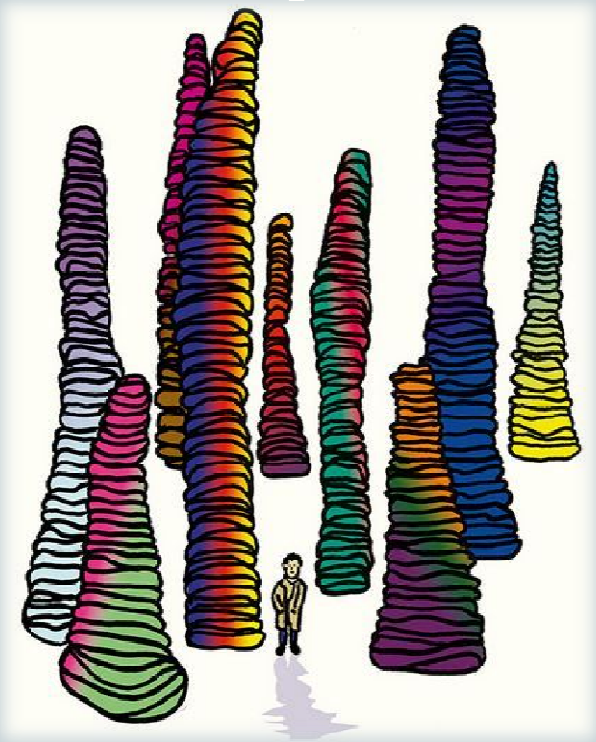
Tessera, outil Big Data



THULEAU Simon
ROTTREAU Elise

Année universitaire 2015-2016
M2 statistique appliquée à l'agroalimentaire et à l'agronomie

Plan

INTRODUCTION

- I- Quelques définitions
  - A) Les Big Data
  - B) Une idée des données
- II- Tessera
  - A) Principe
  - B) Ses packages
- III) Démonstration
  - A) Le jeu de données utilisé
  - B) Présentation R

I- Quelques définitions

A. Les Big Data

- Données complexes et volumineuses
- Difficiles à capturer, traiter, stocker, parcourir, et analyser avec les outils classiques

B. Une idée des données

- **Entreprises**: emails, documents, bases de données…etc.
- **Hors entreprises**: bases de données externes, contenus échangés sur les réseaux sociaux ou publiés en ligne, les historiques…etc.

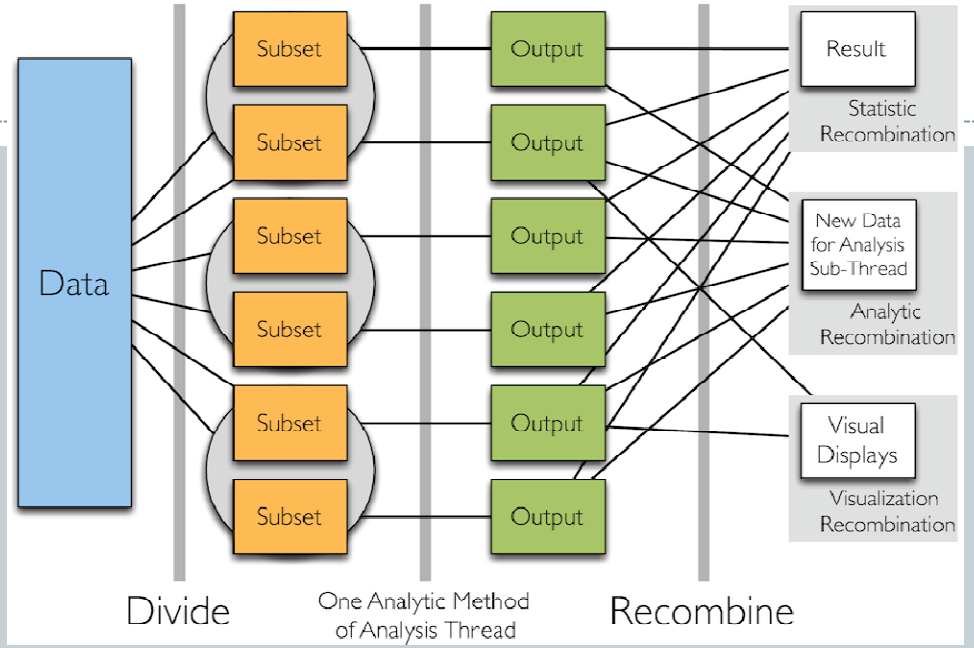
II-Tessera

A. Définition et Principe

- Développé par

- Lancé en novembre 2014
- permet l'exploration rapide de gros et complexes ensembles de données
- Interface R
- Commandes simples
- Résultats à la fois sous forme de graphiques et sous forme de texte / valeurs numériques, afin de mettre en valeur des métriques particulières (moyennes, max, etc…).
- Méthode statistique « D&R »
- Deux importants packages R:
  - Datadr
  - Trelliscope (une interface de visualisation)

Divide and Recombine (D&R)



- Division
  - Données divisées en sous-ensemble
  - Petits jeux de données
  - Clés → accès rapide
- Méthodes analytiques appliquées à chaque sous-ensemble
  - Pas de communication entre les calculs
- Visualisation des méthodes d'analyse
  - Appliquées à chaque subset
  - De très nombreux plot
- Pour chaque méthode
  - Recombinaison
  - Sorties recombinées
  - calcul s parallèles

B. Ses packages

- **Trelliscope**:
  - Visualisation détaillée de gros volumes de données
  - Jeu de données divisé en sous-ensemble
  - Méthode de visualisation appliquée à chaque sous ensemble
  - L'utilisateur peut trier et filtrer les graphiques

- **Datadr:**
  - interface simple pour les opérations de D & R
  - mémoire , disque local
  - Codage effectué entièrement dans R
  - Données représentées sous forme d'objets de recherche

III) Démonstration

A. Le jeu de données utilisé

- *adult*

-inclus dans Tessera
-32561 individus
-16 variables (9 qualitatives, 7 quantitatives)

B. Présentation R

Merci de votre

attention