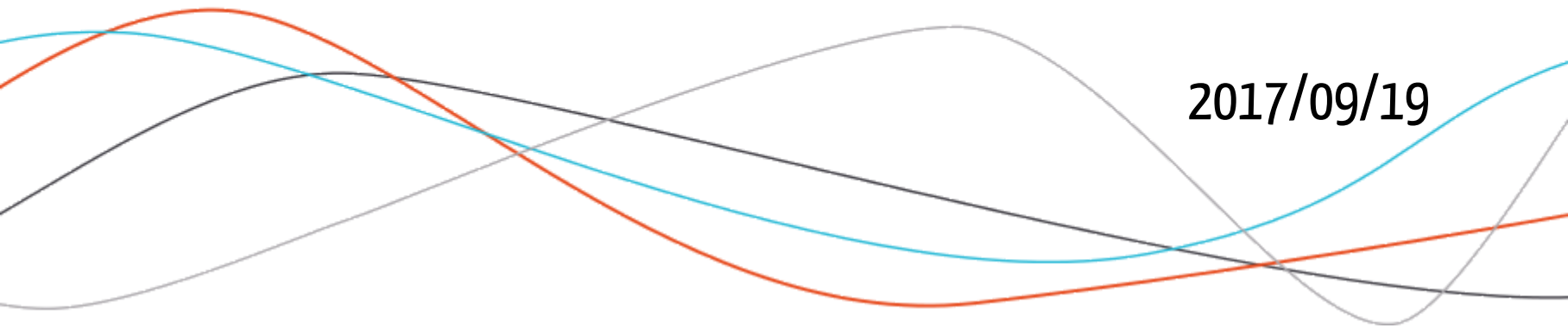
Des données brutes à la dataviz

Colin FAY - ThinkR

2017/09/19



$ whoami

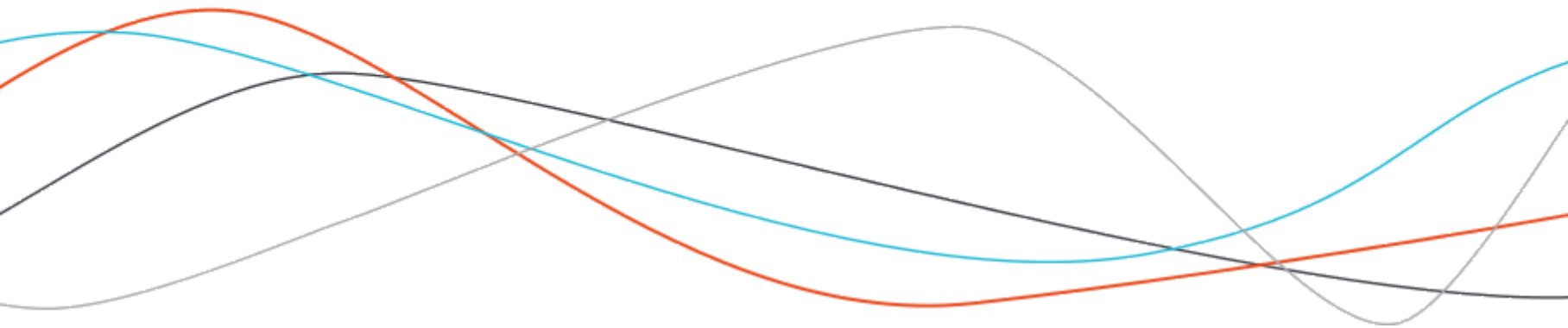
Colin FAY

Data Analyst, formateur R, Social Media Manager chez ThinkR, agence spécialisée en Data Science et en langage R.

Fondateur de Data-Bzh, la première plateforme de data-blogging bretonne.

