EXPLORATION ET VISUALISATION DES DONNÉES

NOTIONS FONDAMENTALES

LES DONNÉES AU 20IÈME SIÈCLE

Au 20e siècle, les problèmes reliés aux données sont surtout des problèmes de

- **génie** (design de machines)
- **science** (formulation de théories)

On résoud ces problèmes **empiriquement**, **théoriquement**, ou en passant par les **calculs** et les **simulations**.

LES DONNÉES AU 20IÈME SIÈCLE

Les machines sont équippées avec des détecteurs ⇒ on utilise les données afin de vérifier si les machines se comportent comme prévu et pour améliorer les designs.

On prépare des expériences ⇒ on utilise les données afin de tester la validité des théories.

- les expériences sont coûteuses et génèrent relativement peu de données.

Les données contiennent de l'informations supplémentaire qui est souvent ignorée.

- Exemple : les données expérimentales de Mendel, analysées par Fisher, s'avèrent trop belles pour être vraies.

LES DONNÉES AU 21 IÈME SIÈCLE

Au 21e siècle, on se retrouve avec:

- **plus de données**
- qui sont surtout **digitales**
- et surtout **observées** (plutôt que engendrées par des expériences conçues)

On résoud ces problèmes à l'aide de méthodes empiriques, théoriques, en passant par les calculs et les simulations, et par l'**exploration** et la **visualisation** des données.

LES DONNÉES AU 21 IÈME SIÈCLE

Empiriquement : observer et décrire ce qui se passe

Théoriquement : généraliser et construire des modèles et des généralisations pour comprendre ce qui se passe

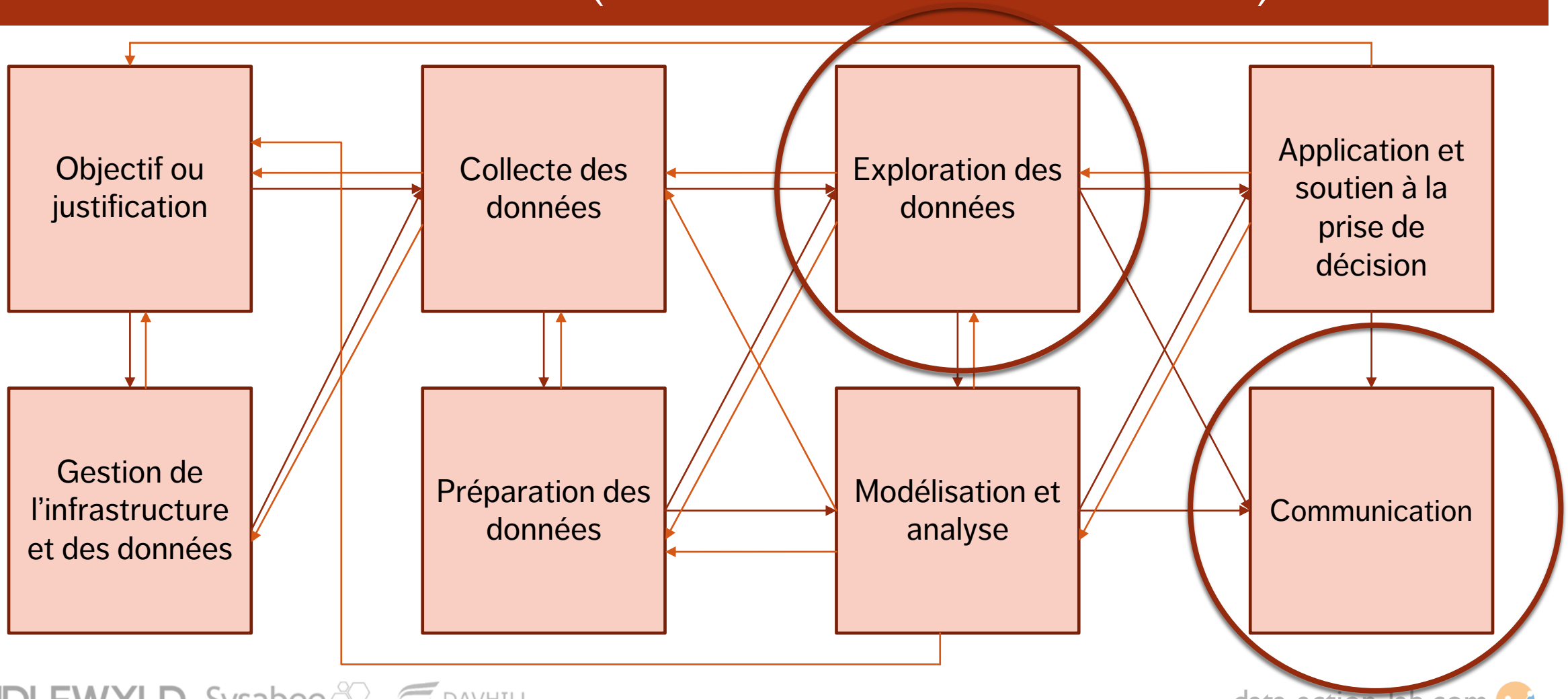
Sur le plan **informatique :** concevoir des simulations informatiques pour mieux comprendre ce qui se passe

Exploration/visualisation des données : la nouvelle approche

« L’horizon des découvertes n’est plus limité par la collecte et le traitement des données, mais plutôt par leur gestion, leur analyse et leur visualisation. »

@DamianMingle

LE PROCESSUS D'ANALYSE (DANS TOUT SON DÉSORDRE)



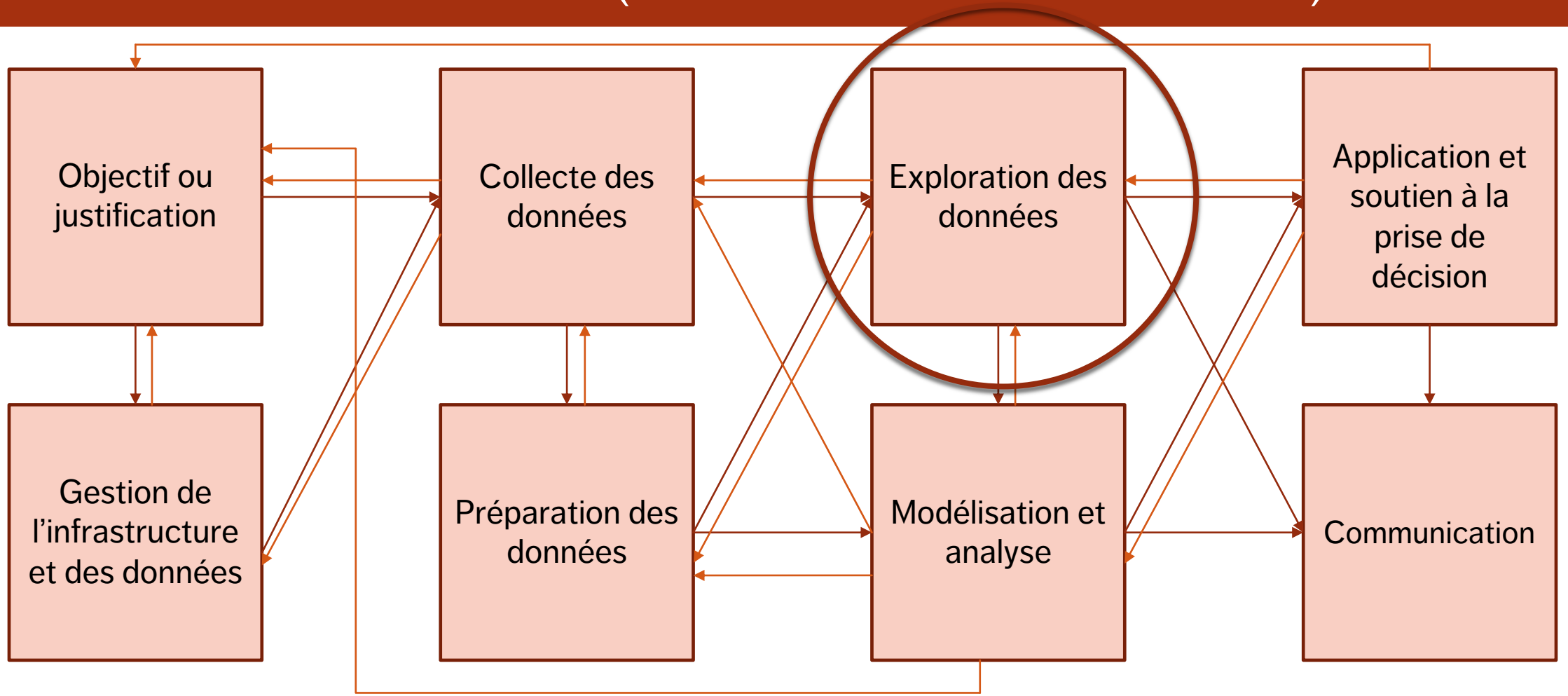
APERÇU

1. Exploration des données
2. Visualisation des données avant l'analyse
3. Visualisation des données après l'analyse
4. Catalogue de visualisations
5. Tableau d'honneur et tableau d'horreur
6. Structure logique des graphiques
7. Introduction aux tableaux de bord

EXPLORATION DES DONNÉES

EXPLORATION ET VISUALISATION DES DONNÉES

LE PROCESSUS D'ANALYSE (DANS TOUT SON DÉSORDRE)



QUESTIONS DE BASE

Quel système est représenté par vos données – objets, caractéristiques, relations?

Comment vos données représentent-elles ce système – quel est son modèle?

Qui a créé le jeu de données? Quand? Dans quel but?

À supposer qu'il s'agit d'un fichier bidimensionnel (fichier plat), que représentent les rangées? Que représentent les colonnes?

Avez-vous toute les informations nécessaires (**métadonnées**) pour répondre à ces questions? Où pouvez-vous obtenir davantage d'information?

SOMMAIRE DES DONNÉES SANS VISUALISATION

```
       Cl              NO3              NH4         
 Min.   :  0.222   Min.   : 0.000   Min.   :    5.00  
 1st Qu.: 10.994   1st Qu.: 1.147   1st Qu.:   37.86  
 Median : 32.470   Median : 2.356   Median :  107.36  
 Mean   : 42.517   Mean   : 3.121   Mean   :  471.73  
 3rd Qu.: 57.750   3rd Qu.: 4.147   3rd Qu.:  244.90  
 Max.   :391.500   Max.   :45.650   Max.   :24064.00  
 NA's   :16        NA's   :2        NA's   :2         

    season    
 Length:340   
 Class :character      autumn spring summer winter
 Mode  :character          80     84     86     90
```

VISUALISATION DES DONNÉES AVANT L'ANALYSE

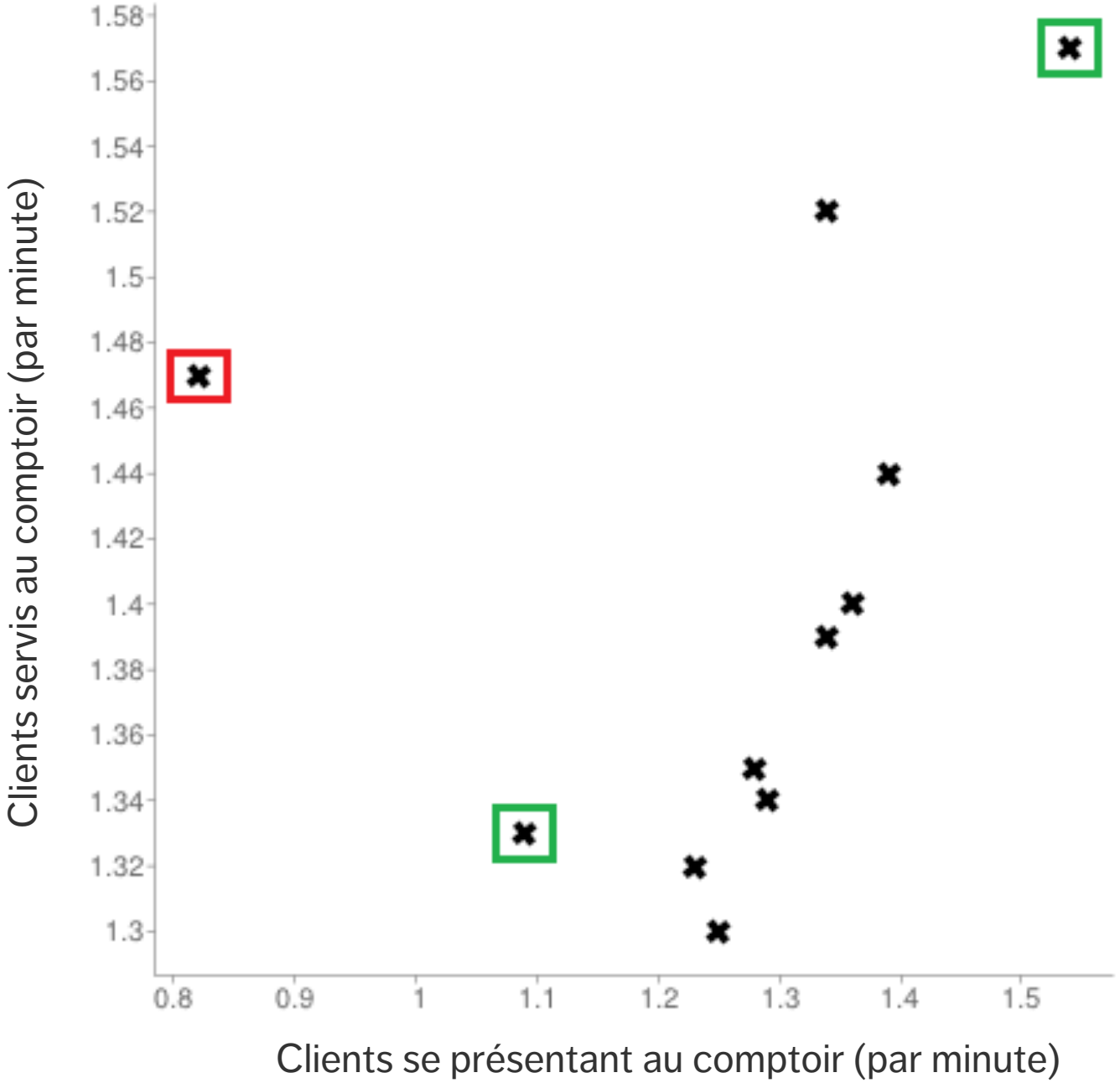
EXPLORATION ET VISUALISATION DES DONNÉES

UTILISATION AVANT L'ANALYSE

La visualisation des données peut être utile pour préparer l'analyse :

- **Détection des anomalies**
  Entrées invalides, valeurs manquantes, données aberrantes
- **Mise en forme des transformations de données**
  Compartimentage, uniformisation, transformations de Box-Cox, transformations de style analyse en composantes principales (ACP)
- **Familiarisation avec les données**
  L'analyse de données est un art, analyse exploratoire
- **Détection de structures de données cachées**
  Agrégation, associations, motifs renseignant la prochaine étape de l'analyse

[Jeu de données personnel]

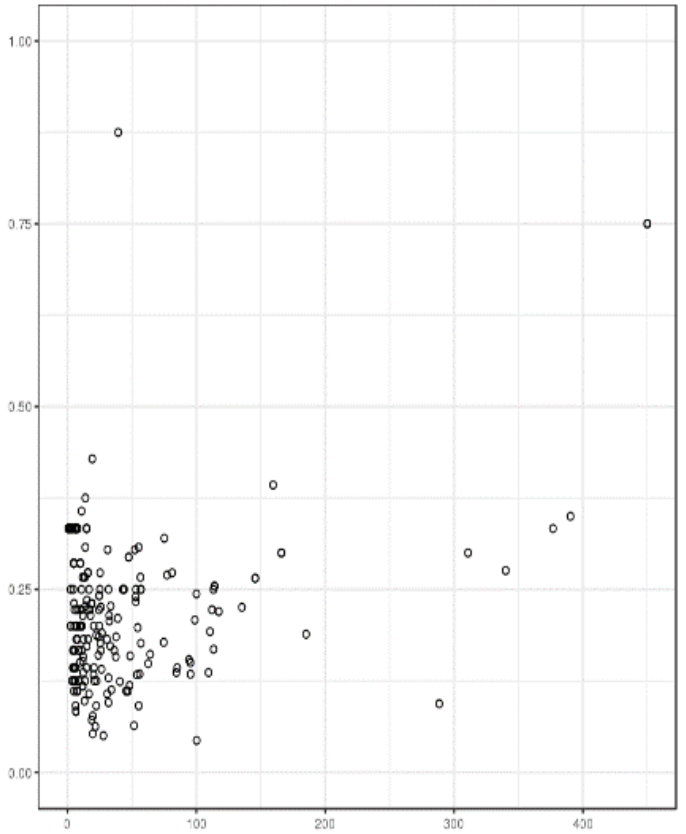


[http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff]

REPRÉSENTATION D'OBSERVATIONS À VARIABLES MULTIPLES

Deux variables peuvent être représentées selon la position sur un plan. Des facteurs additionnels :

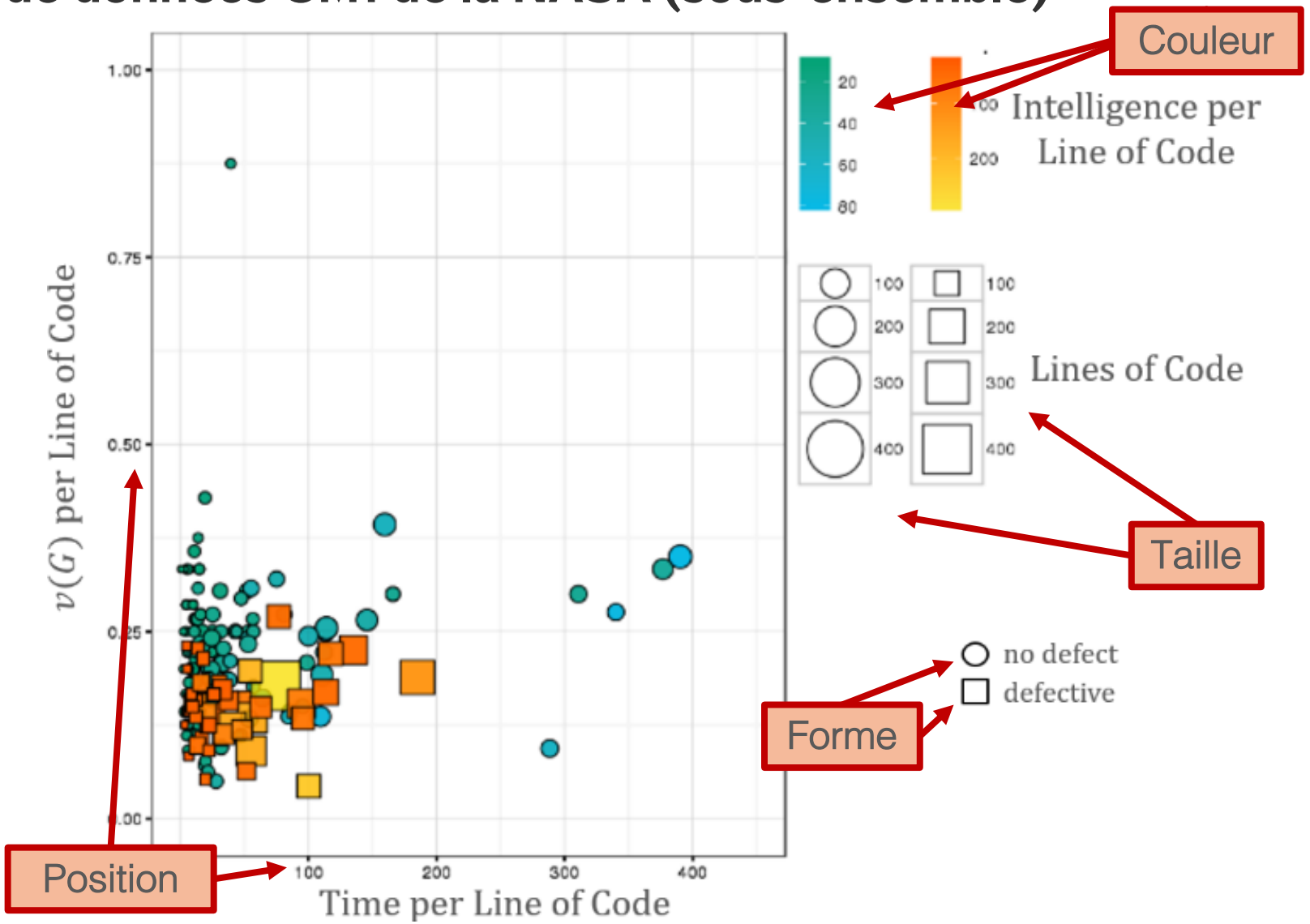
- taille
- couleur
- valeur
- texture
- orientation d'une droite
- forme
- (mouvement?)



