

Dataviz et éthique

Colin FAY - ThinkR

2017/09/20



\$ whoami

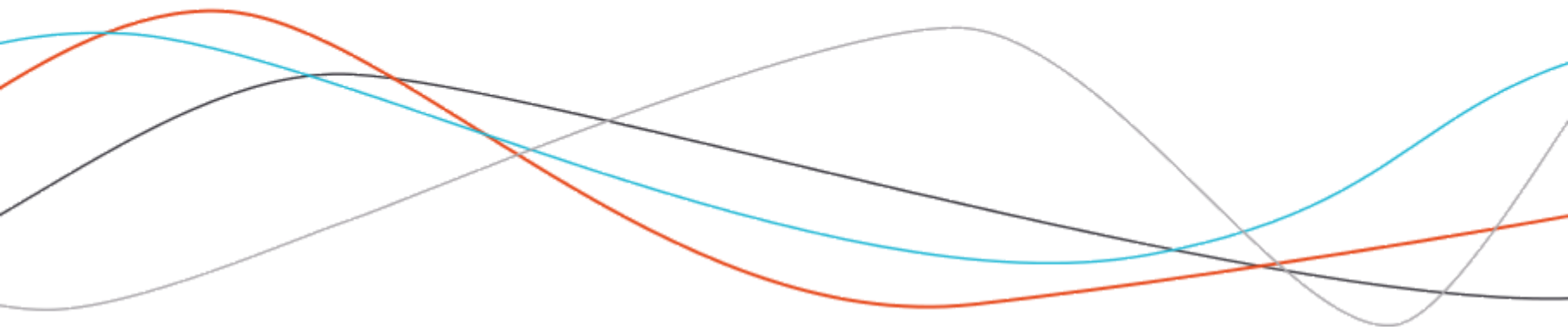
Colin FAY

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://thinkr.fr)
- [http:// data-bzh.fr](http://data-bzh.fr)
- [http:// twitter.com/_colinfay](http://twitter.com/_colinfay)
- [http:// github.com/colinfay](http://github.com/colinfay)

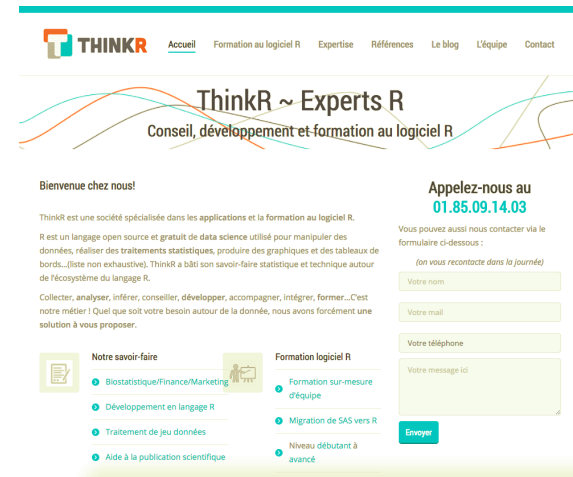
ThinkR



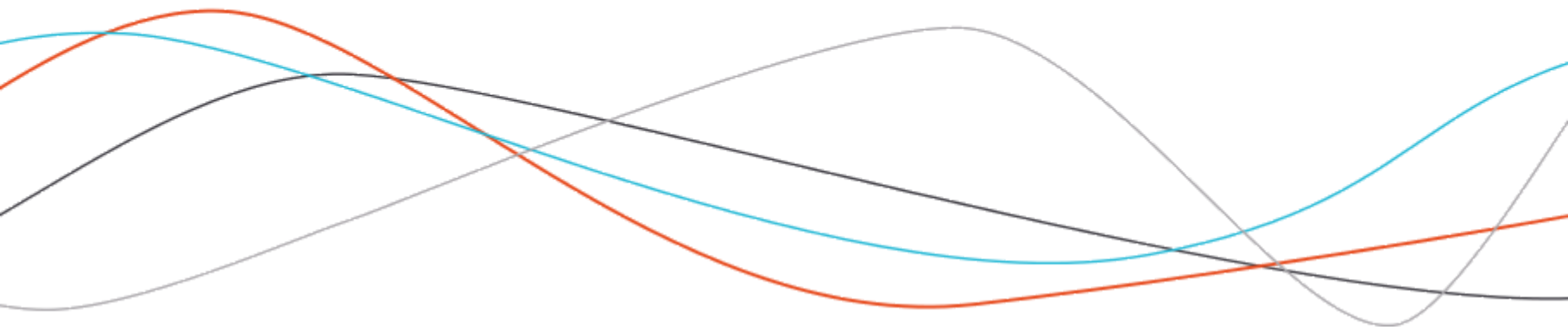
ThinkR

Société spécialisée en Data Science et langage R.