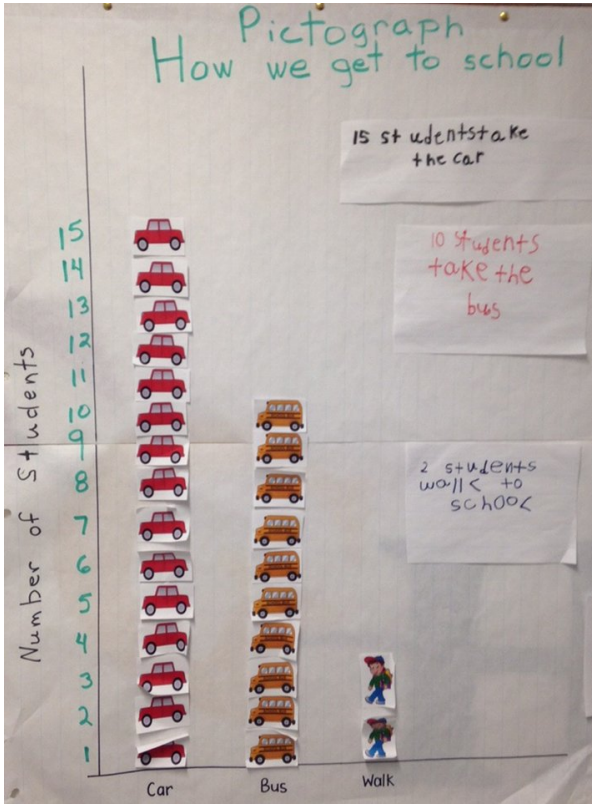
- http://thinkr.fr
- http://data-bzh.fr
- http://twitter.com/_colinfay
- http://github.com/colinfay

De quoi va-t-on parler aujourd'hui ?



De la donnée brute...

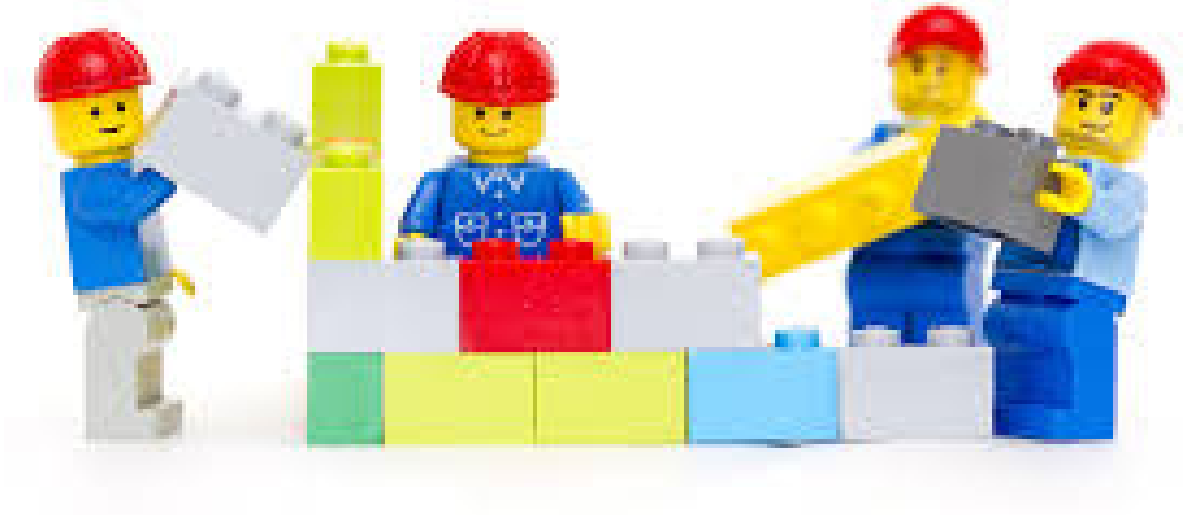
...à la dataviz !


Pictograph
How we get to school
15 students take the car
10 students take the bus
2 students walk to school
Number of Students
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
Car
Bus
Walk

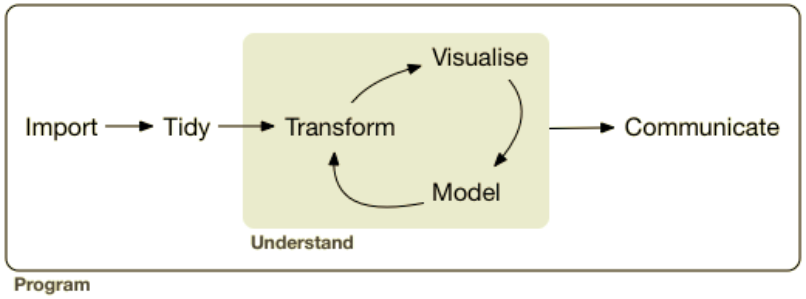
Ce n'est pas (que) de la magie...

... il y a aussi du boulot...

... et de la méthode.



De la méthode



La **dataviz reste avant tout un outil de communication**. Mais avant, il se passe plein de trucs.

