Données et traitement statistique

Christian Delfosse

Introduction (1/2)

Données et traitement = statistique

Branche des mathématiques

Eléments moteurs :

- démographie (17e)
- probabilités (17e : Pascal, Fermat, 18e :Bernoulli)
- industrie et agriculture (19e) puis économie (20e)

Introduction (2/2)

Changements récents :

grandes masses de données
(databases, Tb size)

outils de traitement automatique
(R,SPSS,Hadoop)

maths associées (analyse des correspondances,
en composantes principales)

Types de données

- Catégorielles (issues de comptages (M/F))
- Ordinales (idem, mais catégories ordonnées (réussites par grade))
- Mesures (nb réels, au moins concept. (T° en degrés))
- Mesures absolues (échelle avec zéro (âge, taille))

- Transformations (avec perte de précision,
  mais gain de concision (regroupements))

Fondamental

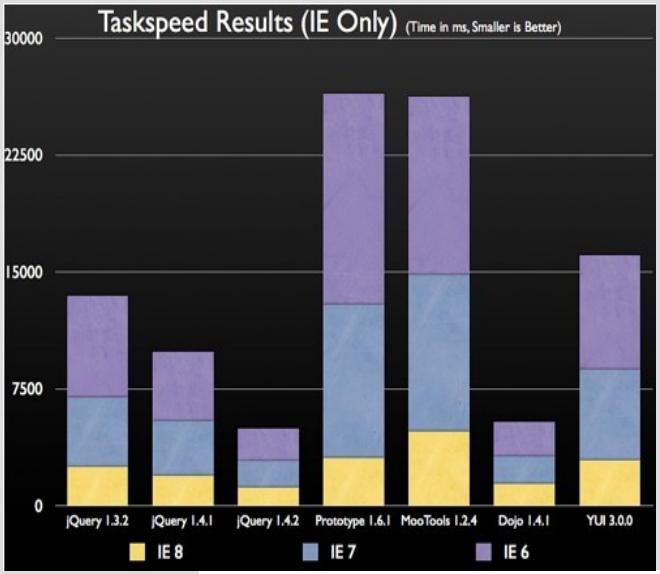
Visualisation = Message

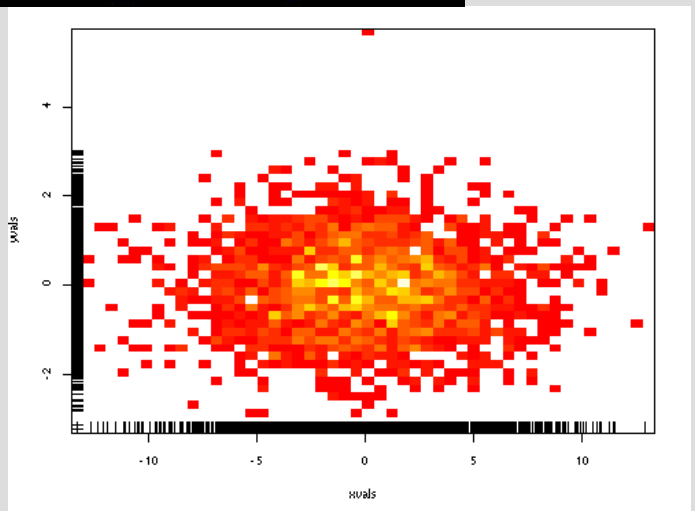
Exigences : honnêteté intellectuelle
sens critique
curiosité de l'existant

Créativité graphique

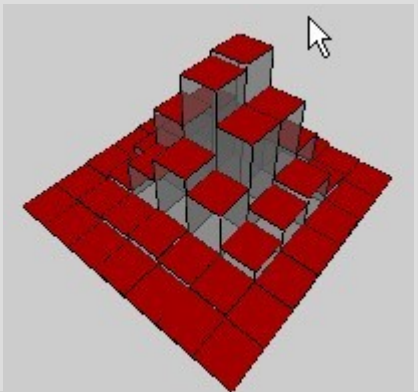
NB : tous les exemples qui suivent se trouvent à https://github.com/mbostock/d3/wiki/Gallery

