Présentation data.table

Meetup R Nantes

9 janvier 2017

Introduction

Philippe Chataignon

- *statisticien* qui s'est toujours intéressé aux outils *informatiques* permettant de traiter les données (*data-scientist*)
- dans mon métier, l'outil de référence est `SAS`
- à titre personnel, j'ai également beaucoup programmé en `Python`
- depuis 2 ans, j'utilise `R` dans mon cadre professionnel, en particulier les packages d'Hadley Wickam (*dplyr*, *ggplot2*, *tidyr*...) et `data.table`

Le métier de statisticien

- pour un statisticien, l'ordinateur est l'*outil* qui lui permet de réaliser les manipulations de données qu'il a en tête
- une qualité d'un langage pour *statisticien* est d'être *concis*, *facile* à mettre en oeuvre, quelquefois au détriment de la *lisibilité* ou de la *maintenabilité*
- différent de la vision d'un *informaticien*

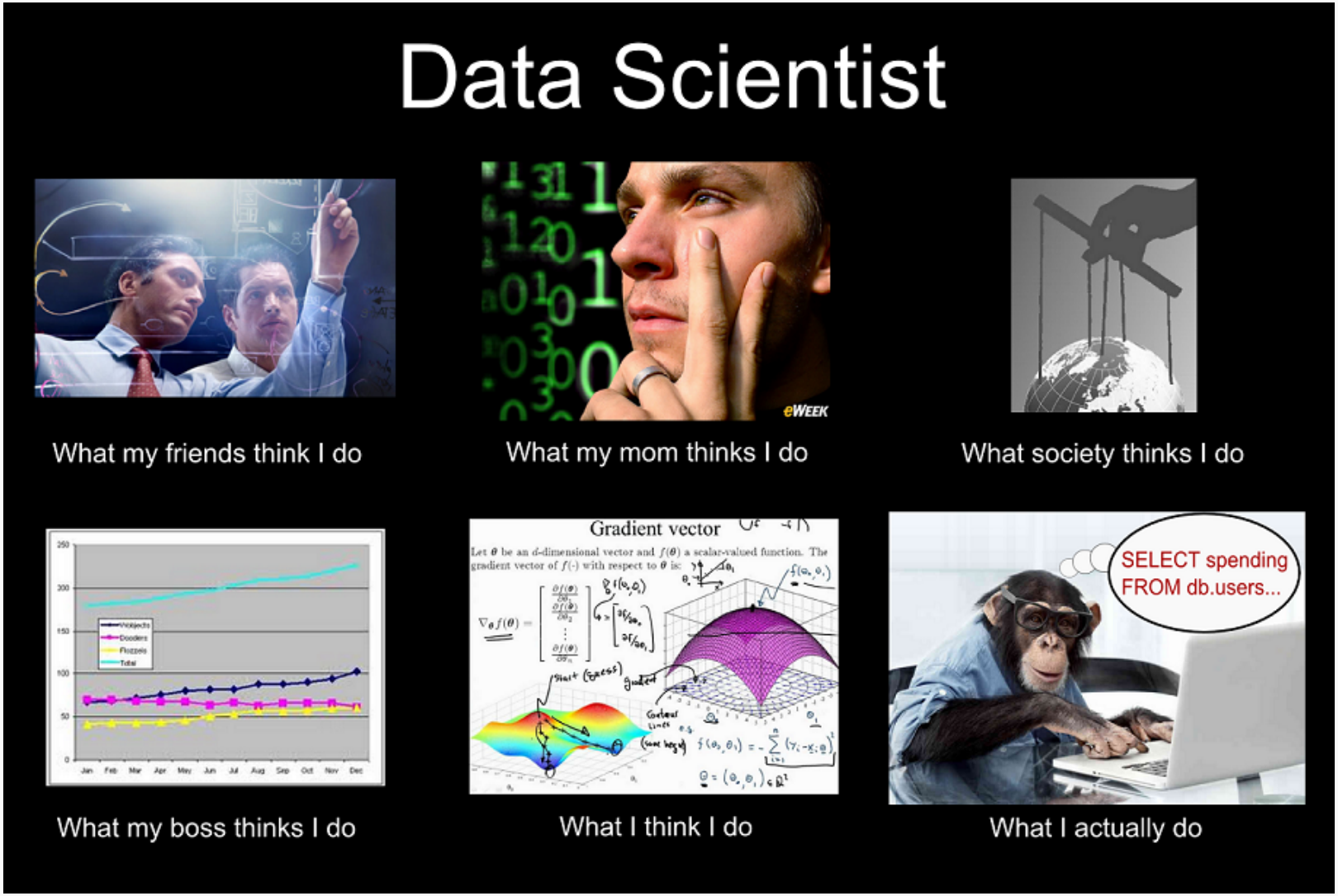
Citation

"The *best* thing about R is that it was written by *statisticians*.

The *worst* thing about R is that it was written by *statisticians*."

Bow Cowgill

Data scientist



data.table

data.table

- `data.table` est un package développé par Matt Doyle qui crée un nouvel objet, une *data.table*, qui est un *data.frame* performant.
- Le *data.frame* est un objet central dans R. Pour R, c'est une *liste de vecteurs* mais on peut considérer qu'il s'agit du même concept qu'une *table SQL* ou qu'un *data SAS* ou qu'un *dataframe Pandas*.
- Dès que le nombre de lignes commence à dépasser le million, on commence à constater des lenteurs

Apports data.table

- une syntaxe relativement simple, compact et très efficace
- la possibilité de définir une *clé* pour *trier/indexer* la table
- la capacité de faire des aggrégations très rapides
- une gestion efficace de la mémoire pour éviter les recopies (utilisation de *référence*)
- la capacité de lire et écrire très rapidement des fichiers `csv`
- la possibilité de faire des traitements par blocs de valeur

La syntaxe de base (data.table)

- DT[I, J, by=]
- la principale évolution de `data.table` est le champ `J` : on peut placer beaucoup de traitements dans ce champ, contrairement aux data.frames.

La syntaxe de base (SQL)

- from[where, select, by]

La syntaxe de base (fonctionnelle)

- DT[*les lignes sélectionnées*,
  ce qu'on fait avec,
  by=*comment on les groupe*]