- Importer (on fait venir les données dans son environnement de travail).
- Nettoyer (on nettoie le bazar mis par quelqu'un d'autre).
- Visualiser (on explore en visualisant).
- et enfin, communiquer (be patient!).

Importer

Importer

Travailler avec des données "in the wild", c'est souvent avoir à faire face à des choses étranges :

- Des formats inconnus.
- Des formats propriétaires.
- Des encodages anarchiques.
- Des nomenclatures inconnues.

... et je vais m'arrêter là qu'on puisse à un moment aller manger.

Nettoyer

Nettoyer

Pourquoi ?

Avant d'être traitées, les données ont souvent besoin d'un petit coup de « cleaning » (ou de burin, ça dépend) :

- Pour formater le texte.
- Pour formater les chiffres.
- Parce qu'il faut savoir quoi faire des données manquantes.
- Parce qu'il y a souvent des lignes / colonnes vides.

...

Importer et nettoyer

Par l'exemple...

Jeu de données : Budget principal du Département de Loire-Atlantique

Rien que pour la "lecture" du jeu de données, j'ai...

- Téléchargé et décompressé le.zip.
- Listé les csv.
- Chargé les 10 csv.
- Travaillé sur l'encodage.
- Assigné la bonne date à chaque csv.
- Supprimé les caractères inexploitables (€).
- Transformé les colonnes en chiffres.
- Joint les tableaux.

Transformer

Transformer

Pourquoi ?

On passe toujours par un petit moment de manipulation de données pour :

- Résumer.
- Créer des nouvelles lignes / colonnes.
- Découvrir des tendances.
- Construire des modèles.

...

Transformer

Par l'exemple...

Jeu de données : Budget principal du Département de Loire-Atlantique

- Supprimer les données manquantes.
- Grouper les observations par année.
- Grouper par type de dépense.
- Calculer la somme par année et par type de dépense.

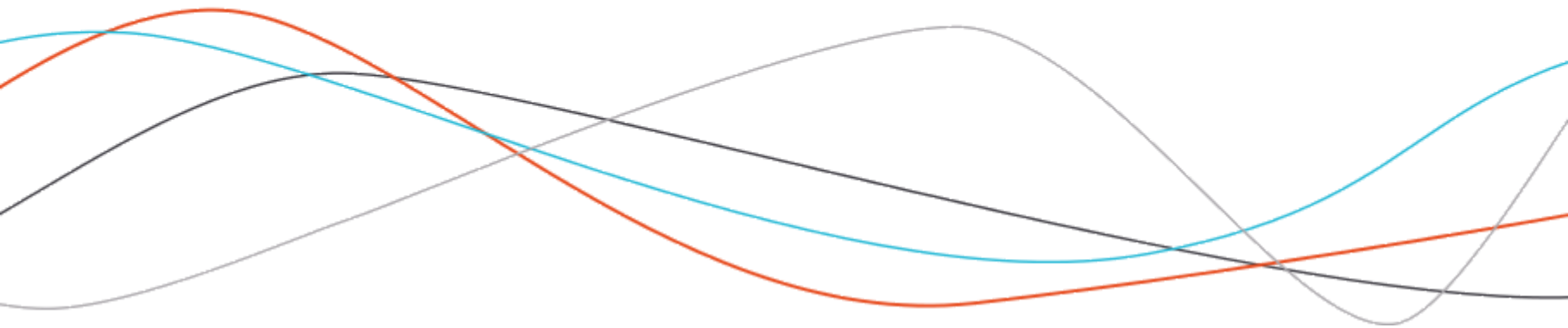
Visualiser

Visualisation : se poser les bonnes questions

- Qu'est-ce qu'on chercher à représenter ? (Un dessin sur du papier permet de se faire une bonne idée de ce que l'on cherche à visualiser)
- Un, deux, trois... variables ?
- Quelles échelles ?
- Quelles formes ?
- Quelles couleurs ?
- ...

