# Anonymisation de données: le cas MTL Trajet

2ème Rencontres Francophones Transport Mobilité

François Bélisle [1] Nicolas Saunier [2] Jérôme Le Ny [2] Mathilde Pelletier [2]

11 juin 2019

[1] MathMobile [2] Polytechnique Montréal


POLYTECHNIQUE MONTRÉAL

UNIVERSITÉ D'INGÉNIERIE

# Plan de la présentation

Contexte

Le problème

La confidentialité différentielle

Conclusion

# Plan de la présentation

# Mouvement des données ouvertes

L'ouverture des données est un "mouvement, une philosophie d'accès à l'information, une politique publique et une pratique de publication de données librement accessibles et exploitables" (Wikipedia)

- fin des années 2000: data.gov aux É.-U.
- 10 critères de la Sunlight Foundation 2010

# Mouvement des données ouvertes

- Montréal Ouvert créé en 2009
- Politique de données ouvertes adoptée en 2011
  - ouverture par défaut
  - sauf si risque pour la sécurité publique ou la protection des données personnelles

# Les études MTL Trajet depuis 2016

- Enquêtes déplacement à l'aide d'un téléphone intelligent, effectuées à l'automne en 2016, 2017 et 2018
- Complémentaires aux enquêtes origine-destination (OD) de la grande région de Montréal effectuées tous les 5 ans
- Ensembles de données de 2016 et 2017 disponibles sur le portail des données ouvertes

# Plan de la présentation

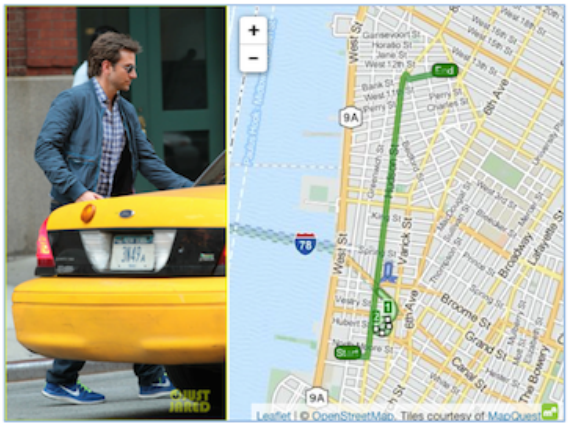
**Peut-on rendre disponibles publiquement des trajectoires “anonymisées”?**

# Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset

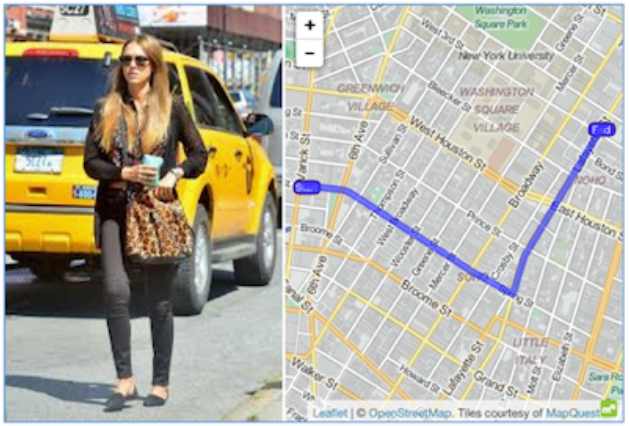
## Stalking celebrities

First things first. How might I track a person? Well, to zone in on a particular trip, I can use any combination of known characteristics that appear in the dataset, such as the pickup or drop-off coordinates or datetime, the medallion or license number, or even the fare amount from a receipt. Being the avid fanboy that I am (*note: sarcasm*), I thought it might be interesting to find out something new about some of the celebrities who had been seen in New York in 2013. In particular, where did they go to / come from, and how much did they tip?

In order to do this, I spent some of the most riveting hours of my professional career searching through images of "celebrities in taxis in Manhattan in 2013" to find enough information to identify the correct record in the database. I had some success – combining the below photos of Bradley Cooper and Jessica Alba with some information from celebrity gossip blogs allowed me to find their trips, which are shown in the accompanying maps.




Bradley Cooper (Click to Explore)




Jessica Alba (Click to Explore)

In Brad Cooper's case, we now know that his cab took him to Greenwich Village, possibly to have dinner at Melibea, and that he paid $10.50, with *no recorded tip*. Ironically, he got in the cab to escape the photographers! We also know that Jessica Alba got into her taxi outside her hotel, the Trump SoHo, and somewhat surprisingly also did not add a tip to her $9 fare. Now while this information is relatively benign, particularly a year down the line, I have revealed information that was *not previously in the public domain*. Considering the speculative drivel that usually accompanies these photos (trust me, I know!), a celebrity journalist would be thrilled to learn this additional information.

# Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset

## A few innocent nights at the gentlemen's club

But OK, perhaps you're not convinced. After all, this dataset is (thankfully) not real-time. How about we leave the poor celebrities alone and consider something a little more provocative. Larry Flynt's Hustler Club is in a fairly isolated location in Hell's Kitchen, and no doubt experiences significant cab traffic in the early hours of the morning. I ran a query to pull out all pickups that occurred outside the club after midnight and before 6am, and mapped the drop-off coordinates to see if I could pinpoint individuals who frequented the establishment. The map below shows my results – the yellow points correspond to drop-offs that are closely clustered, implying a frequent customer.


Click to Explore

The potential consequences of this analysis cannot be overstated. Go ahead, zoom in. You will see that the GPS coordinates are terrifyingly precise. Using this freely-obtainable, easily-created map, one can find out where many of Hustler's customers live, as there are only a handful of locations possible for each point. Add a little local knowledge, and, well, it's not rocket science. *"I was working late at the office"* no longer cuts it: Big Brother is watching.

Even without suspicions or knowledge of the neighborhood, I was able to pinpoint certain individuals with high probability. Somewhat shockingly, just googling an address reveals all kinds of information about its inhabitants. Take the following example:

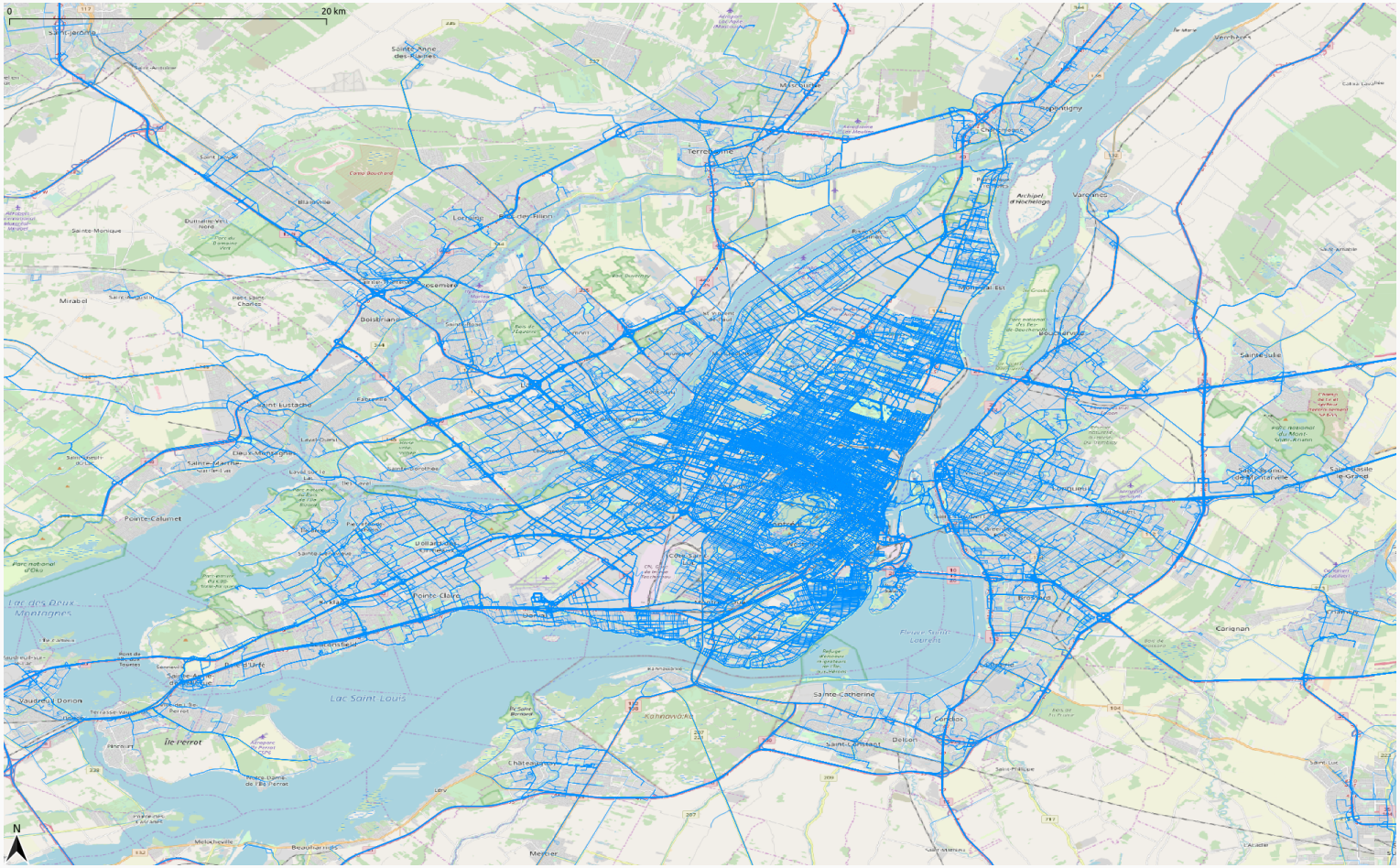
> *Examining one of the clusters in the map above revealed that only one of the 5 likely drop-off addresses was inhabited; a search for that address revealed its resident's name. In addition, by examining other drop-offs at this address, I found that this gentleman also frequented such establishments as "Rick's Cabaret" and "Flashdancers". Using websites like Spokeo and Facebook, I was also able to find out his property value, ethnicity, relationship status, court records and even a profile picture!*

# Et MTL Trajet?

*“L’ensemble représente les données recueillies des trajets effectués par les utilisateurs. Cependant, par souci de confidentialité les données sensibles des utilisateurs ont toutes été enlevées. Le début ainsi que la fin des trajets ont aussi été traités afin d’éliminer la possibilité de retrouver le domicile ou encore le lieu de travail des usagers. L’extrémité de chaque trajet (origine et destination) est tronquée à l’intersection empruntée la plus proche.”*

(Portail des données ouvertes de la Ville de Montréal)

# Exploration du nombre de trajectoires uniques dans MTL Trajet 2017




60436 déplacements dans la semaine du 2 octobre 2017

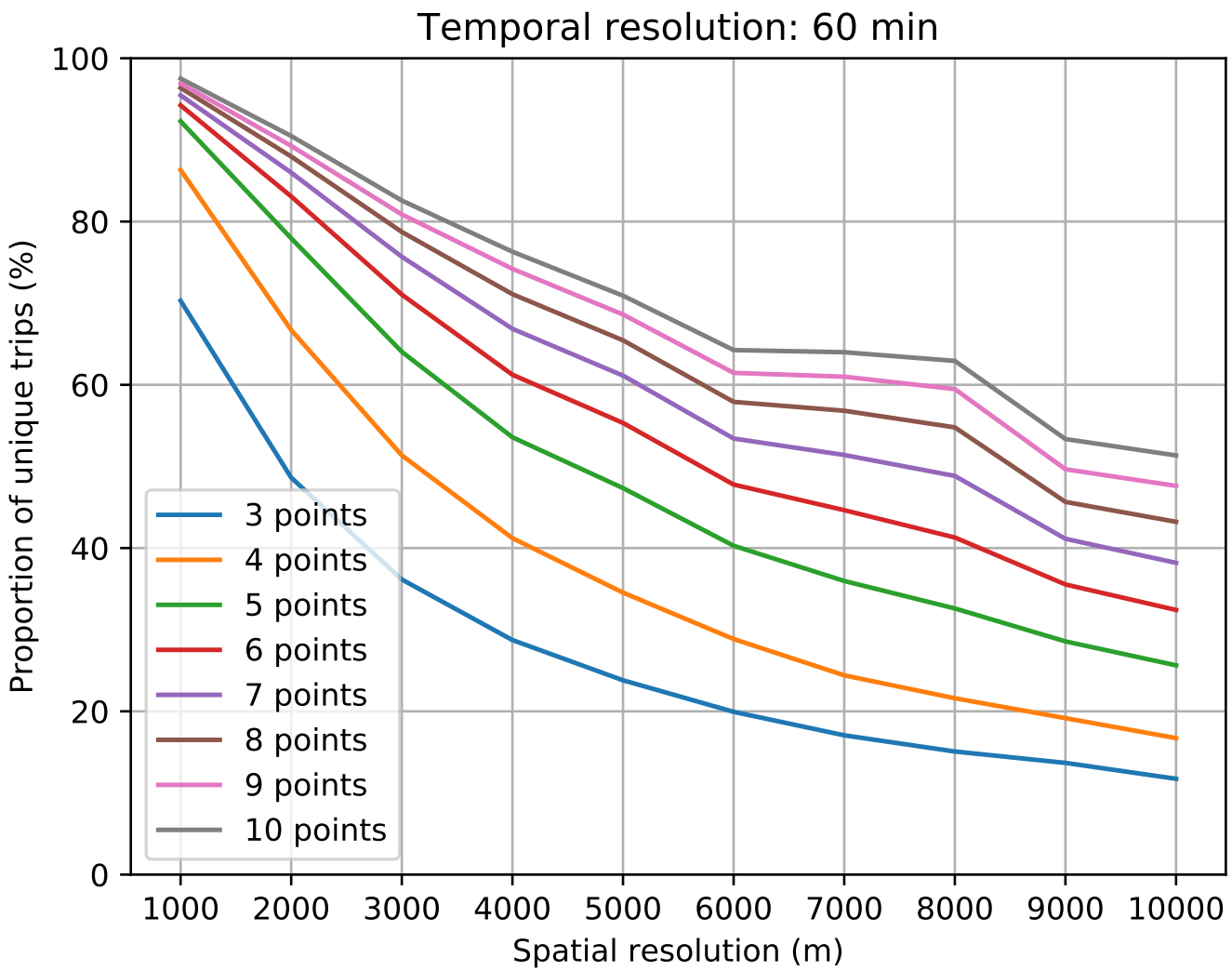
# Exploration du nombre de trajectoires uniques dans MTL Trajet 2017

- Chaque déplacement est composé de $n$ points $(x_i, y_i)$ et instants $t_i$
- On projette chaque point dans une grille de taille donnée et chaque instant dans des intervalles de temps de durée donnée
- Chaque déplacement est représenté par les points origines et destinations et un nombre $r - 2$ de points tirés au hasard, tous discrétisées spatialement et temporellement