Jeu de données CM1 de la NASA (sous-ensemble)

[http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff]

Jeu de données CM1 de la NASA (sous-ensemble)



VISUALISATIONS COURANTES POUR L'EXPLORATION DES DONNÉES

Graphique linéaire/graphique à traits/droite numérique

Histogramme

(Diagramme à moustaches)

Graphique linéaire

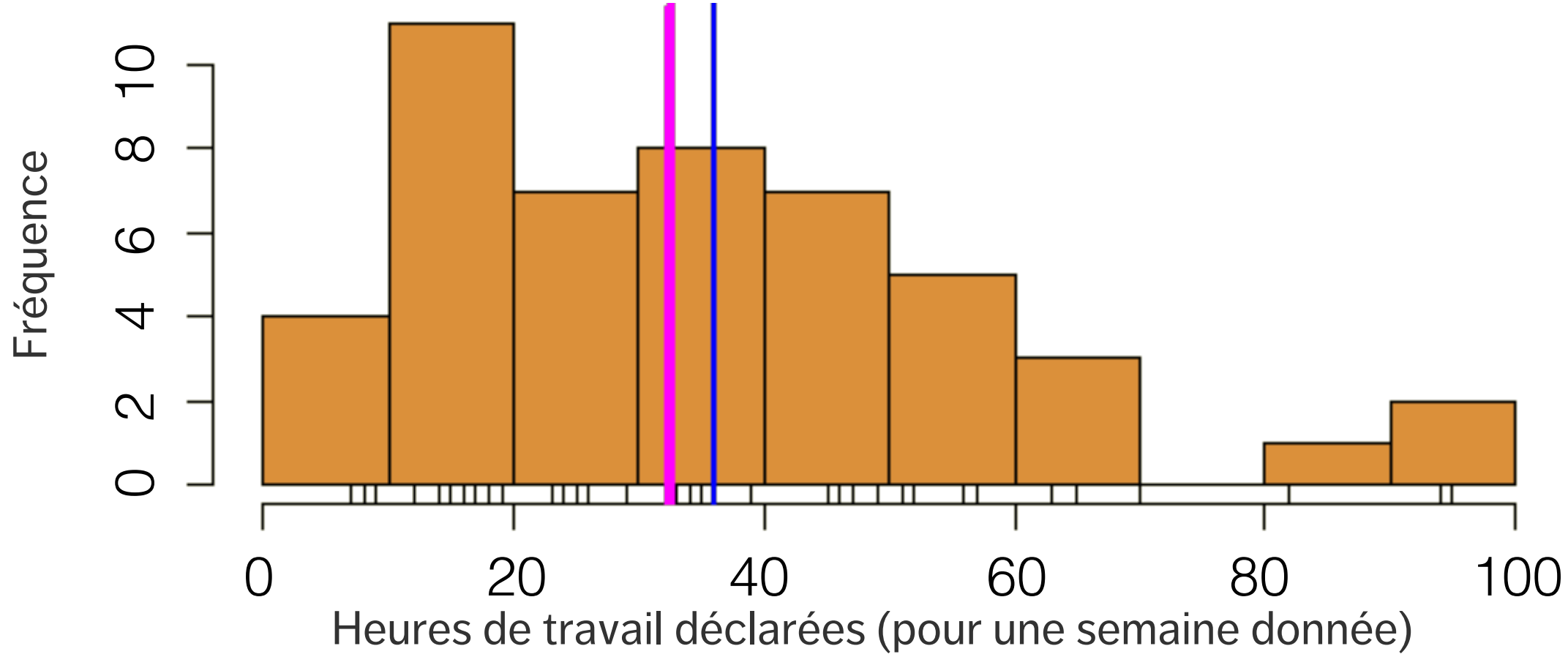
Diagramme en bâtons

Nuage de points

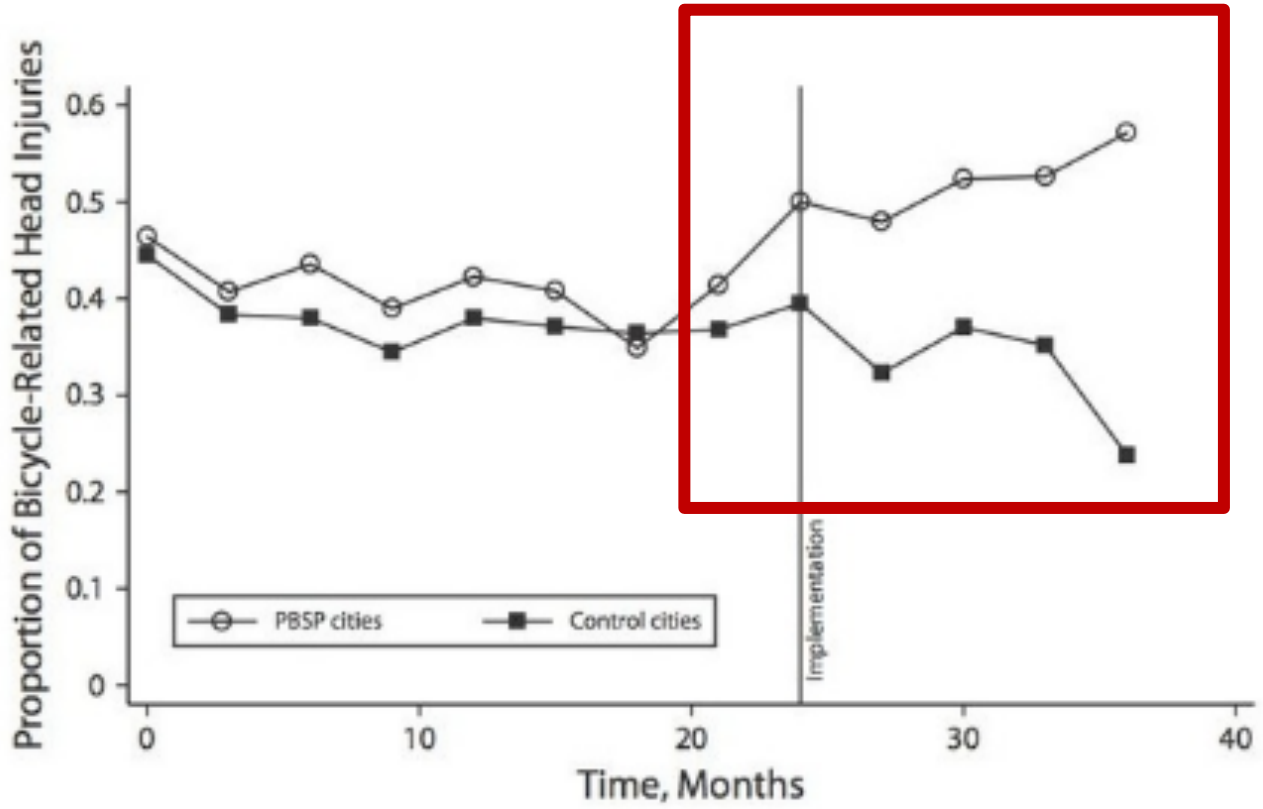
GRAPHIQUE À TRAITS

- Les trous le long de la droite numérique indiquent l'absence de ces valeurs dans le jeu de données.
- Rappel : Ce graphique ne représente pas nécessairement les données du jeu de données dans l'ordre – il s'agit d'une droite numérique, elle n'indique donc que les valeurs présentent dans le jeu.
- Les valeurs identiques se chevauchent.

HISTOGRAMME



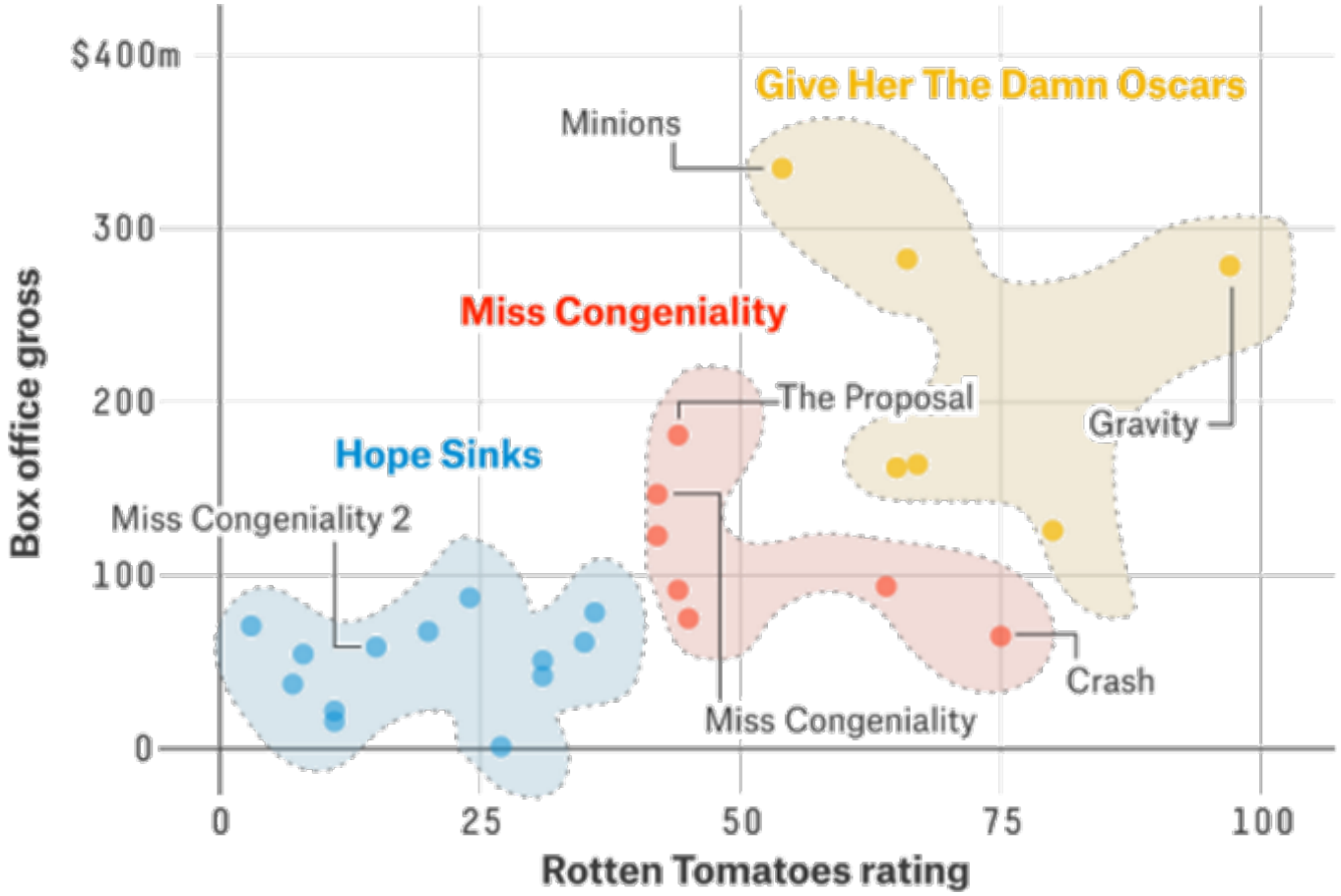
GRAPHIQUE LINÉAIRE

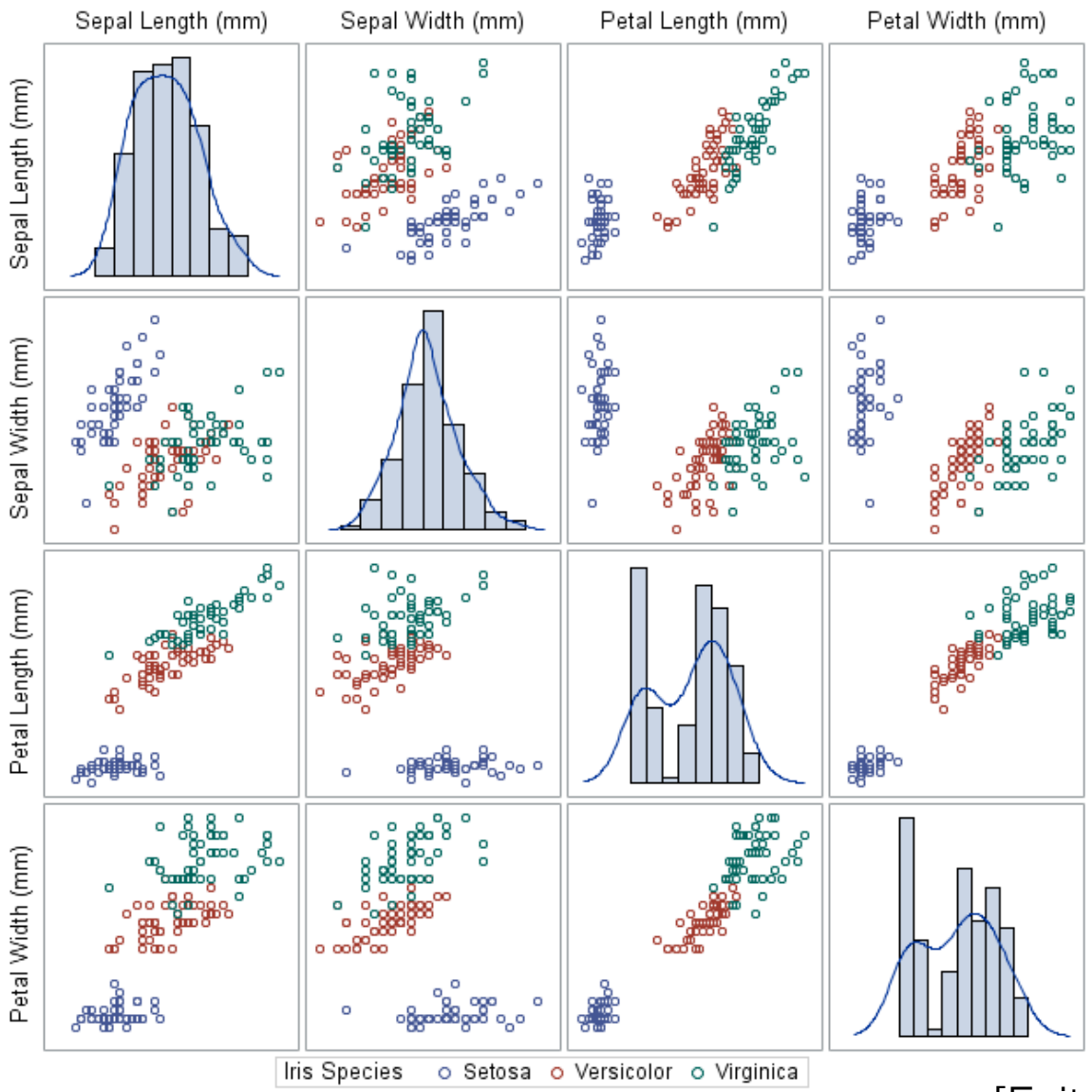


Proportion des blessures à la tête parmi l'ensemble des blessures liées au cyclisme dans les villes possédant un service de vélopartage et les villes de référence, en fonction de la date de l'intervention (ligne verticale); Amérique du Nord.