Admettons qu'ici, on veuille se représenter l'évolution des recettes et dépenses du Département. Mais avant...

interlude::on()



Si vous ne deviez retenir qu'un slide de ma présentation :

(... hormis les legos, bien sûr)

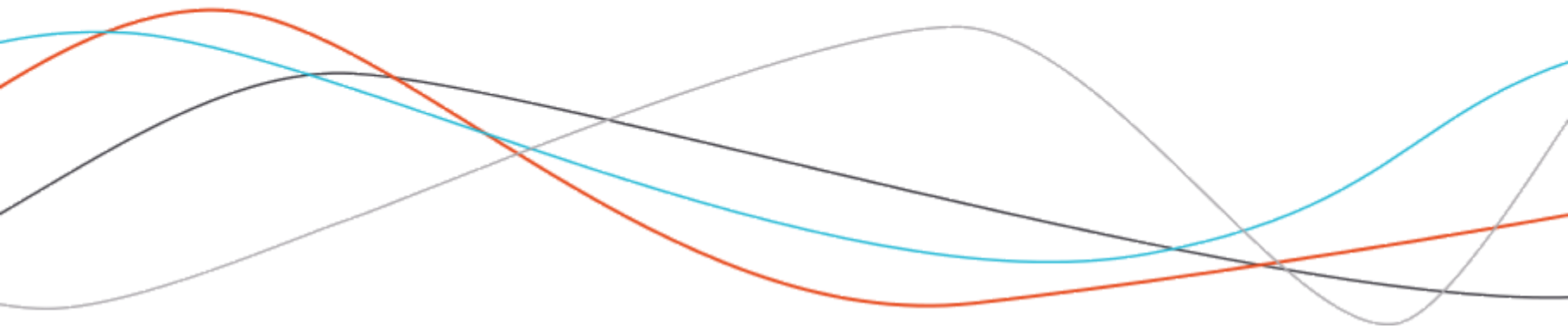
> Une dataviz contient **la juste dose d'éléments**. Pas plus, pas moins.

> Avant d'ajouter chaque nouvel élément, il est **indispensable** de se poser la question : est-ce que j'ajoute cela parce que c'est beau, ou parce que ça a du sens ?

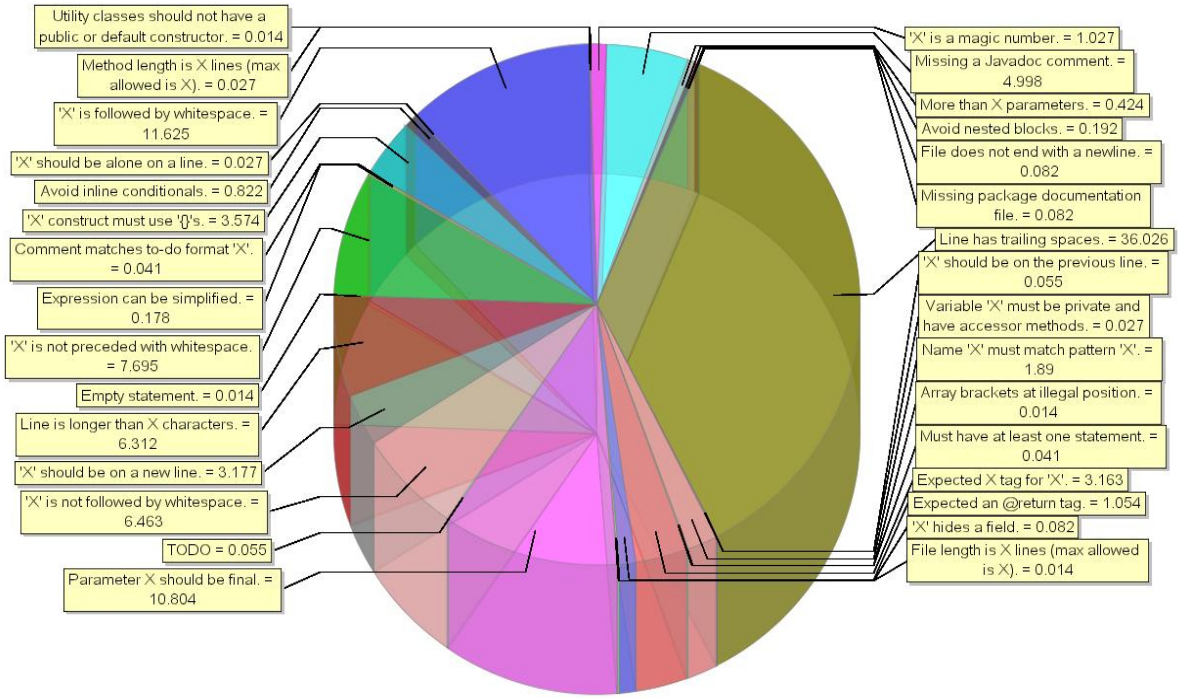
> Une **dataviz se doit d'être informative**, pas "belle" : ça presque, on s'en fout (même si c'est mieux).