Histogramme

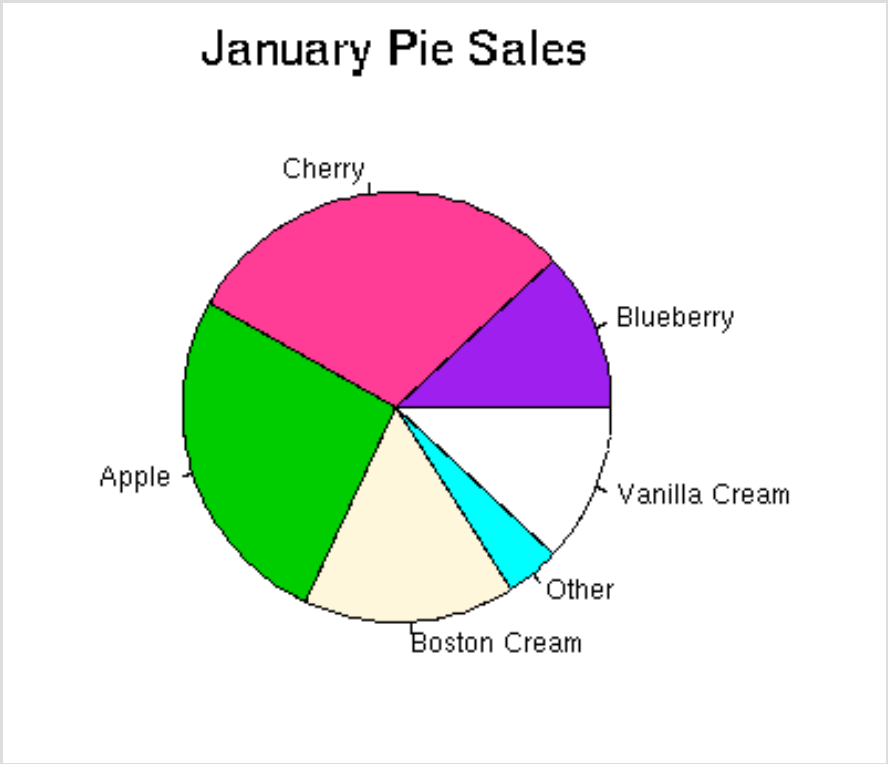




- Bien connu
- Généralement bien compris
- Facile à faire (tableur, Highcharts)
- D3.js : Showreel
- Aussi 2D et 3D



Camemberts (pie-charts)