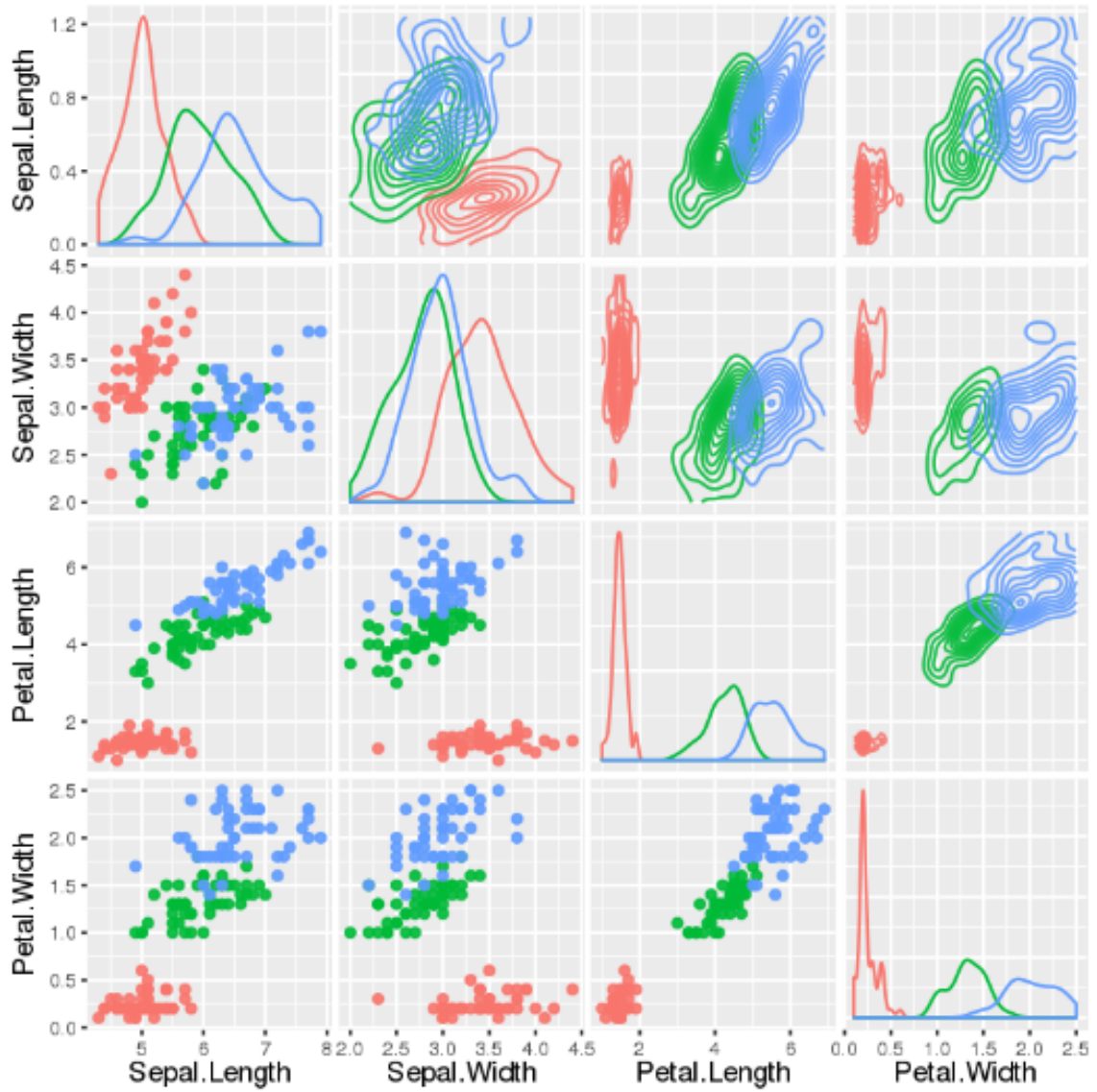
[Graves et coll., *Am.J.Phys.Health*, 2014]

NUAGE DE POINTS

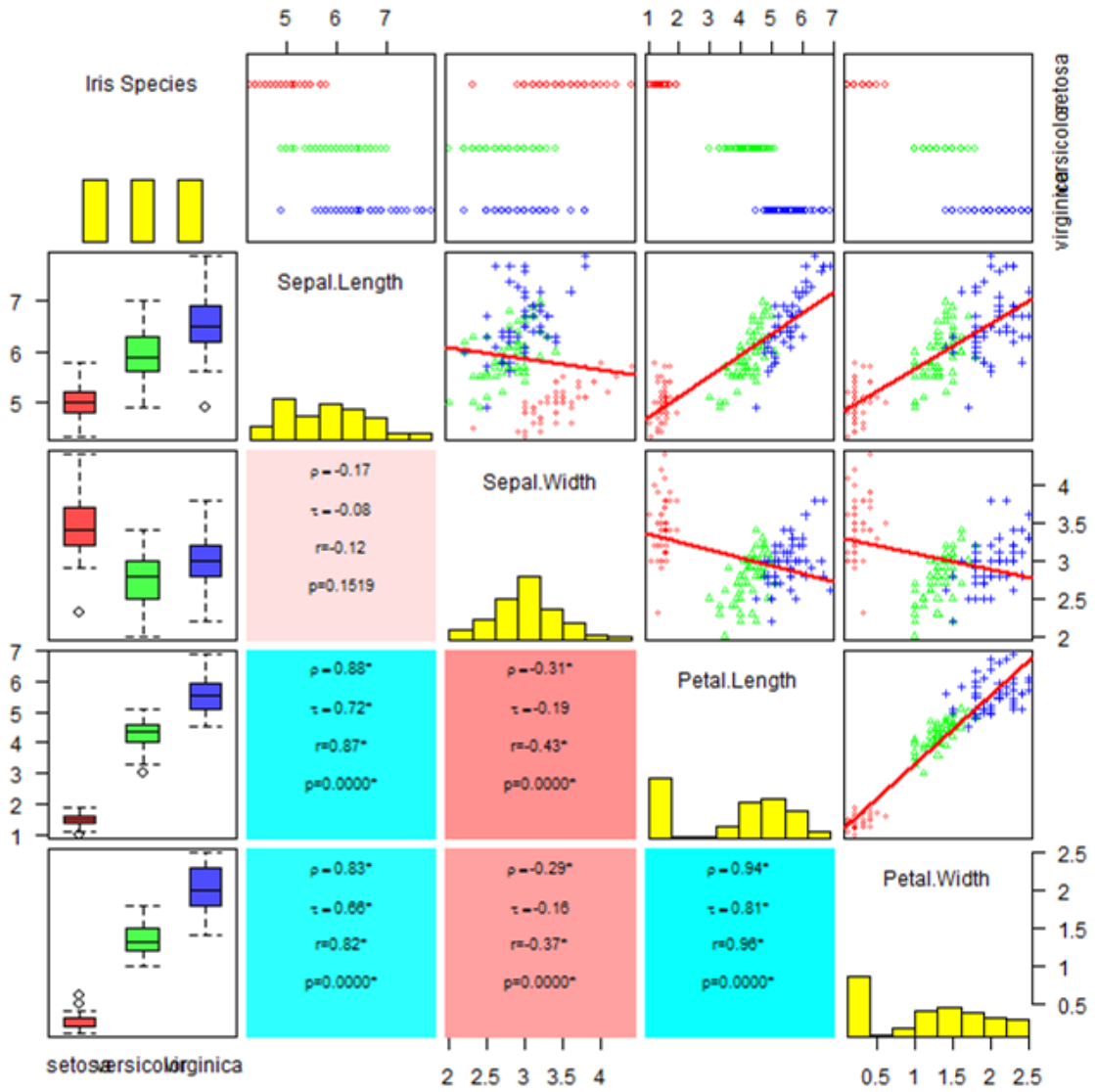




[Fait avec la procédure sgscatter de SAS]



[Fait avec la commande ggpairs de R]



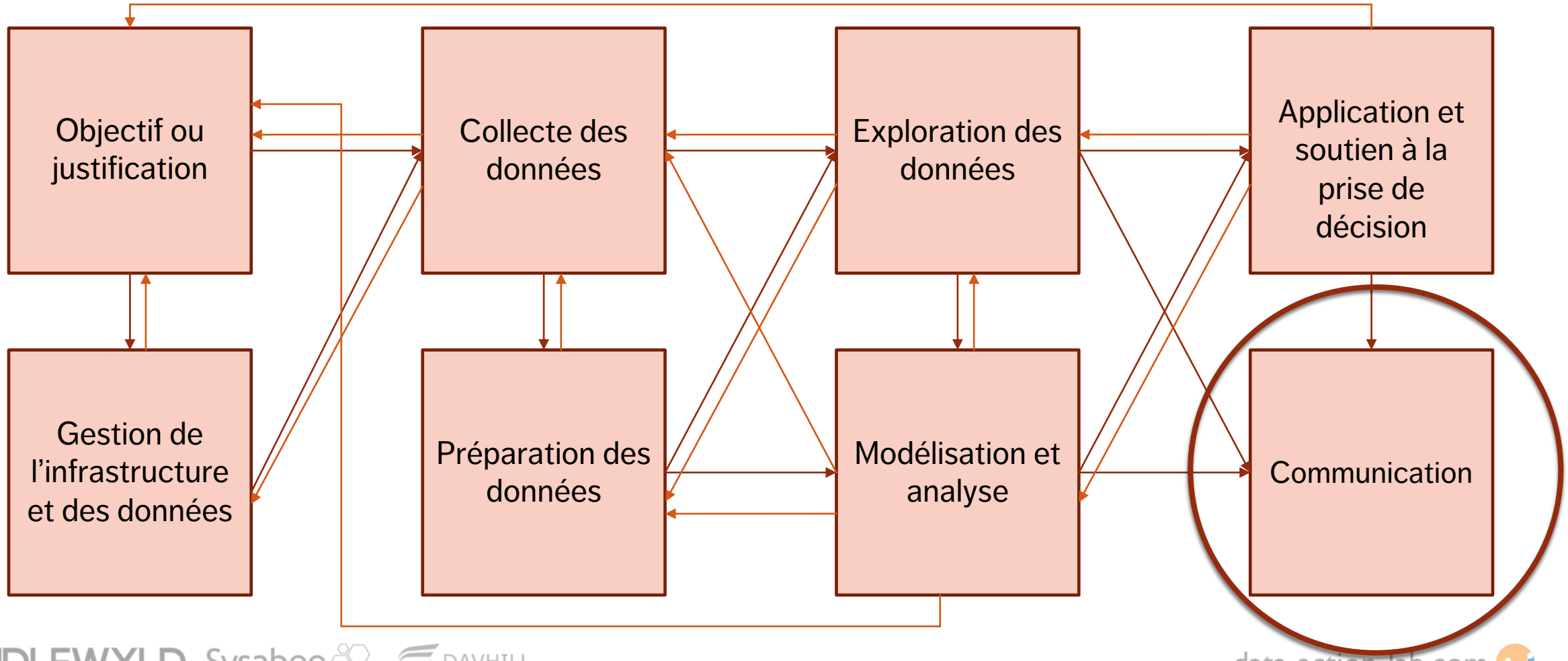
Ce graphique commence-t-il à être trop chargé?

[Auteur inconnu, trouvé en ligne]

VISUALISATION DES DONNÉES APRÈS L'ANALYSE

EXPLORATION ET VISUALISATION DES DONNÉES

LE PROCESSUS D'ANALYSE (DANS TOUT SON DÉSORDRE)



[E. Tufte, *Beautiful Evidence*]

PRINCIPES FONDAMENTAUX DU DESIGN ANALYTIQUE

Le raisonnement et la communication sont inter-reliés dans nos vies et notre univers causal, dynamique, et mulivarié.

La **symétrie** dans les visualisations : les consommateurs devraient rechercher exactement ce que les producteurs offrent, soit :

- des comparaisons pertinentes
- des réseaux causaux et leur structure sous-jacente
- des relations multivariées
- des données intégrées et pertinentes
- une documentation transparente
- un accent sur le contenu

ACCESSIBILITÉ

On peut traduire un tableau en braille assez facilement, mais ce n'est pas toujours possible pour un graphique.

L'une des solutions peut être de décrire les caractéristiques et les structures de la visualisation… **à condition de pouvoir les repérer**.

Les analyses doivent produire des visualisations claires et pertinentes, mais ils doivent également les décrire d'une façon qui permet d'en « saisir » la portée.

ACCESSIBILITÉ

Les analystes doivent avoir compris tous les éléments d'information transmis, ce qui n'est pas nécessairement réaliste.

Perception des données :

- représentations texturées
- conversion texte-parole
- utilisation de sons ou de musique
- représentations odorantes ou axées sur le goût (?!?)

[J.P. Blackard]

INFOGRAPHIE

Crée pour raconter une **histoire** (**subjectivité**)

Cible un public **précis**

Autonome et indépendante

La conception graphique est un aspect clé

Ne peut généralement pas être réutilisée avec d'autres données

Peut comprendre de l'information **impossible à quantifier**

VISUALISATION DES DONNÉES

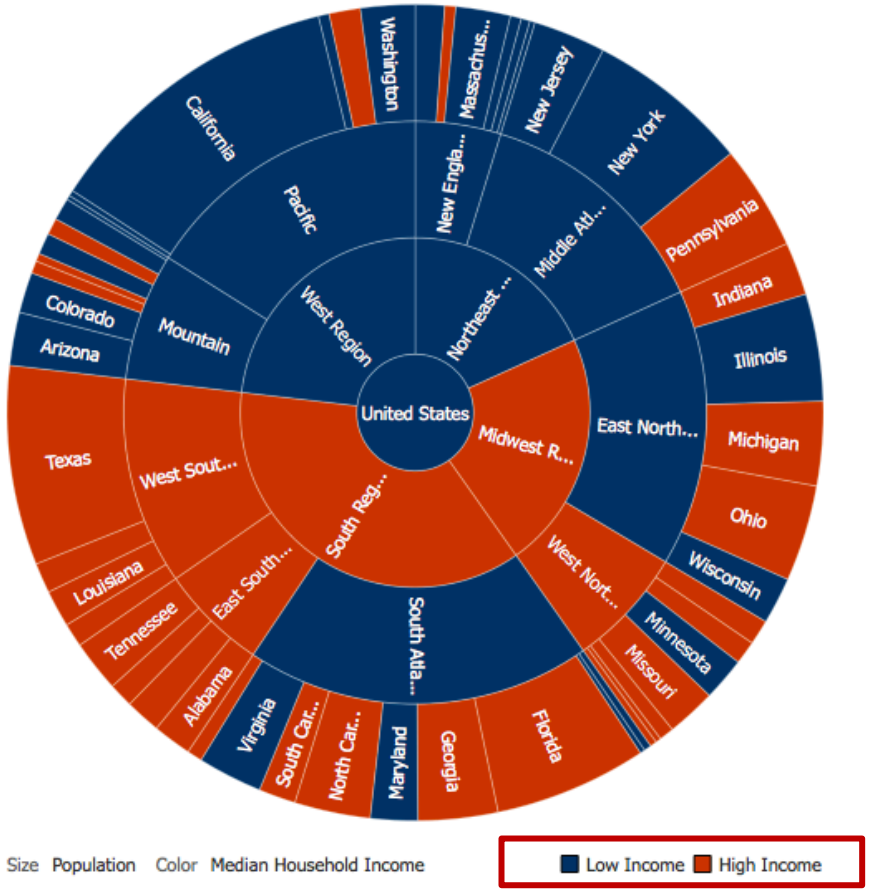
Une **méthode** et un objet à la fois (**objectivité**)

Met généralement l'accent sur des données **quantifiables**

Sert à extraire le sens des données ou à les rendre **accessibles** (les jeux de données peuvent être imposants et difficiles à manipuler)

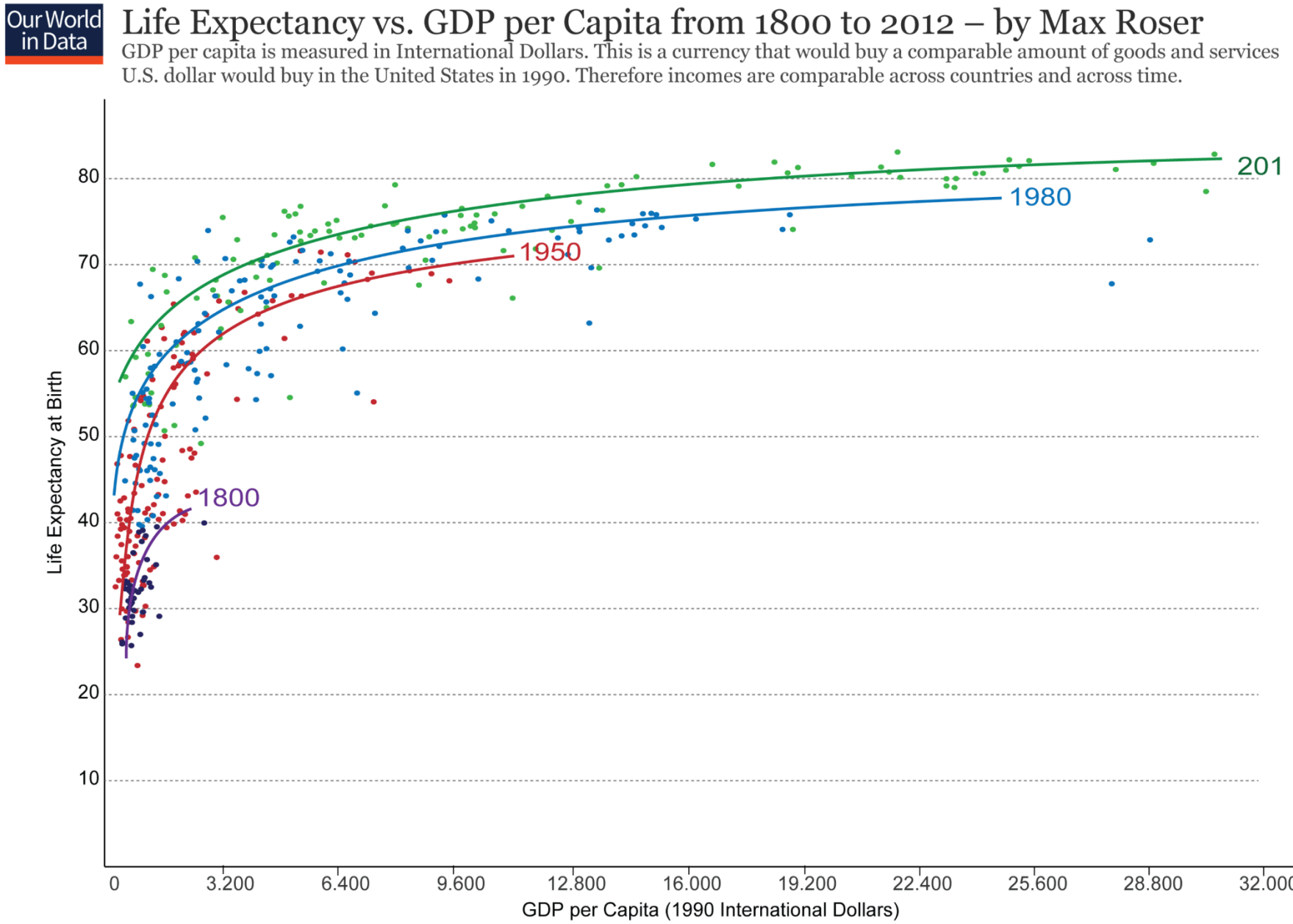
Peut être générée automatiquement

L'apparence est moins importante que **l'information** transmise par les données



?