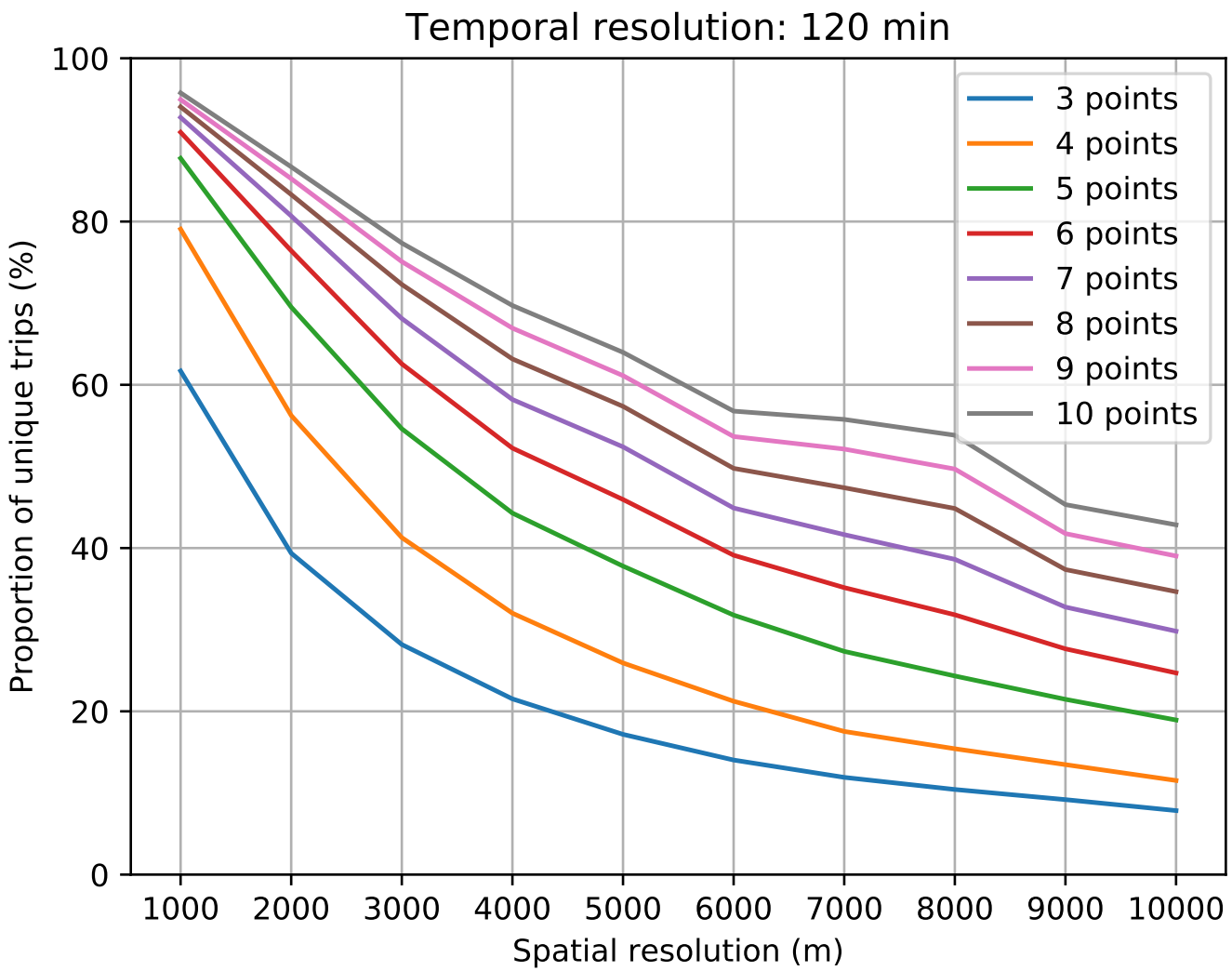
$$[(x_1^d, y_1^d, t_1^d), (x_2^d, y_2^d, t_2^d), ..., (x_r^d, y_r^d, t_r^d)]$$

- On calcule la proportion de déplacements dont la représentation discrétisée est unique

# Exploration du nombre de trajectoires uniques dans MTL Trajet 2017



# Exploration du nombre de trajectoires uniques dans MTL Trajet 2017



# Plan de la présentation

# Quelles solutions?

- Retirer les identifiants uniques et les données personnelles est insuffisant
- Discrétiser (agréger) les données est insuffisant
- Des traitements comme la k-anonymité sont vulnérables aux attaques
- Une solution existe: la confidentialité différentielle (Dwork et al. 2006)

# La confidentialité différentielle

- La confidentialité différentielle porte sur un mécanisme (algorithme) et non sur l'ensemble de données
- Le mécanisme a accès à un ensemble de données contenant des données personnelles et donne des réponses bruitées à des questions (requêtes)
    - le niveau de bruit est soigneusement choisi pour satisfaire une définition formelle de la confidentialité différentielle

# La confidentialité différentielle

Soit $\epsilon$ un réel positif et $\mathcal{A}$ un algorithme probabiliste qui prend pour entrée un ensemble de données. Soit $\mathrm{im}\mathcal{A}$ l'image de $\mathcal{A}$. L'algorithme $\mathcal{A}$ est dit $\epsilon$-différentiellement confidentiel, si, pour tous ensembles de données $D_1$ et $D_2$ qui diffèrent d'un seul élément (l'information à propos d'une seule personne) et pour tout sous-ensemble $S$ de $\mathrm{im}\mathcal{A}$,

$$\Pr[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \times \Pr[\mathcal{A}(D_2) \in S]$$

où la probabilité est fondée sur l'aléa introduit par l'algorithme

# La confidentialité différentielle

- Intuitivement, cela signifie que pour deux ensembles de données voisins, un algorithme différentiellement confidentiel se comportera à peu près de la même façon sur les deux ensembles
- La définition donne une garantie solide que la présence ou l'absence d'un individu dans l'ensemble n'affectera pas significativement la sortie finale de l'algorithme

# La confidentialité différentielle

Soit un ensemble de données $D_1$ (exemple Wikipedia)

| Nom | A visité Polytechnique Montréal (X) |
|---|---|
| Vincent | 1 |
| Agathe | 1 |
| Martin | 0 |
| Hélène | 0 |
| Sophie | 1 |

Supposons: un utilisateur malintentionné veut savoir si Sophie a visité Polytechnique et connait quelle ligne lui correspond
Cet utilisateur est uniquement autorisé à demander les données à travers une requête $F_i$ qui renvoie la somme partielle des $i$ premières lignes de la colonne X de l'ensemble de données

# La confidentialité différentielle

| Nom | A visité Polytechnique Montréal (X) |
|---|---|
| Vincent | 1 |
| Agathe | 1 |
| Martin | 0 |
| Hélène | 0 |
| Sophie | 1 |

Comment savoir si Sophie a visité Polytechnique?

# La confidentialité différentielle

| Nom | A visité Polytechnique Montréal (X) |
|---|---|
| Vincent | 1 |
| Agathe | 1 |
| Martin | 0 |
| Hélène | 0 |
| Sophie | 1 |

Comment savoir si Sophie a visité Polytechnique?

$$F_5(D_1) - F_4(D_1)$$

# La confidentialité différentielle

| Nom | A visité Polytechnique Montréal (X) |
|---|---|
| Vincent | 1 |
| Agathe | 1 |
| Martin | 0 |
| Hélène | 0 |
| Sophie | 0 |

Si on construit un second ensemble $D_2$ en remplaçant (Sophie, 1) par (Sophie, 0): si l'utilisateur recevait les valeurs $F_i$ via un algorithme $\epsilon$-différentiellement confidentiel, pour un $\epsilon$ suffisamment petit, alors il serait incapable de faire la différence entre $D_2$ et $D_1$

# Un mécanisme $\epsilon$-différentiellement confidentiel

- Ajouter au résultat de notre requête du bruit suivant la loi de Laplace

# Un mécanisme $\epsilon$-différentiellement confidentiel

- Ajouter au résultat de notre requête du bruit suivant la loi de Laplace
- Combien de bruit? Ca dépend

# Un mécanisme $\epsilon$-différentiellement confidentiel

- Ajouter au résultat de notre requête du bruit suivant la loi de Laplace
- Combien de bruit? Ca dépend
  - du paramètre $\epsilon$

# Un mécanisme $\epsilon$-différentiellement confidentiel

- Ajouter au résultat de notre requête du bruit suivant la loi de Laplace
- Combien de bruit? Ca dépend
  - du paramètre $\epsilon$
  - du risque pour l'individu le plus différent d'avoir ses renseignements personnels découverts, soit la sensibilité $\Delta f = \max_{D,D'} ||f(D) - f(D')||_1$, la différence maximale entre deux résultats de la requête $f$ pour deux ensembles de données qui diffèrent simplement d'un élément