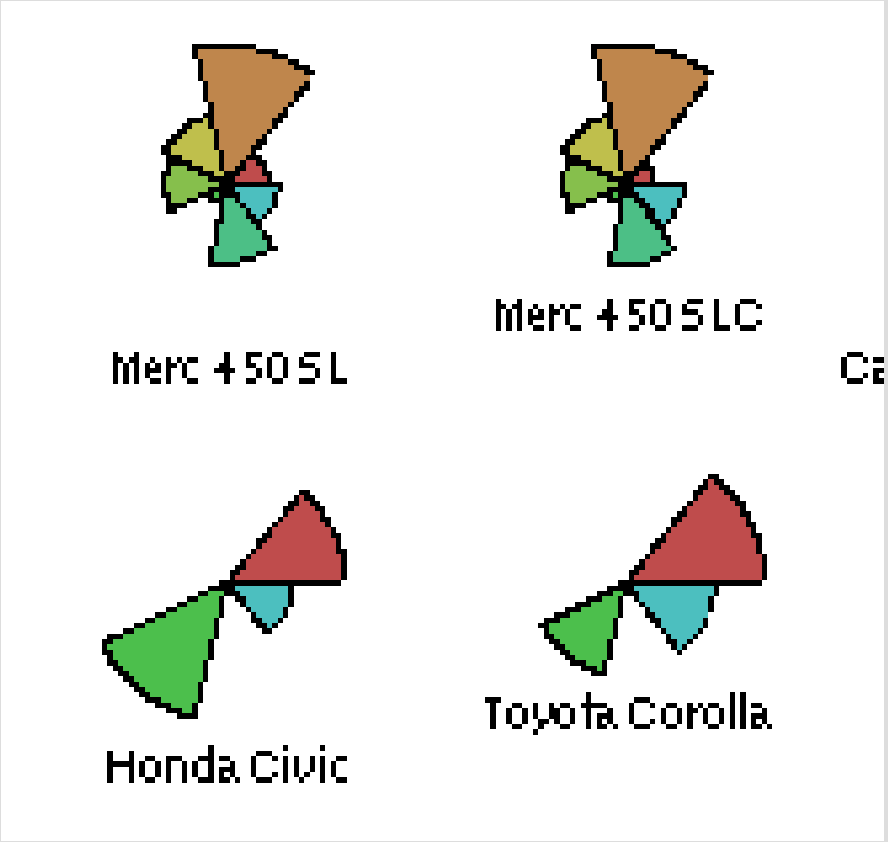
Assez bien connu

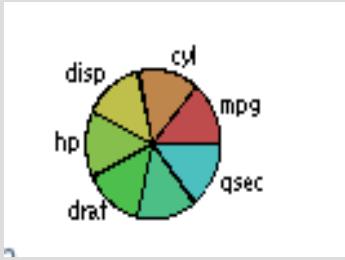
Facile à faire
(tableur, Highcharts)

Variantes :
sortir quartier

D3.js : Interactive
Sales Data Pie Chart