[https://ourworldindata.org/life-expectancy]



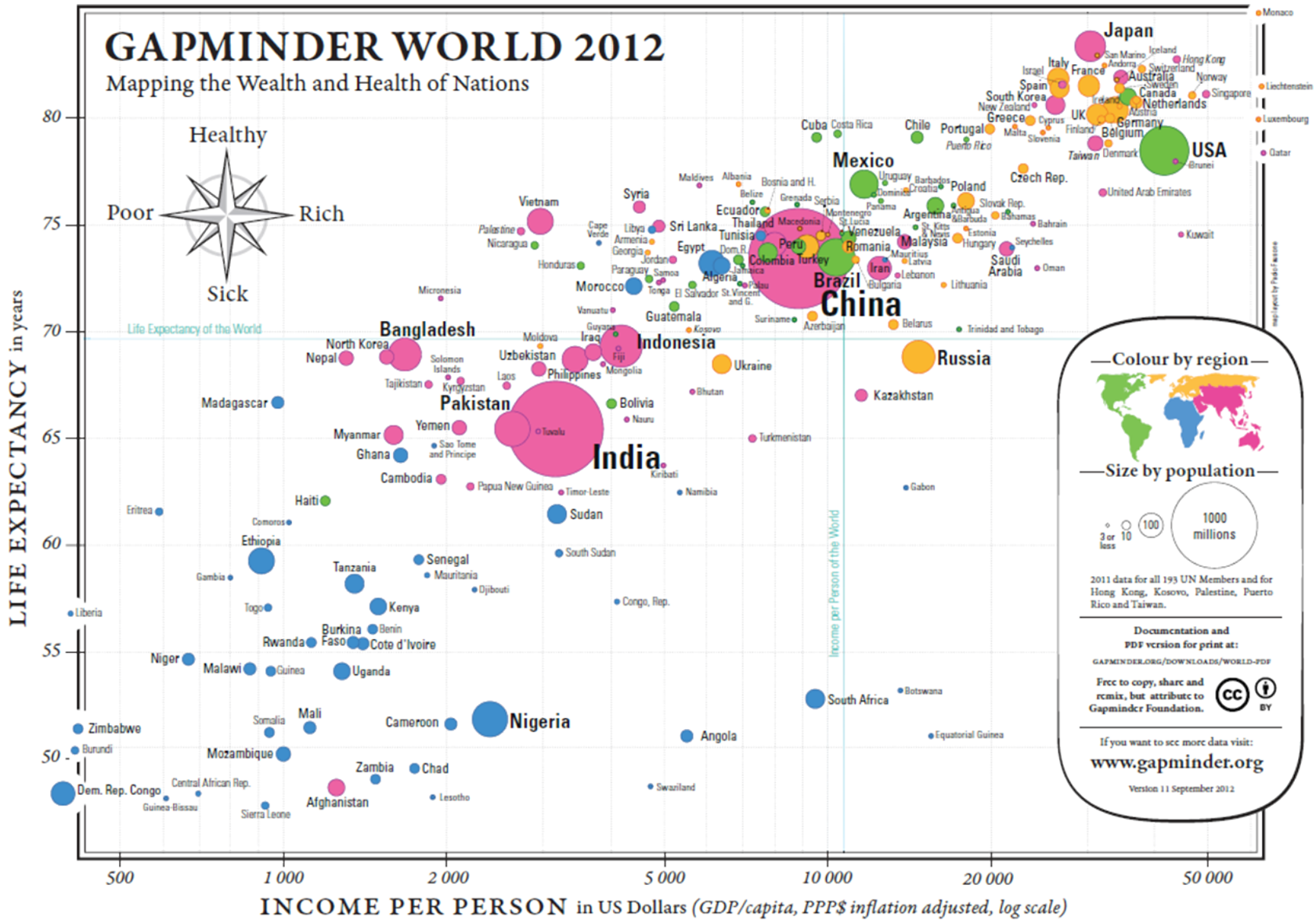
Ce graphique représente la relation entre l'espérance de vie et le PIB par habitant.

En général, plus le PIB d'un pays élevé, meilleure est son espérance de vie.

La corrélation semble suivre une courbe **logarithmique** : l'augmentation de l'espérance de vie par unité additionnelle de PIB est de moins en moins importante à mesure que le PIB augmente.

Data sources: Data on life expectancy are from Gapminder.org; data on GDP per capita are from the 'New Maddison Project Database'.
The interactive data visualisation is available at OurWorldinData.org. There you find the raw data and more visualisations on this topic.

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



PRÉSENTATION DES RÉSULTATS DE L'ANALYSE

Les graphiques devraient être **clairs** et **attrayants**.

Ce ne sont pas toutes les jolies images qui ont une histoire à raconter, mais s'il est impossible de raconter une histoire à l'aide d'une jolie image, peut-être qu'il est temps de revoir l'histoire…

De nouvelles techniques de représentation graphique apparaissent régulièrement – il est trop tôt pour déterminer lesquelles résisteront à l'épreuve du temps.

Il ne faut pas avoir peur d'essayer quelque chose de nouveau tant que cela permet de **transmettre l'information souhaitée**.

TRAITEMENT VISUEL

La perception est **fragmentée** – les yeux sont constamment en mode balayage.

Les centres de traitement visuel sont constamment à la recherche de motifs.

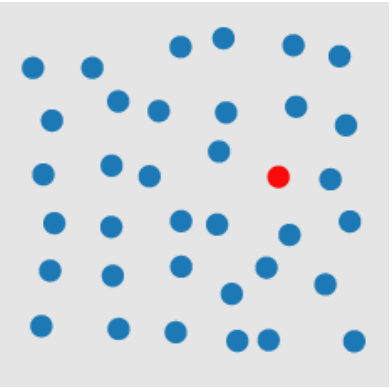
- **Traitement préattentif :** rapide , instinctif, efficace, superficiel, collecte d'information et détection de motifs.

  caractéristiques → motifs → objets

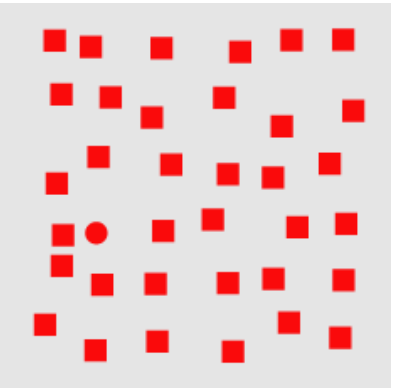
- **Traitement attentif :** lent, délibéré, focalisé, découverte de caractéristiques à l'intérieur des motifs.

  objets → motifs → caractéristiques

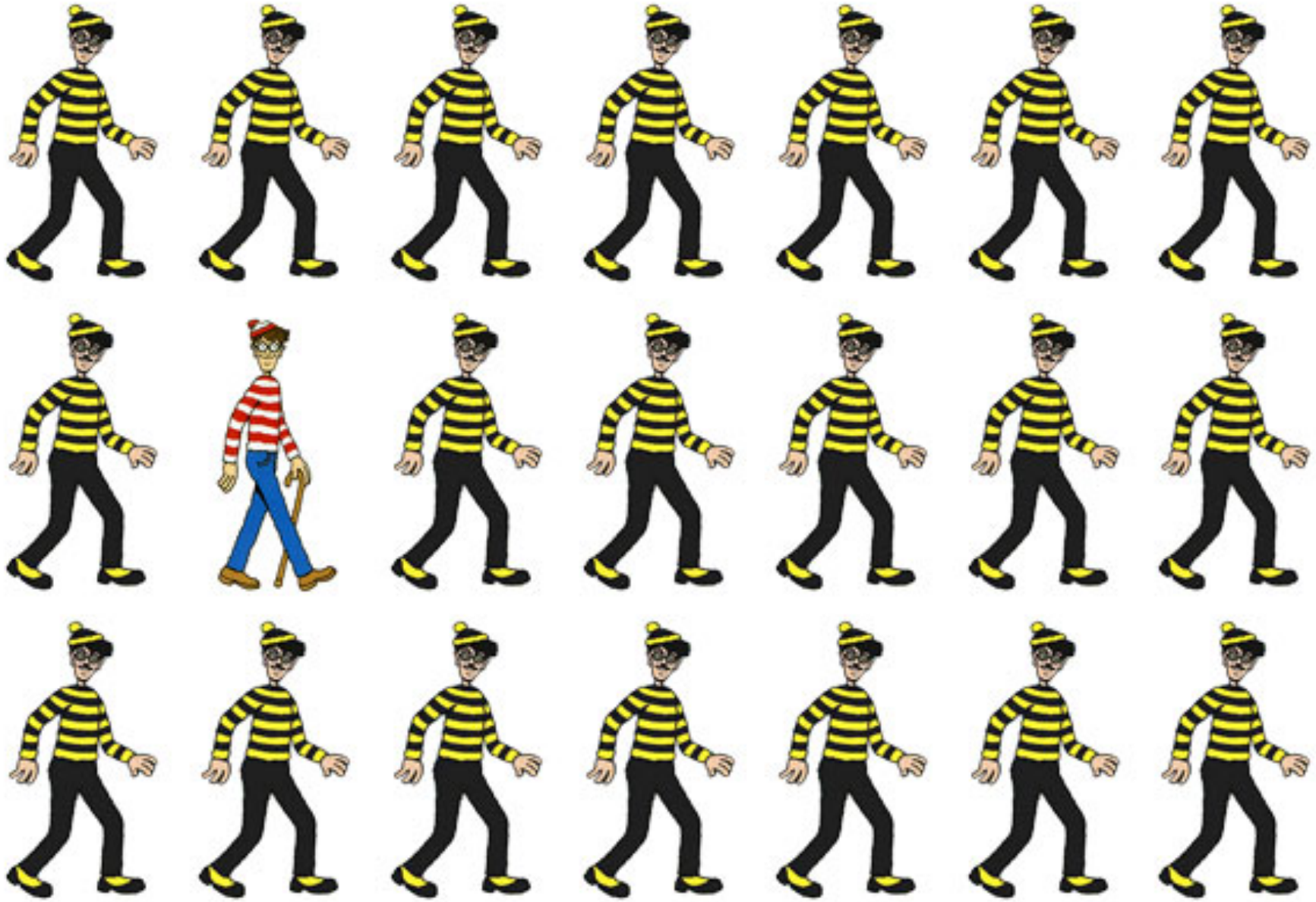
Vision préattentive



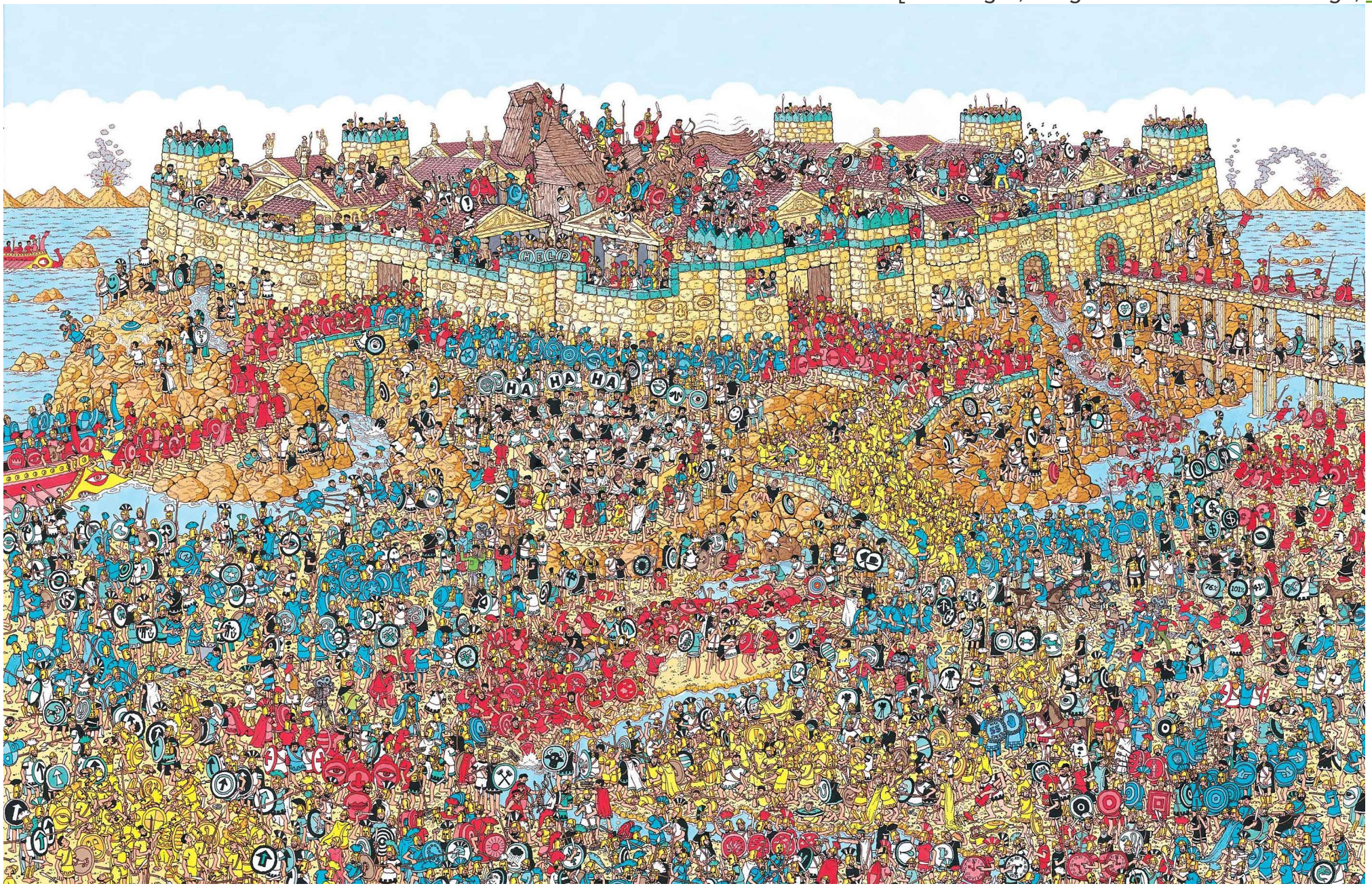
Vision attentive



[L. Groeger, *A Big Article About Wee Things*, ProPublica.com]



[L. Groeger, *A Big Article About Wee Things*, ProPublica.com]



[http://flowingdata.com/tag/pitfalls]

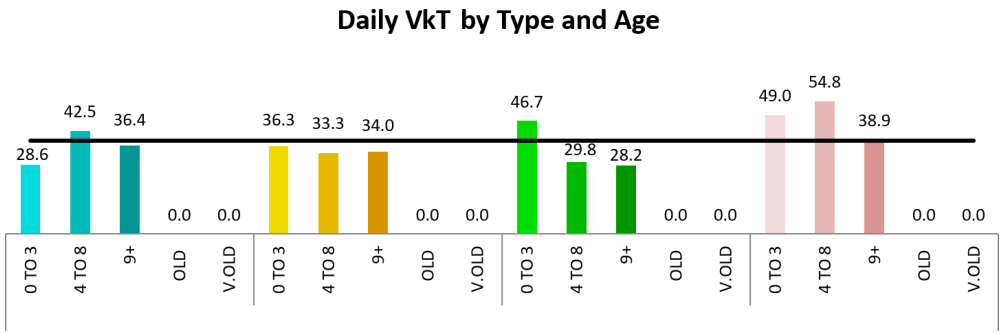
RÈGLES DE BASE

1. Examiner les données

Aberrations, pics, anomalies.

2. Expliquer l'encodage

Ne pas présumer que le lecteur comprend la signification de tous les éléments.



3. Étiqueter les axes

Il est important d'afficher l'échelle.

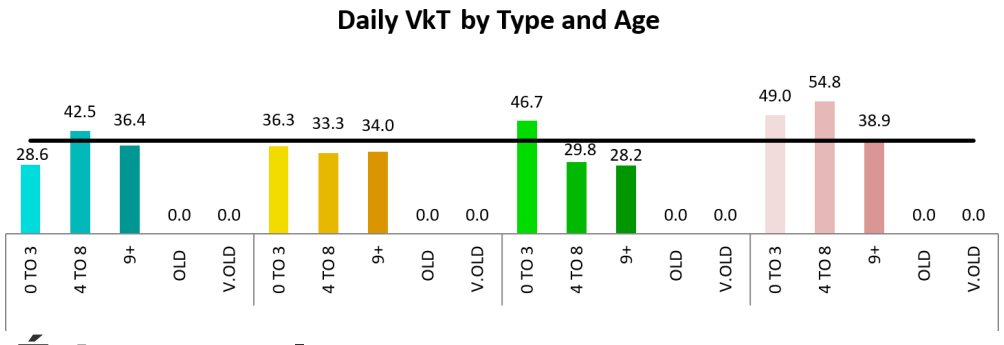
RÈGLES DE BASE

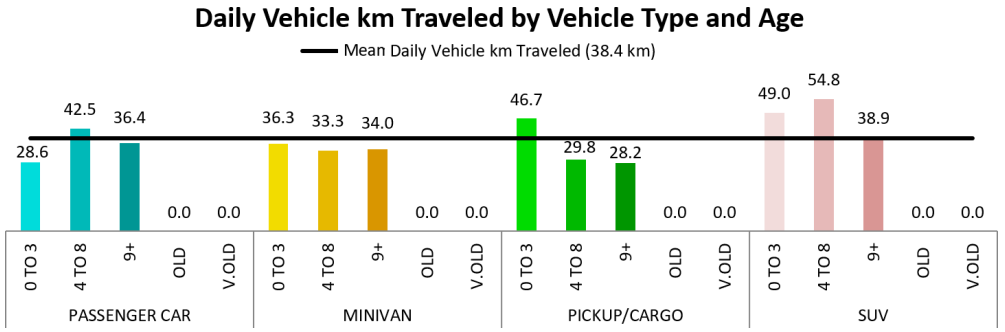
1. Examiner les données

Aberrations, pics, anomalies.