- Formation
- Analyse
- Développement
- Consultance



Data-Bzh



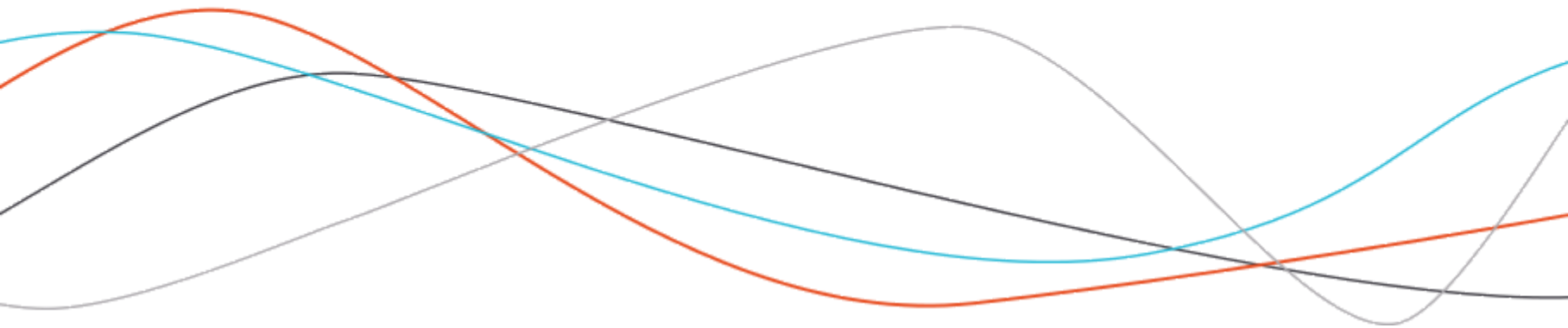
Data-Bzh

Première plateforme de data-blogging en Bretagne.

- Dataviz
- Analyse
- Conférences
- Tutoriels



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



Éthique et Dataviz

La question à mille dollars :

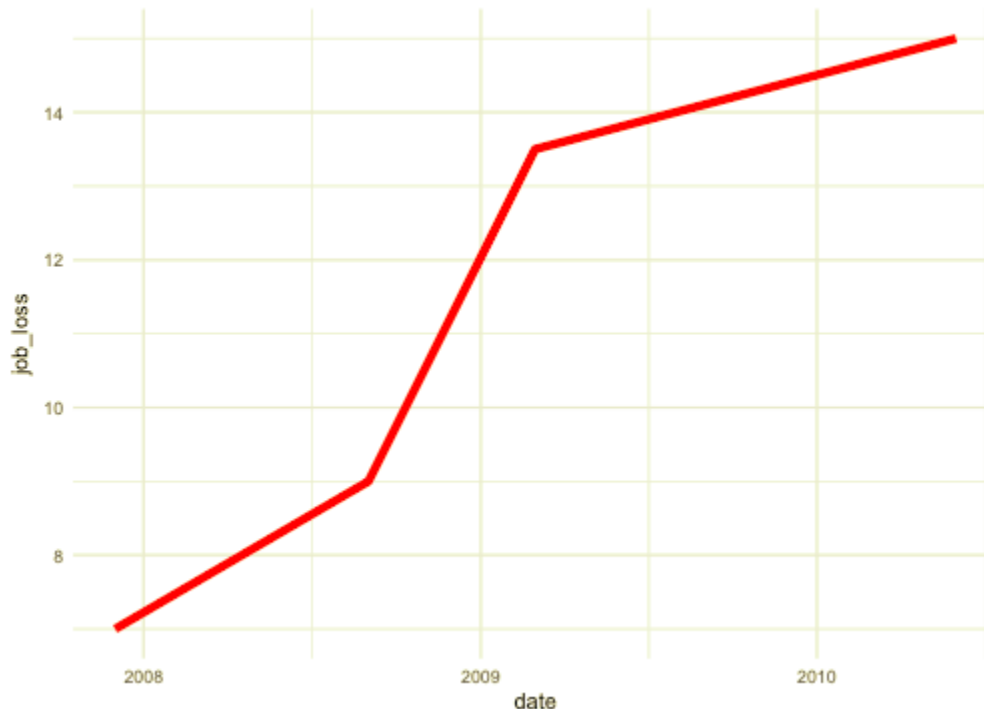
Pourquoi c'est important de se poser la question de l'éthique en dataviz ?



Éthique et Dataviz

La question à mille dollars :

Pourquoi c'est important de se poser la question de l'éthique en dataviz ?



C'est quoi l'éthique ?

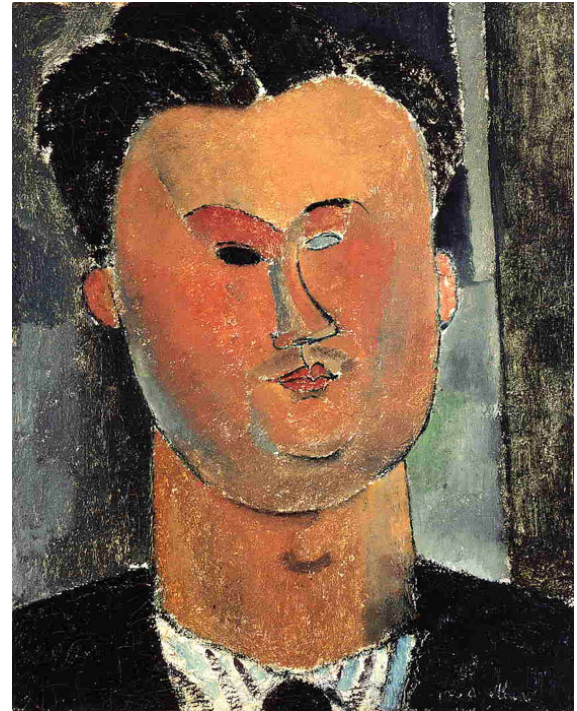
"L'éthique c'est l'esthétique du dedans."