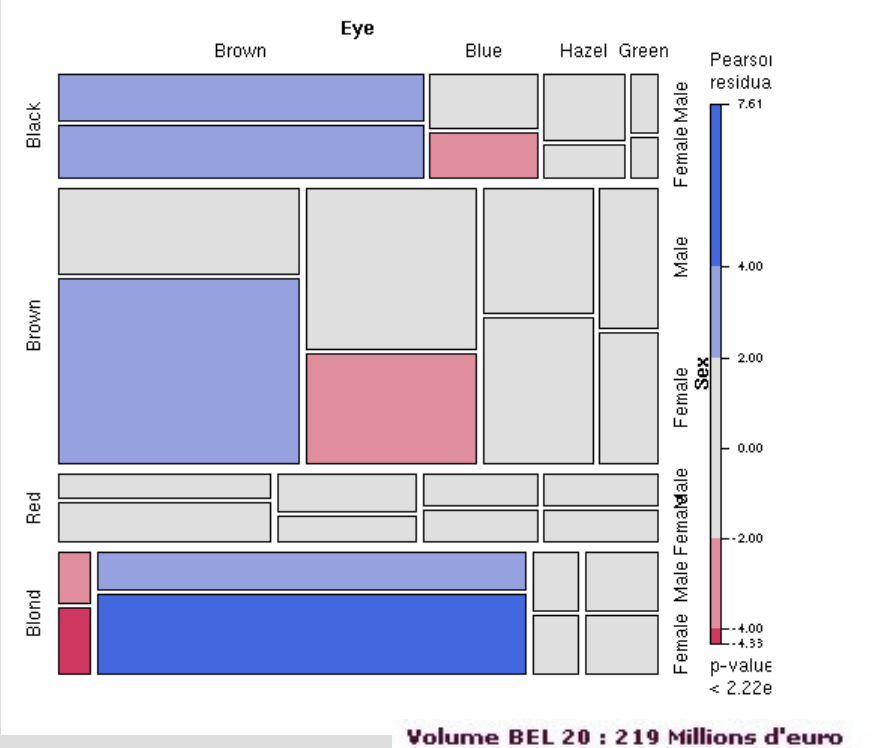
Etoiles





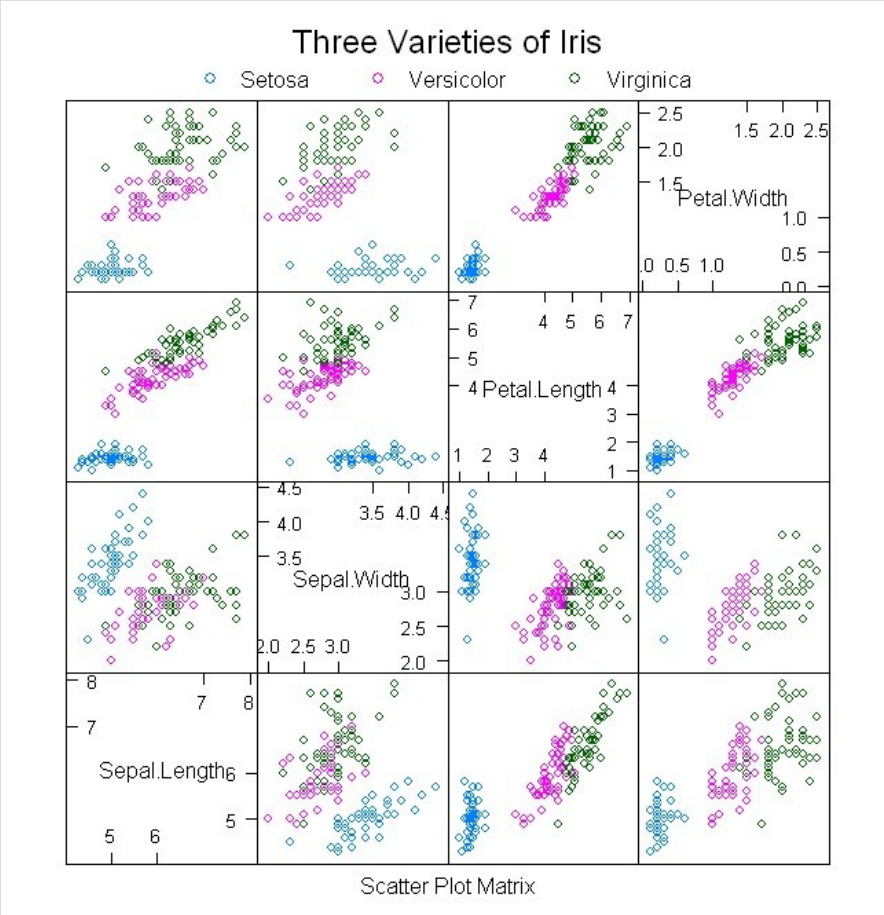
- Camemberts pour données multivariées
- Besoin outil plus avancé (R)
- D3.js : Wind History