2. Expliquer l'encodage

Ne pas présumer que le lecteur comprend la signification de tous les éléments.





3. Étiqueter les axes

Il est important d'afficher l'échelle.

[http://flowingdata.com/tag/pitfalls]

RÈGLES DE BASE

4. Afficher les unités

Ne pas forcer le lecteur à faire des suppositions.

Rayon de 3,5

superficie de 3,5

5. Respecter les principes géométriques

L'échelle des cercles et des formes en deux dimensions est définie par leur superficie, celle des diagrammes à bâtons, par leur longueur.

6. Indiquer les sources

Éviter tout risque d'accusation de plagiat et permettre aux lecteurs d'en apprendre plus.

7. Penser au public

Une affiche peut contenir plus de texte, mais un diaporama devrait être concis.

[http://flowingdata.com/tag/pitfalls]

RÈGLES DE BASE

4. Afficher les unités

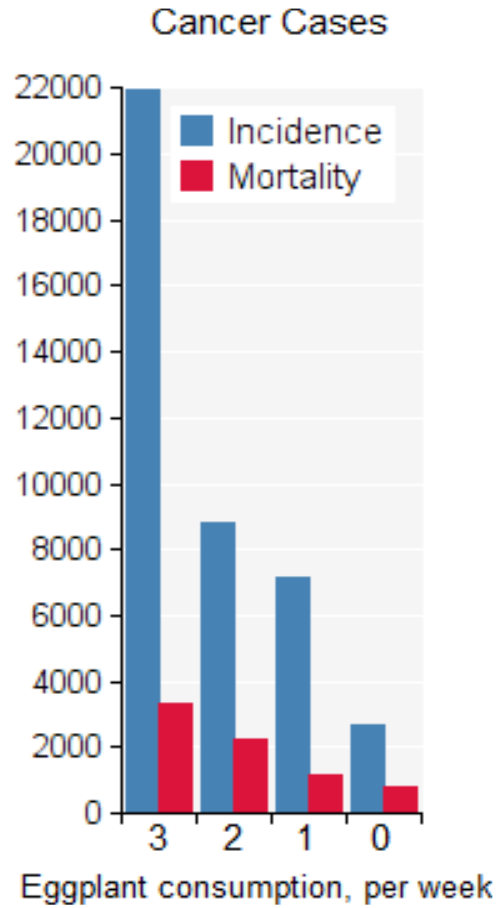
Ne pas forcer le lecteur à faire des suppositions.

Rayon de 3,5

superficie de 3,5

5. Respecter les principes géométriques

L'échelle des cercles et des formes en deux dimensions est définie par leur superficie, celle des diagrammes à bâtons, par leur longueur.



6. Indiquer les sources

Éviter tout risque d'accusation de plagiat et permettre aux lecteurs d'en apprendre plus.

7. Penser au public

Une affiche peut contenir plus de texte, mais un diaporama devrait être concis.

[http://flowingdata.com/tag/pitfalls]

RÈGLES DE BASE

4. Afficher les unités

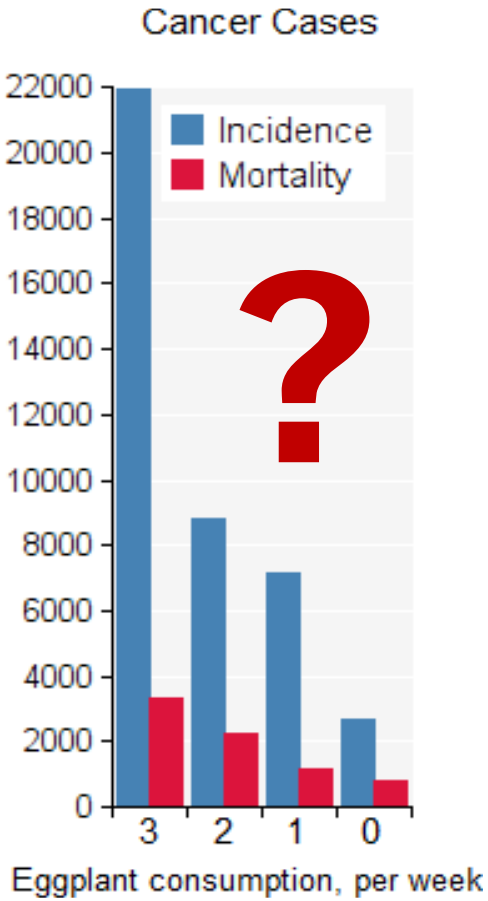
Ne pas forcer le lecteur à faire des suppositions.

Rayon de 3,5

Superficie de 3,5

5. Respecter les principes géométriques

L'échelle des cercles et des formes en deux dimensions est définie par leur superficie, celle des diagrammes à bâtons, par leur longueur.



6. Indiquer les sources

Éviter tout risque d'accusation de plagiat et permettre aux lecteurs d'en apprendre plus.

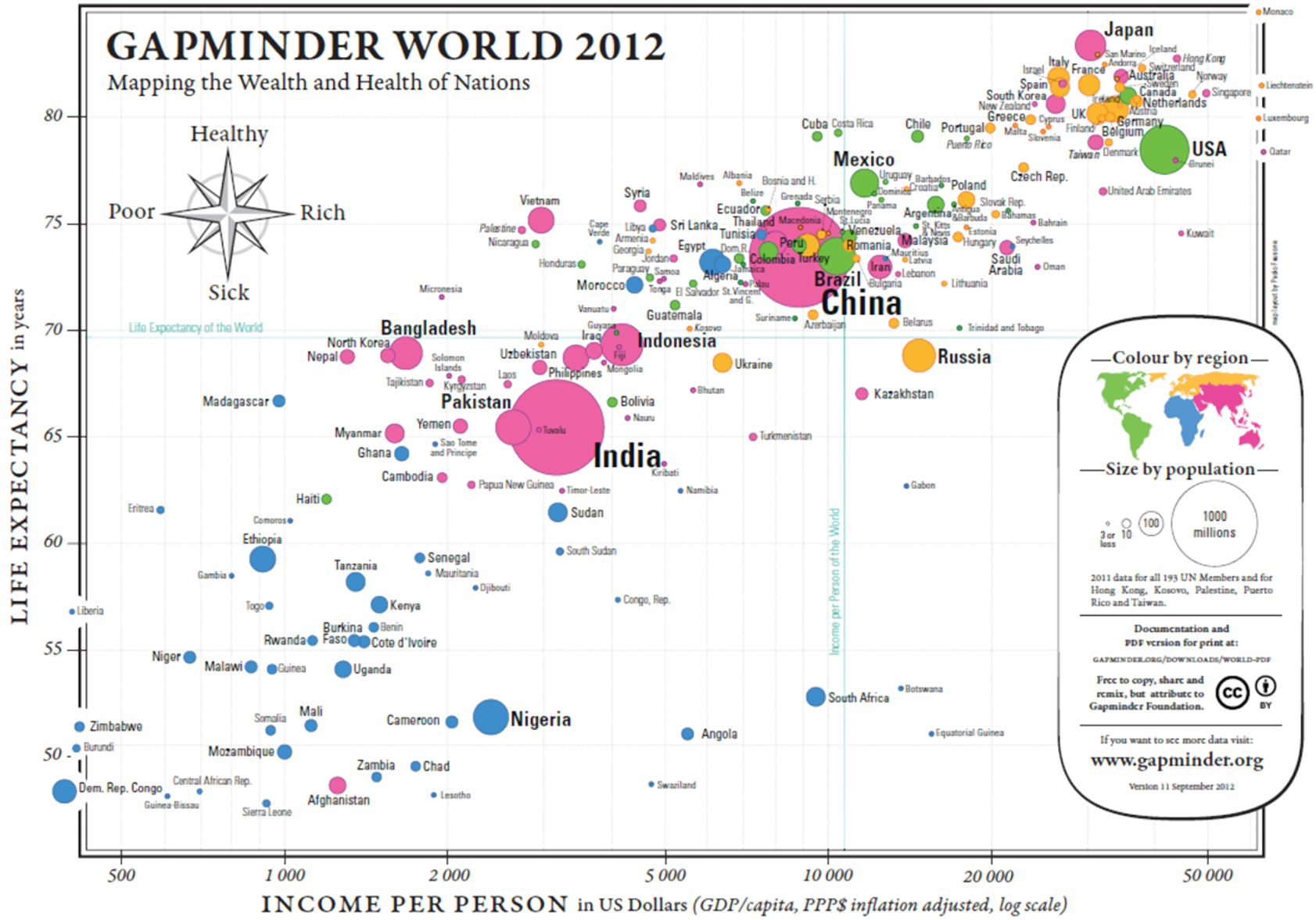
7. Penser au public

Une affiche peut contenir plus de texte, mais un diaporama devrait être concis.

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]

Exercice de groupe

- Comment ce visuel aide-t-il le public à comprendre les données?
- Repérez-vous des motifs intéressants dans ce visuel?



DISCUSSION

Le message passe-t-il? L'intégration des données contribue à transmettre l'information importante.

Dans *La sémiologie graphique*, Bertin affirme que **les variables rétiniennes n'ont pas toutes le niveau d'efficacité** pour relayer ou représenter de l'information. Il peut être nécessaire de faire des essais pour trouver le meilleur choix dans un contexte donné.

L'addition de certain éléments conceptuels peut améliorer la compréhension des données.

La façon dont nous percevons les motifs influence notre interprétation de la représentation des données.

Les représentations de données ne devraient pas reposer sur une méthode de visualisation choisie au hasard. Le résultat variera selon la structure des données et la combinaison des questions étudiées.

CATALOGUE DE VISUALISATION

EXPLORATION ET VISUALISATION DES DONNÉES

[J. Nelson, IDV Solutions]

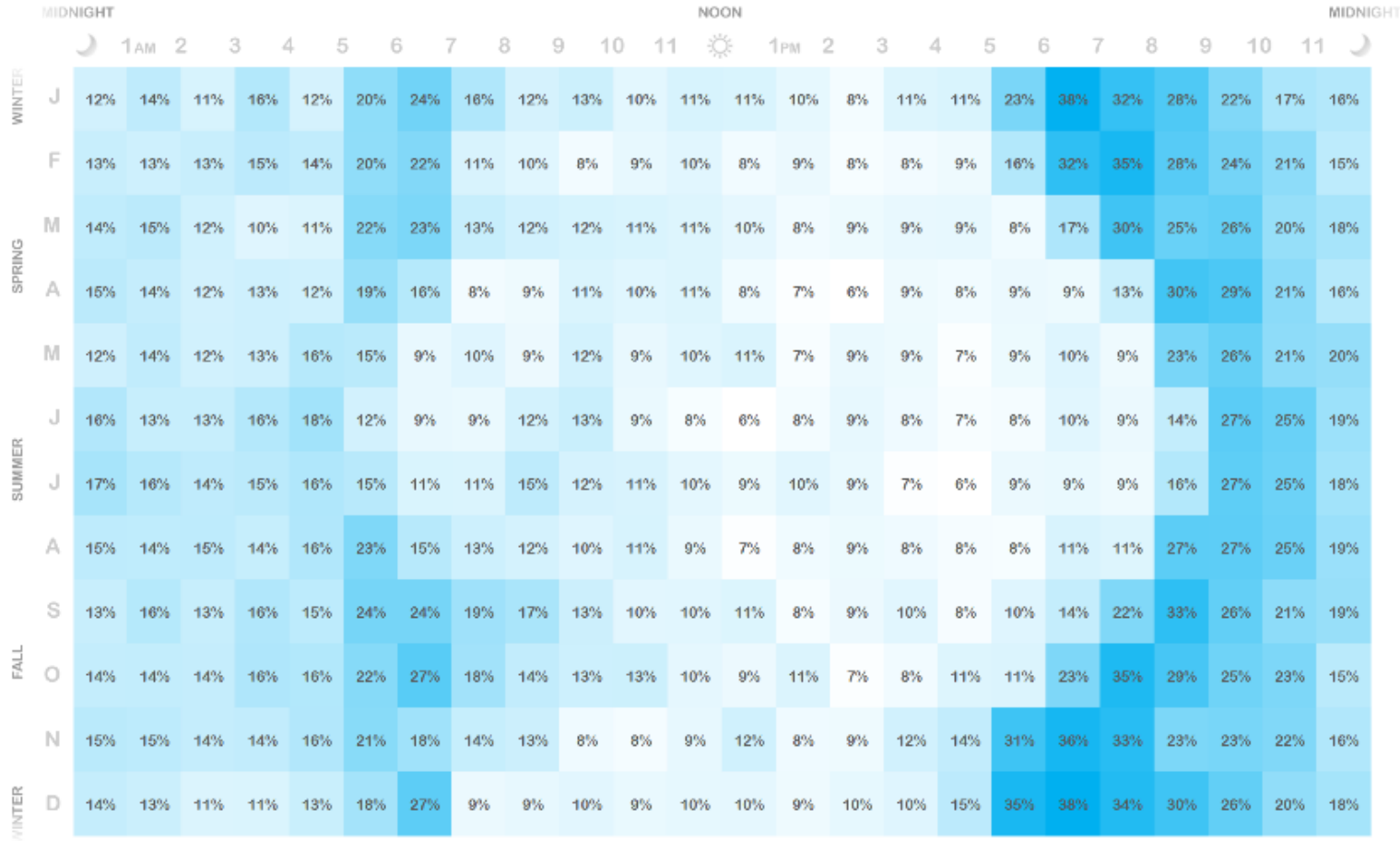
CARTES DE DENSITÉ

L'horizon du risque pour les piétons

Il s'agit du taux d'accidents de la route mortels impliquant des piétons, selon l'heure de la journée, tout au long de l'année.

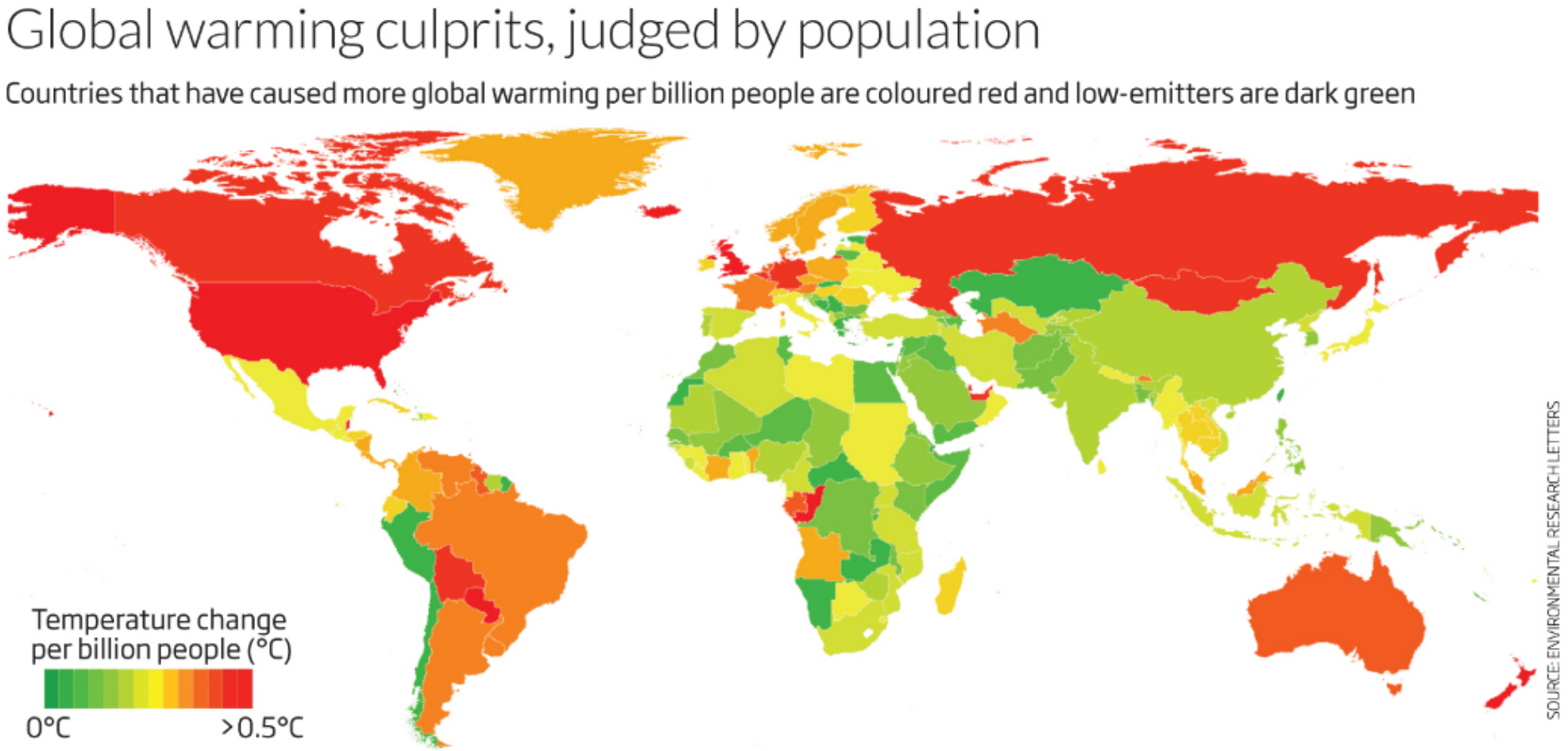
Le déplacement du soleil au fil des saisons se reflète par une courbe de risque élevé qui fait écho à la courbature de la terre même (**Remarque :** ???).

Source : Fatality Analysis Reporting System (NHTSA 2006-2010).



