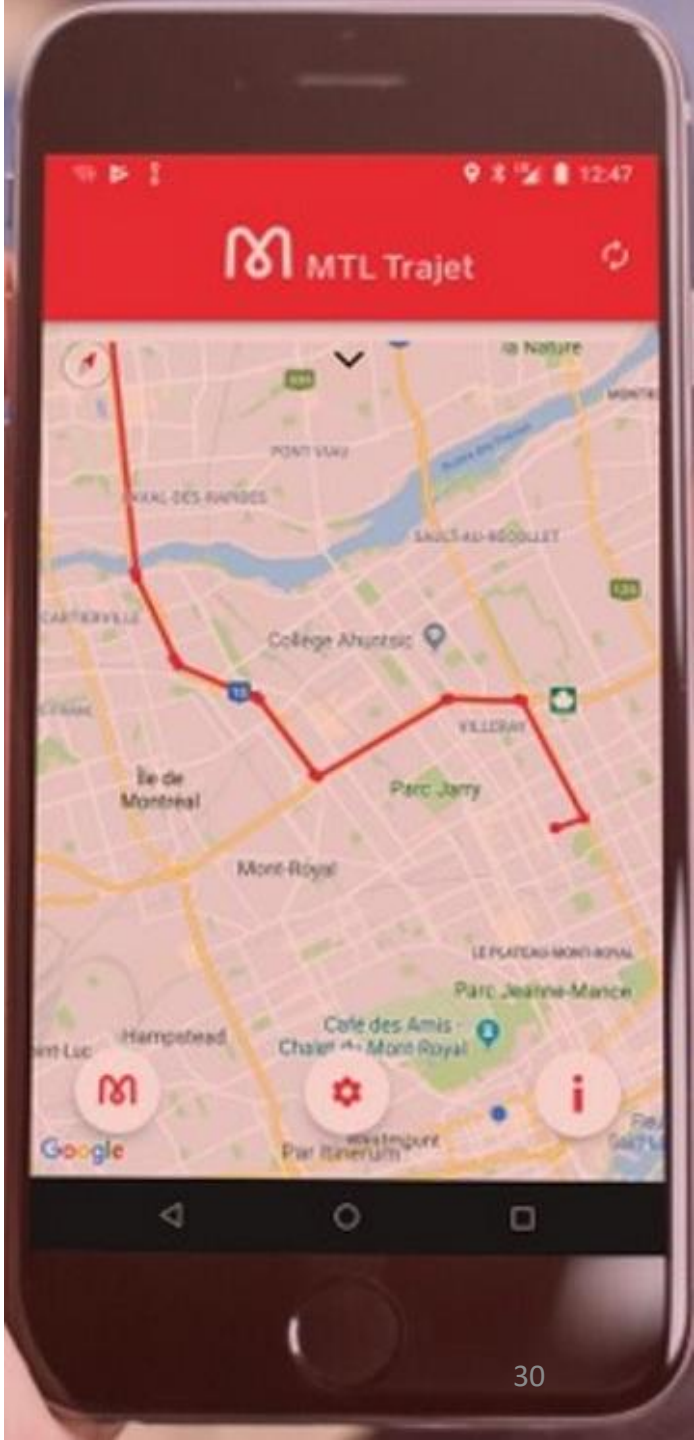
Utilité - Vie privée



Le cas d´étude de MTL-Trajet

Les données MTL-Trajet

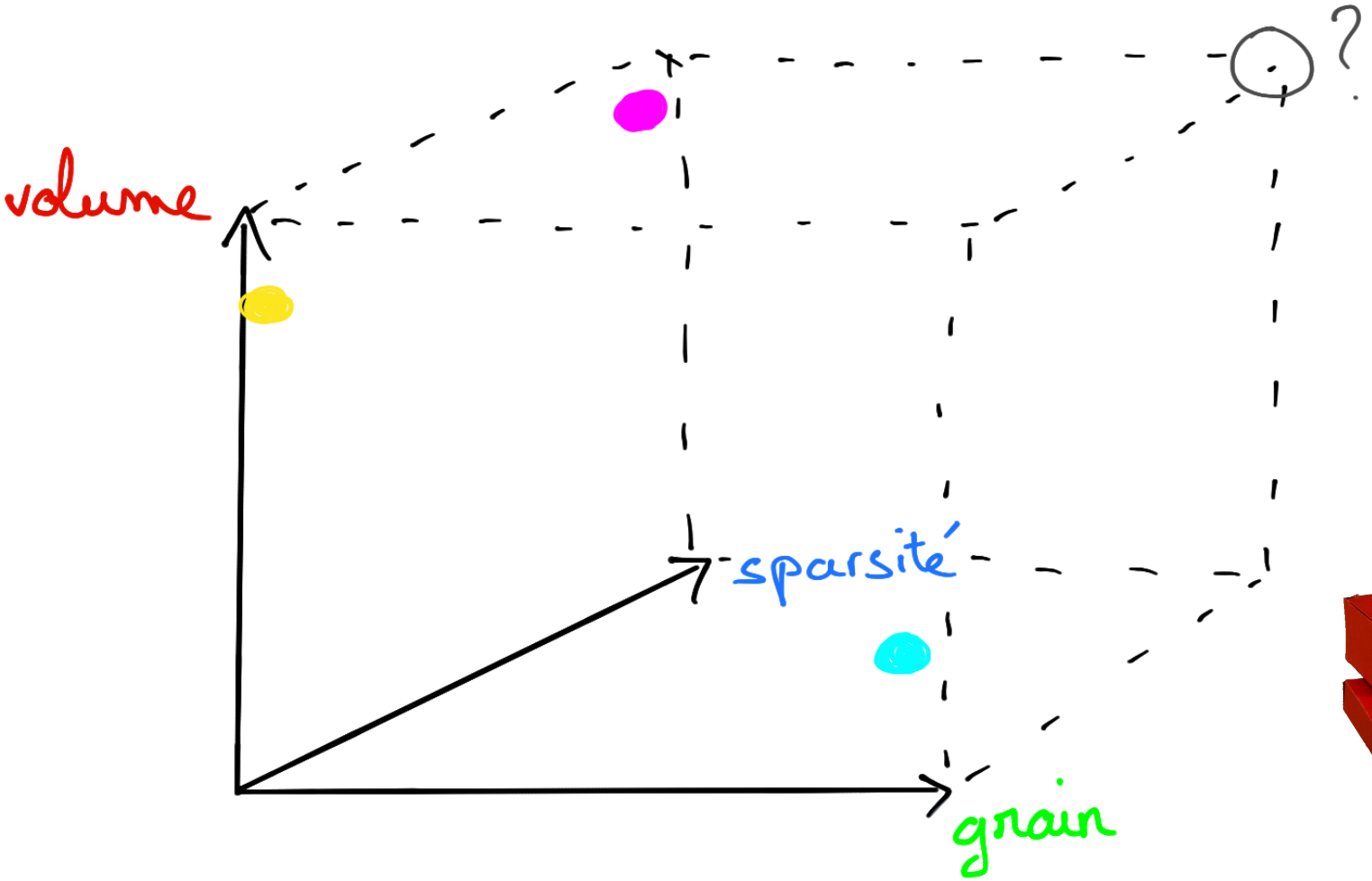
- Données collectées en 2017
- Contient la mobilité de 4072 utilisateurs sur une période d'un mois
- Propriétés des trajets :
    1. Sont dissociés de l'utilisateur par l'utilisation d'un pseudonyme par trajet
    2. Contiennent seulement des informations de déplacements
    3. Le début et la fin sont déplacés à l'intersection la plus proche



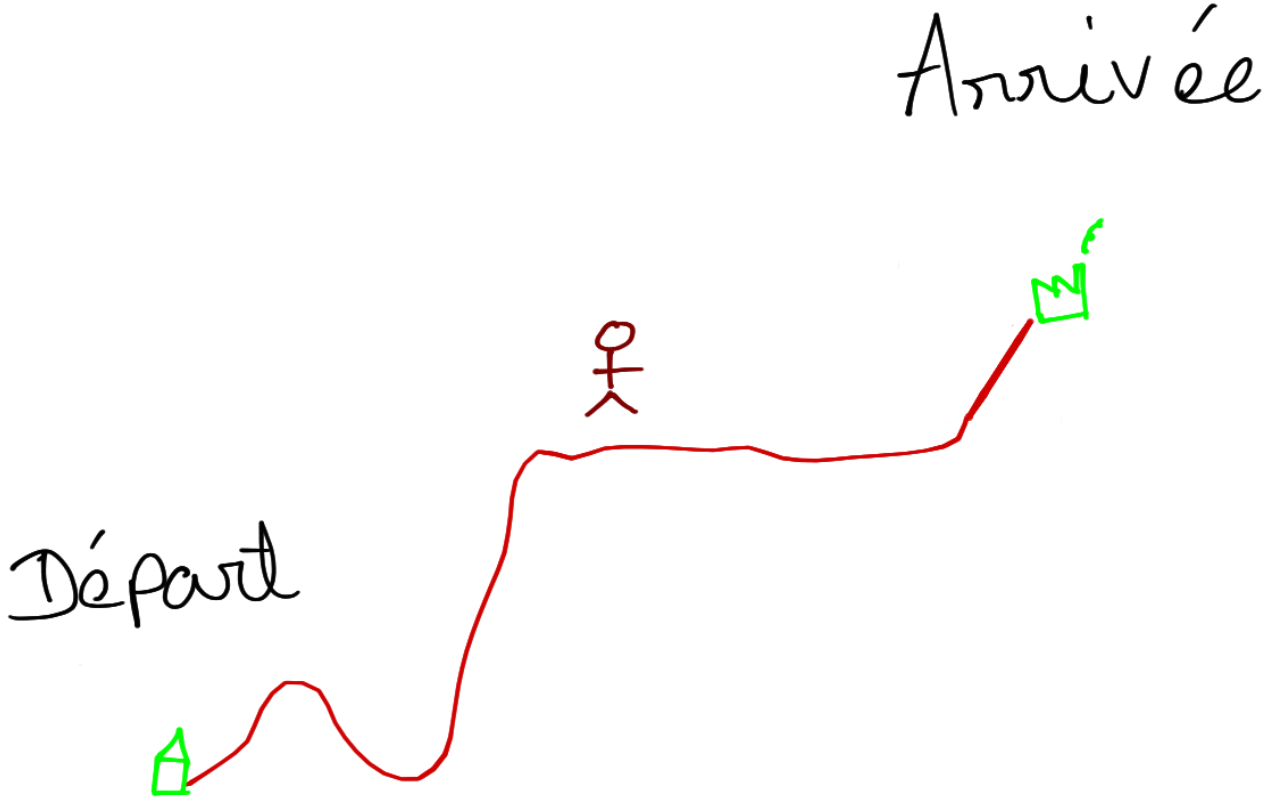
Contraintes d'utilités

1. Conserver une granularité spatiale et temporelle fine qui permette l'analyse de micro-trajets.
   - « Combien de déplacements (c'est-à-dire de trajets) ont lieu sur l'avenue Saint-Laurent ? »
2. Conserver le lien entre les points de départ et d'arrivée de chaque trajet.
   - « D'où partent les trajets qui arrivent au centre-ville à 14h ? ».
3. Conserver la durée et les informations temporelles des trajets.
   - « Quelle est la durée moyenne des trajets effectués le lundi ? »