> Quitte à choisir, **autant avoir une dataviz moche mais informative, qu'une "belle" dataviz où on ne comprend rien**.

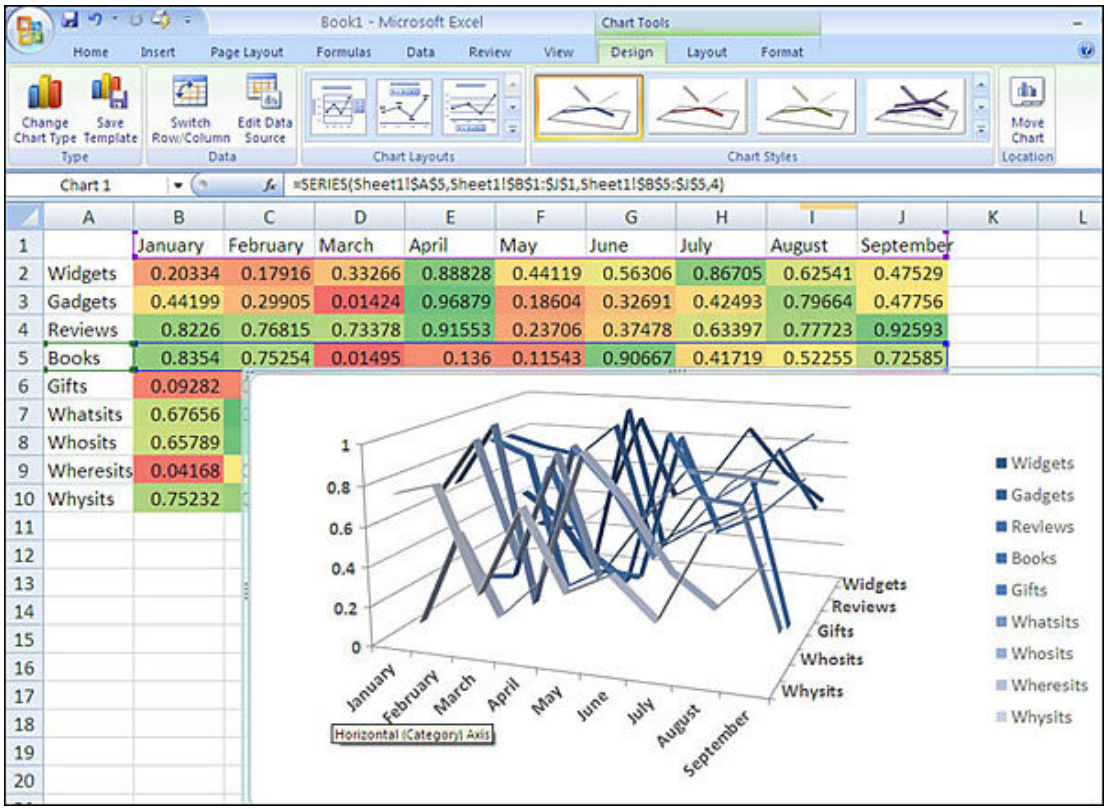
Pourquoi je vous dis ça ?