[New Scientist, janvier 2014]

CARTES GÉOGRAPHIQUES



[New Scientist, janvier 2014]

CARTES GÉOGRAPHIQUES

[Upside Down Maps]

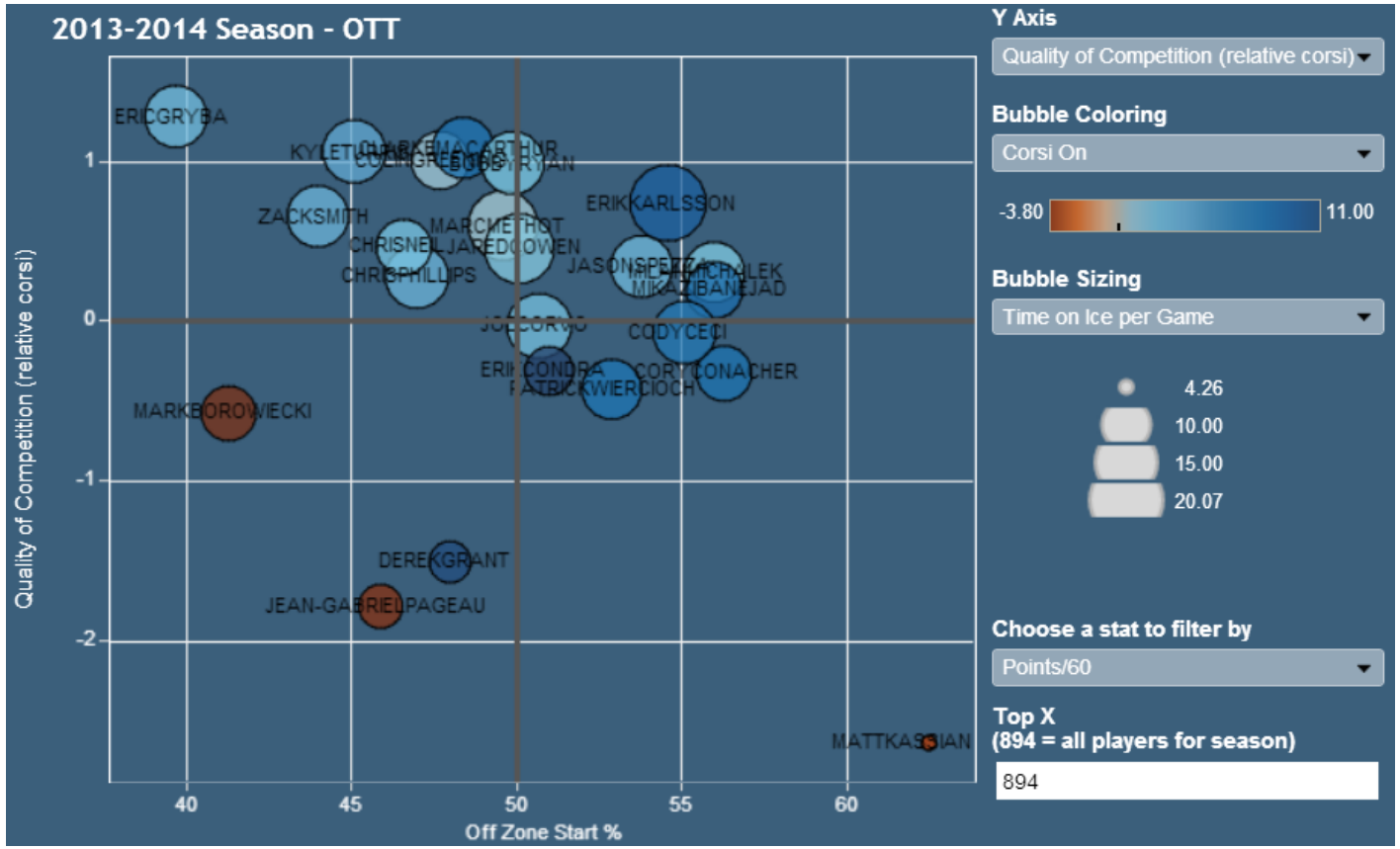
CARTES

La plupart des gens ont l'habitude d'utiliser des cartes géographiques, elles sont donc généralement faciles à interpréter.

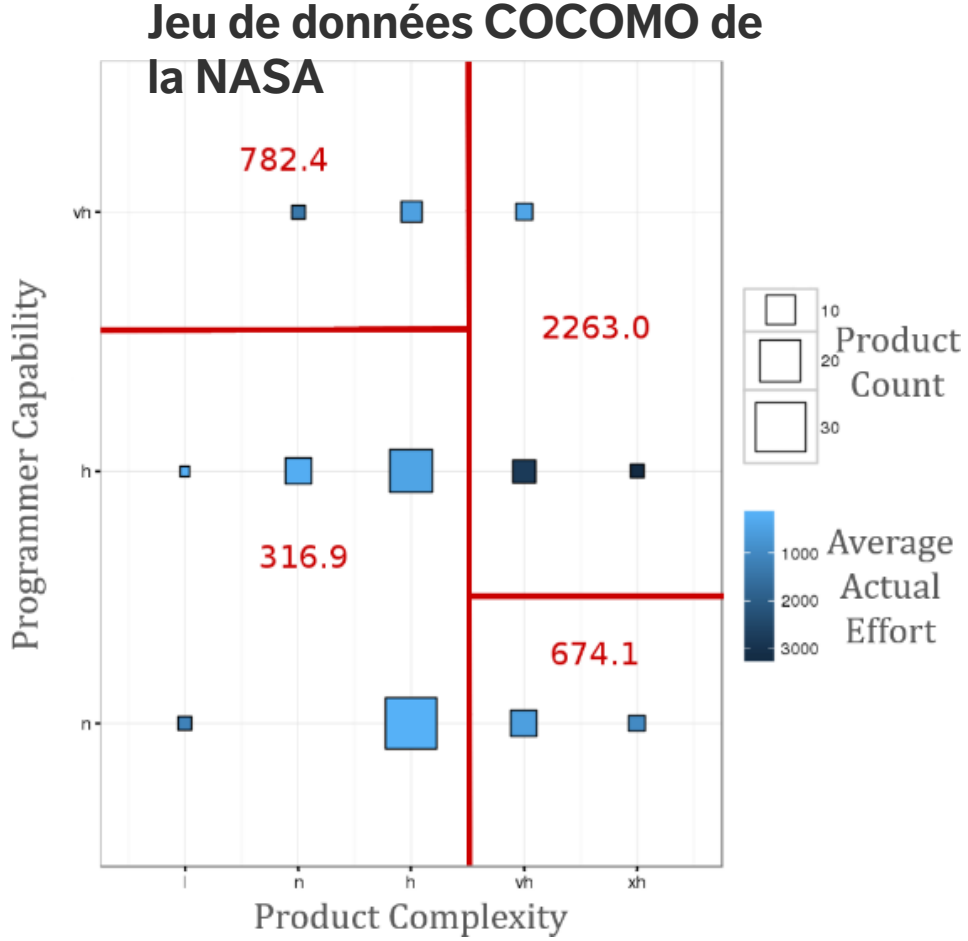
Ces cartes peuvent avoir un effet marquant quand la visualisation des données produit un **résultat inattendu**

- qui masque des renseignements importants
- ou qui correspond à un manque de renseignements importants
- ou qui change la perspective du lecteur.

GRAPHIQUE À BULLES

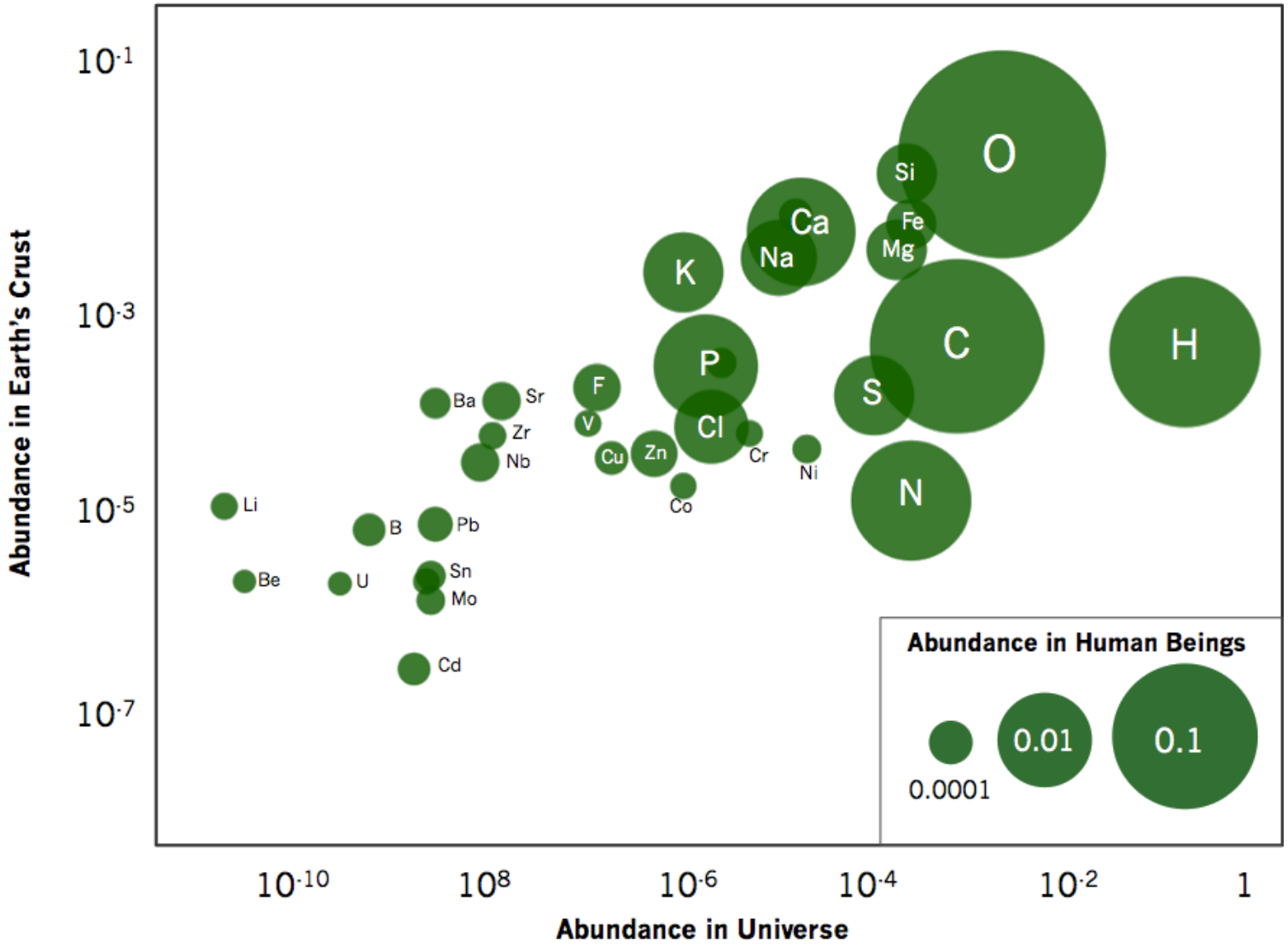


Utilisation des joueurs (Sénateurs d'Ottawa)



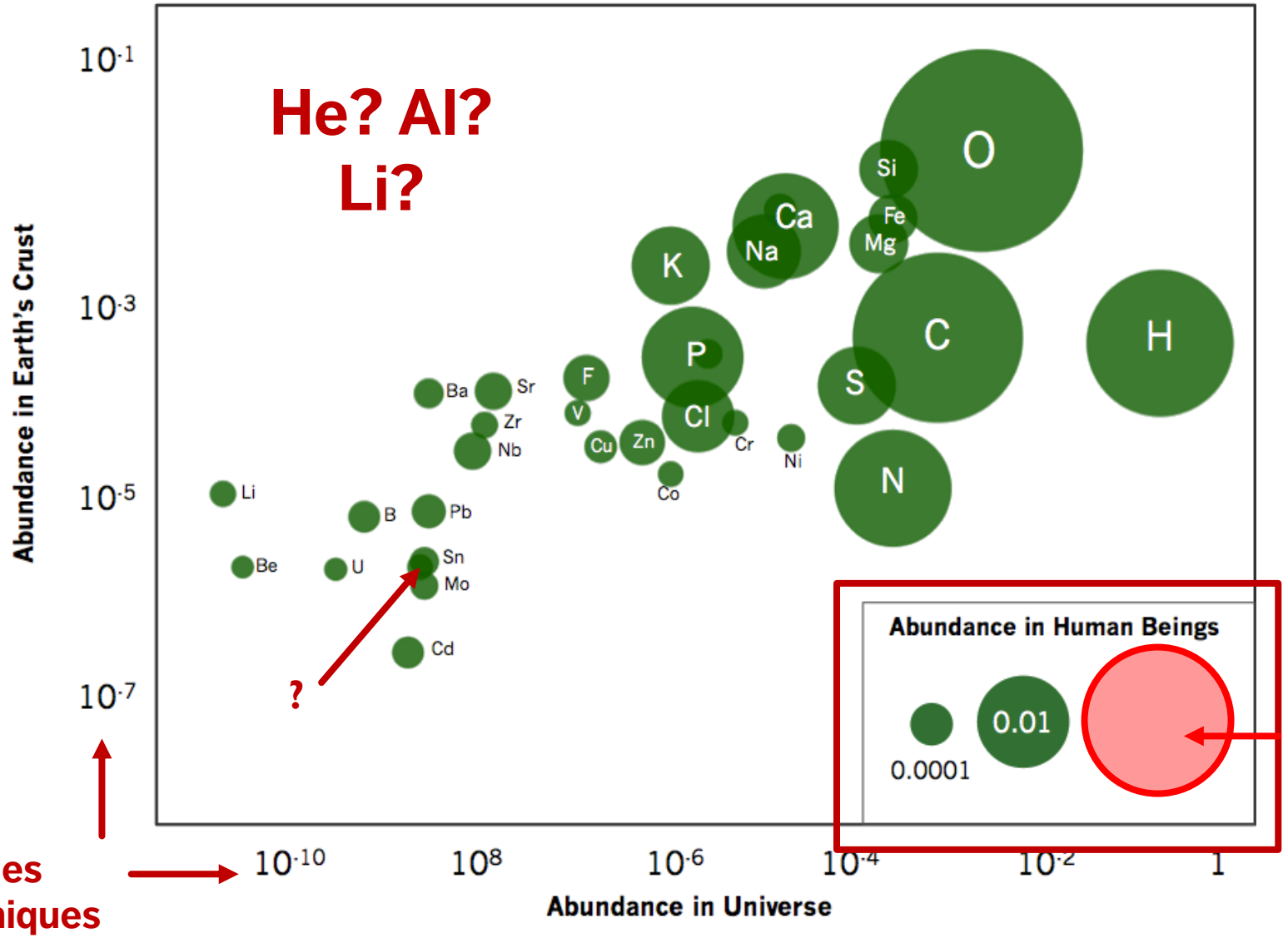
[M. Twardos, The Information Diet]

Abondance des éléments chimiques



[M. Twardos, The Information Diet]

Abondance des éléments chimiques



source ?

Échelle logarithmique ?

Échelles logarithmiques

GRAPHIQUE À BULLES

La couleur et **la géométrie** nous permettent de représenter (au moins) deux variables additionnelles sur un nuage à point en deux dimensions.

Il peut être nécessaire de revoir l'échelle ou de compartimenter les données.

L'utilisation d'une vidéo permet d'ajouter une variable ordinale.

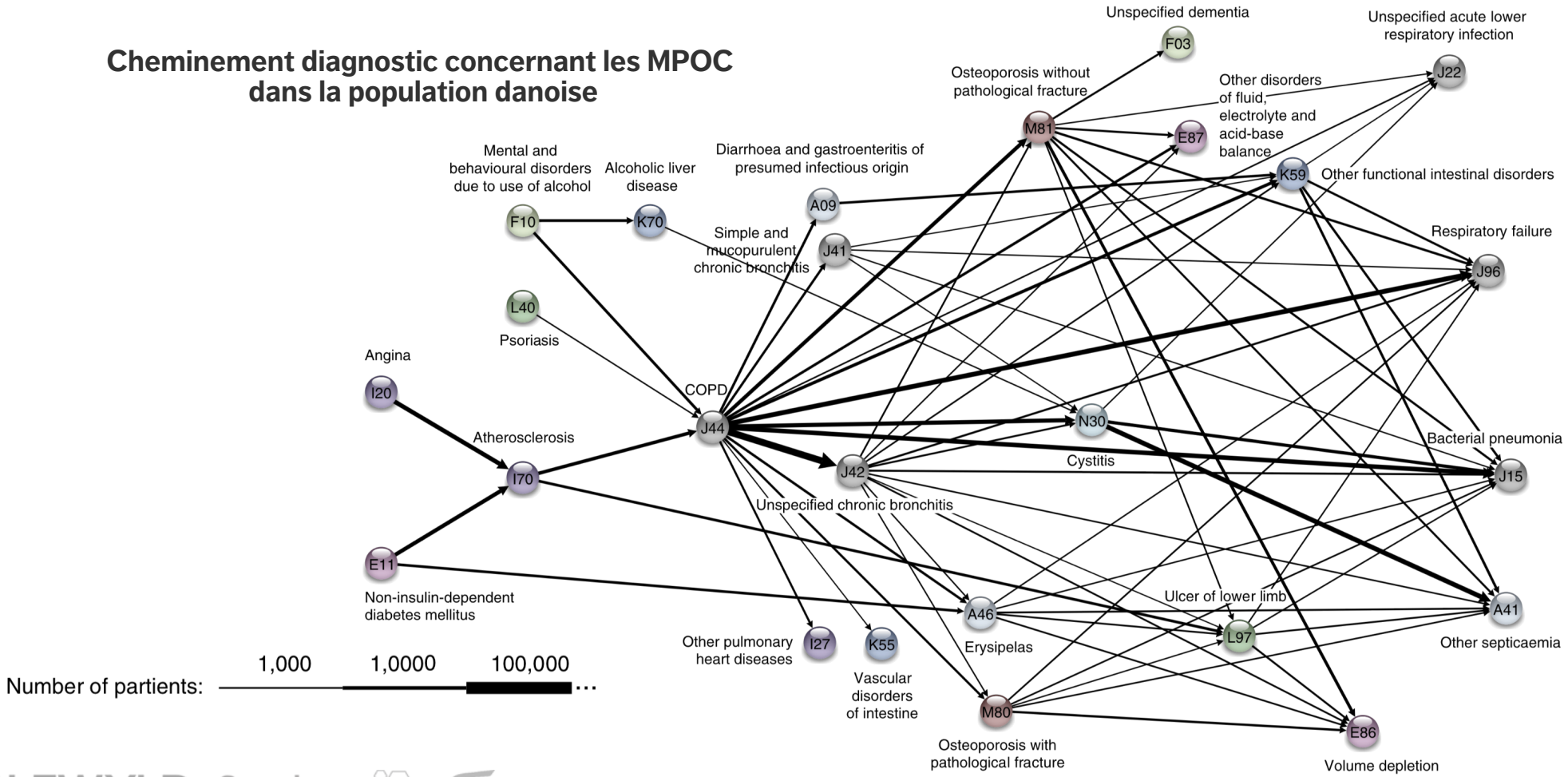
Il est aussi possible d'ajouter du texte pour représenter une variable nominale additionnelle.