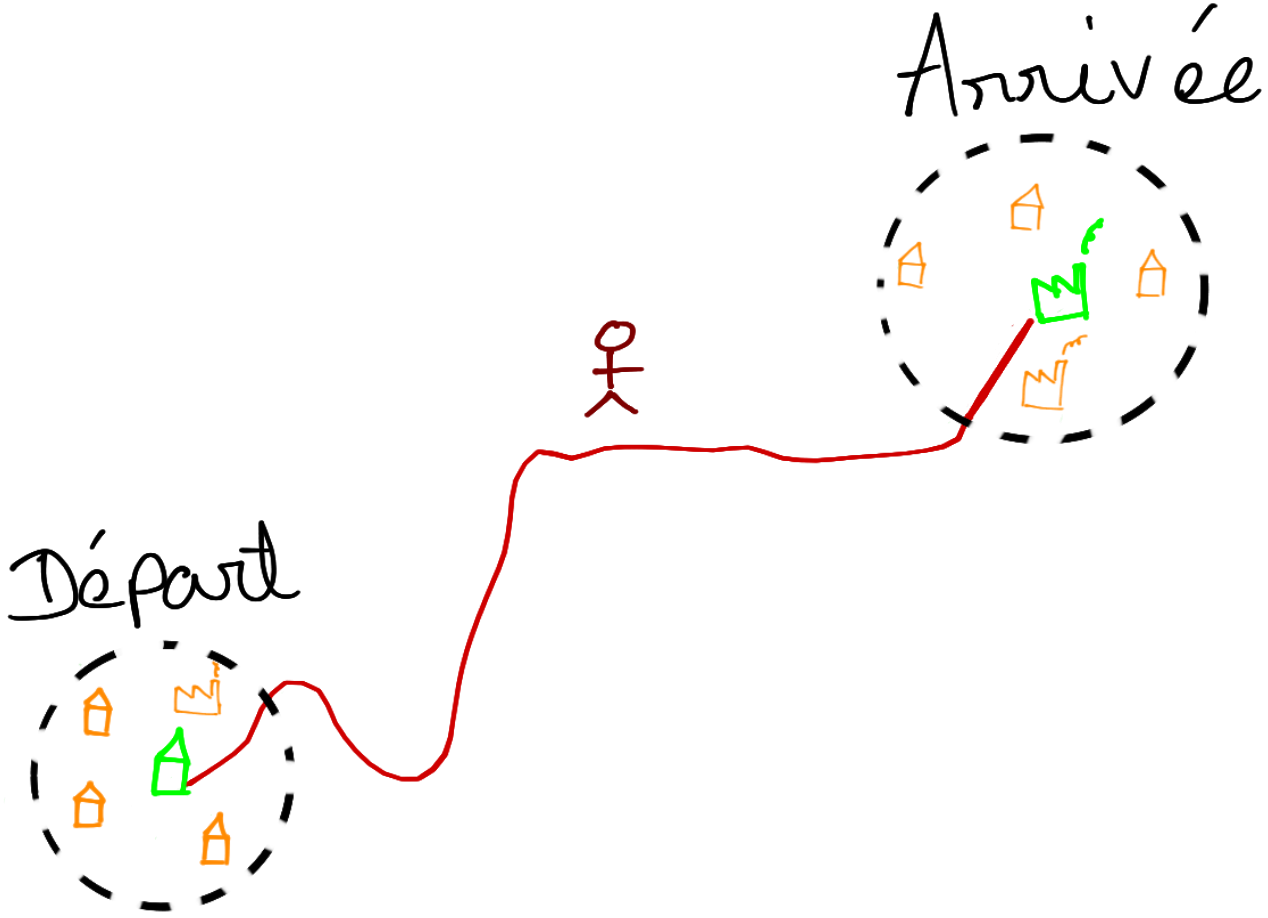
Visualisation de ces contraintes



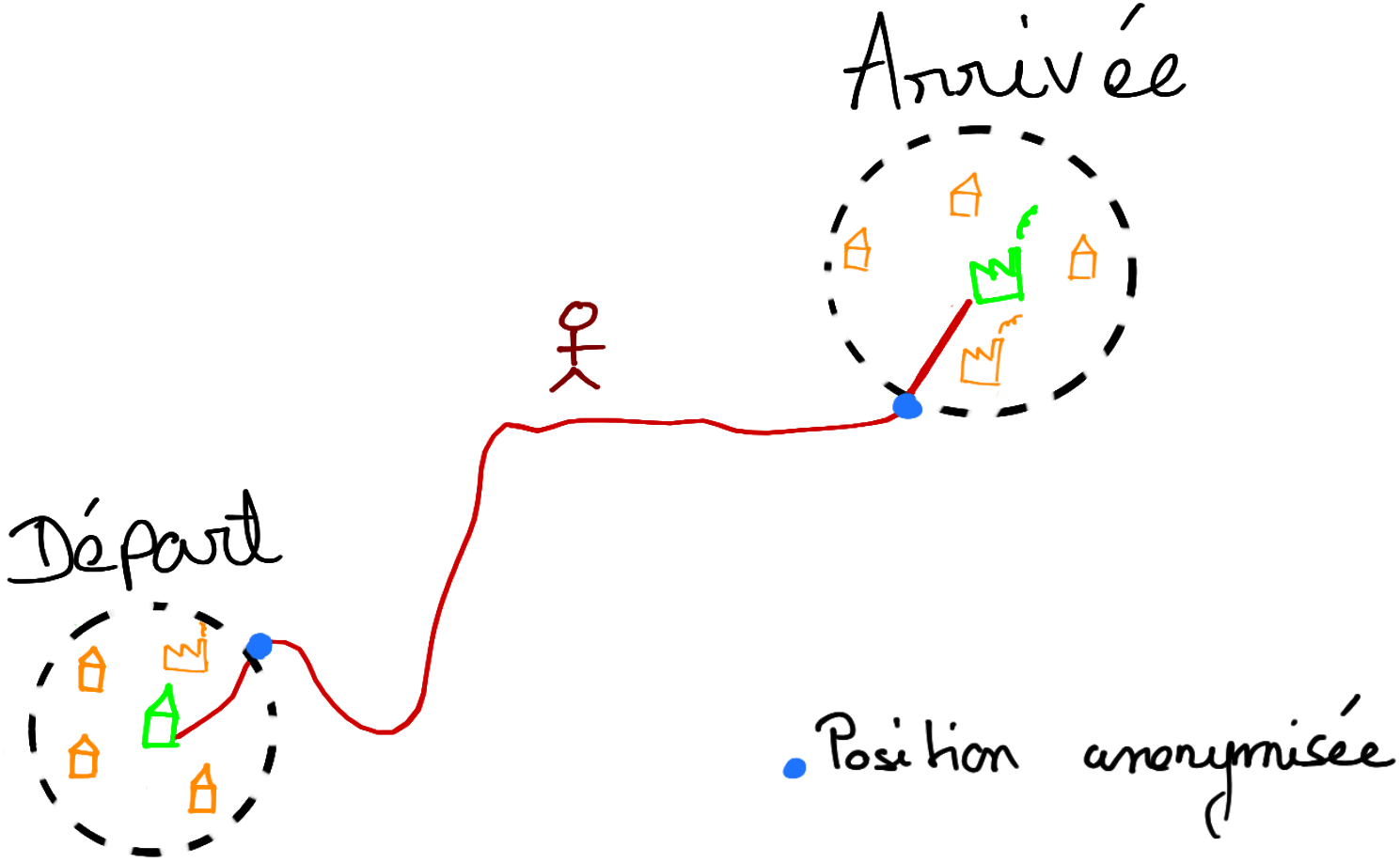
Proposition d'anonymisation



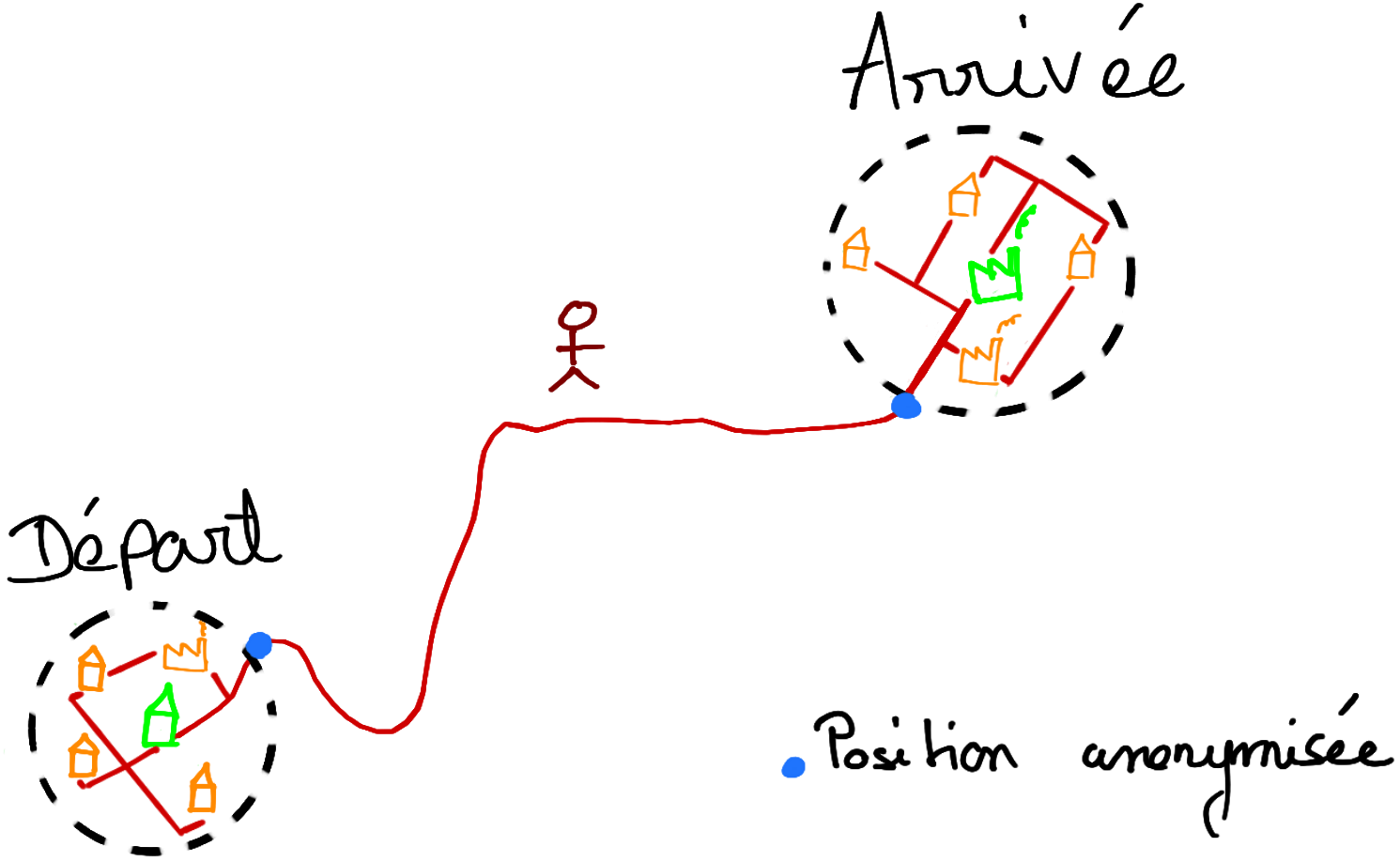
Proposition d'anonymisation



Proposition d'anonymisation



Proposition d'anonymisation

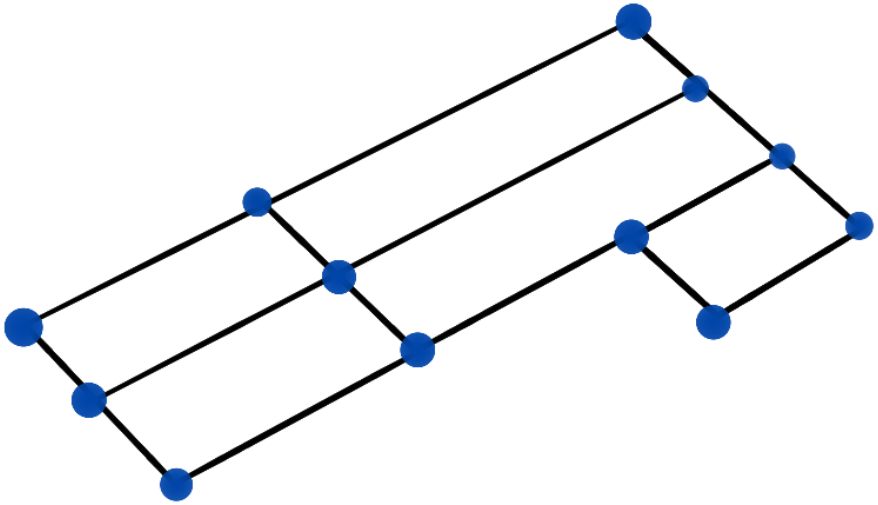


Proposition d'anonymisation

1. Créer des zones de Δ bâtiments connectés ensembles
2. Couper le début et la fin des trajets à la sortie/à l'entrée de la première/dernière zone

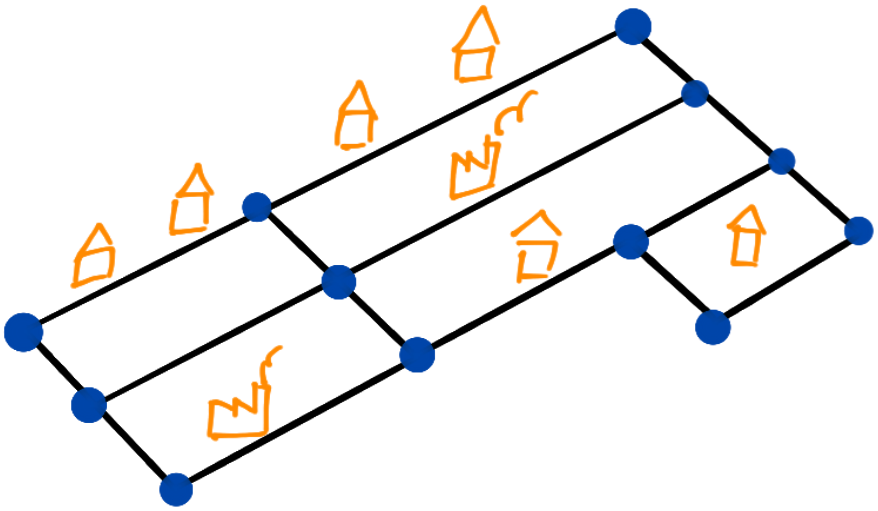