Pierre Reverdy

Trop souvent, créer une dataviz se résume à se concentrer sur l'esthétique du dehors, au détriment de l'esthétique du dedans.

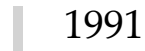
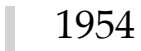
Quand peut-on dire qu'une dataviz est "belle du dedans" ?

Créer une dataviz, c'est respecter à la fois les données et son audience.





Top thousand ways to lie



... beyond dataviz : cherry picking

[Alerte Pratique douteuse]

Cherry picking : ne choisir que les résultats favorables.

Si nous savons par exemple qu'un traitement est inefficace, nous pouvons réaliser 40 tests, et nous avons au moins deux chances de trouver des résultats significatifs.

Il suffit ensuite de ne publier que les données significatives.



... beyond dataviz : Biased samples

[Alerte Pratique douteuse]

Biased sample : réaliser un test statistiques sur un échantillon non représentatif.

Si nous voulons montrer que les français sont petits, nous pourrions réaliser un échantillonnage à la sortie d'une école primaire. Ou l'inverse, en nous concentrant sur les joueurs de basket.



We will find the average height of Americans based on a sample of NBA players.

Selection Bias?

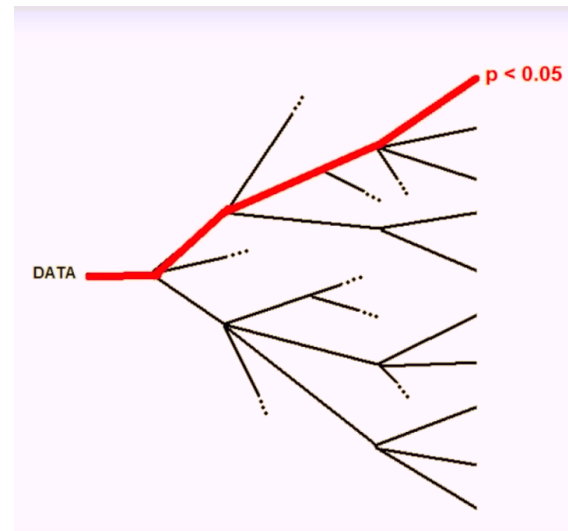
... beyond dataviz : p hacking

■ If you torture the data long enough, they will confess.

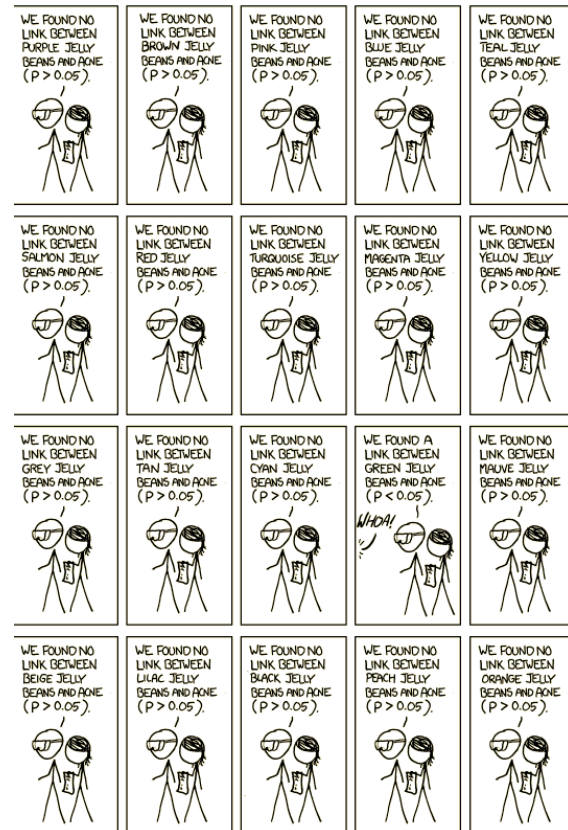
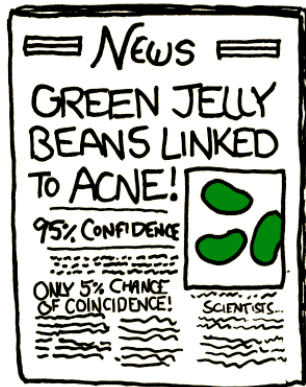
[Alerte Pratique douteuse]

p hacking : En testant suffisamment, on trouve toujours une valeur significative.

Quand un test n'est pas significatif et qu'on a pas beaucoup d'éthique, on peut faire du *p hacking*, c'est-à-dire torturer les données pour avoir un p significatif, en retirant certaines données / créant des sous groupes / changeant de test...