À son efficacité maximale quand le graphique **n'est pas trop chargé.**

Un de mes **graphiques préférés** – bon équilibre entre caractéristiques modernes et traditionnelles.

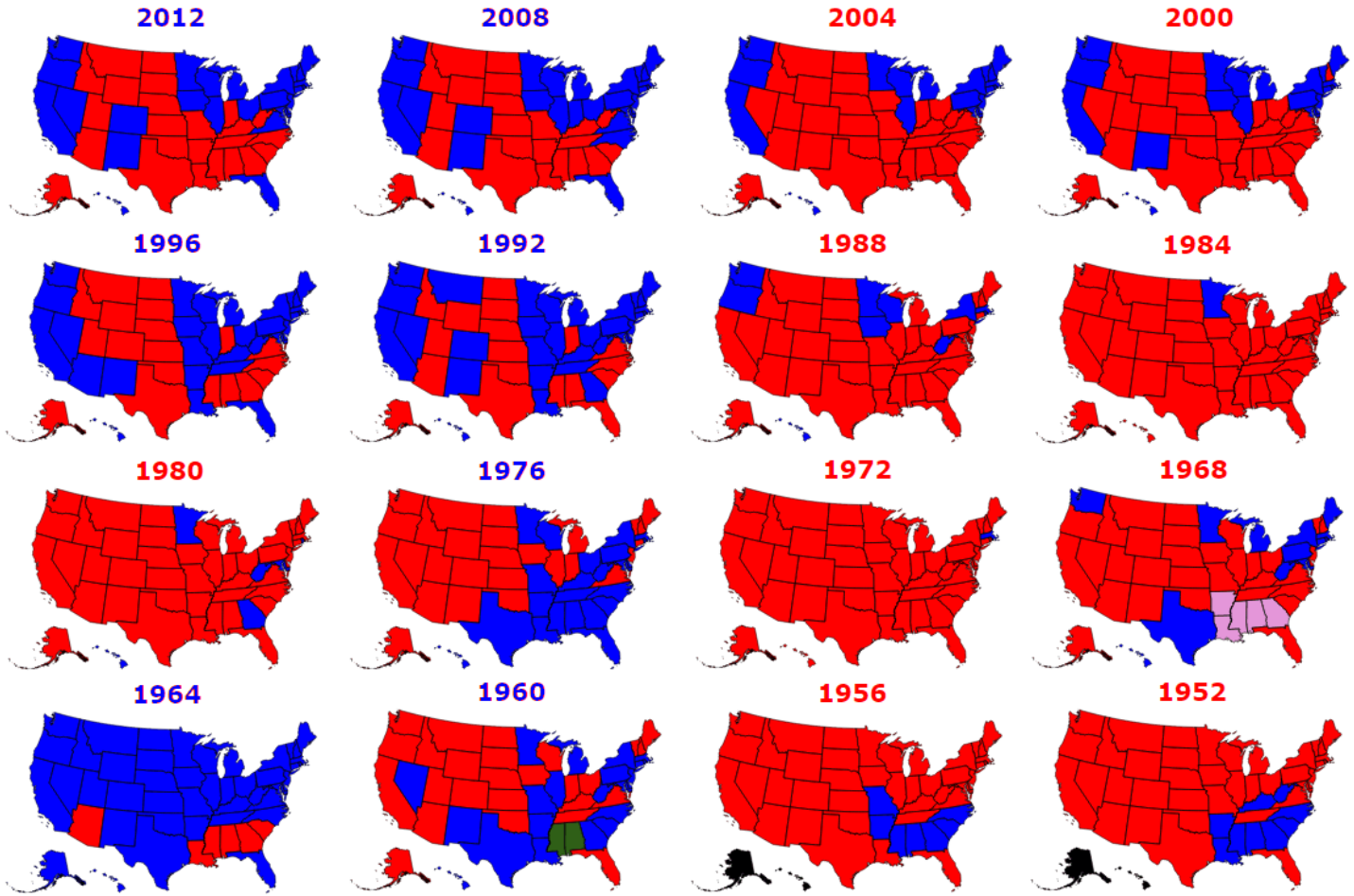
[A.B. Jensen et coll., *Nature Communications*]

DIAGRAMME DE RÉSEAU



[Metrocosm.com]

MINIATURES



Résultats du collège électoral des É.-U. de 1952 à 2012

DISCUSSION ET POINT À RETENIR

« Si certains types de visualisation deviennent dominant, il y a toujours un risque que les questions qui se prêtent particulièrement bien à ces visualisations deviennent dominantes, ce qui aurait une incidence sur les méthodes de collecte de données, l'accessibilité des données, le choix des pistes de recherche et ainsi de suite. » (P.Boily)

Une visualisation animée **n'est pas toujours** la meilleure solution.

Quelles renseignements peuvent être transmis grâce à l'interactivité? Tout dépend des données et de la visualisation.

À retenir : explorer les données; essayer plusieurs méthodes.

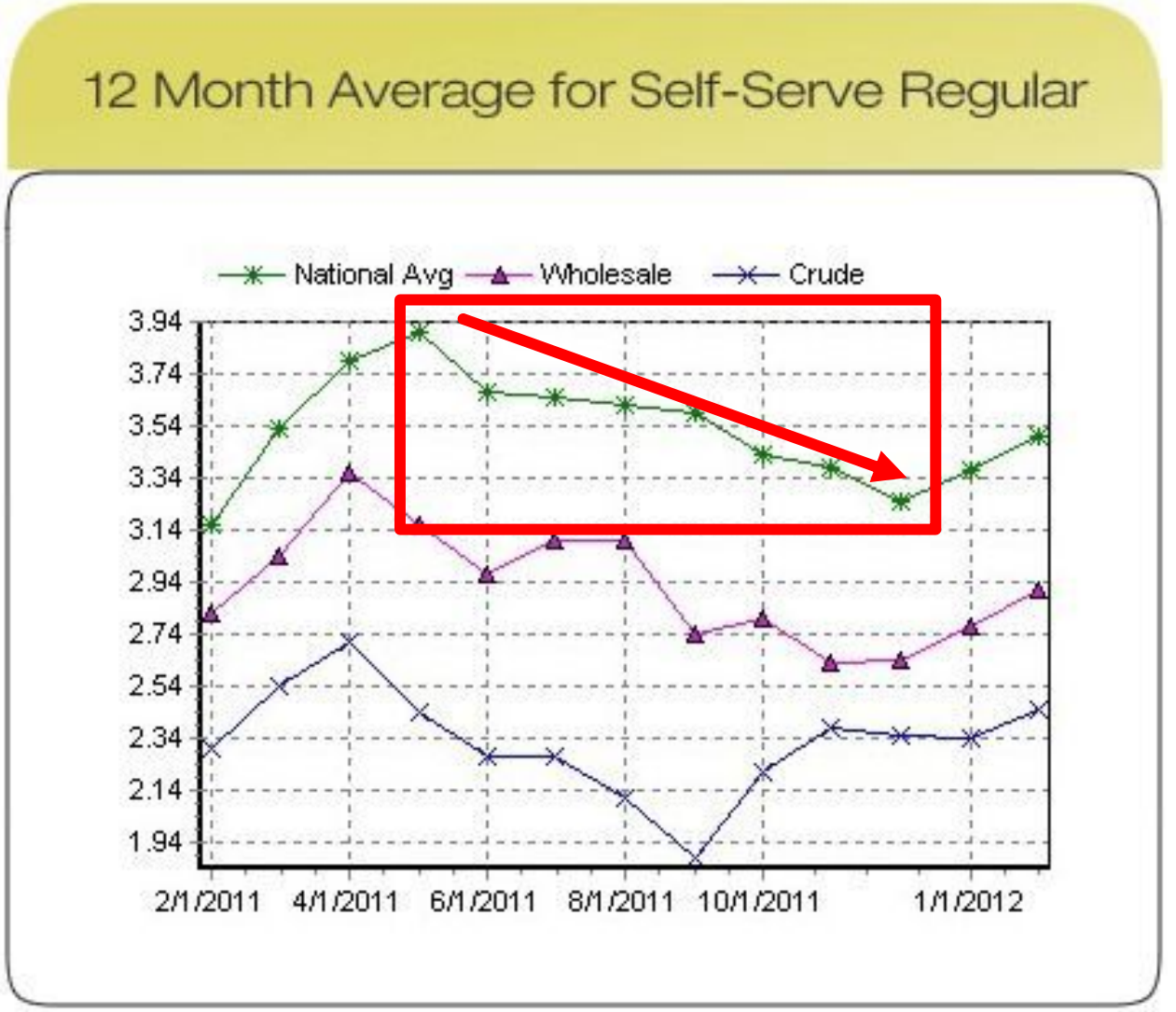
TABLEAU D'HONNEUR ET TABLEAU D'HORREUR

EXPLORATION ET VISUALISATION DES DONNÉES

[Fox News, 20 février 2012, mediamatters.org, simplystatistics.org, badgraphs.tumblr.com]

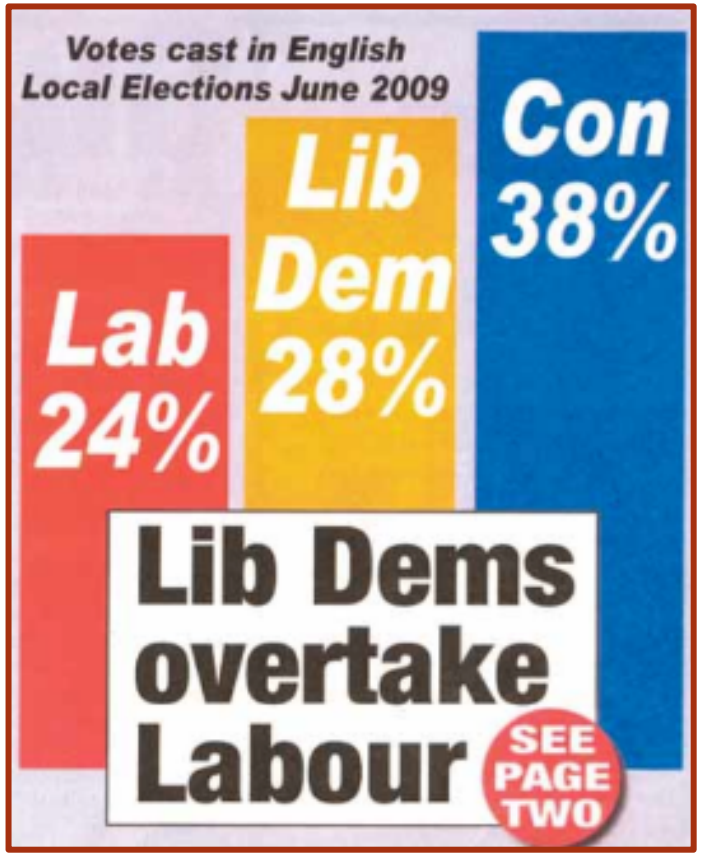
GRAPHIQUES TROMPEURS

[Prix de l'essence selon l'AAA, trouvé sur mediamatters.org]



[Norwich North Electoral District, mediamatters.org, simplystatistics.org, badgraphs.tumblr.com]

GRAPHIQUES TROMPEURS



GRAPHIQUES TROMPEURS

Problèmes : information fallacieuse, sélective ou traitée de façon incompétente.

Solutions :

- Échelles et unités de mesure uniformes
- Séries chronologiques complètes
- Ne pas choisir arbitrairement la fourchette de données
- La troncation d'un axe peut exagérer certains effets
- Les nombres doivent être sensés

À SURVEILLER

Certaines méthodes produisent des graphiques impressionnants, mais trompeurs.

Se méfier :

- **de la manipulation des axes** et **des échelles linéaires**;
- **des effets d'échelle**, lorsque des données sont représentées par des formes ou des volumes;
- **des choix arbitraires** permettant d'omettre certaines observations.

Pour les jeux de données dont le nombre de dimensions est réduit, un **tableau** peut être aussi informatif et comporter moins de risque de mésinterprétation.

À SURVEILLER

Différentes manières d'évaluer le caractère trompeur d'un graphique :

- **Facteur de mensonge :** rapport entre la taille de l'effet affichée dans le graphique et la taille de l'effet dans les données.
- **Densité des données :** rapport entre le nombre d'observations et la superficie du graphique.
- **Rapport de bric-à-brac graphique :** rapport entre la superficie nécessaire pour transmettre l'information et la superficie du graphique.

On souhaitera habituellement que le facteur de mensonge et le rapport de bric—à—brac graphique se rapprochent autant que possible de 1, tandis que la densité des données devrait être « élevée » (dans la limite du raisonnable).

À VOUS DE JUGER

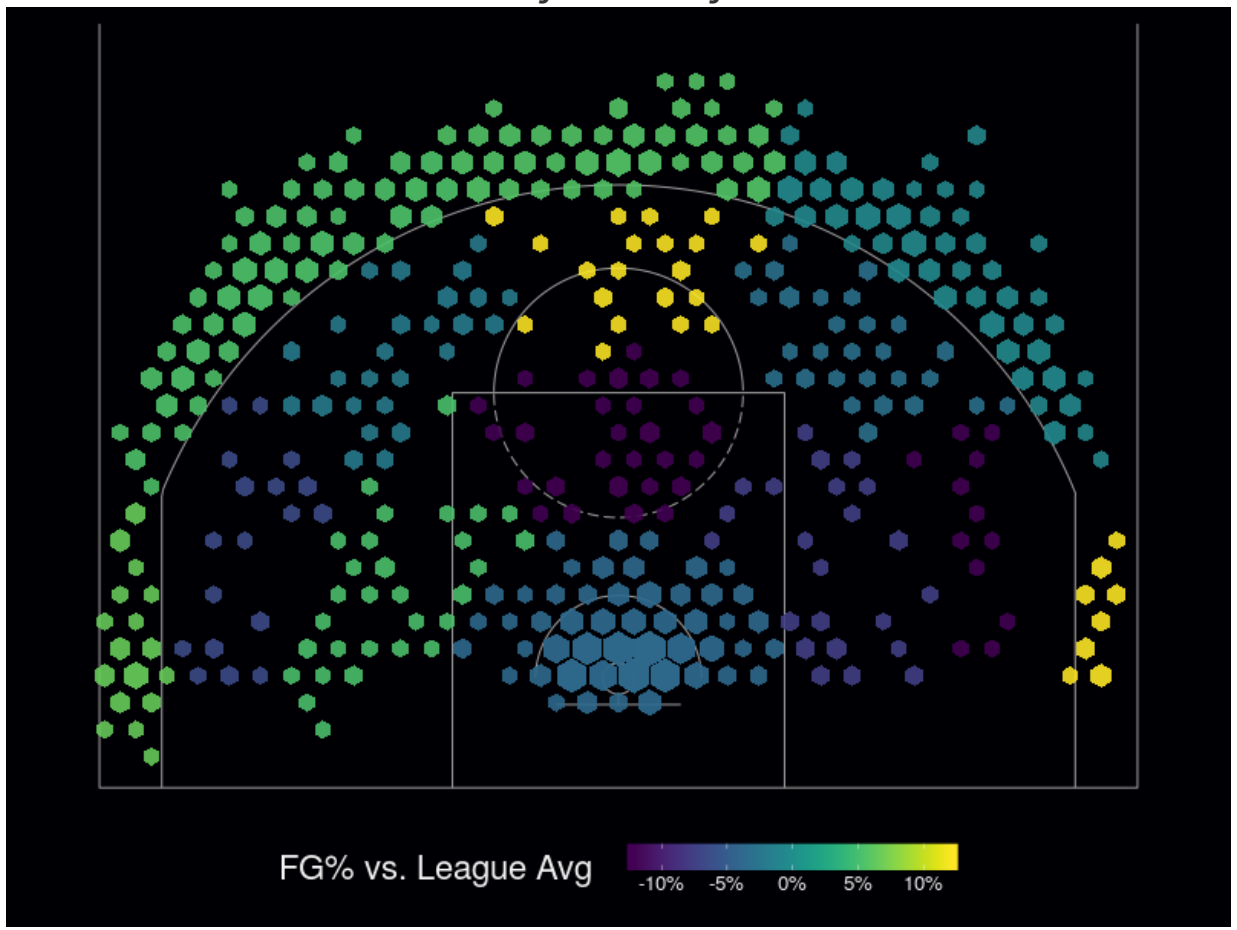
Certaines des visualisations suivantes (peuvent être considérées comme) bonnes. Mais pas toutes!

Dans quelle catégorie tombe chacune? À vous de juger...