# Un mécanisme $\epsilon$-différentiellement confidentiel

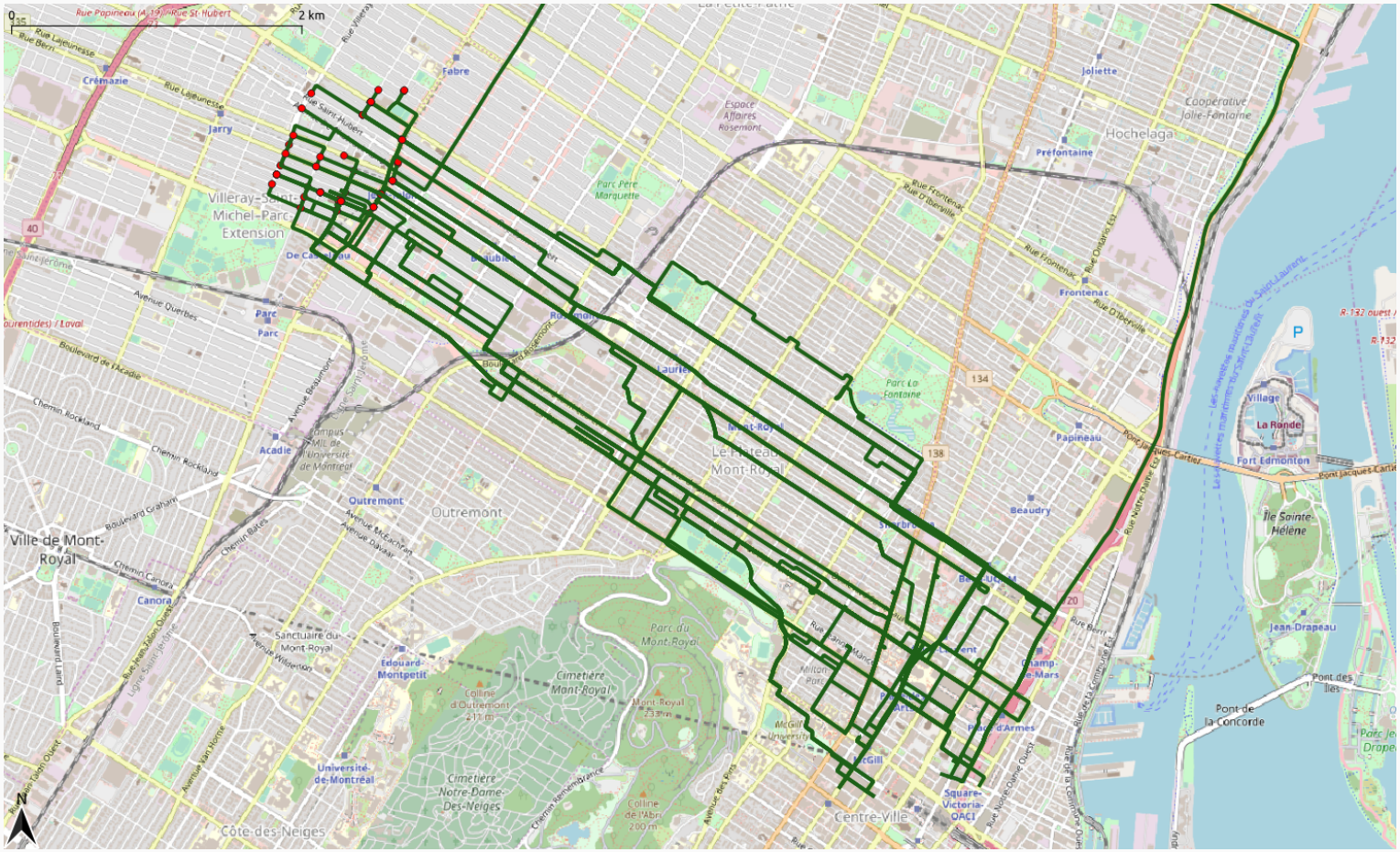
- Ajouter au résultat de notre requête du bruit suivant la loi de Laplace
- Combien de bruit? Ca dépend
  - du paramètre $\epsilon$
  - du risque pour l'individu le plus différent d'avoir ses renseignements personnels découverts, soit la sensibilité $\Delta f = \max_{D,D'} ||f(D) - f(D')||_1$, la différence maximale entre deux résultats de la requête $f$ pour deux ensembles de données qui diffèrent simplement d'un élément
- Ajout d'un bruit suivant la loi de Laplace de paramètre $\Delta f / \varepsilon$ (écart-type)

# Un mécanisme $\epsilon$-différentiellement confidentiel

- Ajouter au résultat de notre requête du bruit suivant la loi de Laplace
- Combien de bruit? Ca dépend
  - du paramètre $\epsilon$
  - du risque pour l'individu le plus différent d'avoir ses renseignements personnels découverts, soit la sensibilité $\Delta f = \max_{D,D'} ||f(D) - f(D')||_1$, la différence maximale entre deux résultats de la requête $f$ pour deux ensembles de données qui diffèrent simplement d'un élément
- Ajout d'un bruit suivant la loi de Laplace de paramètre $\Delta f / \varepsilon$ (écart-type)
- Si on autorise $n$ requêtes, il faut considérer un budget de confidentialité $\varepsilon_{total} = \sum_{i=1}^{n} \varepsilon_i$

# Exemple pratique: temps de parcours moyen

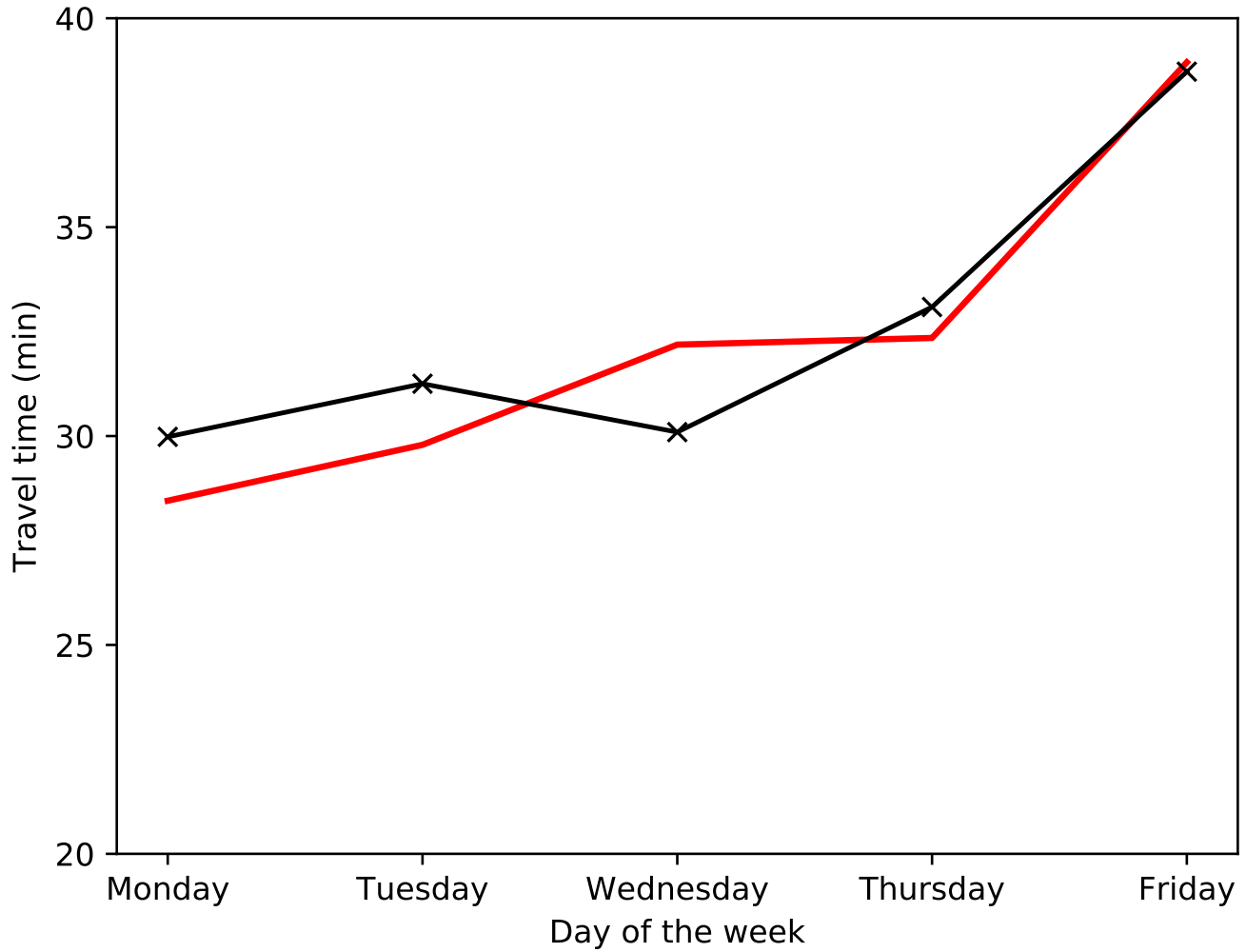
2 zones avec 72 déplacements relevés dans MTL Trajet 2017