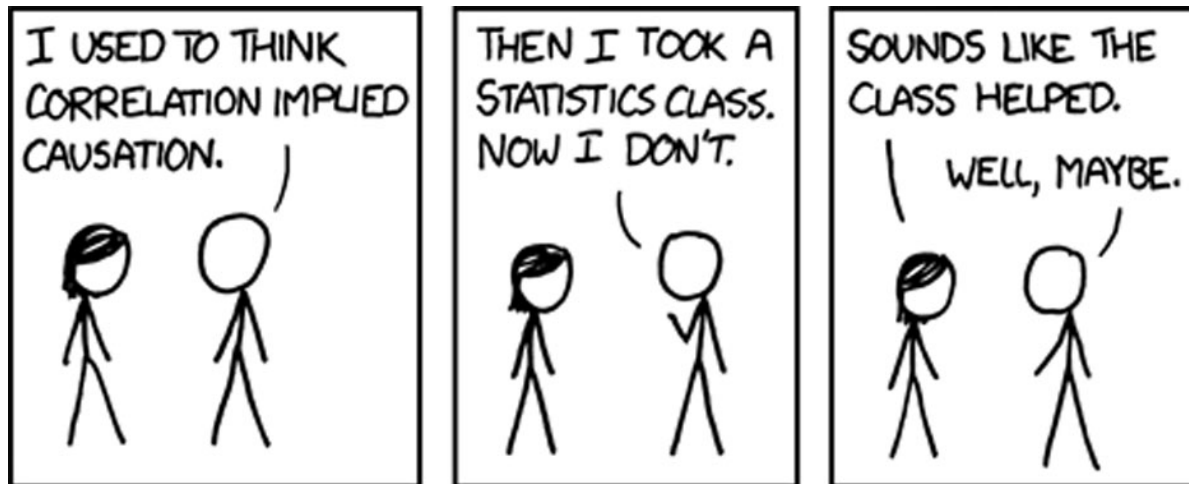
p-hacking



... beyond dataviz : false causality

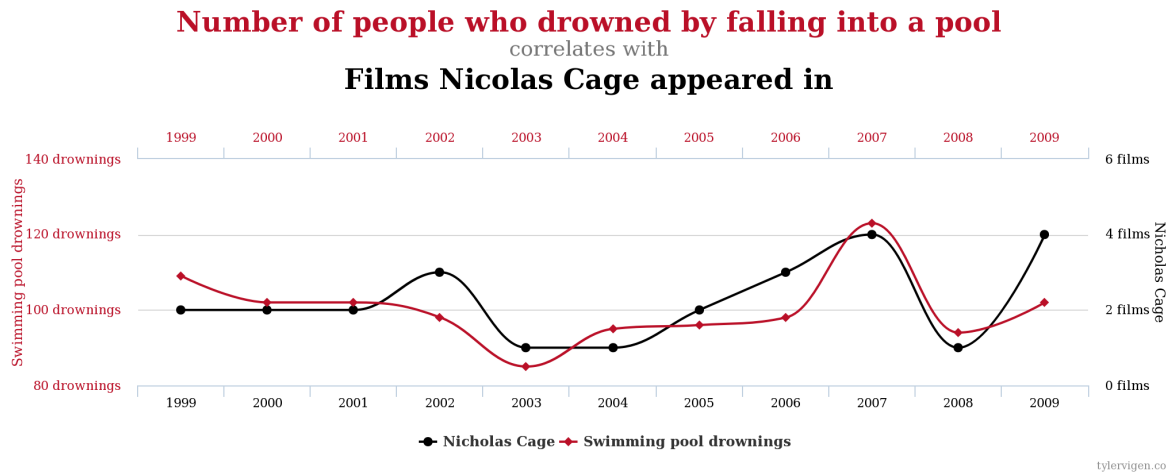
"Cum hoc ergo propter hoc" - Correlation does not imply causation.

Ce n'est pas parce que deux éléments sont corrélés qu'il existe une relation de cause à effet.



... beyond dataviz : false causality

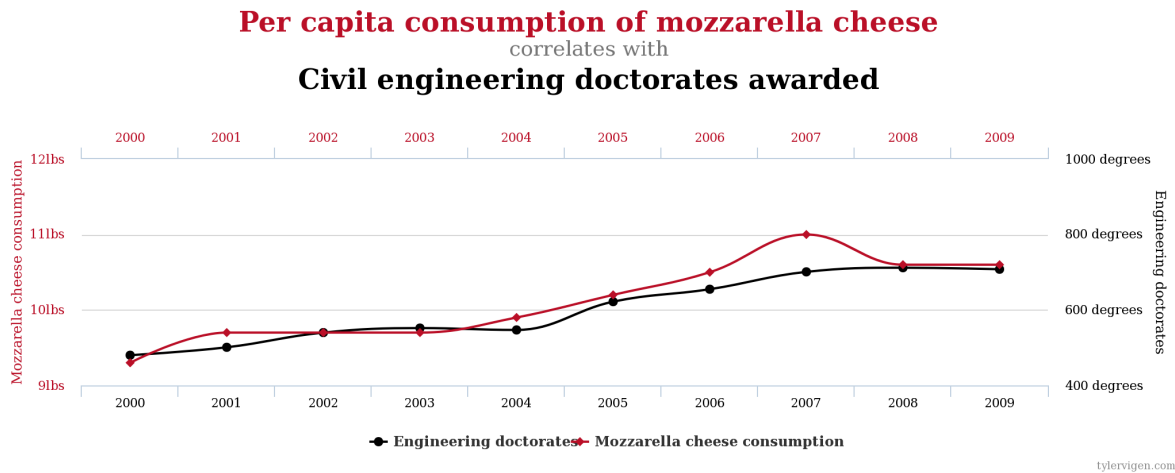
"Cum hoc ergo propter hoc" - Correlation does not imply causation.



via : <http://www.tylervigen.com/spurious-correlations>

... beyond dataviz : false causality

"Cum hoc ergo propter hoc" - Correlation does not imply causation.



via : <http://www.tylervigen.com/spurious-correlations>

Pourquoi s'intéresser à l'éthique en dataviz ?

