

**Etudiant :** [M2 MIAGE SID] BERTRAND Guillaume

# Rapport projet Big Data – crimes à Chicago

M2 MIAGE SID

<b>Accès aux sources</b>	<b>3</b>
<b>Accès aux résultats</b>	<b>3</b>
<b>Analyse des résultats</b>	<b>3</b>
Q1	3
Q2	4
Q3	4
Q4	4
Q5	4
<b>Performances</b>	<b>4</b>
Temps de réponses (local - pseudo distribué)	4
Temps	4
Temps	5
Explication	5
Temps	5
Explication	5
Temps	5
Explication	5
Temps	5
Explication	5
<b>Conclusion</b>	<b>6</b>

Modalités de test :

- **tout le dataset en mode pseudo-distribué** - - en local
- utilisation d' Apache **Spark**

**Pourquoi tester tout le dataset en local ?**

- éviter le risque d'erreur sur le cluster **en cas de situations non prévues sur l'échantillon**
- bien entendu on ne procède pas ainsi en situation réelle

## Accès aux sources

github:

<https://github.com/fubrasp/projet-big-data>

## Accès aux résultats

<http://www.bertrandguillaume.fr/q1.html>

OU (cf. repository github)

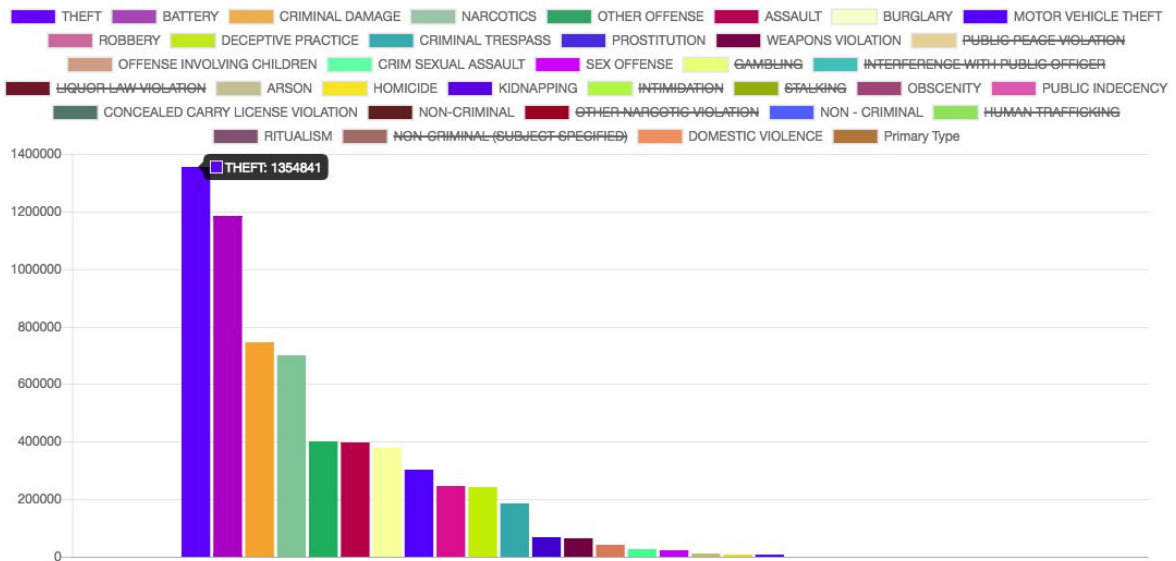
en faisant

1. la commande **npm install**
2. en **lançant le programme Main en scala** (penser à supprimer les dossier générés)
3. puis **en générant les pages de résultats** en **allant dans le dossier representations**, en tapant la commande suivante: **php PageConstructor.php**
4. puis en ouvrant avec un serveur (open in browser depuis intellij par exemple) les pages générées

# Analyse des résultats

Q1

## Question 1 - Donnez le classement décroissant des catégories de crimes



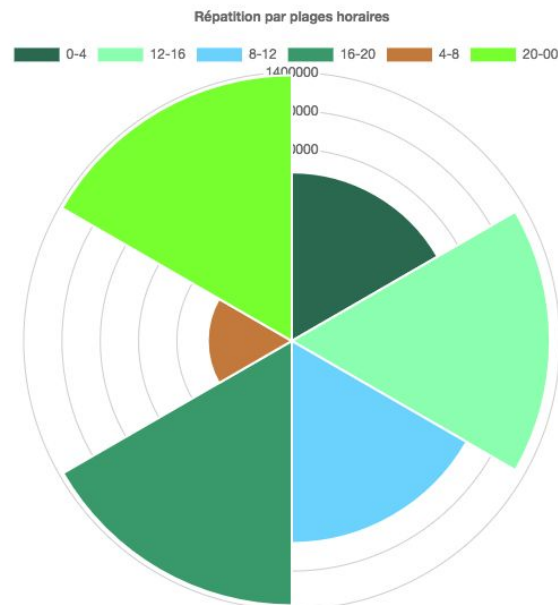
Projet M2SID 2018 - BERTRAND Guillaume

Les **vol**s violents ou non sont la raison principale des crimes, suivi par notamment :