Création de zones : Représentation du domaine

$G = \langle A, S \rangle$



Création de zones : Représentation du domaine

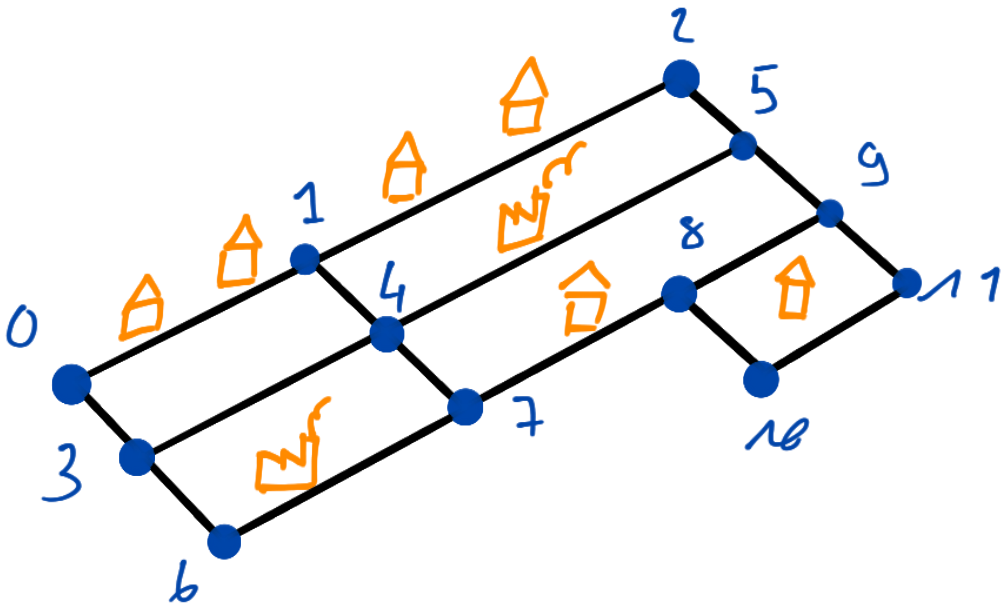
$G = \langle A, S \rangle$



Création de zones : Représentation du domaine

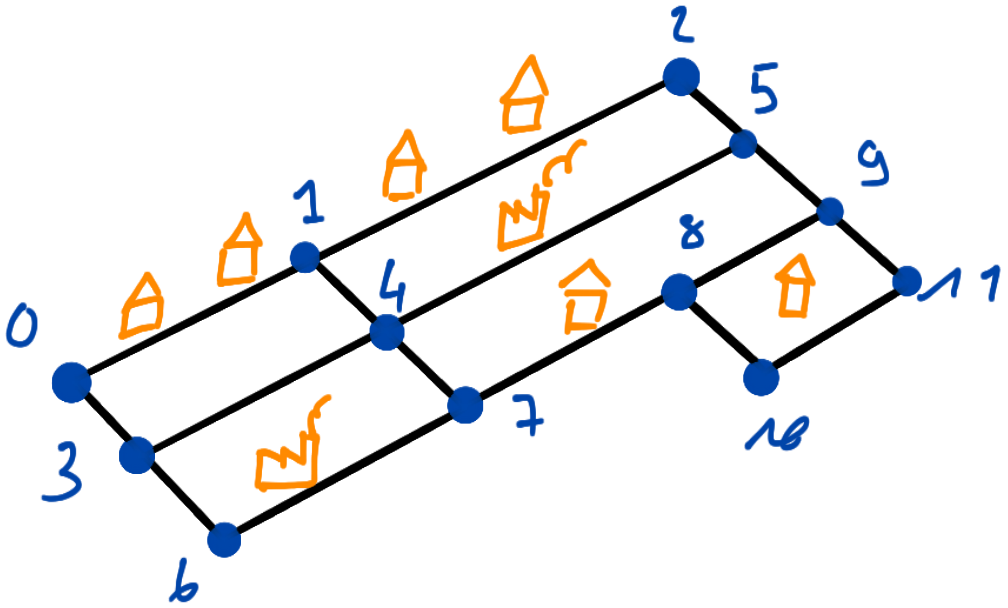
$G = \langle A, S \rangle$

$w_{01} = 2$



Création de zones : Représentation du domaine

$G = \langle A, S \rangle$



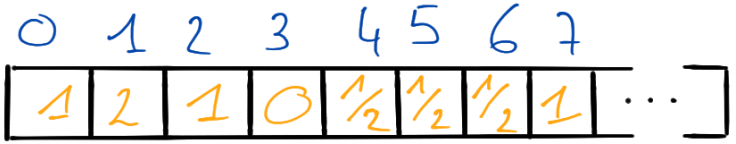
$w_{01} = 2$

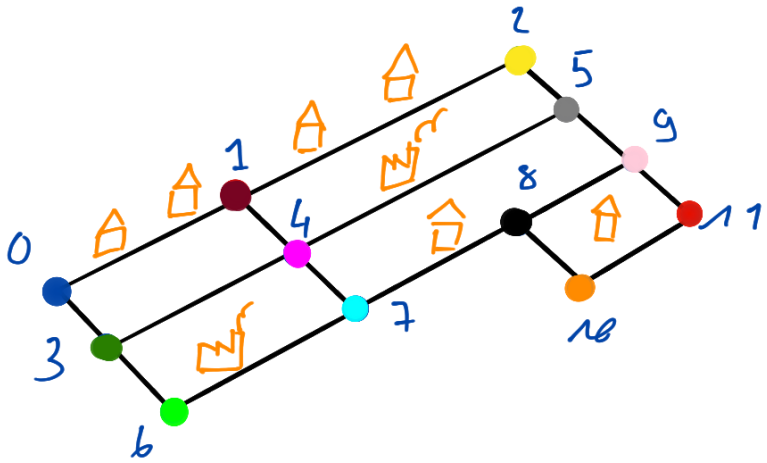
$\delta_0 = \frac{2}{2} + \frac{0}{2}$

$\delta_0 = 1$ bât.