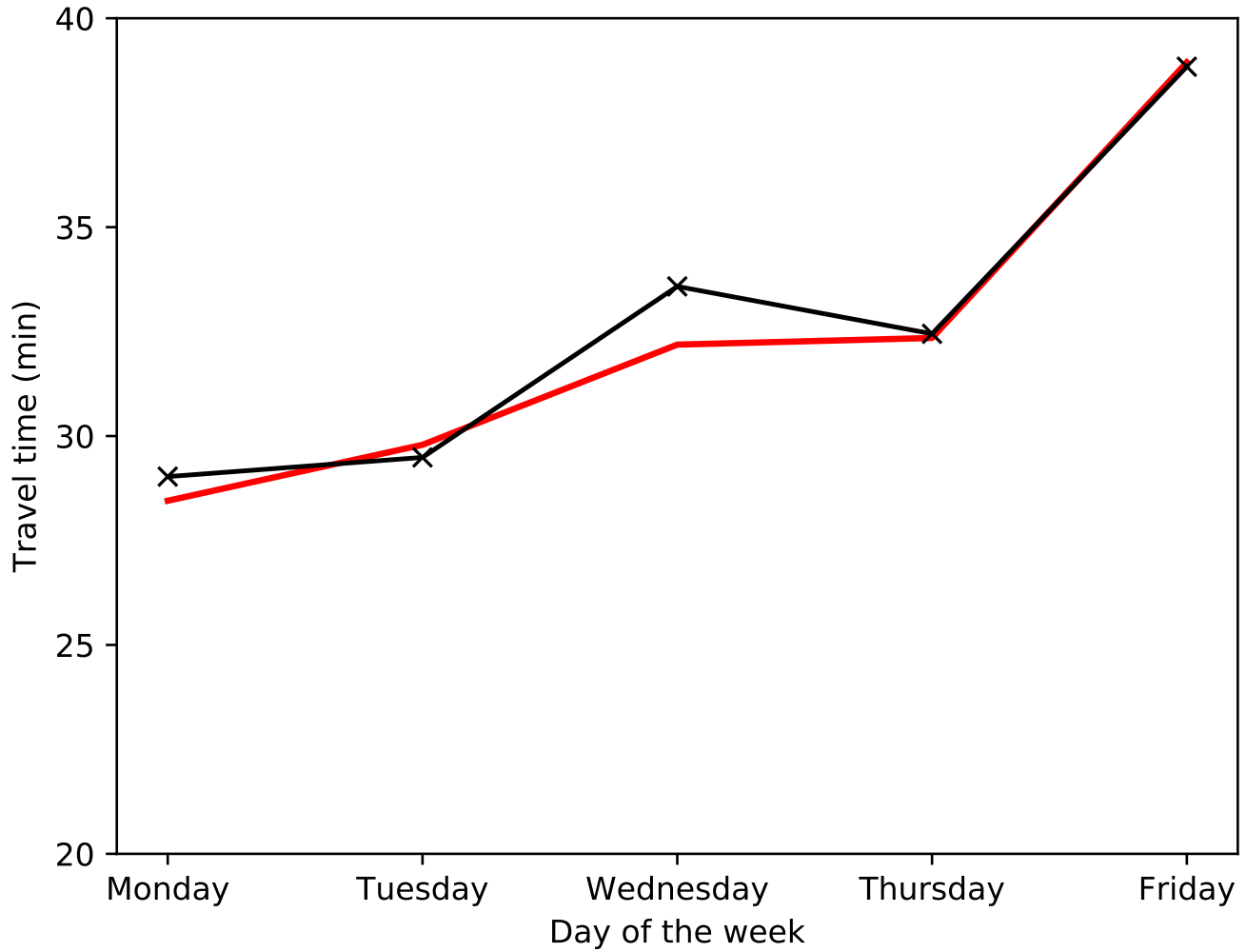
# Exemple pratique: temps de parcours moyen

- On voudrait que les résultats soient les mêmes pour deux ensembles de données dont on a modifié un temps de parcours de $\pm\rho/2$ min, par ex. $\pm$20 min
- On vise une précision (écart-type du bruit suivant la loi de Laplace) de 1 min
- Alors $\epsilon = 2.36$

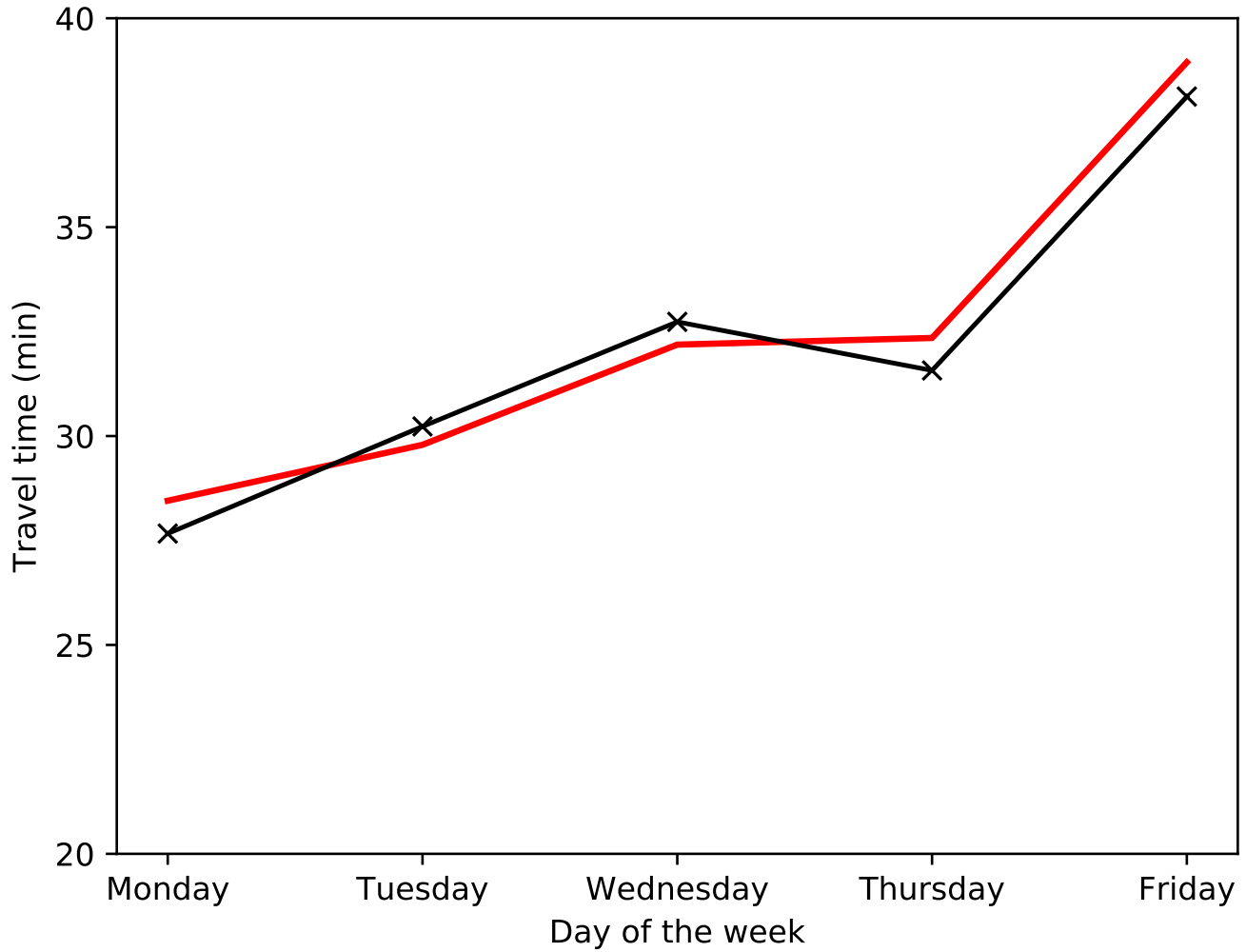
# Exemple pratique: temps de parcours moyen