Mosaiques



- Peu connu
- Répartition de critères
- Besoin outil plus avancé (R)
- D3.js : Zoomable Treemap

Matrice d'association

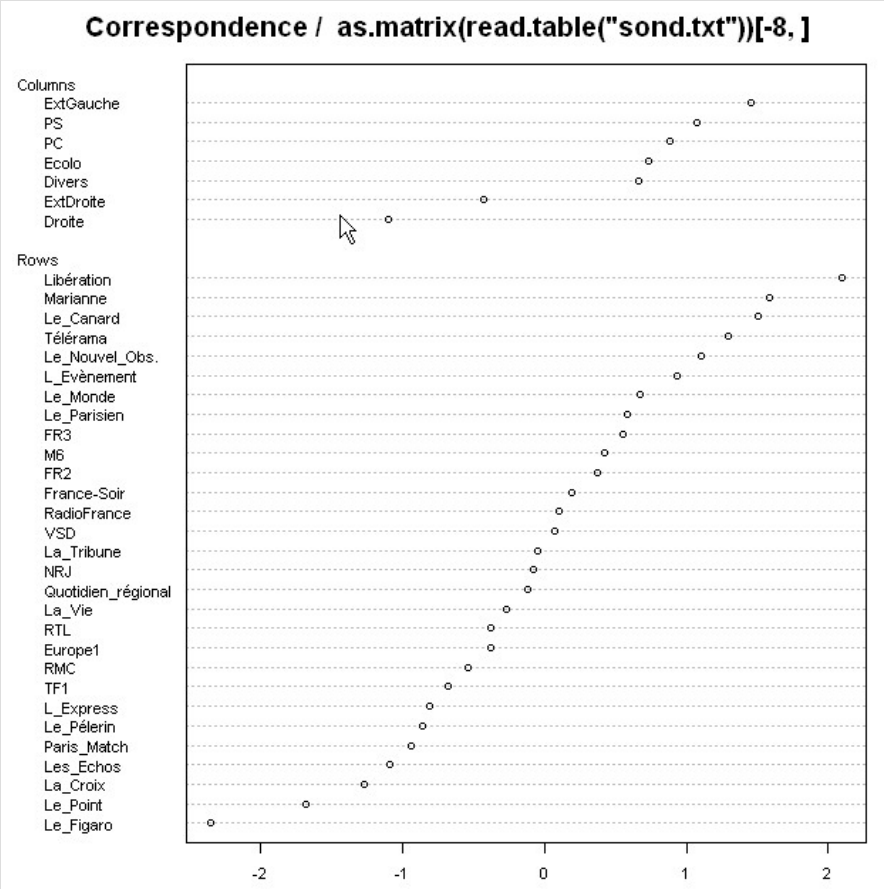


Données : Iris

Dispersion selon mesures prises par paires

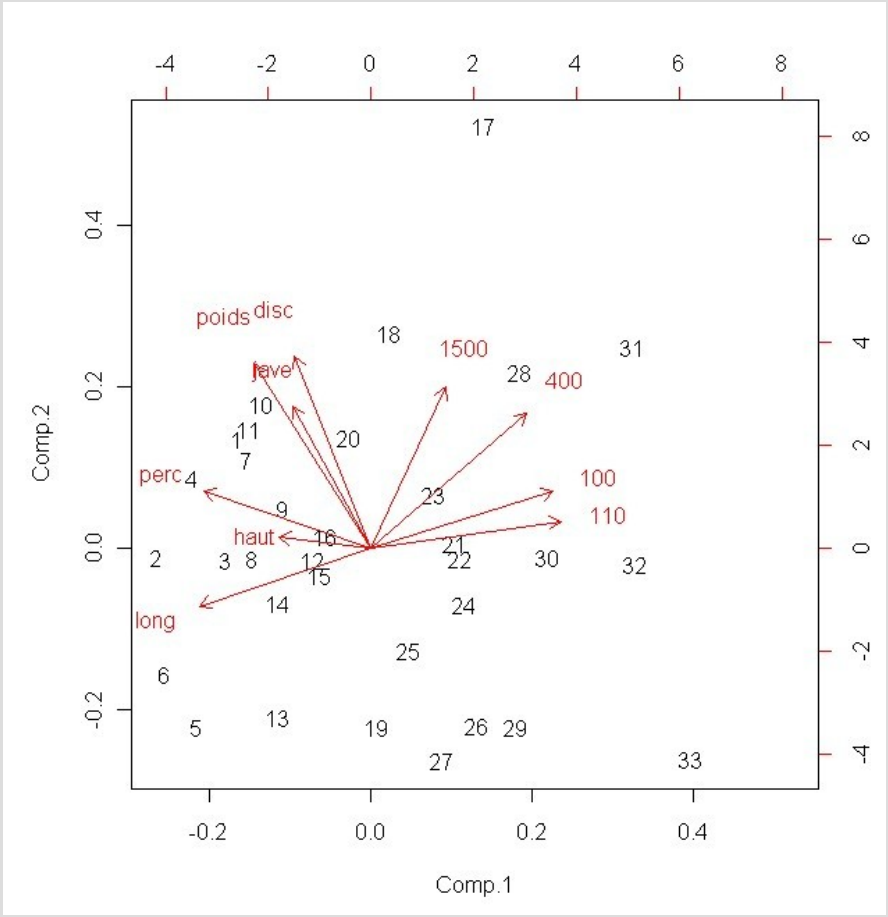
D3.js : Scatterplot Matrix

Analyse des correspondances



- Données catégorielles
- Commentaire
  - données (sortie des urnes, source d'info)
  - graphique
- Besoin de
  - pre-processing par un outil statistique