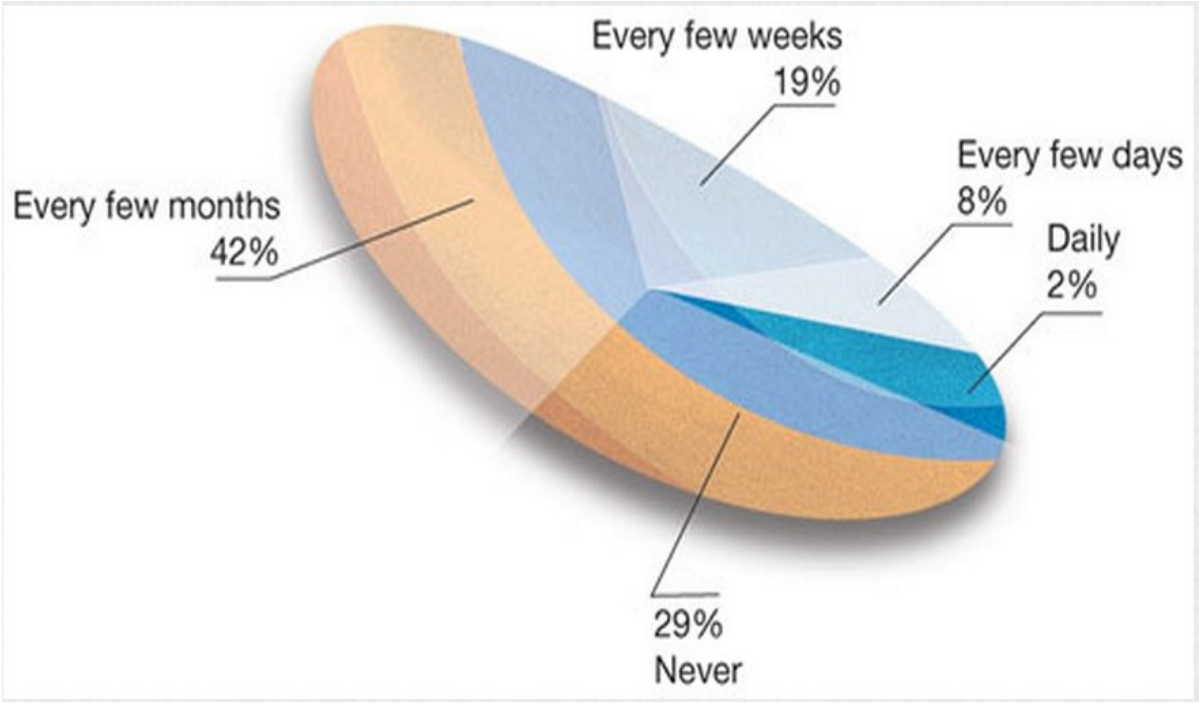
Pourquoi je vous dis ça :



Pourquoi je vous dis ça :



Pourquoi je vous dis ça :



Alors que bon...

Pourquoi je vous dis ça :

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------
-----------

## filter(): dplyr, stats
## lag():    dplyr, stats
```

Pourquoi je vous dis ça :

Le pire dans cette histoire, c'est que le même graphe m'a pris 2 minutes à créer :