Création de zones : Clustering

$G = \langle A, S \rangle$

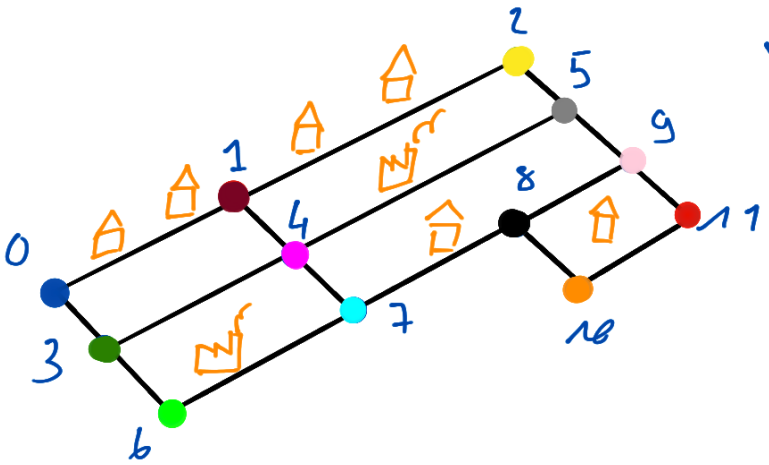


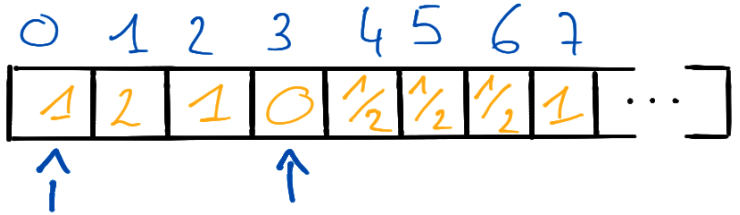


$\Delta = 1$

Création de zones : Clustering

$G = \langle A, S \rangle$

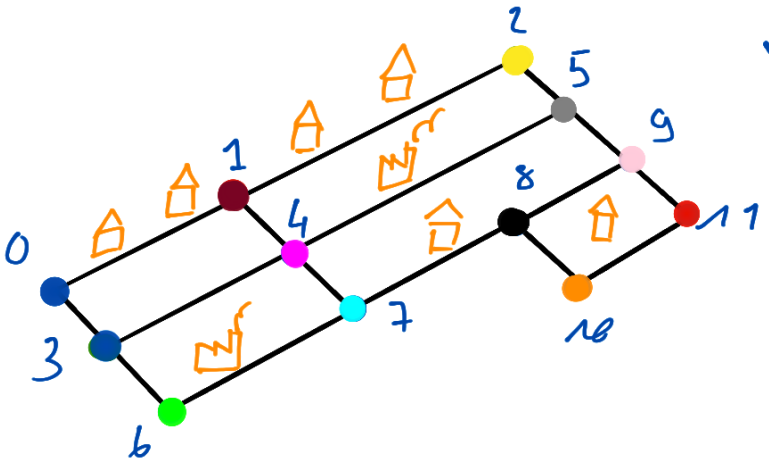


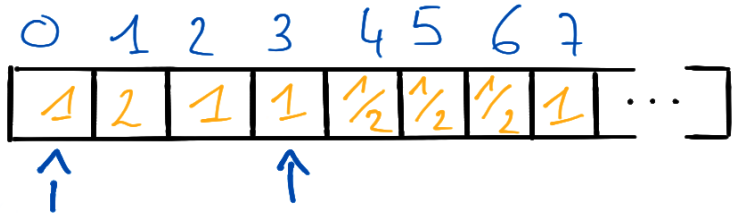


$\Delta = 1$

Création de zones : Clustering

$G = \langle A, S \rangle$

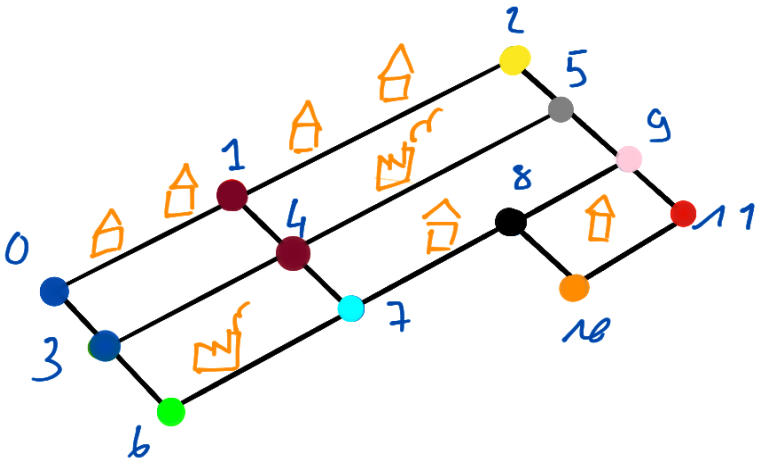


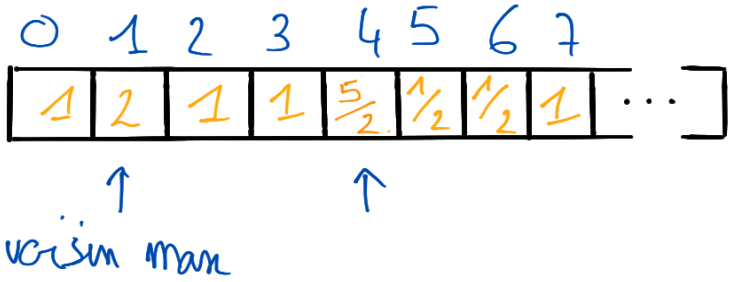


$\Delta = 1$

Création de zones : Clustering

$G = \langle A, S \rangle$

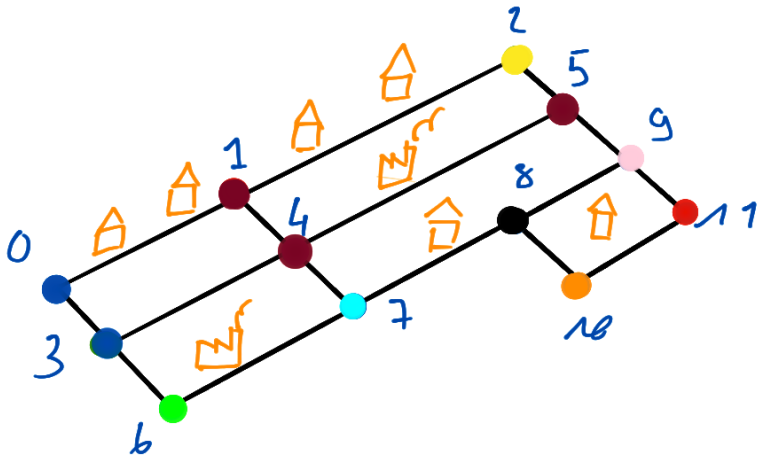


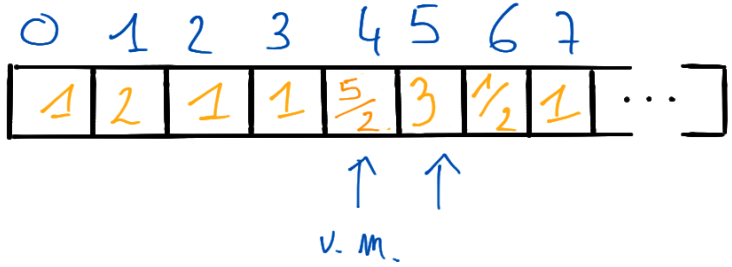


$\Delta = 1$

Création de zones : Clustering

$G = \langle A, S \rangle$

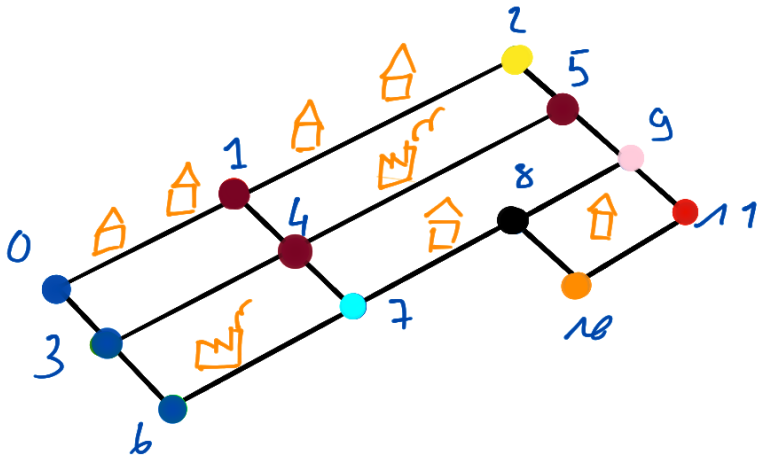


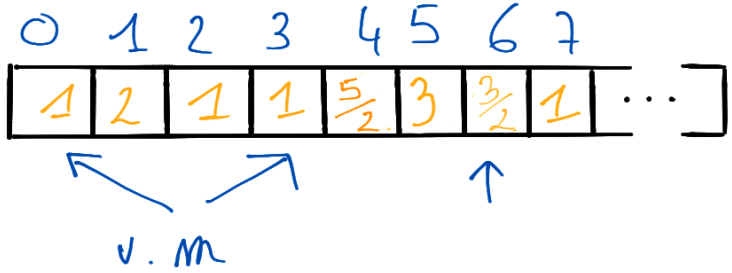


$\triangle = 1$

Création de zones : Clustering

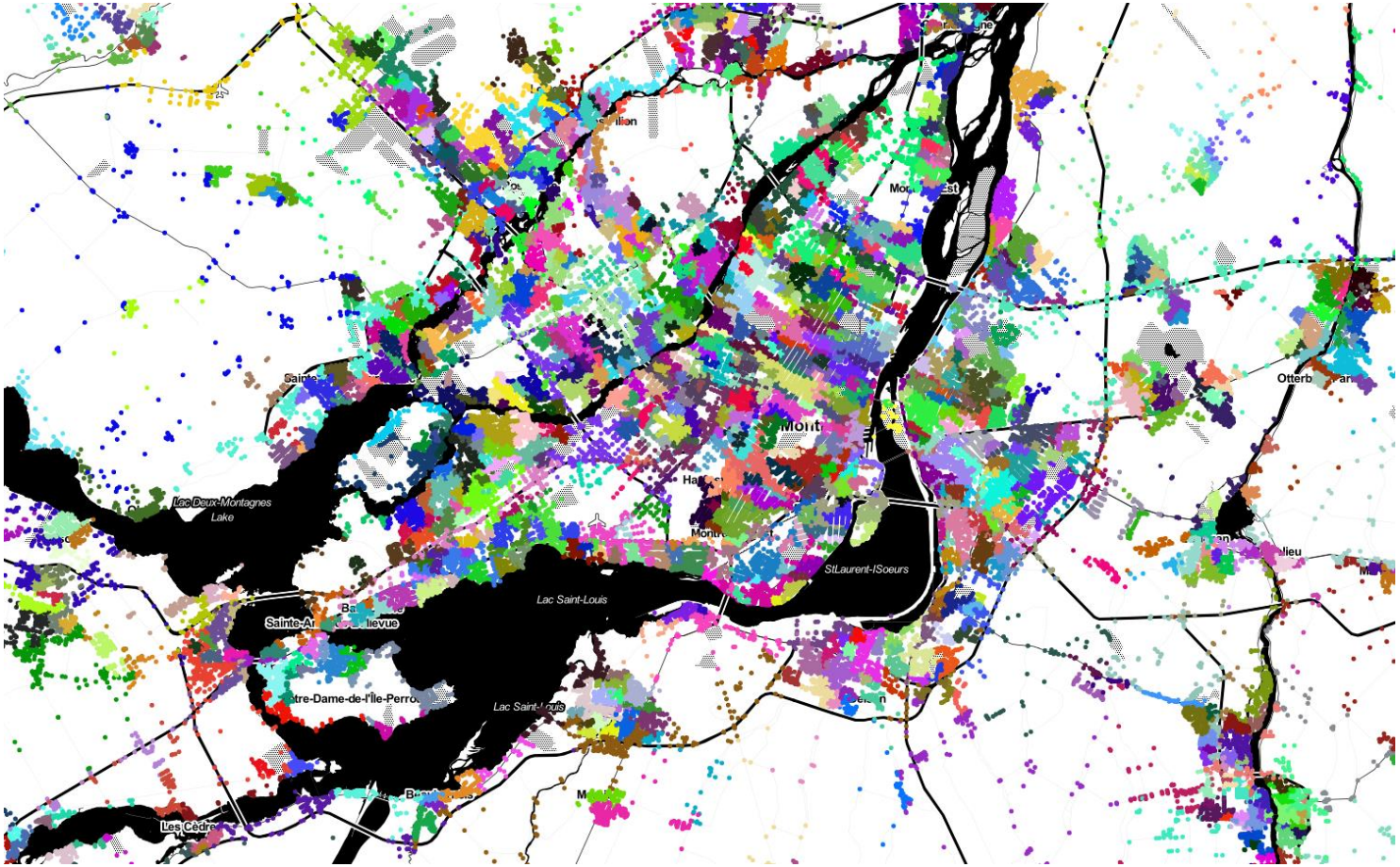
$G = \langle A, S \rangle$





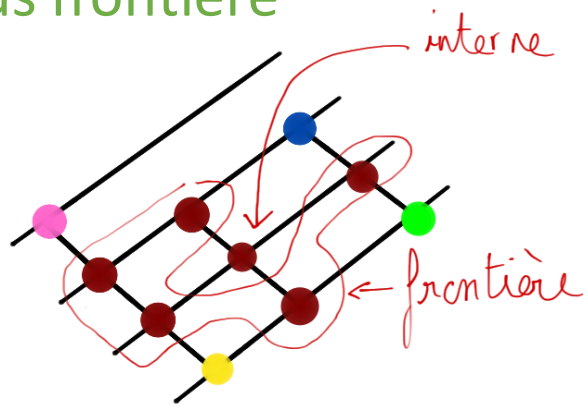
$\Delta = 1$

Clustering : exemple en image



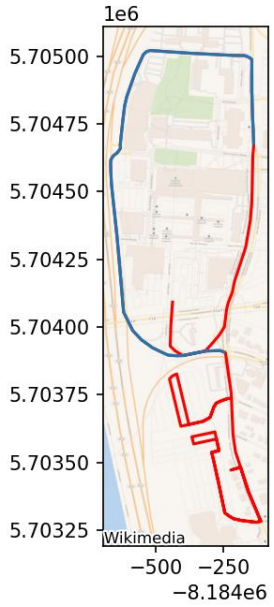
Anonymisation de trajet

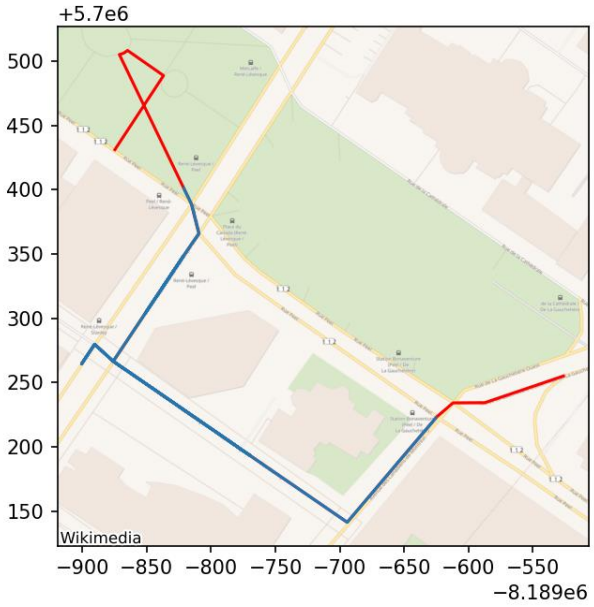
1. Filtrer les nœuds frontière