Analyse en composantes principales

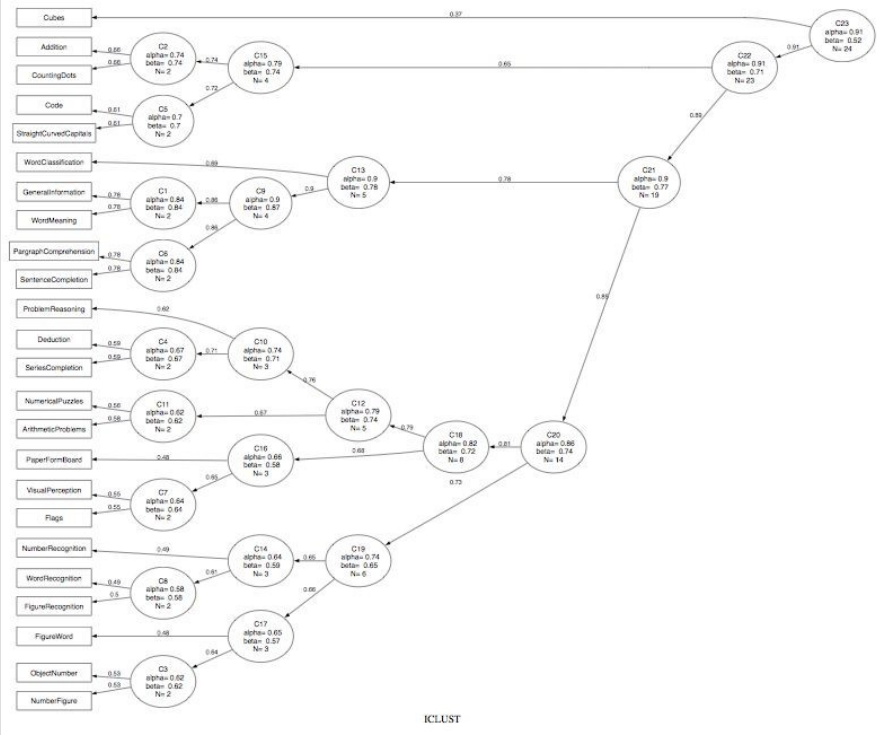


Données mesurées

Commentaires
- données (décathlon)
- graphique

Besoin de
- pre-processing par un outil statistique

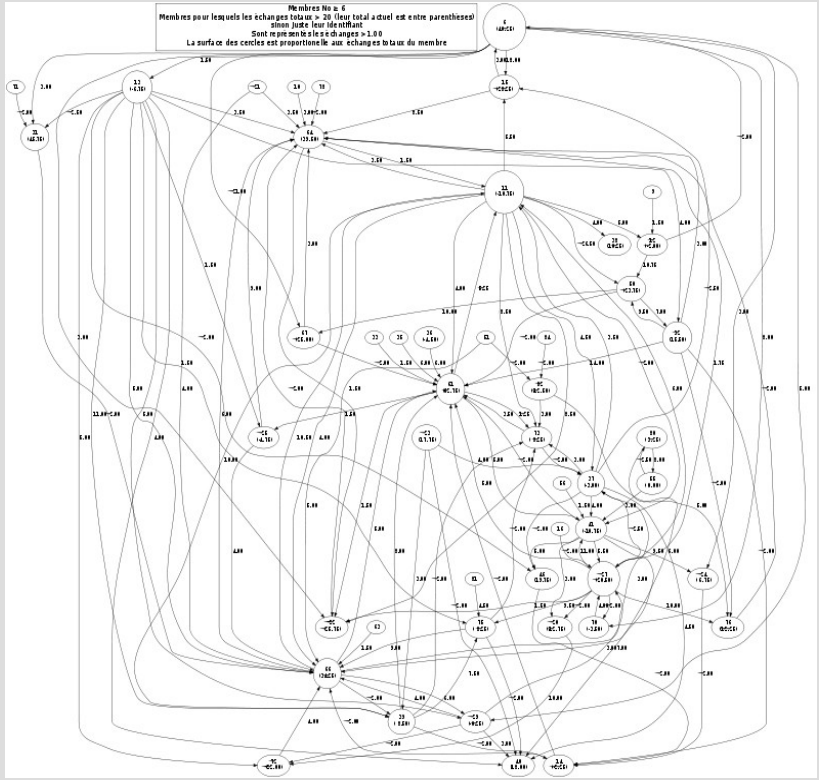
Clustering



- Le partitionnement de données
- = regroupement
- Besoin outil plus avancé (R)
- Ne pas confondre avec un affichage d'arborescence (D3.js : Cluster Dendrogram)

Graphes de liaisons

- http://www.graphviz.org/
- Génération automatique
- Syntaxe simple
- Peut générer pdf, png mais aussi map(html)



Références

http://www.r-project.org/ (R)

http://addictedtor.free.fr/graphiques/thumbs.php

(graphiques dans R)

http://www.infoq.com/presentations/Distributed-Data-Analysis-with-Hadoop-and-R

(Hadoop, R)

https://dl.dropboxusercontent.com/u/20439275/CD_Pres_2014.pdf

https://github.com/mbostock/d3/wiki/Gallery (exemples de D3.js)

Application Twitter

A fait l'objet du dernier cours avant le congé.

On peut adresser les requêtes à :
http://christian-delfosse.infographie-heaj.eu/Test/TestTwitter.html?query=...

Les requêtes déjà effectuées sont disponibles sous :
http://christian-delfosse.infographie-heaj.eu/Test/AllQueries.html?database=twitterResults

Dans les deux cas, voir la console.

Application Mobilité BXL

Voir https://github.com/dwmaj/Data-visualisation-workshop

(répertoire jsonTOjsonp)

On peut avoir en temps réel des données sur les embarras de circulation à Bxl à

http://christian-delfosse.infographie-heaj.eu/Test/Brussels.html

On peut avoir les données depuis vendredi midi à

http://christian-delfosse.infographie-heaj.eu/Test/AllQueries.html?database=mobility

Dans les deux cas, voir la console.

Pour la visualisation (github répertoire d3_js#json#update#filtering):

http://christian-delfosse.infographie-heaj.eu/Test/showJams.html