```
library(tidyverse)
df <- data.frame(What = factor(c("Every few Months", "Every few
weeks",
                                 "Every few days","Daily", "Never"),
levels = c("Every few Months", "Every few weeks",
                                 "Every few days","Daily", "Never")),
                    How_much = c(42, 19, 8, 2, 29))
ggplot(df, aes(What, How_much)) +
  geom_col(fill = "#E2673C") +
  labs(title = "A better plot",
       x = "What",
       y = "How much, in %") +
  theme_wsj()
```

Moral

Si en sortant d'ici...

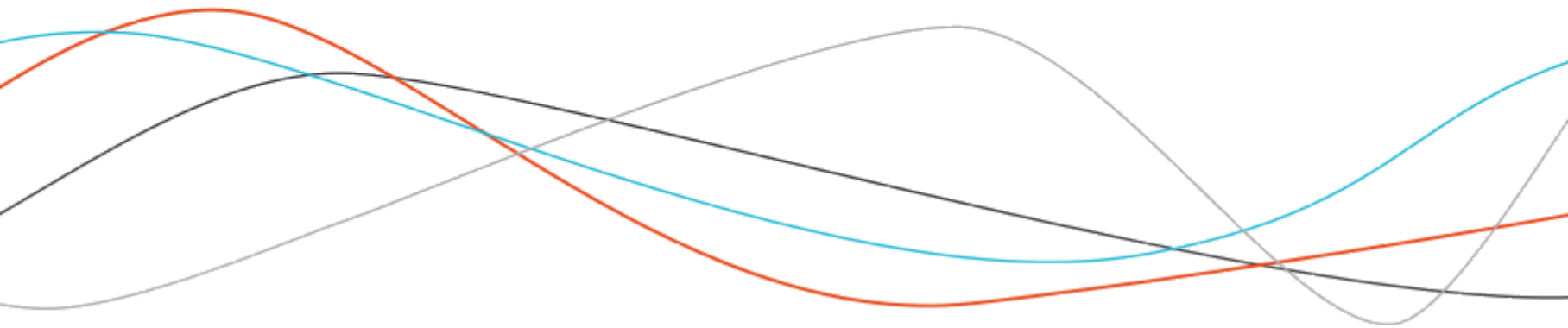
> ... vous faites un graphique en camembert, je viendrais vous hanter pendant votre sommeil.

> ... vous faites un graphique camembert en 3D, je vous promets que je fais une crise cardiaque (chiche !).

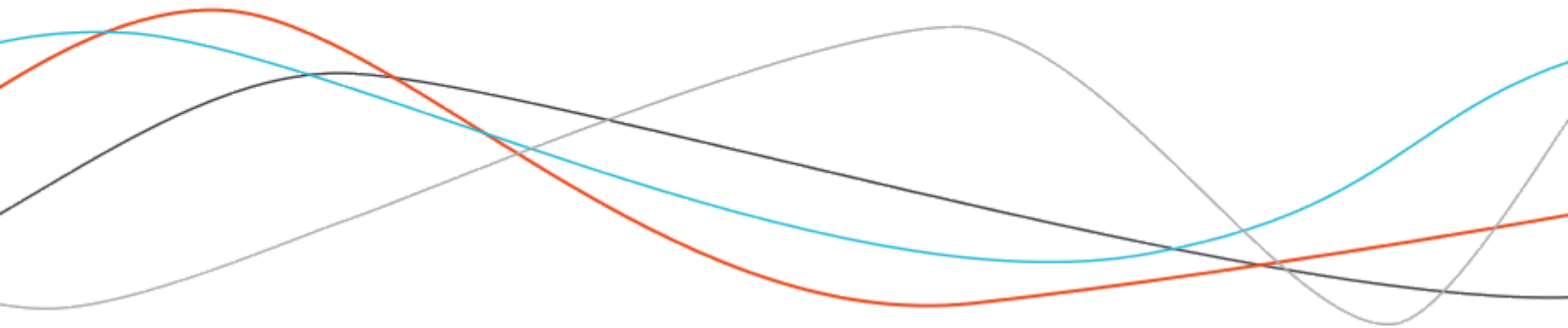
Bref :

> La première question à vous poser ne devra JAMAIS être "est-ce que ça va être joli" ? / "est-ce que mon graphique 3D qui tourne va impressionner Julien du service Marketing ?"

interlude::off()