2. Trouver les intersections frontières de début et fin de chaque trajet.
3. Déplacer le début et la fin de chaque trajet sur des nœuds filtrés.
4. Discrétiser les points aux 15 minutes.

Exemple de trajets anonymisés



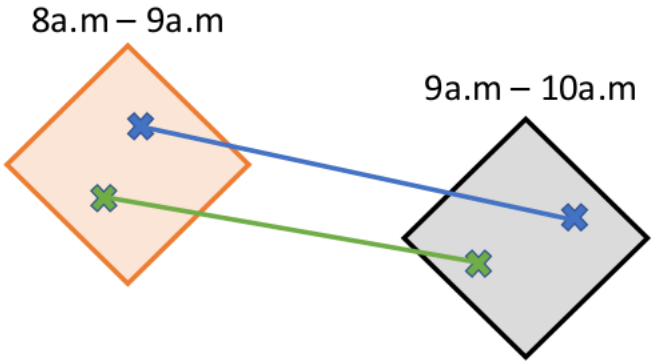


Évaluation de la proposition

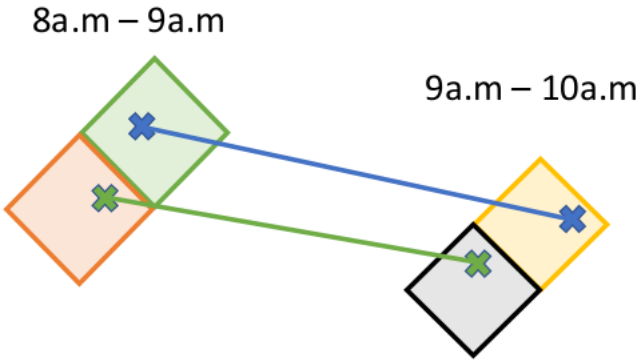
- Utilité : pas de perte au niveau du grain spatial, et discrétisation temporelle aux 15 minutes.
- Vie privée : Cadriciel d'évaluation de la vie privée avec 4 sondes
  1. Attaque d'unicité (ré-identification)
  2. Attaque d'appartenance
  3. Attaque de chainage
  4. Attaque d'inférence de POIs (dérivée du chainage)

Unicité dans une base de données de trajet

- Définition d'unicité : pour un ensemble d'éléments (ex: trajets ou points) et une certaine caractéristique (ex: lieu et heure de départ/arrivée), si l'on fait une requête sur cette caractéristique, l'unicité correspond à la taille de l'ensemble retourné.
- On considère un trajet comme se composant d'un point de départ et d'un point d'arrivée.



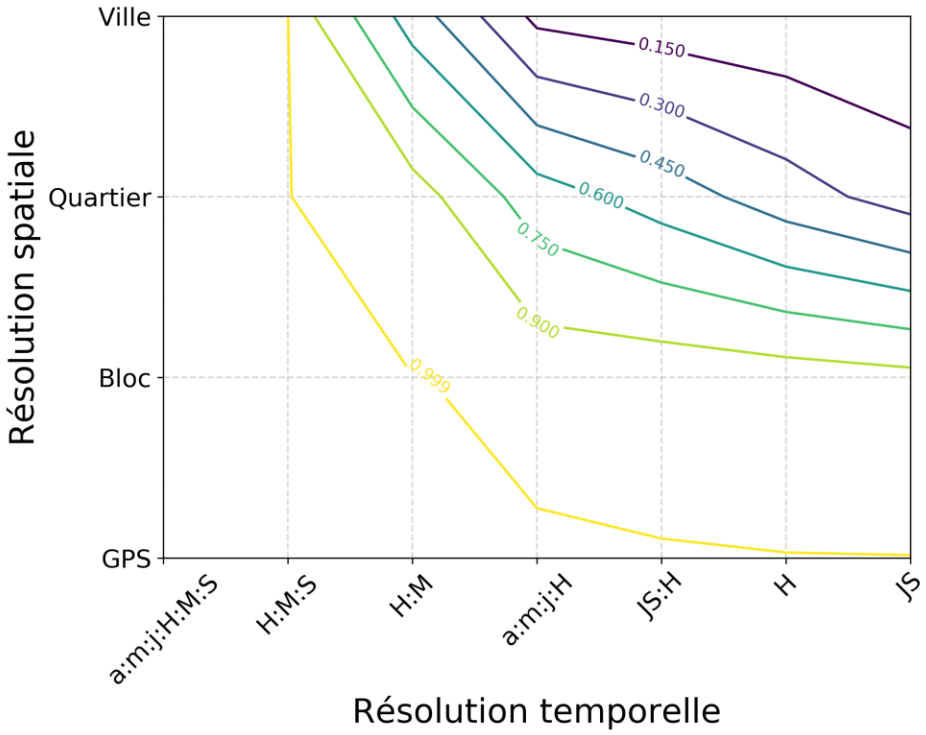
Trajet vert et trajet bleu ne sont pas uniques