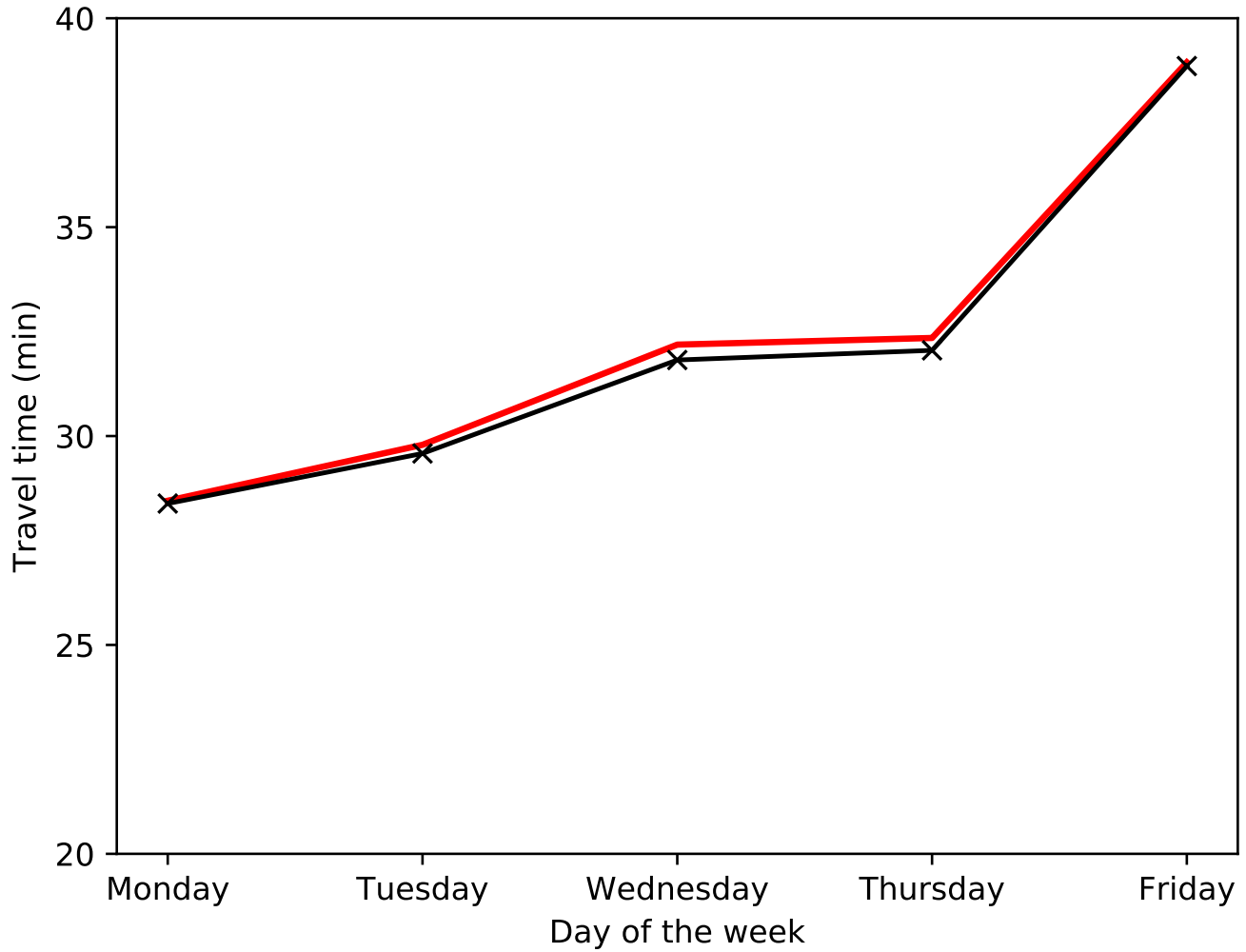
# Exemple pratique: temps de parcours moyen



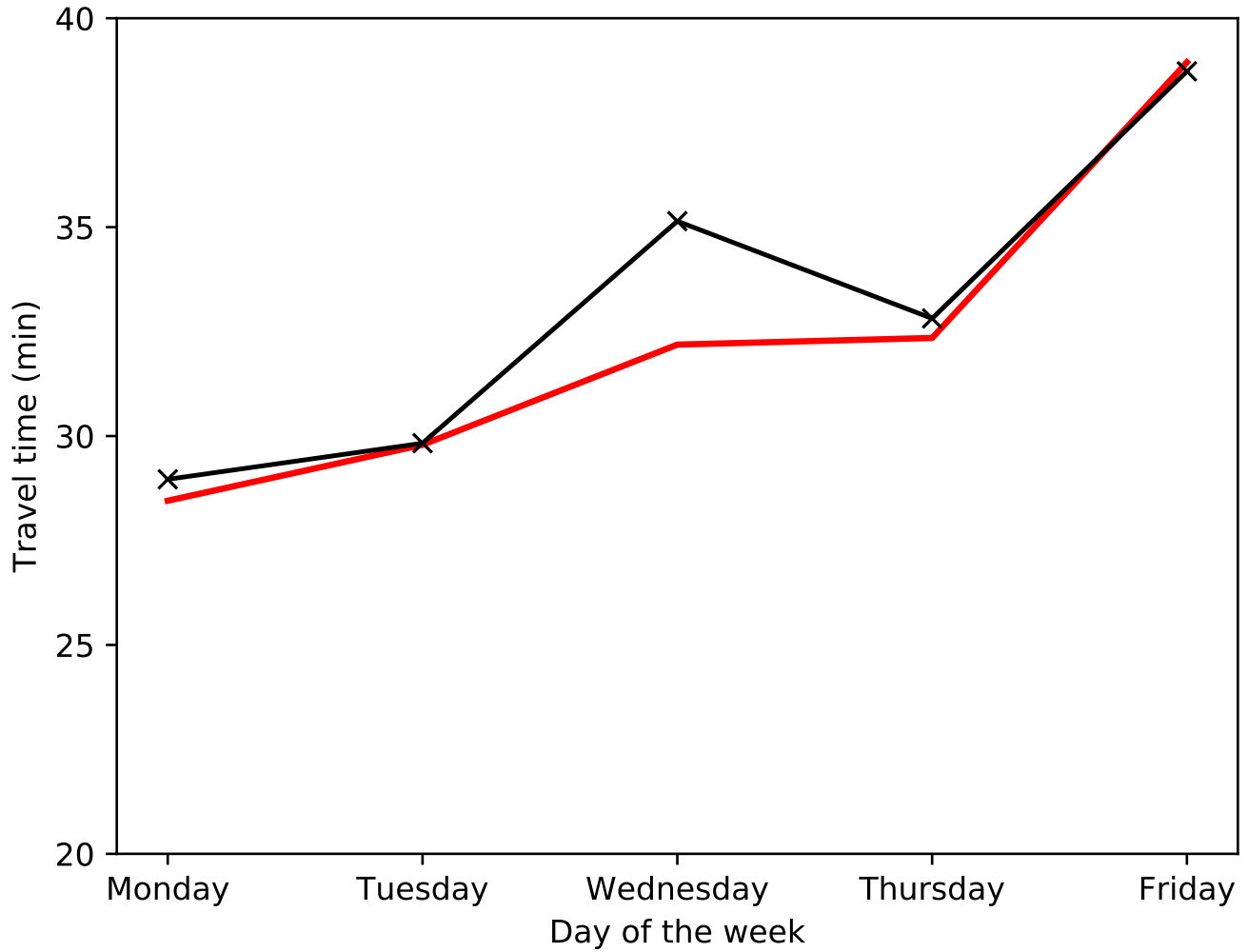
# Exemple pratique: temps de parcours moyen