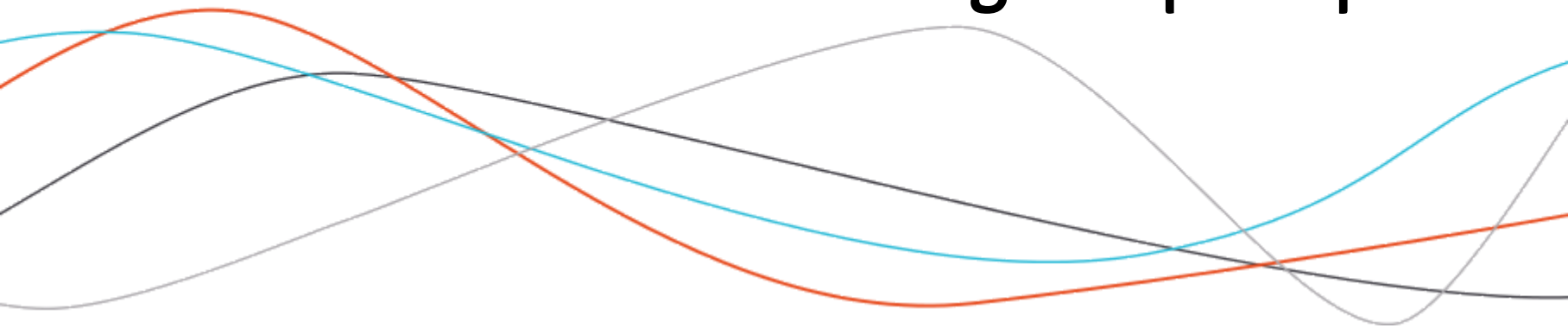
- Savoir reconnaître un "faux".
- Questionner ses propres pratiques.
- Ne pas succomber à la tentation.
- Sensibiliser.

One of the best ways to learn how to detect bias in data visualization is to intentionally manipulate a chart, and tell two (or more) opposing stories with the same data.

How to Lie with Charts - Jack Dougherty

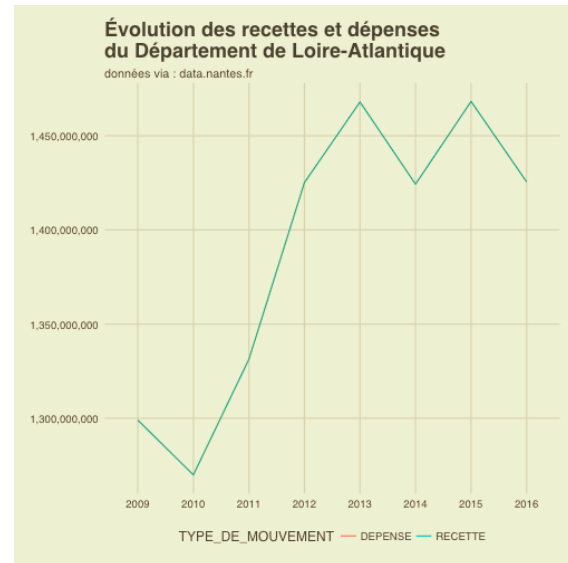
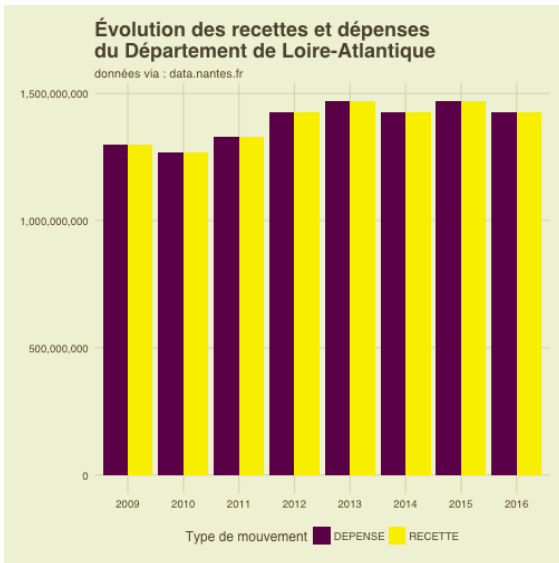
How to lie with dataviz

Petit guide pratique



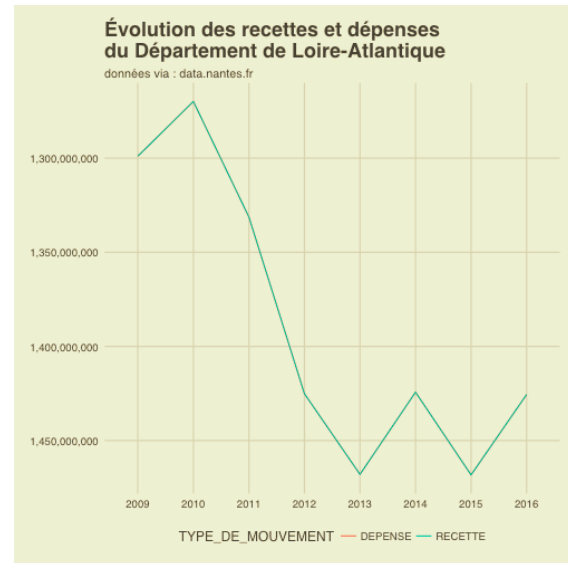
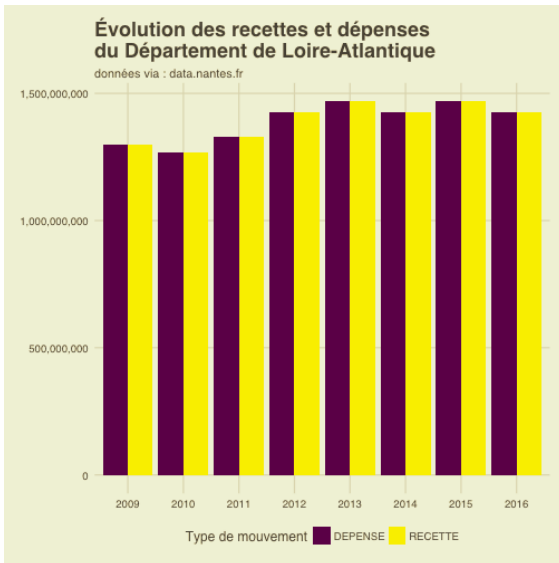
How to lie with dataviz