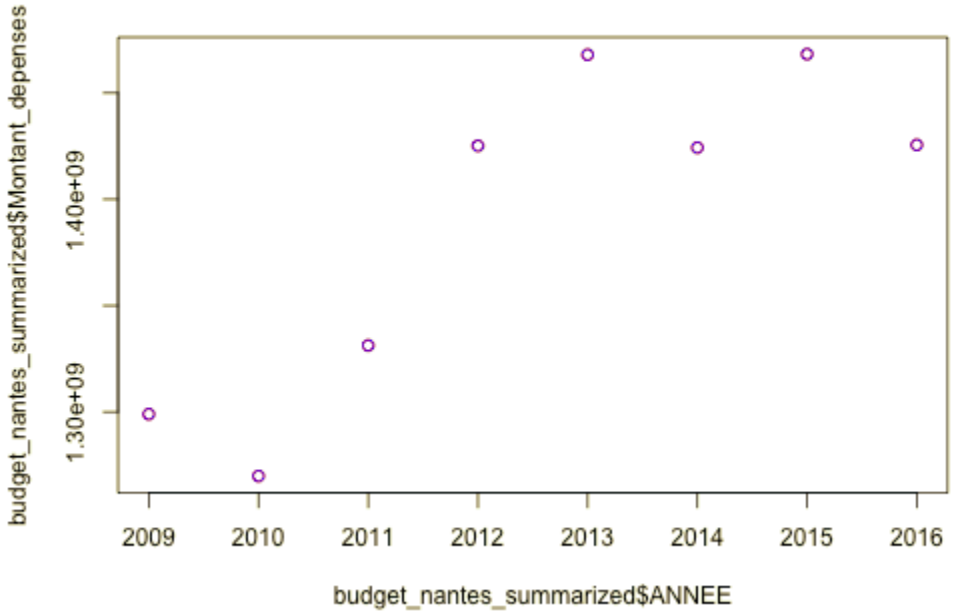
Créer une dataviz, quelques étapes



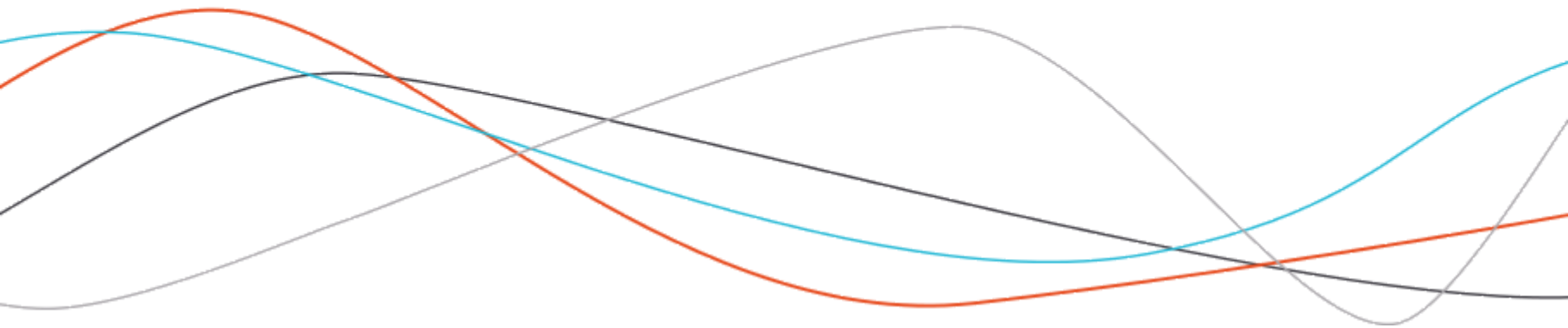
Étape 1 : un tableau rapide, moche, et informatif

TYPE_DE_MOUVEMENT	ANNEE	Montant_depenses
DEPENSE	2009	1.299e+09
DEPENSE	2010	1.27e+09
DEPENSE	2011	1.331e+09
DEPENSE	2012	1.425e+09
DEPENSE	2013	1.468e+09
DEPENSE	2014	1.424e+09

Étape 2 : une dataviz rapide, moche, et informative :

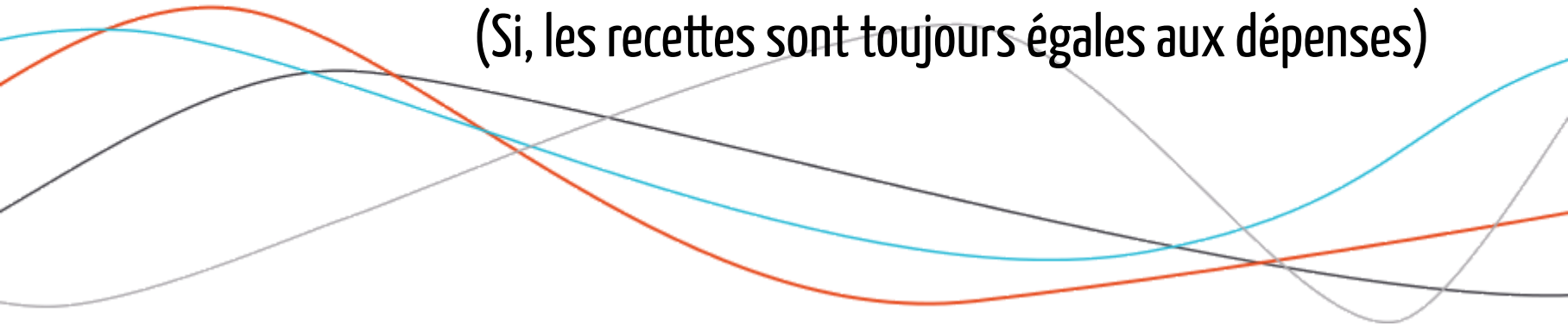


Wait... il n'y a rien qui vous surprenne ?



Wait... il n'y a rien qui vous surprenne ?

(Si, les recettes sont toujours égales aux dépenses)



Bref, enfin, on peut enfin penser à mettre les paillettes :

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

Merci !

des questions ?

Retrouvez-moi sur les internets :

(je parle principalement de données)

- colin@thinkr.fr
- http://twitter.com/_colinfay
- http://twitter.com/thinkr_fr
- https://github.com/ColinFay

J'écris des trucs sur les internets :

(et ça parle principalement de données)

- https://thinkr.fr/
- http://colinfay.me/
- http://data-bzh.fr/