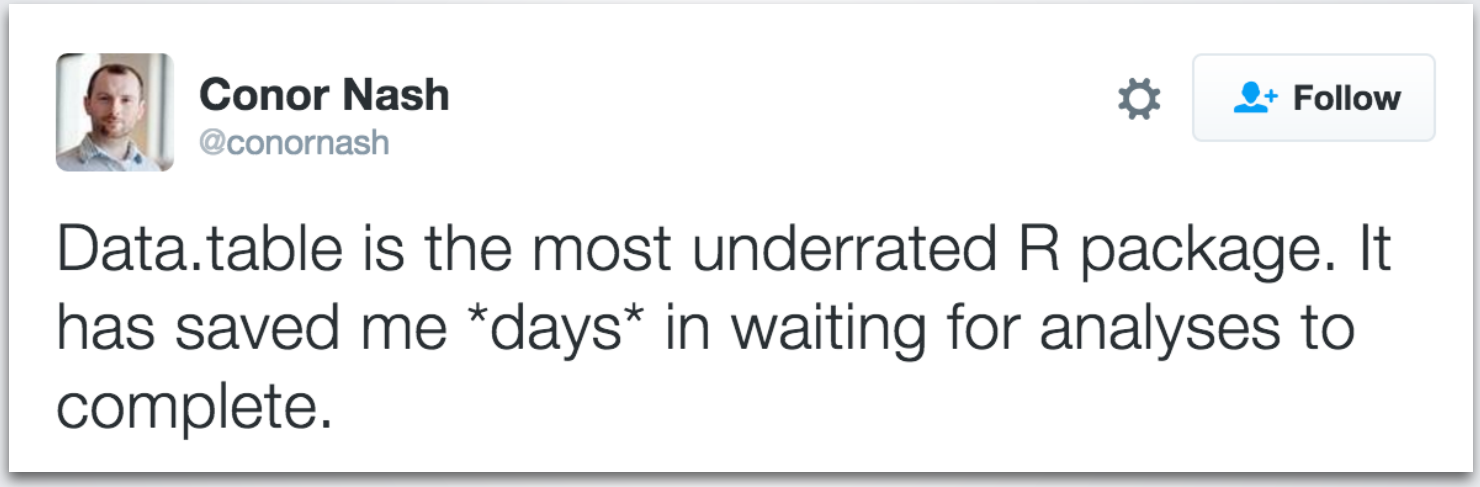
Utilisation basique de data.table

Exemple d'utilisation sur données Justice

Conclusion

Un package méconnu

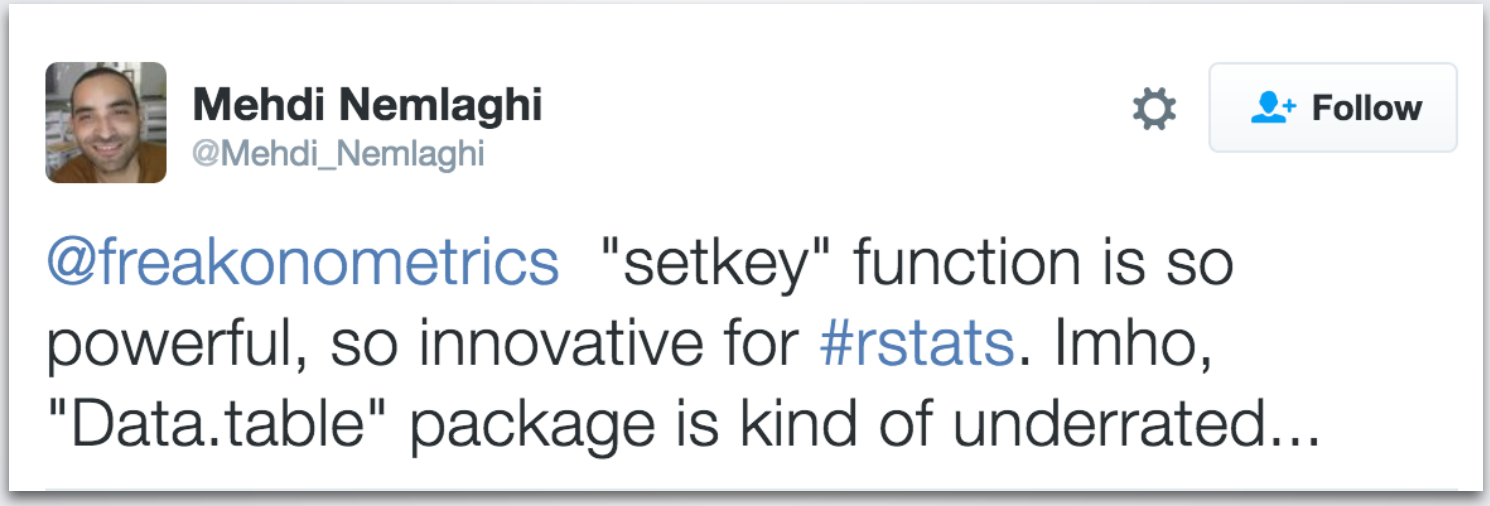
MOST UNDERRATED PACKAGE



August 2016

Un package méconnu

MOST UNDERRATED PACKAGE



May 2015

Un package très utile au quotidien

- on dit souvent que la préparation des données représente 80% du travail en *data science* : c'est vrai
- un package comme `data.table` permet de gagner beaucoup de temps pour les tâches *utiles* et *fréquentes*, même si elles sont peu *spectaculaires*.
- `data.table` fournit une structure de données très efficace, et même *indispensable* pour utiliser R
- il complète très bien les packages d'Hadley Wickam, en particulier `dplyr`, qui sait utiliser `data.table` avec le package `dtplyr`.

Questions