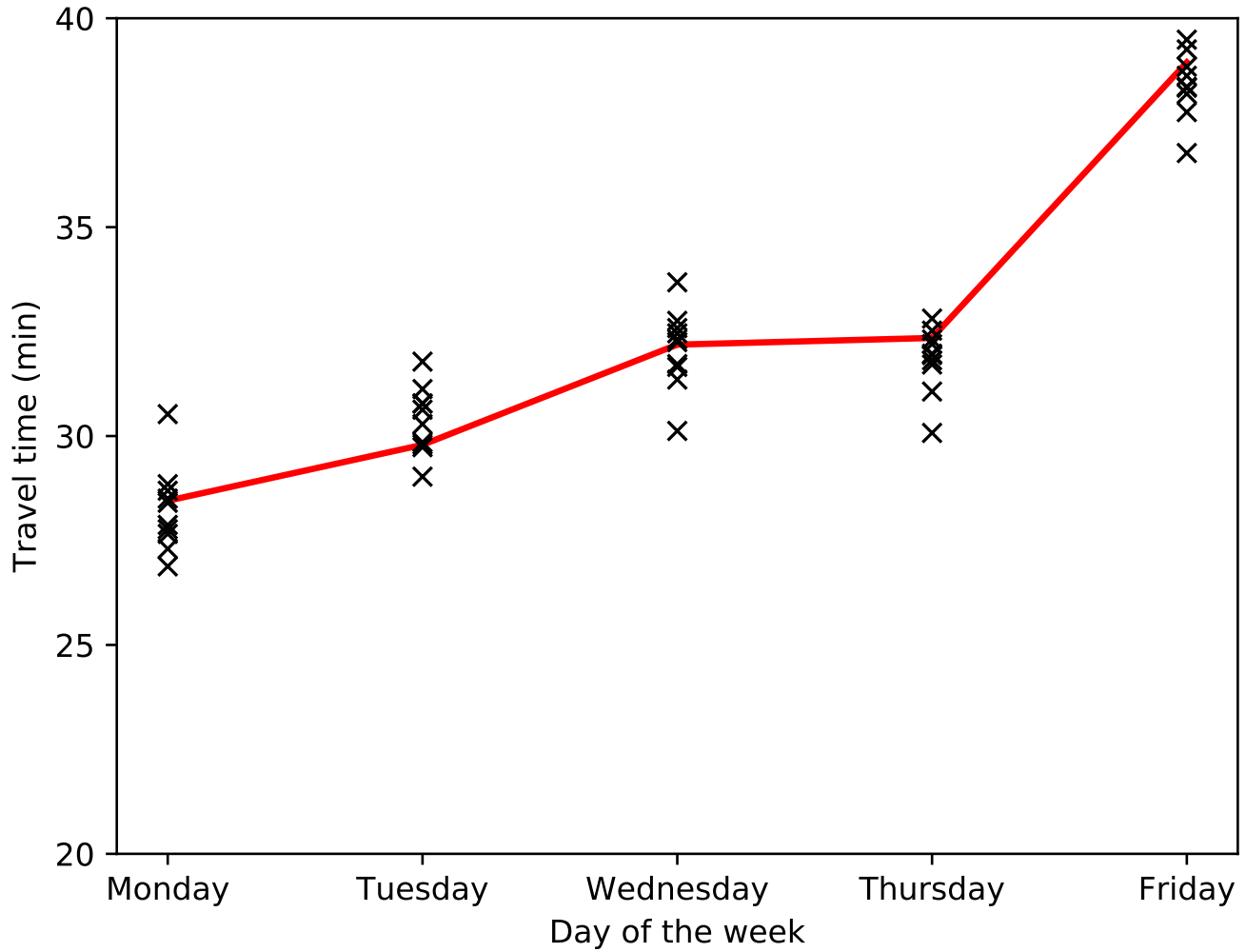
# Exemple pratique: temps de parcours moyen



# Exemple pratique: temps de parcours moyen



# Exemple pratique: temps de parcours moyen



# Plan de la présentation

# Conclusion

- Les données ouvertes sont une bonne chose

# Conclusion

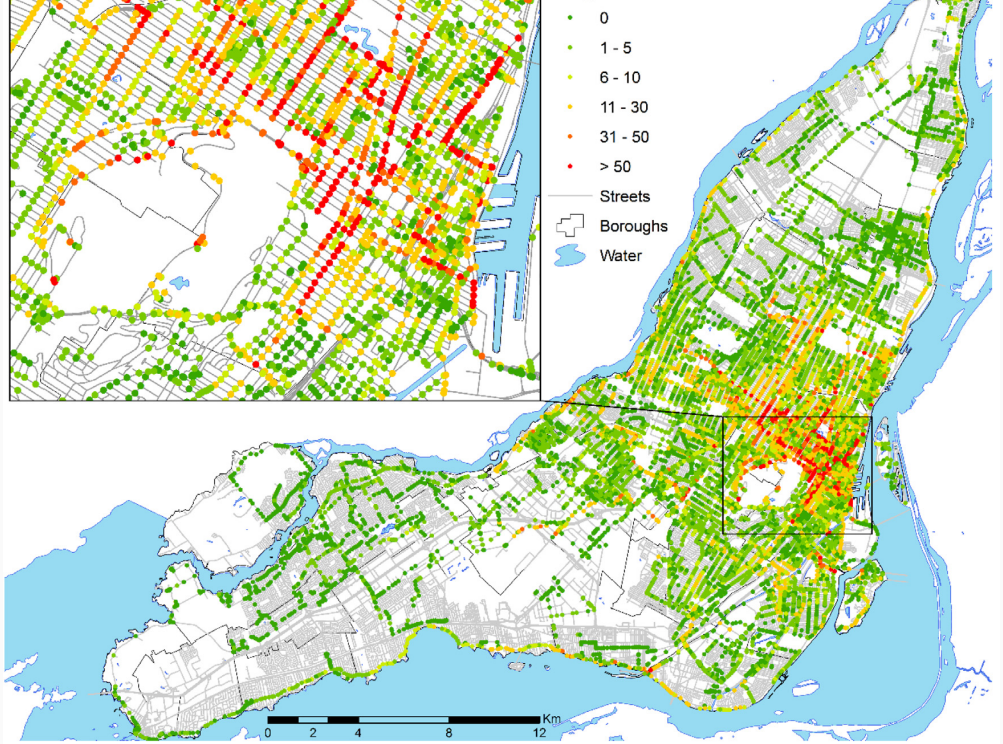
- Les données ouvertes sont une bonne chose
- Certaines données personnelles, en particulier les données de trajectoires, ne peuvent être disponibles publiquement sous forme brute