Jouer avec l'échelle



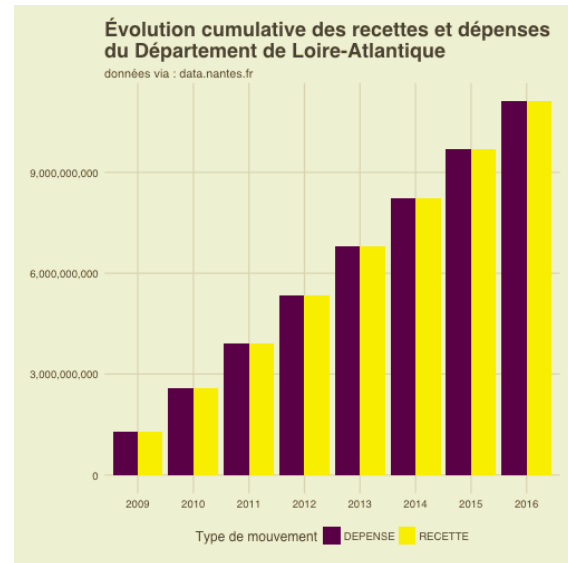
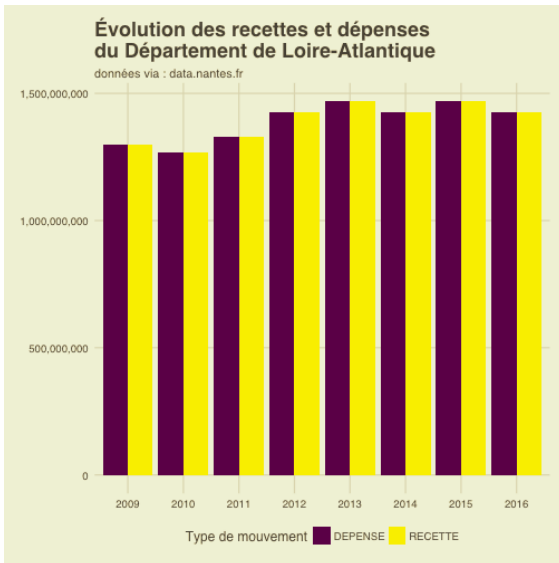
How to lie with dataviz

Renverser l'axe



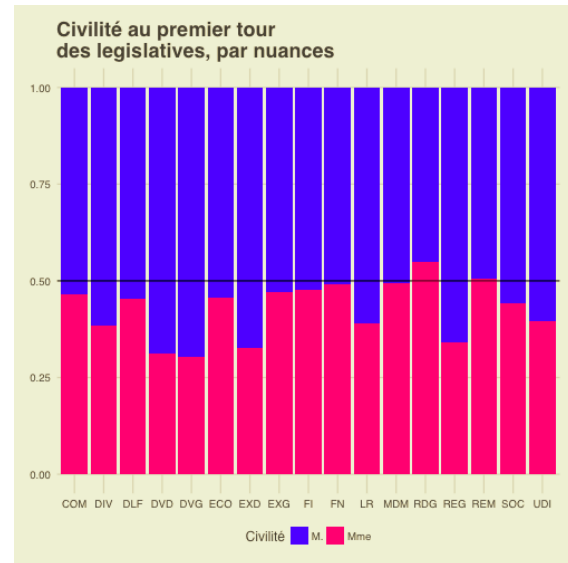
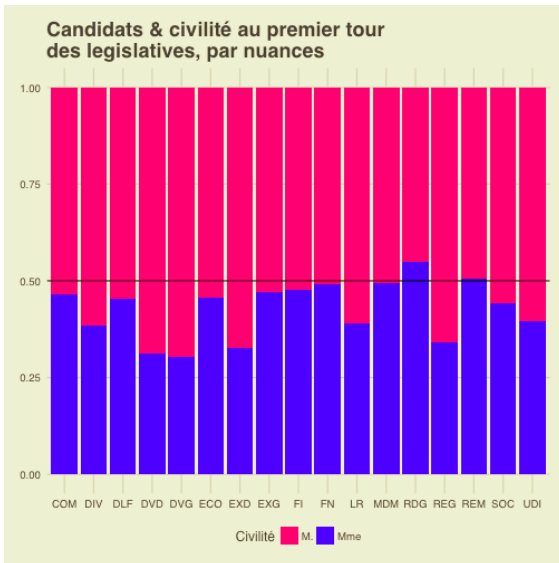
How to lie with dataviz

