TRCUZTURE GOURPE 2

# INTRO

### SALUT SALUT

Cet été, j’ai eu le plaisir d’aller a New York. Un très joli voyage. Une très jolie statue. De très beaux taxis ? Moi, je ne savais pas avant d’aller, mais saviez-vous que New York comporte plus de 13'500 taxis ? C’est LE moyen de transport par defaut pour une grande partie de la population et des touristes.

### PROBLEME

Mais il est assez difficile d’estimer les temps de trajet, la congestion, surtout pour quelqu’un qui ne vient pas du coin. N’aurais-je pas meilleur temps d’y aller a pied ? Et c’est a cette question que nous avons essayé de répondre avec « Cab ou pas Cab ».

### RESSOURCES

Nous disposons d’une grande base de données des trajets fait par des taxis new yorkais de 2010 jusqu’à aujourd’hui, et nous avons essayer d’en faire un estimateur du temps de trajet en taxis entre deux points a une heure et date données.

# CHEMINEMENT

### SOLUTION INITIALE

Nous avons commencé par regarder a quel point il était faisable de juste regarder dans notre immense base de données les trajets les plus proches de ce qu’on voulait estimer. Si l’année passée, une personne a fait quasiment le même trajet a quelques metres prêt, on peut supposer que les temps de trajets seront similaires. Mais la base de donnée est tout bonnement gargantuesque et les queries prenaient beaucoup de temps. Or on ne peut pas laisser l’utilisateur attendre 5 minutes que l’estimation se fasse.

### SOLUTION 2

Nous nous sommes ensuite mis en tete de modéliser la relation entre nos données d’entrée et le temps de trajet. Sans rentrer dans les details, quelques visualisations sur des extraits de données nous avait laisser penser que le probleme serait relativement simple a modéliser, mais les résultats étaient assez decevants.

### SOLUTION FINALE

Nous avons donc décidé de faire un compromis entre nos deux idées. Knn est un algorithme qui implemente exactement notre idée de depart, on regarde dans un set de données pour les samples les plus proches de nos données de prediction et on moyenne leurs sorties. Mais au lieu de prendre toutes les données, nous avons pris des samples aleatoires dans la base de donnée et fait des tests pendant un certain temps, jusqu’à ce que la precision se stabilise – jusqu’à ce que notre mini-dataset soit representatif du dataset complet.

# CONCLU

### DEMO

Intro L

API google C

Cheminement algo S

Visualisation carte L

Presentation du code complet C

Conclusion S

Demo C