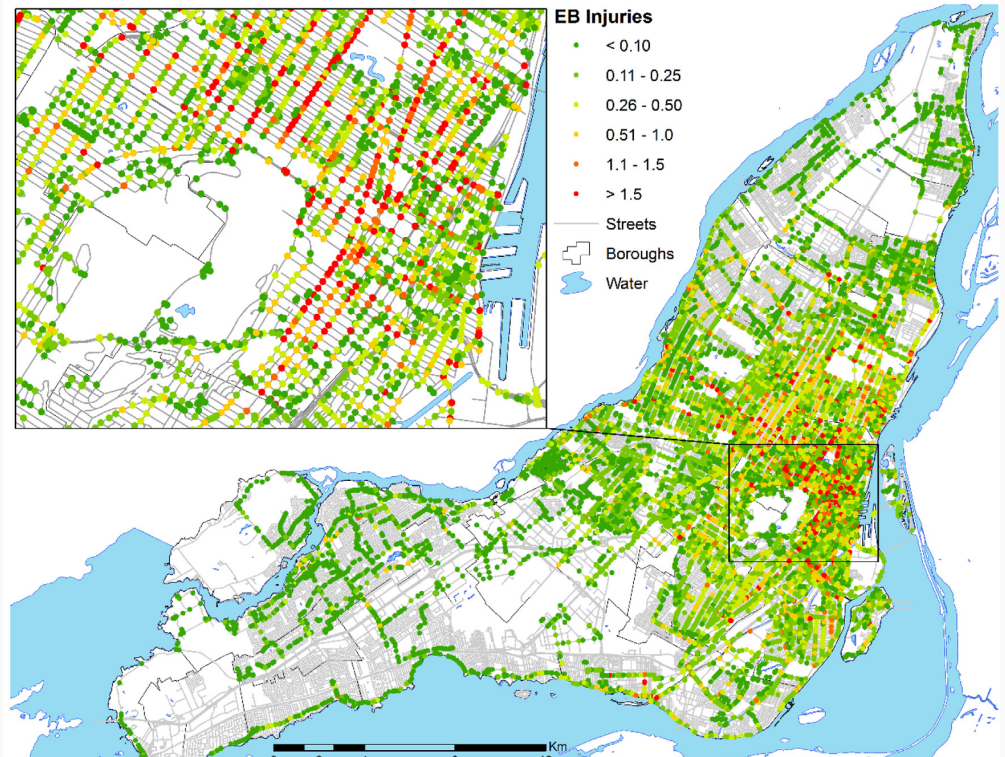
# Conclusion

- Les données ouvertes sont une bonne chose
- Certaines données personnelles, en particulier les données de trajectoires, ne peuvent être disponibles publiquement sous forme brute
- La solution est de créer une plateforme de requêtes $\epsilon$-différentiellement confidentielles

# Conclusion

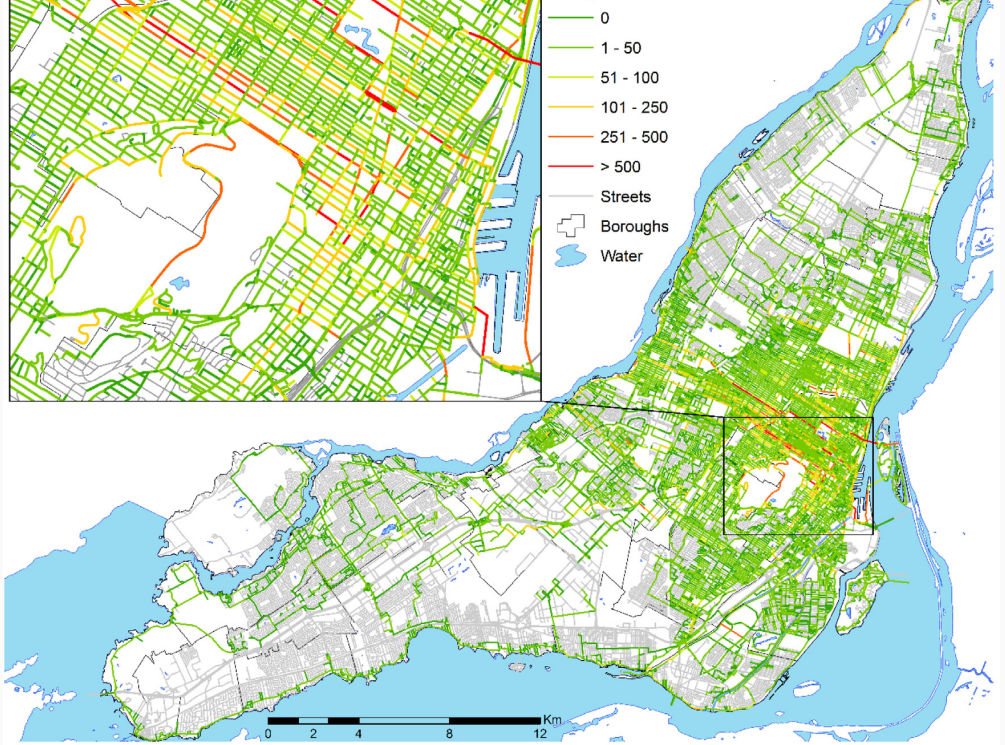
- Les données ouvertes sont une bonne chose
- Certaines données personnelles, en particulier les données de trajectoires, ne peuvent être disponibles publiquement sous forme brute
- La solution est de créer une plateforme de requêtes $\epsilon$-différentiellement confidentielles
    - applications au diagnostic de la circulation et de la sécurité routière à l'aide de données de véhicules sonde

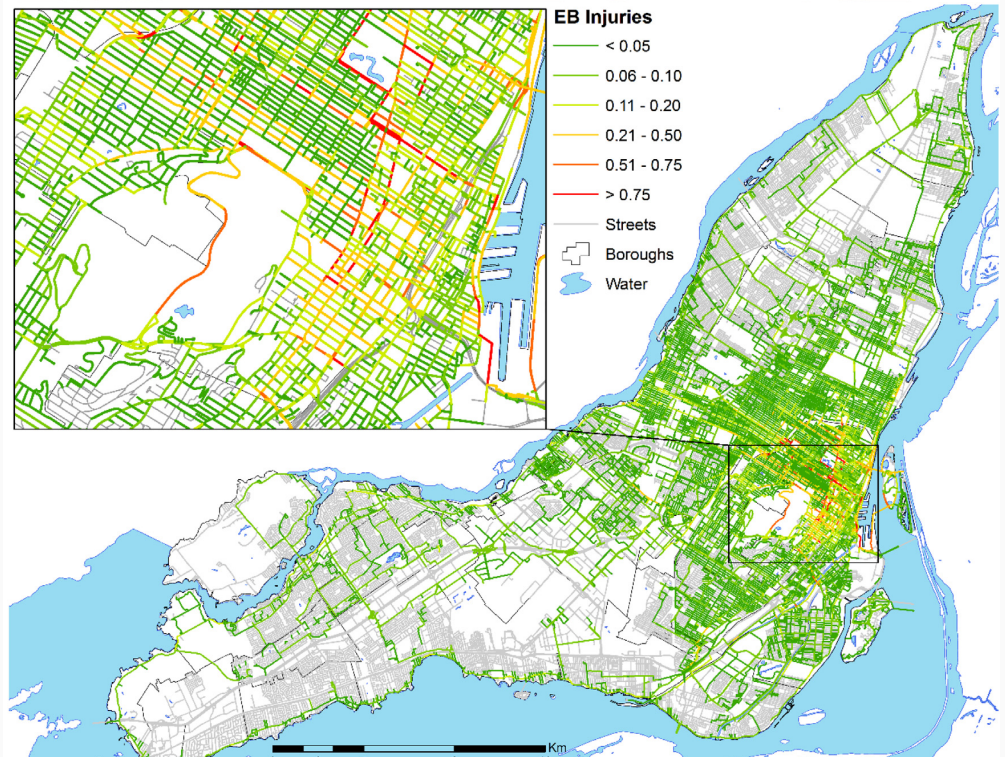



## Données cyclistes (sonde)

Corrélation du nombre de décélérations brusques avec l'estimé (méthode bayésienne empirique) du nombre de blessés et morts aux carrefours: 0.6 et 0.53 pour les carrefours avec et sans feux resp.




## Données cyclistes (sonde)

Corrélation de 0.57 du nombre de décélérations brusque avec l'estimé (méthode bayésienne empirique) du nombre de blessés et morts sur les segments

# Questions?