Créer des colonnes cumulatives



How to lie with dataviz

Jouer avec les couleurs

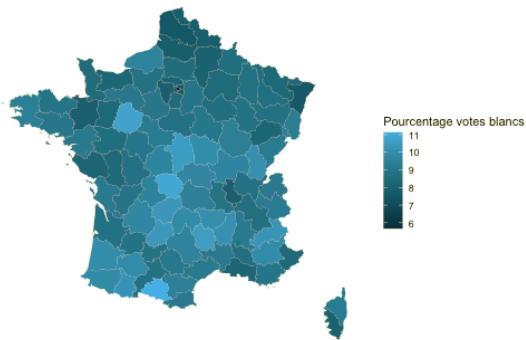


How to lie with dataviz

Jouer avec les couleurs

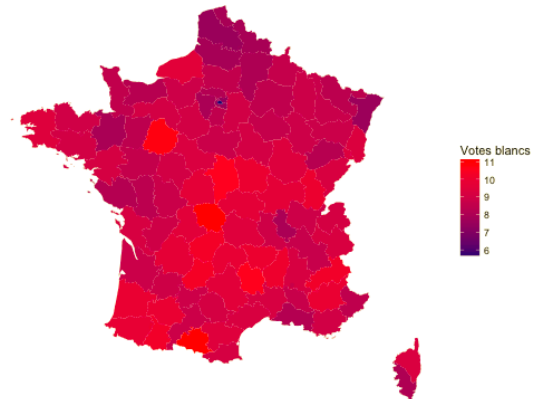
Pourcentage de votes blancs
au second tour des présidentielles 2017