Trajet vert et trajet bleu sont uniques

Attaque d'unicité

Données ouvertes



Notre solution

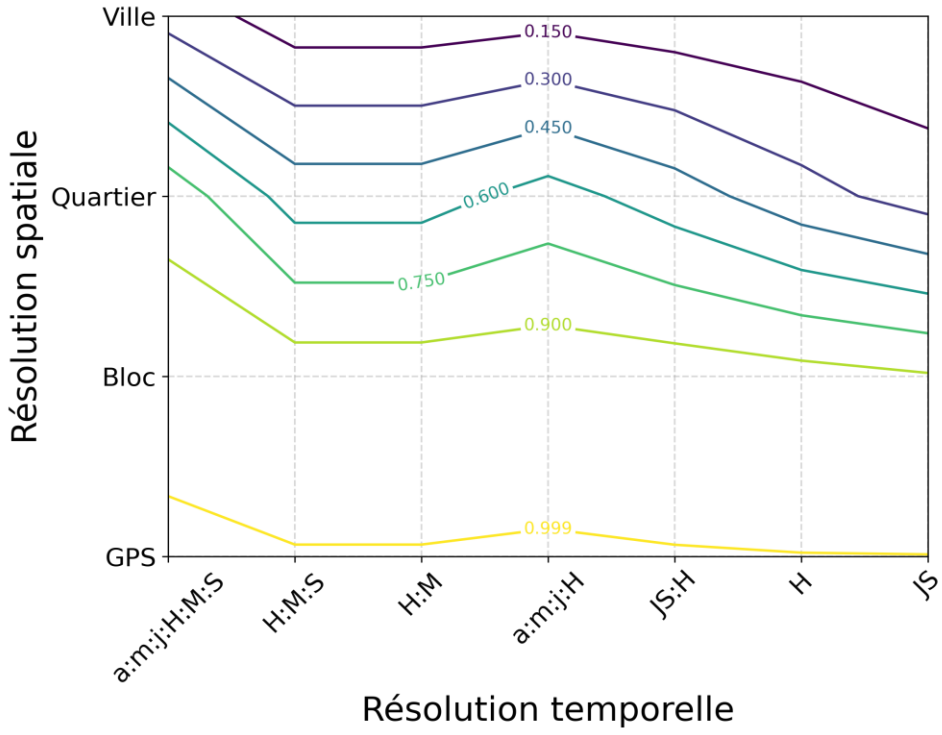
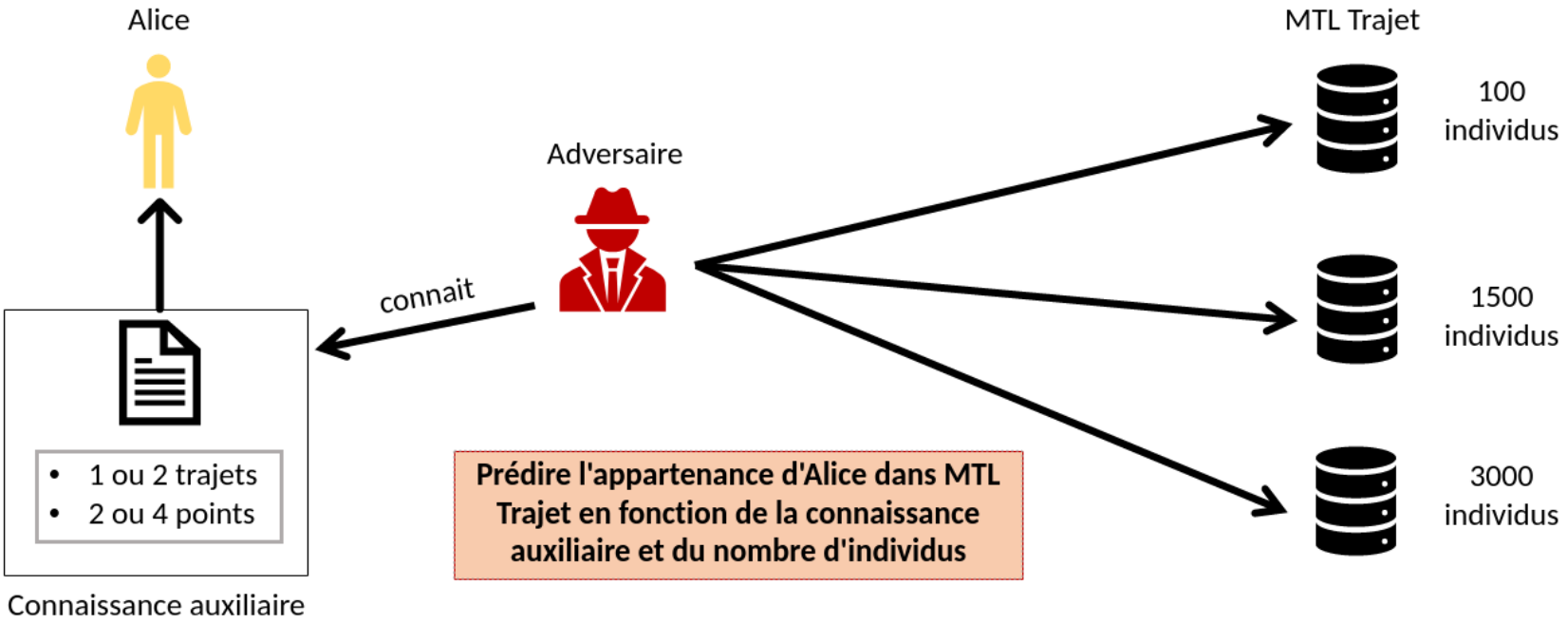
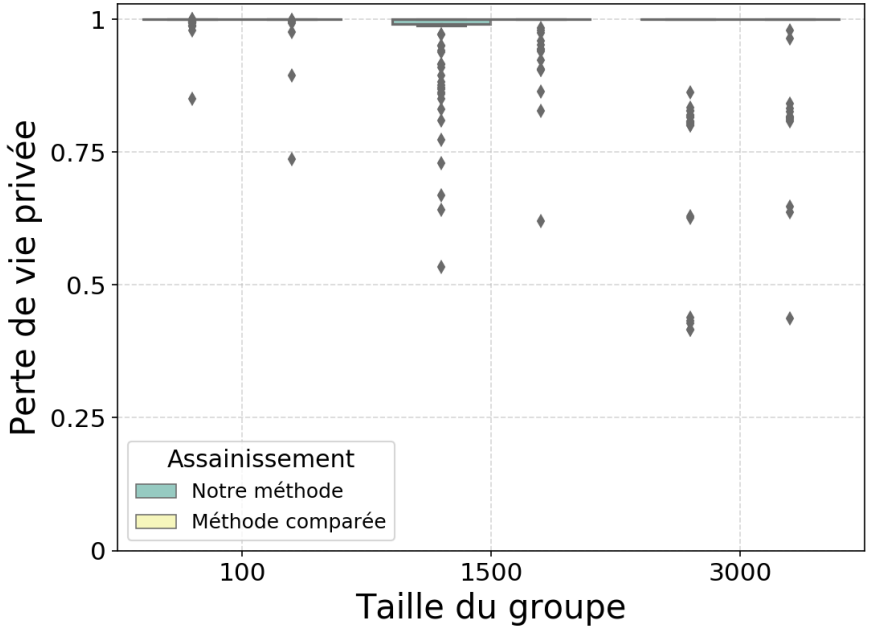


Schéma de l'attaque d'appartenance

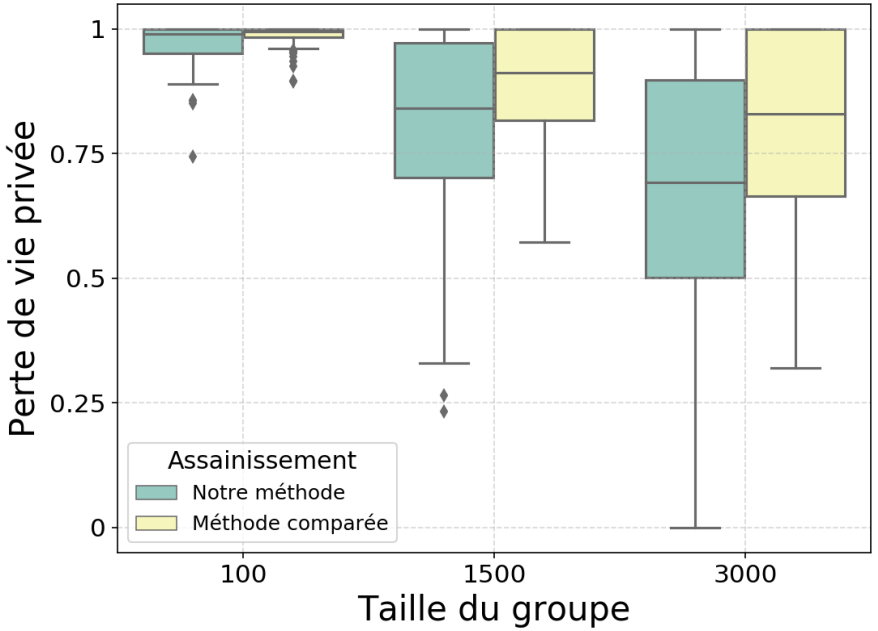


Résultats de l'attaque d'appartenance

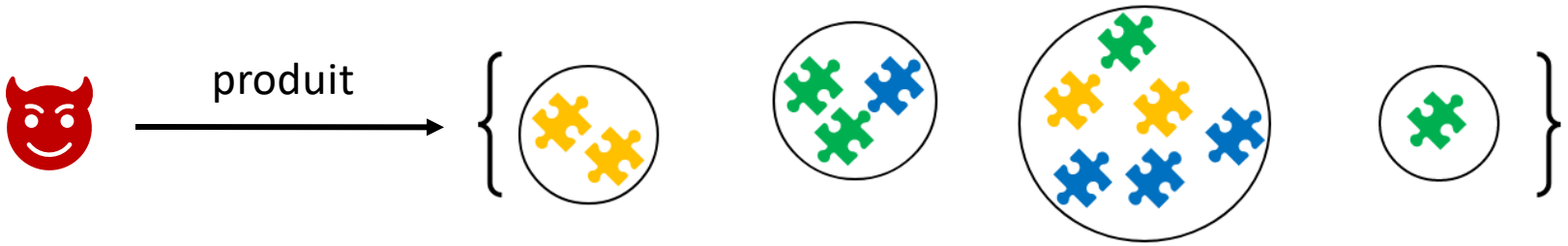
Connait 2 trajets



Connait 1 trajet



Évaluation de la chainabilité



Précision

<< Pureté du groupe par rapport à un individu >>

$$P(\ ,\ ) = \frac{|\ \cap\ |}{|\ |}$$

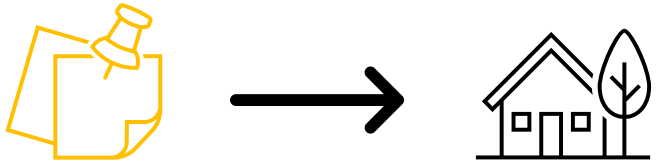
Rappel

<< Complétude du groupe par rapport à un individu >>

$$R(\ ,\ ) = \frac{|\ \cap\ |}{|\ |}$$

Heuristiques de chainages

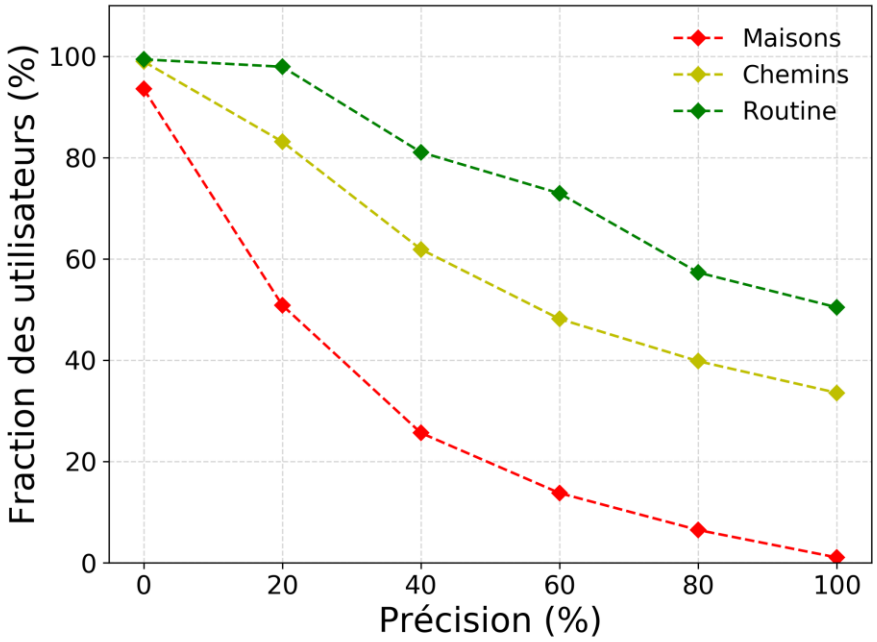
1. Domicile est central aux déplacements: L'adversaire connait le domicile des utilisateurs



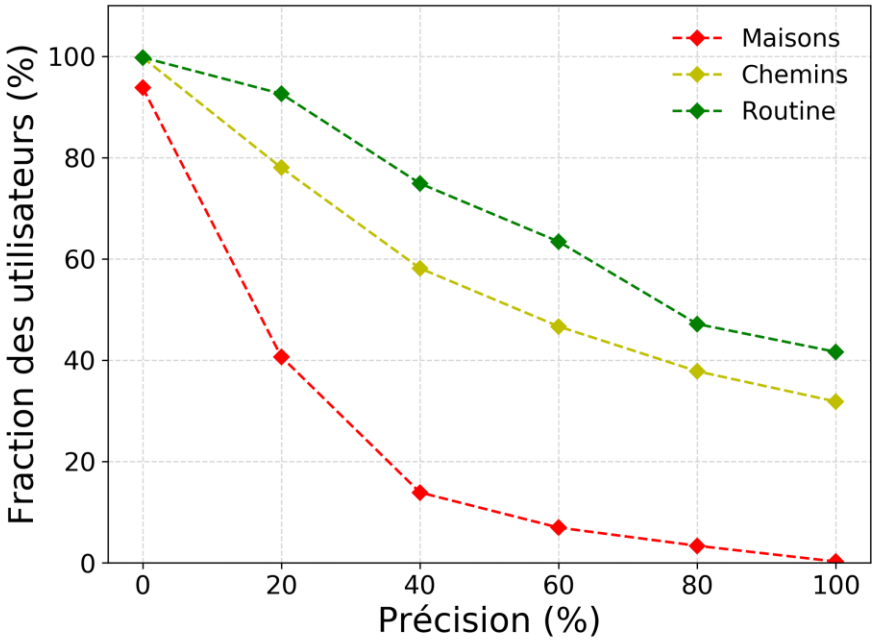
2. Habitude de déplacement : L'adversaire se base sur le fait qu'un même individu passe par les mêmes endroits pour des trajets similaires
3. Routine quotidienne : Les utilisateurs partent du même endroit à la même heure **et/ou** arrivent au même endroit à la même heure.