- les **attouchements**
- les **dégradations**
- les **drogues**

## Q2

Question 2 - Donnez le nombre de crimes en fonction de 6 plages horaires

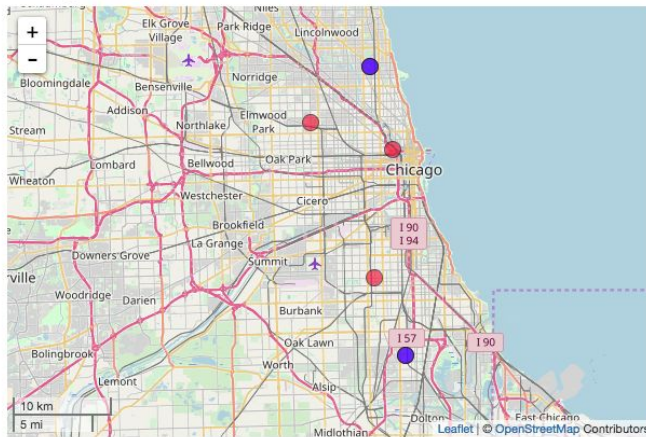


Projet M2SID 2018 - BERTRAND Guillaume

Les **crimes** ont lieu plutôt en **fin de matinée**, **après-midi** et **soirée**.

## Q3

### Question 3 - Donnez les 3 zones les plus dangereuses et les zones les moins dangereuses



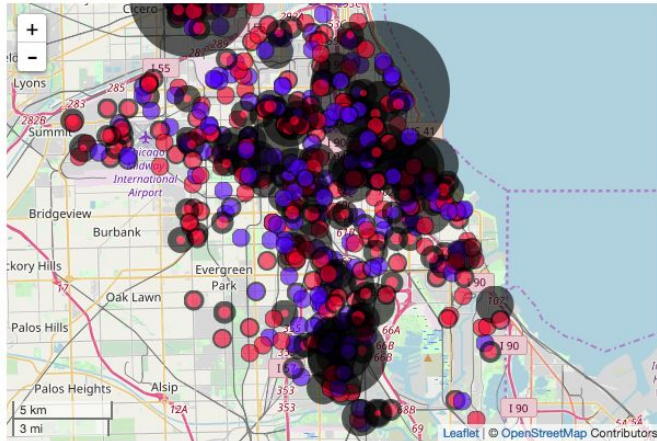
Projet M2SID 2018 - BERTRAND Guillaume

Une zone aberrante est présente, nous n'en tiendrons pas compte.

Il est relativement **difficile de tirer des conclusions.**

## Q4

### Question 4 - Donnez la répartition géographique des crimes commis/élucidé



Le noir symbolise l'incidence des crimes sur un point similaire.



On constate une **quantité de crime invressemblable** dont de **nombreux non élucidé**, comme en France on peut se douter que les vols correspondent à la plupart des crimes non élucidés.

Sans surprise, il apparaît de nombreux crimes aux mêmes endroits:

- **quartiers défavorisés**
- **zones touristiques**

**Et ce de manière récurrente** parfois jusqu'à 112 fois au même endroit !

## Q5

Question 5 - Donnez le top 3 des mois les plus concernés par les cas de crime



Projet M2SID 2018 - BERTRAND Guillaume

Il apparaît relativement **difficile d'établir un mois avec plus de criminalité** qu'un autre **les différences entre les 3 premiers mois sont très minimes.**

## Performances

Temps de réponses (local - pseudo distribué)

Temps

Q1 - 15939 ms, 16s

Temps

Q2 - 58043 ms, 58s

Explication

Ce temps est améliorable avec un dispatch par tableau pour les plages horaires



## Temps

Q3 - 148370 ms, 148s, 2min 28s

## Explication

Le process des Kmeans est relativement lourd

de plus pour rappelle le process entier pour la question 3

- on détermine le cluster de chaque donnée
- on compte le nombre de données pour chaque cluster
- on part du postulat où une zone dangereuse l'est par le nombre de crimes (et non par la gravité : meurtres etc)
- on prend respectivement les 3 clusters les plus dangereux et les moins dangereux.
- on trace autour des centroids une zone de 2km

## Temps

Q4 - 172699 ms, 173s, 2min 53s

## Explication

- dataset entier a manipuler
- à grouper (on compte le nombre de crimes au même endroit, cela évite les doublons de points non visibles sur la carte et cela permet de quantifier réellement les crimes)

## Temps

Q5 - 183038, 183s, 3min 3s

## Explication

Pas de remarques supplémentaires à ajouter.

## Conclusion

Ce projet m'a permis de me familiariser à la problématique d'optimisation d'algorithme et son impact dans le cadre de traitements "Big Data". La finalité de celui-ci en elle-même est particulièrement intéressante.