Données via data.gouv

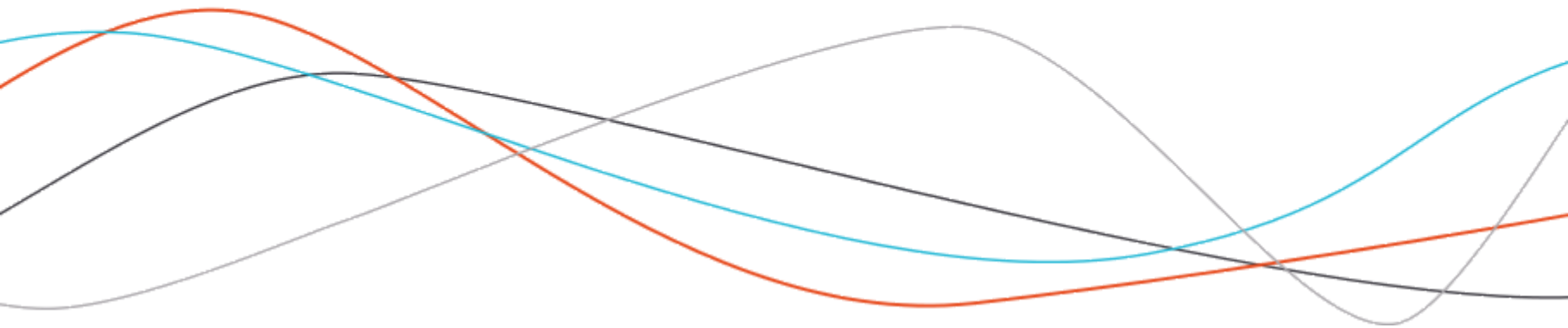


Pourcentage de votes blancs
au second tour des présidentielles 2017

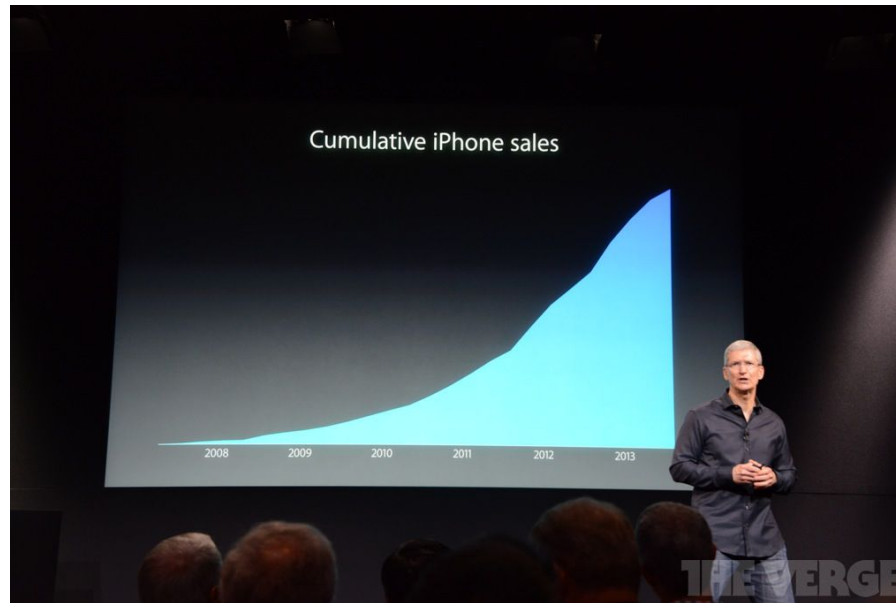
Données via data.gouv



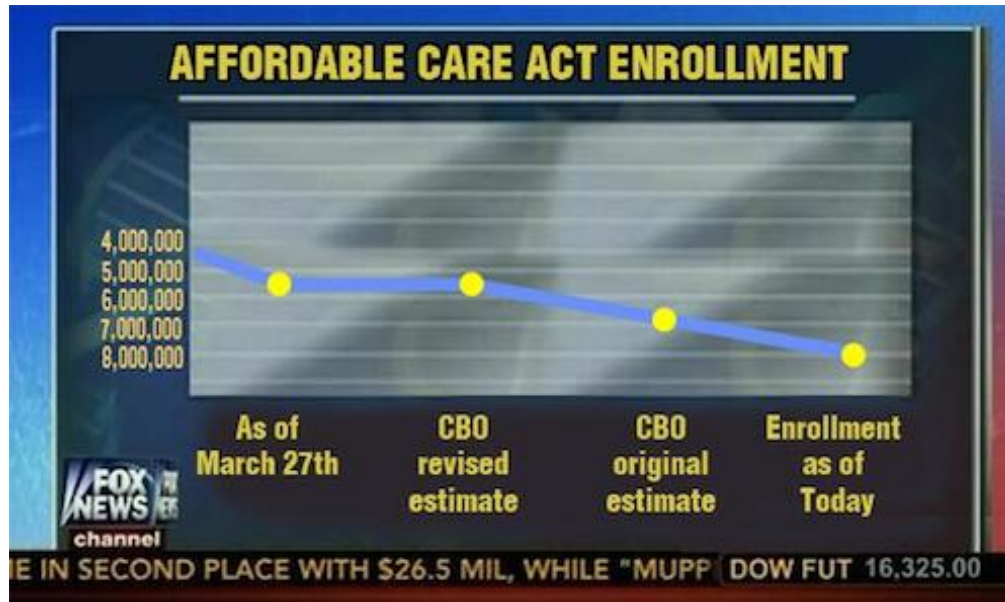
**Non mais dans la vraie vie, ça
n'existe pas...**



Dans la vraie vie



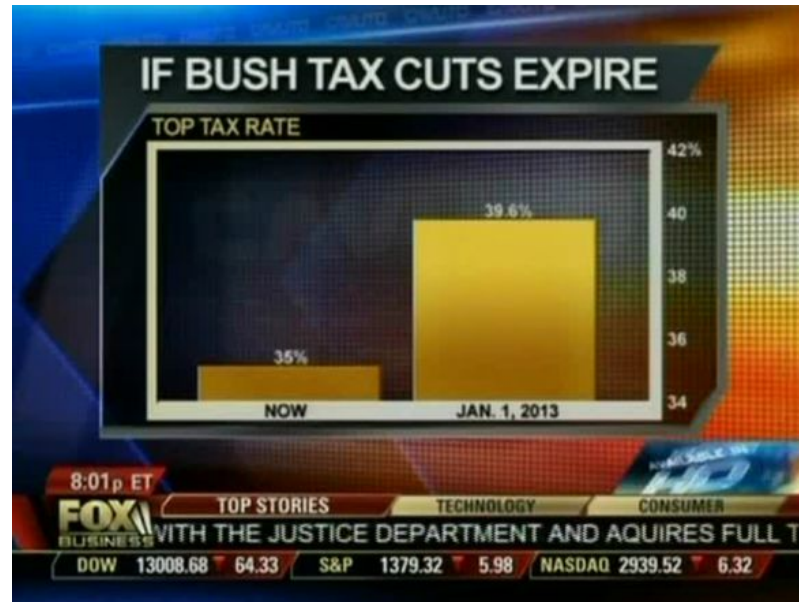
Dans la vraie vie



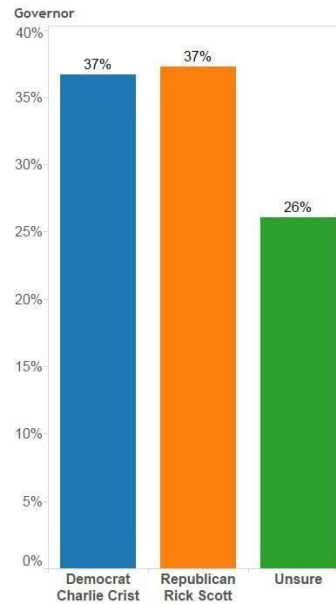
Dans la vraie vie



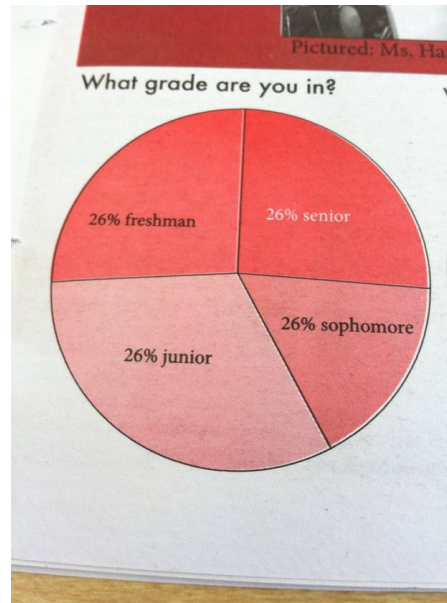
Dans la vraie vie



Dans la vraie vie



Dans la vraie vie





Et maintenant, on fait quoi ?

- On y réfléchit.
- On en parle.
- On pense à Colin à chaque fois qu'on fait une dataviz.
- On utilise des logiciels de statistiques pour faire des dataviz.



Merci !

des questions ?

Retrouvez-moi sur les internets :

(je parle principalement de données)

- colin@think.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/>