Attaque de chainage

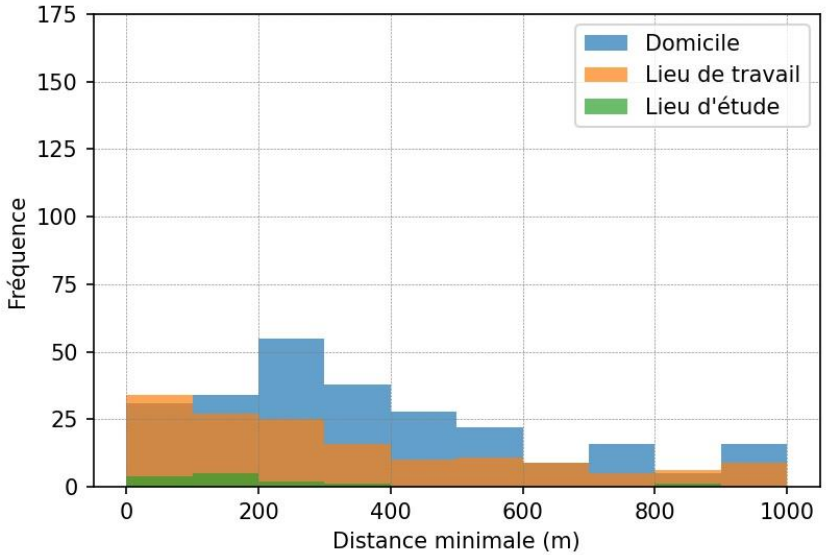
Données ouvertes

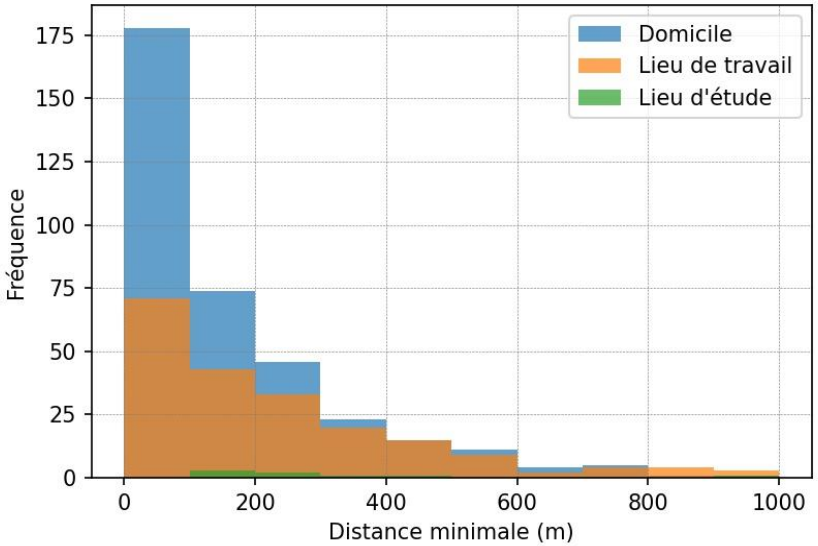


Notre solution

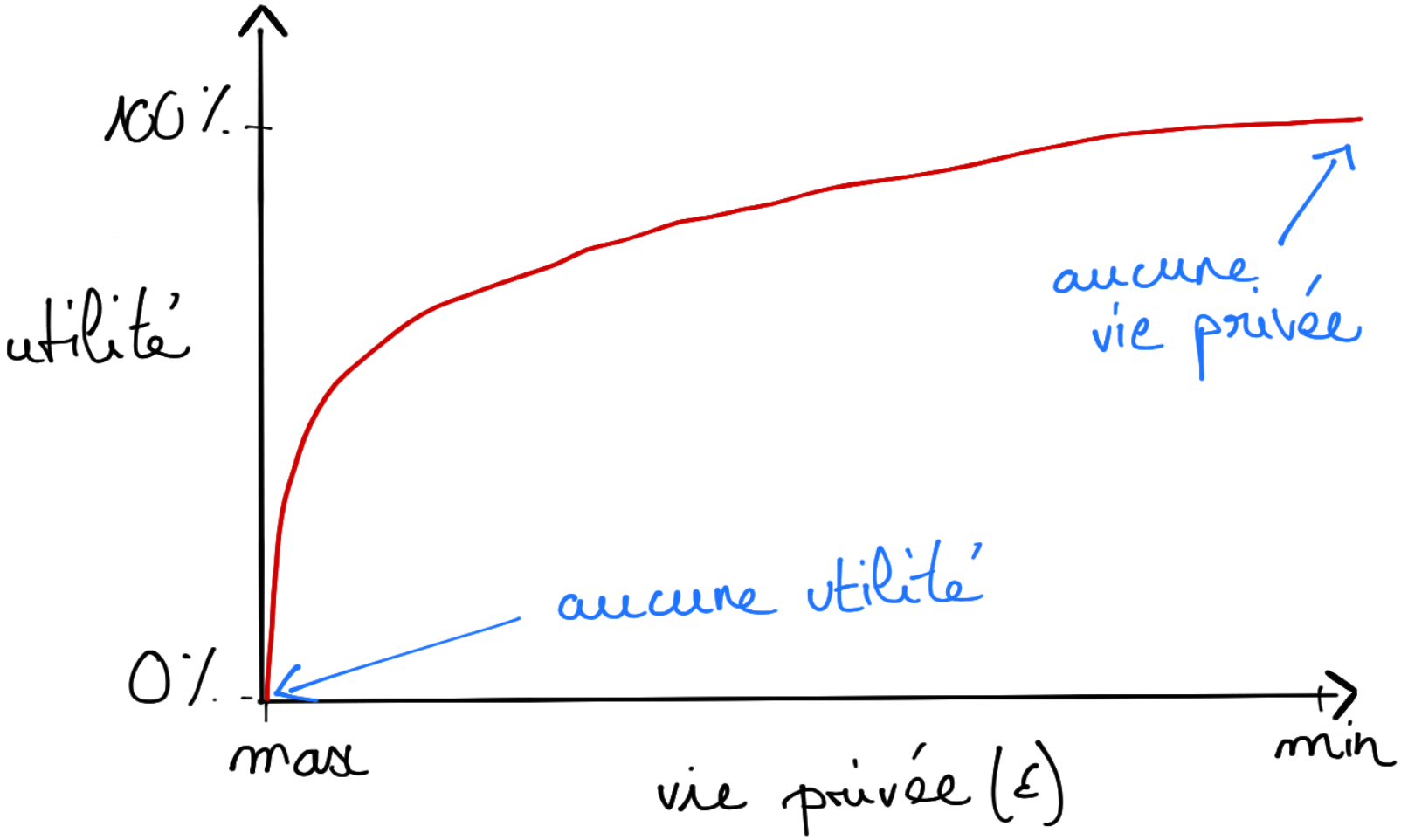


Inférence de POIs





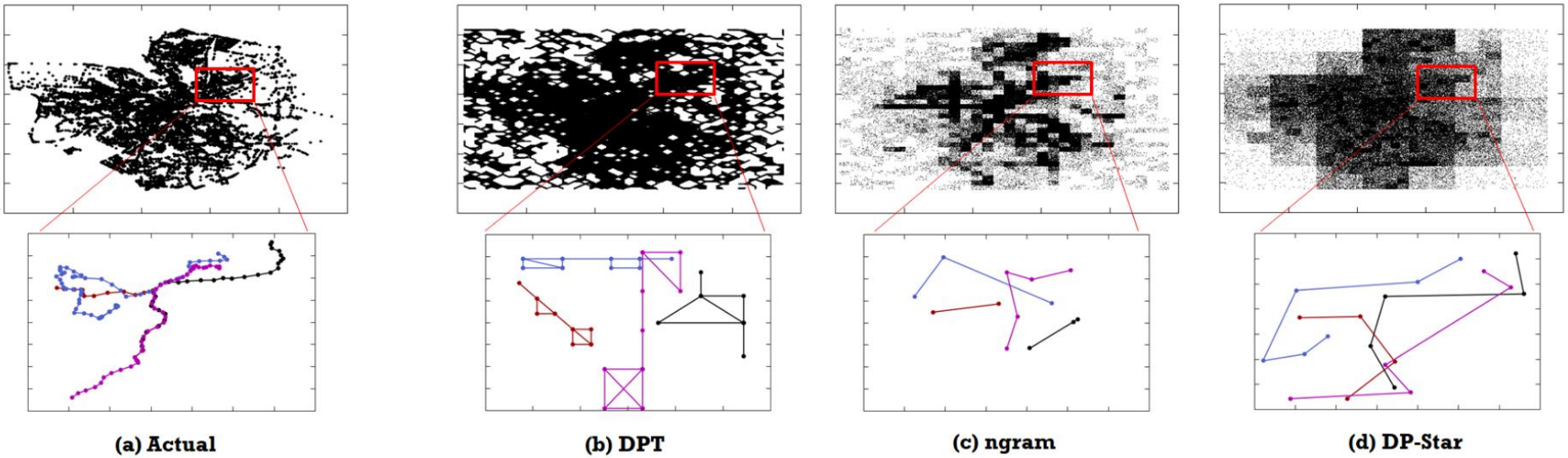
Conclusion – Baisser l'utilité



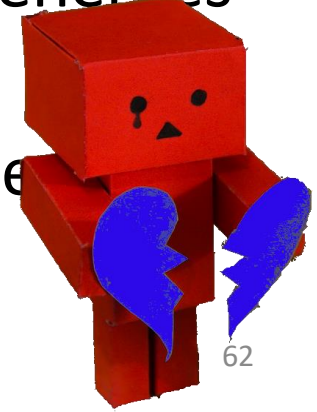
Conclusion - Génération de trajectoires

- Conservent les propriétés de leur modèles génératif
  - Probabilité de transition
  - Matrice Origine-Destination
  - Lieux les plus fréquents
  - ...
- À l'heure actuelle ne permettent pas de générées des données fines et <<Humaine>>
- L'aspect temporel n'est pas pris en compte.

Conclusion - Génération de trajectoires



- À l'heure actuelle ne permettent pas de générées des données fines et <<Humaine>>
- L'aspect temporel n'est pas pris en compte



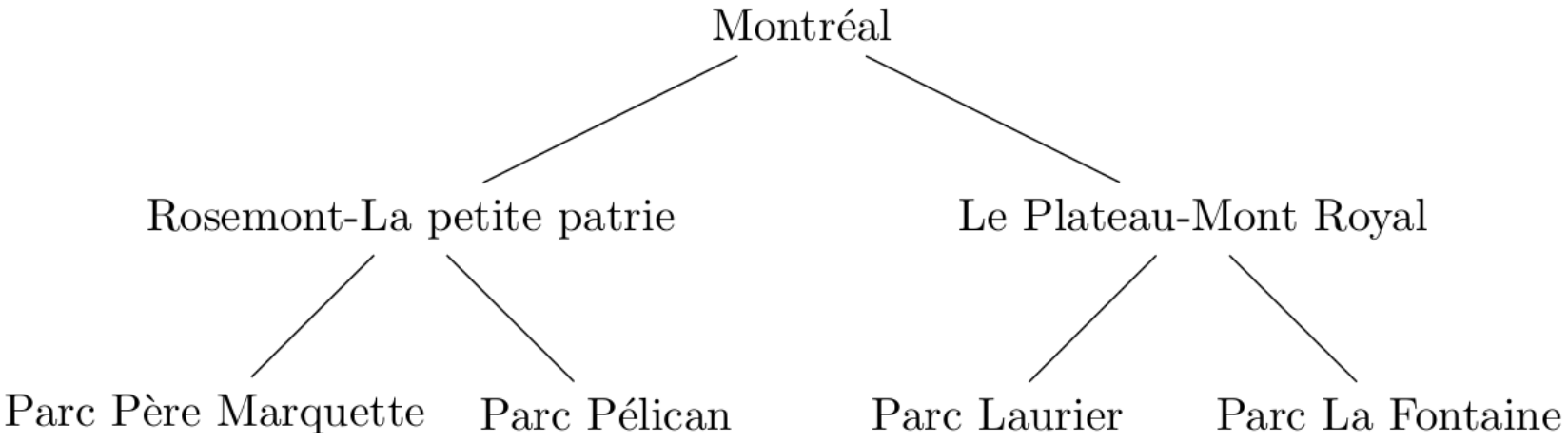
Message à retenir

- Ouverture de données de trajets sans modèle de vie privée.
- Utilité trop élevée pour l'utilisation des travaux de l'état de l'art.
- Proposition d'une méthode d'anonymisation : risques encore présents (MTL-Trajet).
- Ouverture : génération trajectoire, mieux calibrer l'utilité.

Back-up slides

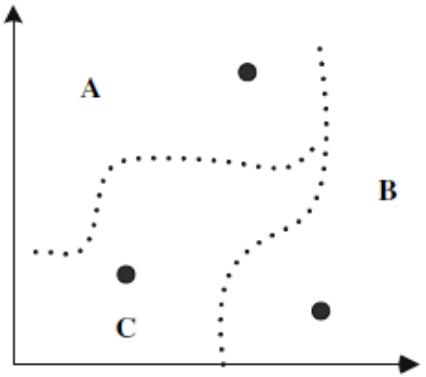
Quelques méthodes d'assainissement

- Généralisation : Perte de précision sur un point/la sémantique d'un point

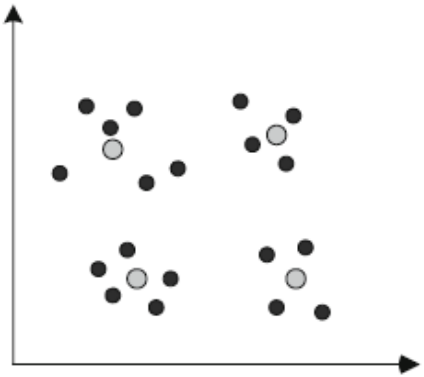


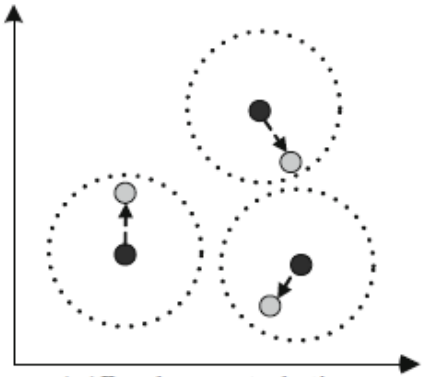
Quelques méthodes d'assainissement

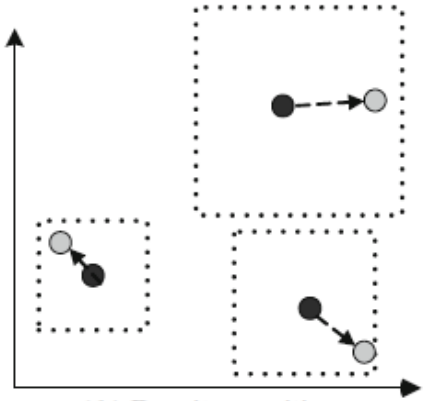
- Généralisation : Perte de précision sur un point/la sémantique d'un point
- Agrégation : Fusionne plusieurs points proches en un seul point représentatif
- Perturbation : Décaler une position d'une distance et direction aléatoire
- Suppression : Supprimer un point



(a.) Aereal aggregation


(b.)Point aggregation


(c.)Random perturbation


(d.) Density sensitive

K-anonymat (Sweeney, 2002)

- QID : Attributs qui ne sont pas identifiants mais prit ensemble le deviennent.
- Garantie : Chaque utilisateurs est indistingable d'au moins k-1 autres dans la base de données par rapport aux QID.

	Quasi-identifiant			Sensible
	Code Postal	Âge	Sexe	Maladie
1	13053	28	Homme	Maladie au cœur
2	13068	29	Homme	Maladie au cœur
3	13068	21	Femme	Infection virale
4	13053	23	Homme	Infection virale
5	14853	50	Femme	Cancer
6	14853	55	Femme	Maladie au cœur
7	14850	47	Femme	Infection virale
8	14850	49	Homme	Infection virale
9	13053	31	Femme	Cancer
10	13053	37	Homme	Cancer
11	13068	36	Homme	Cancer
12	13068	35	Homme	Cancer

(a) Données originales

	Quasi-identifiant			Sensible
	Code Postal	Âge	Sexe	Maladie
1	130**	<30	*	Maladie au cœur
2	130**	<30	*	Maladie au cœur
3	130**	<30	*	Infection virale
4	130**	<30	*	Infection virale
5	1485*	>40	*	Cancer
6	1485*	>40	*	Maladie au cœur
7	1485*	>40	*	Infection virale
8	1485*	>40	*	Infection virale
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

(b) Données 4-anonymisées

K-anonymat (Sweeney, 2002)

- QID : Attributs qui ne sont pas identifiants mais prit ensemble le deviennent.
- Garantie : Chaque utilisateurs est indistingable d'au moins k-1 autres dans la base de données par rapport aux QID.

	Quasi-identifiant			Sensible
	Code Postal	Âge	Sexe	Maladie
1	13053	28	Homme	Maladie au cœur
2	13068	29	Homme	Maladie au cœur
3	13068	21	Femme	Infection virale
4	13053	23	Homme	Infection virale
5	14853	50	Femme	Cancer
6	14853	55	Femme	Maladie au cœur
7	14850	47	Femme	Infection virale
8	14850	49	Homme	Infection virale
9	13053	31	Femme	Cancer
10	13053	37	Homme	Cancer
11	13068	36	Homme	Cancer
12	13068	35	Homme	Cancer

(a) Données originales

	Quasi-identifiant			Sensible
	Code Postal	Âge	Sexe	Maladie
1	130**	<30	*	Maladie au cœur
2	130**	<30	*	Maladie au cœur
3	130**	<30	*	Infection virale
4	130**	<30	*	Infection virale
5	1485*	>40	*	Cancer
6	1485*	>40	*	Maladie au cœur
7	1485*	>40	*	Infection virale
8	1485*	>40	*	Infection virale
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

(b) Données 4-anonymisées

Définition formelle de la DP (Dwork, 2006)

- Définition : Un mécanisme f satisfait la e-DP ssi pour toute paire de bases de données D et D0 qui diffèrent au maximum d'un enregistrement, et pour n'importe quel sortie S de f, il est vrai que

$$Pr[f(D) = S] \leq e \times Pr[f(D0) = S]$$

- Ajout de bruit (Laplacien) aux requêtes en fonction de la contribution d'un individu dans D (la sensibilité)
- Epsilon (e) : le paramètre de vie privée.

Génération (e-DP) de trajectoires

- Idée : générer des trajectoires proches de l'ensemble original mais qui ne correspondent pas des individus réels.

- Procédé :
  1. Représenter les données avec une structure de données générative (ex: arbre préfixes).
  2. Bruiter cette structure afin de la rendre e-DP.
  3. Générer des trajectoires depuis la structure.

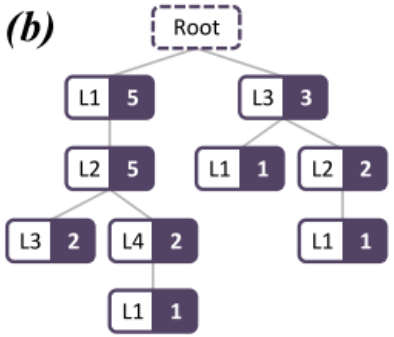
Génération (e-DP) de trajectoires

- Idé
  l'er
  des
  pas
- Pro
  1.
  2.
  3.

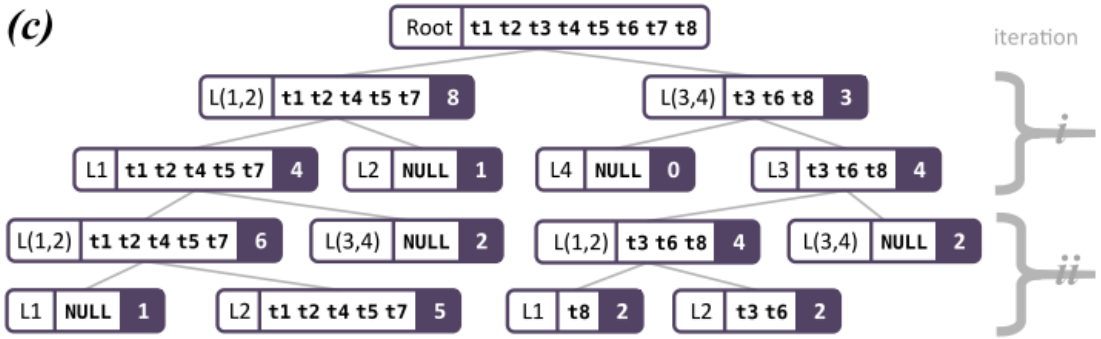
(a)

Pseudo-identifier	Spatial locations
t1	L1 → L2 → L3
t2	L1 → L2
t3	L3 → L2 → L1
t4	L1 → L2 → L4
t5	L1 → L2 → L3
t6	L3 → L2
t7	L1 → L2 → L4 → L1
t8	L3 → L1

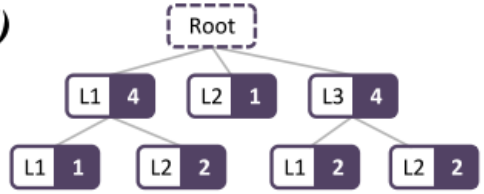
(b)



(c)



(d)



(e)

Pseudo-identifier	Spatial locations
t1	L1
t2	L1 → L2
t3	L1 → L2
t4	L3 → L1
t5	L3 → L2

Tirée de (Fiore et al., 2019)