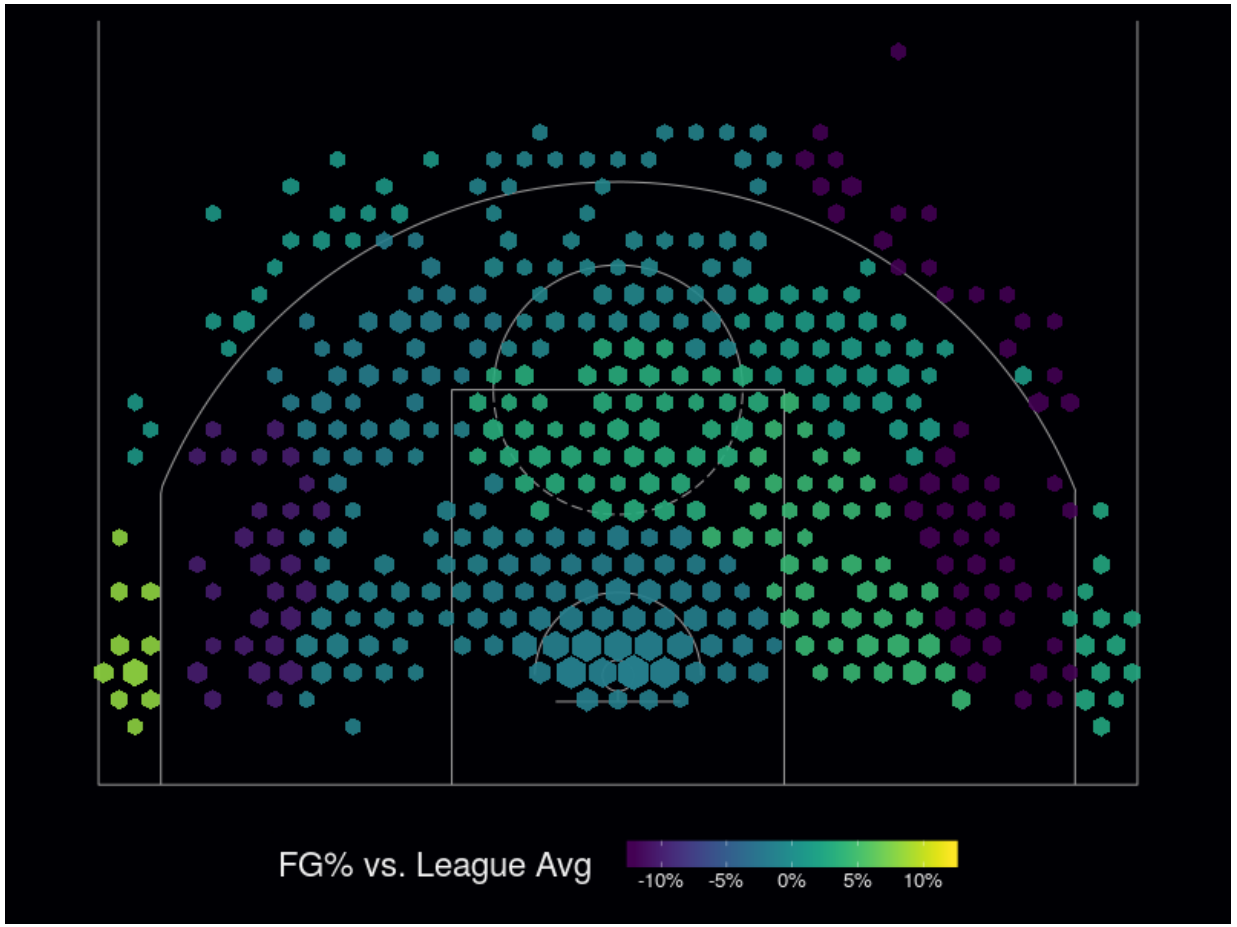
Paniers marqués (%) dans la NBA par rapport à la moyenne de la ligue (2015–2016)

[T.W. Schneider]

Kyle Lowry



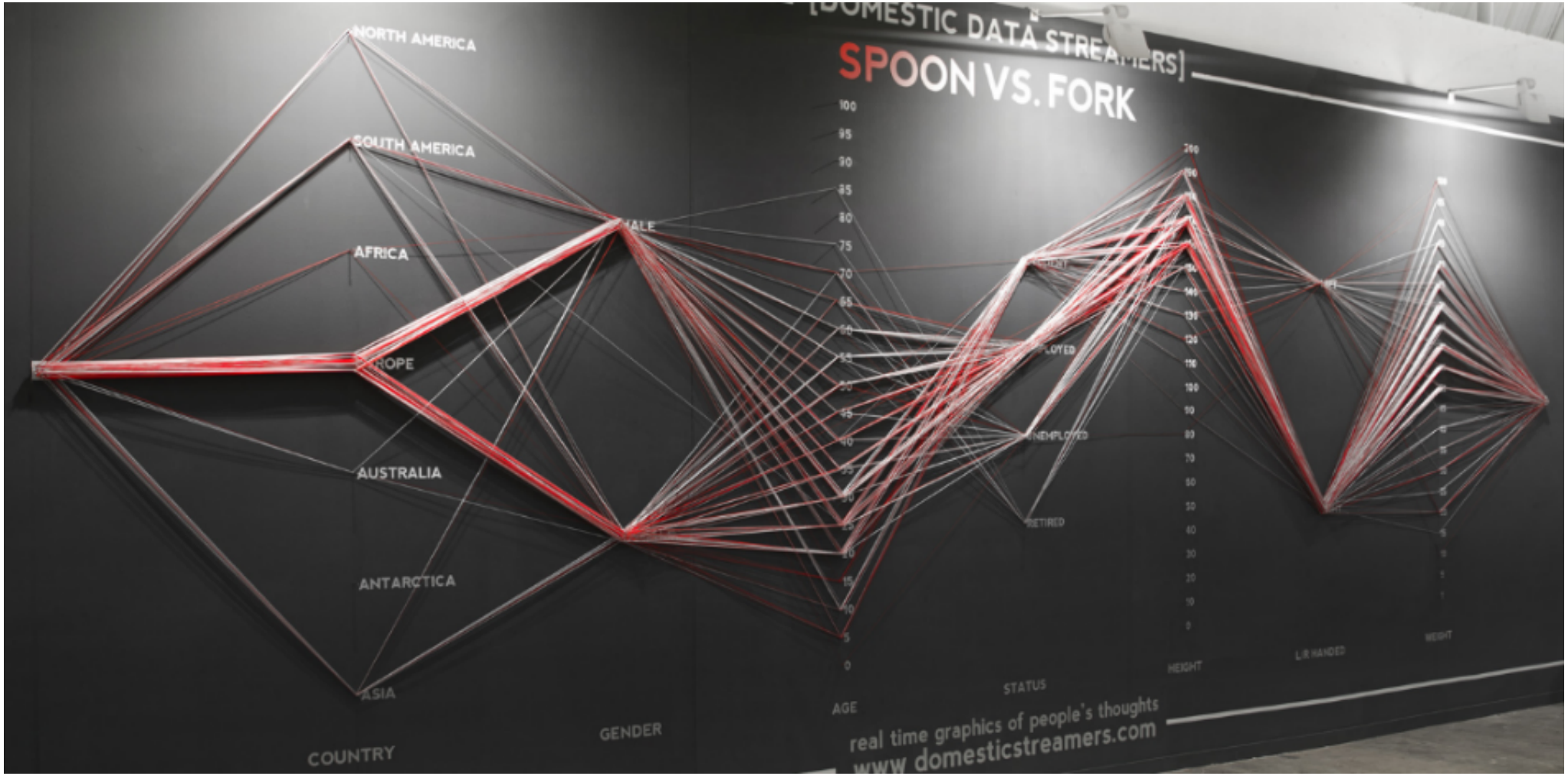
DeMar DeRozan



Quelles comparaisons pouvez-vous faire? Comprenez-vous l'encodage? Le contexte?

Cuillère ou fourchette?

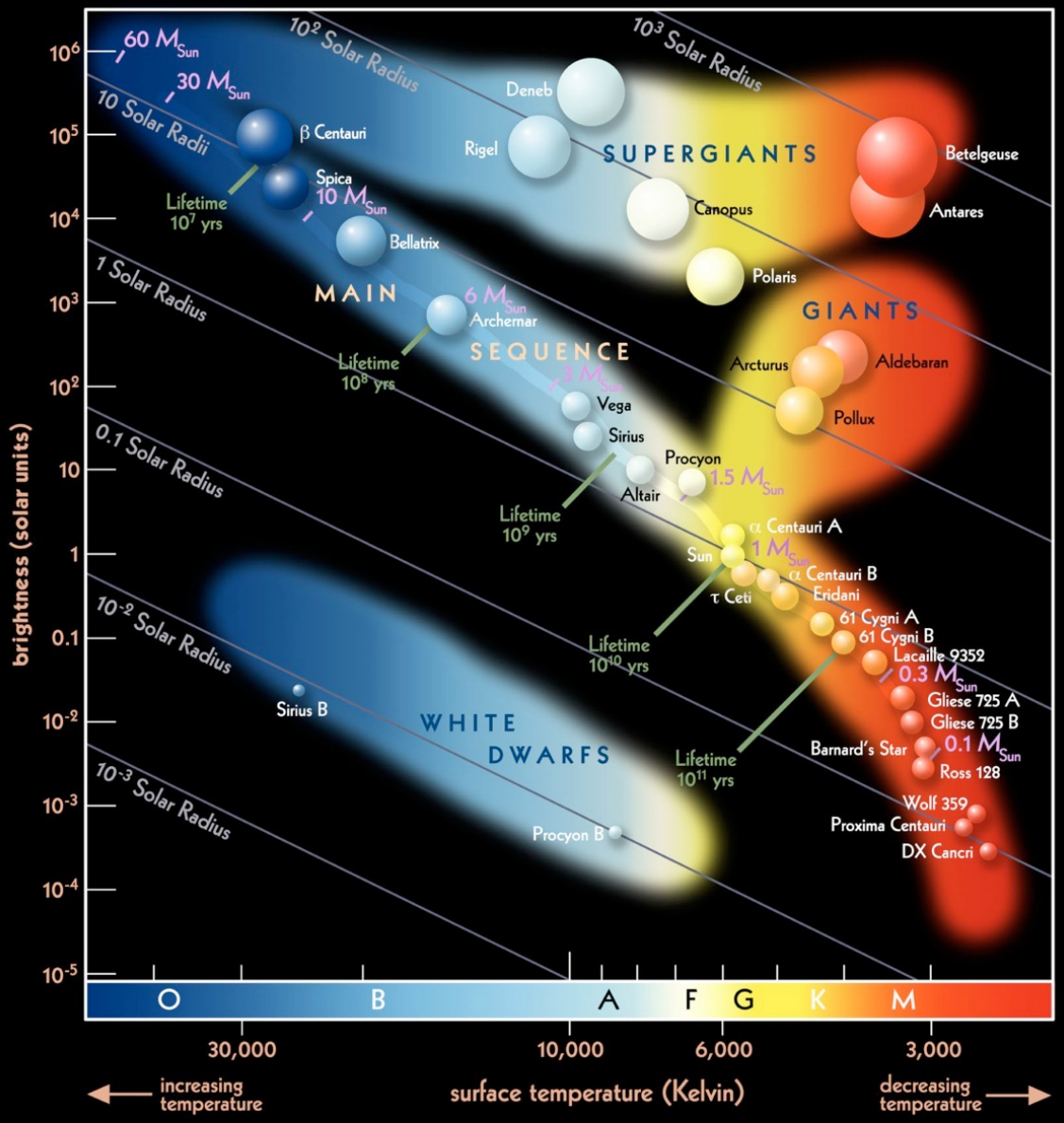
[domesticstreamers.com]



Existent-ils des problèmes de collecte de données? Où pensez-vous que cette activité s'est déroulée? La compétition « cuillère ou fourchette » est-elle une distraction?

[European Southern Observatory]

Diagramme Hertzprung-Russell



Éléments de données

- Rayon des étoiles (x 2)
- Température à la surface (x 2)
- Classe spectrale
- Luminosité
- Masse
- Durée de vie
- Nom

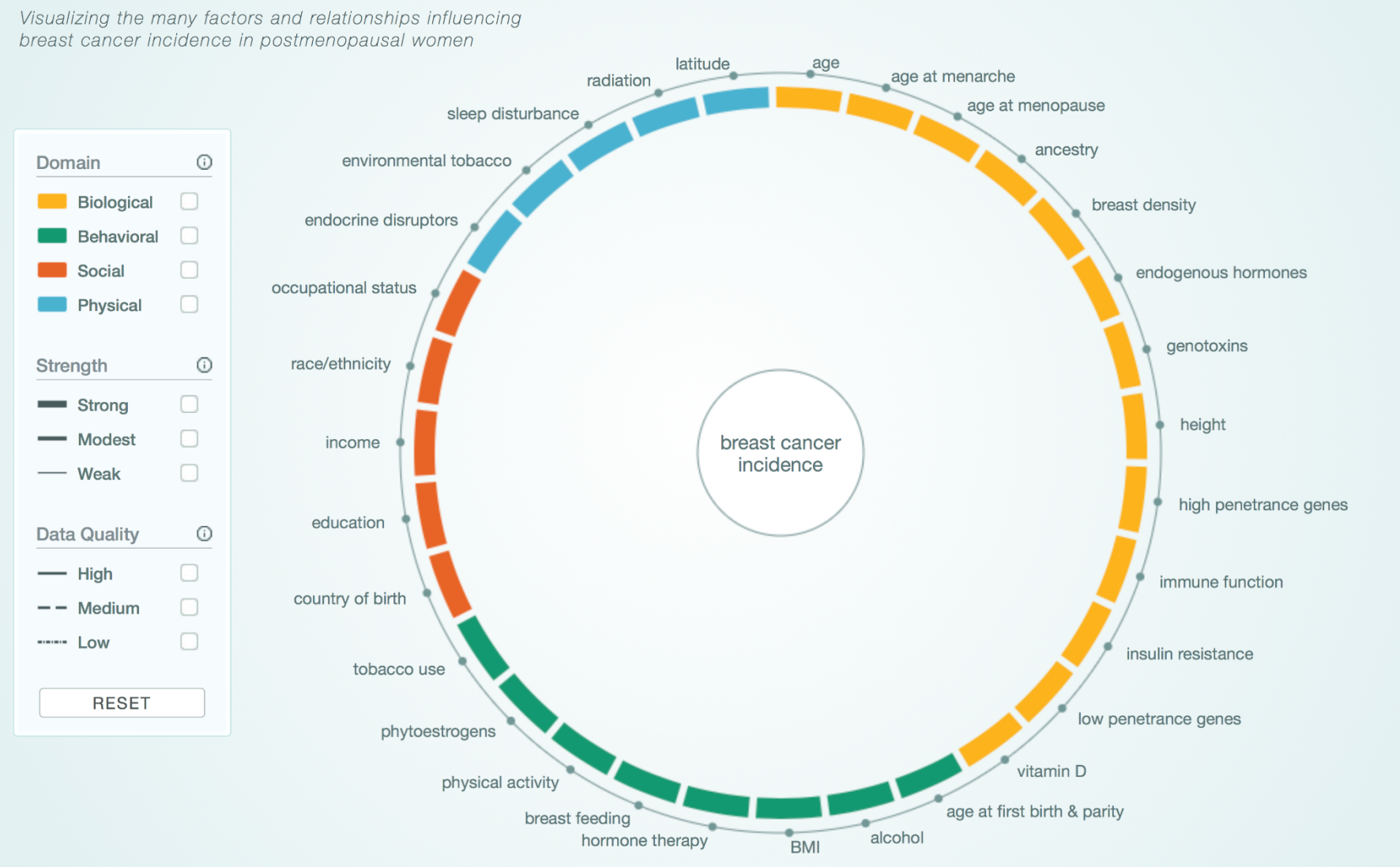
Structure sous-jacente

- 4 regroupements
- La durée de vie, la masse et le rayon sont liés à la luminosité et à la température à la surface sur la séquence principale.

Le diagramme ne montre qu'un sous-ensemble des étoiles.

[http://www.cbcrp.org/causes/index.php]

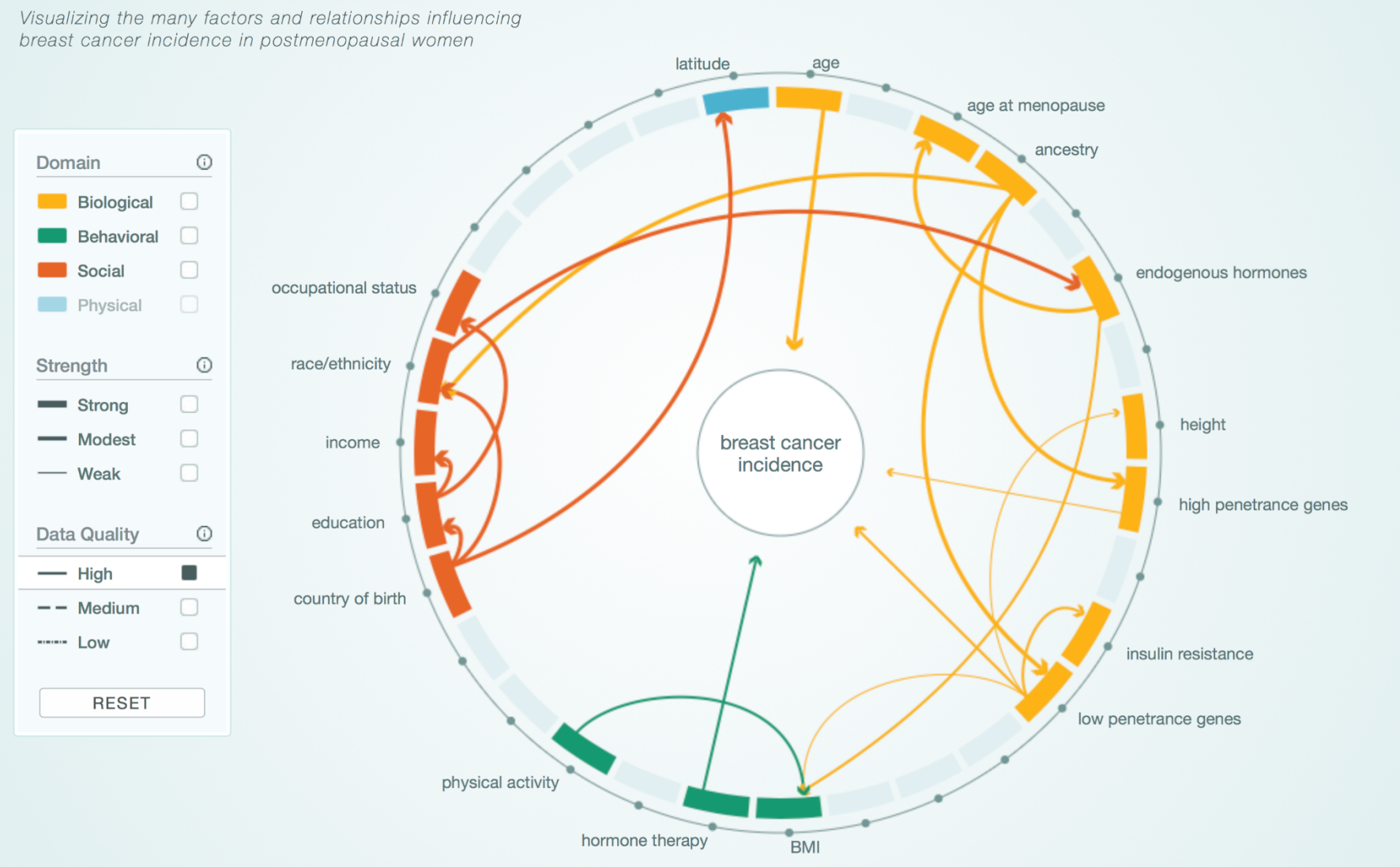
Modèle des causes du cancer du sein



Pouvez-vous inférer une relation de cause à effet de ce diagramme?

[http://www.cbcrp.org/causes/index.php]

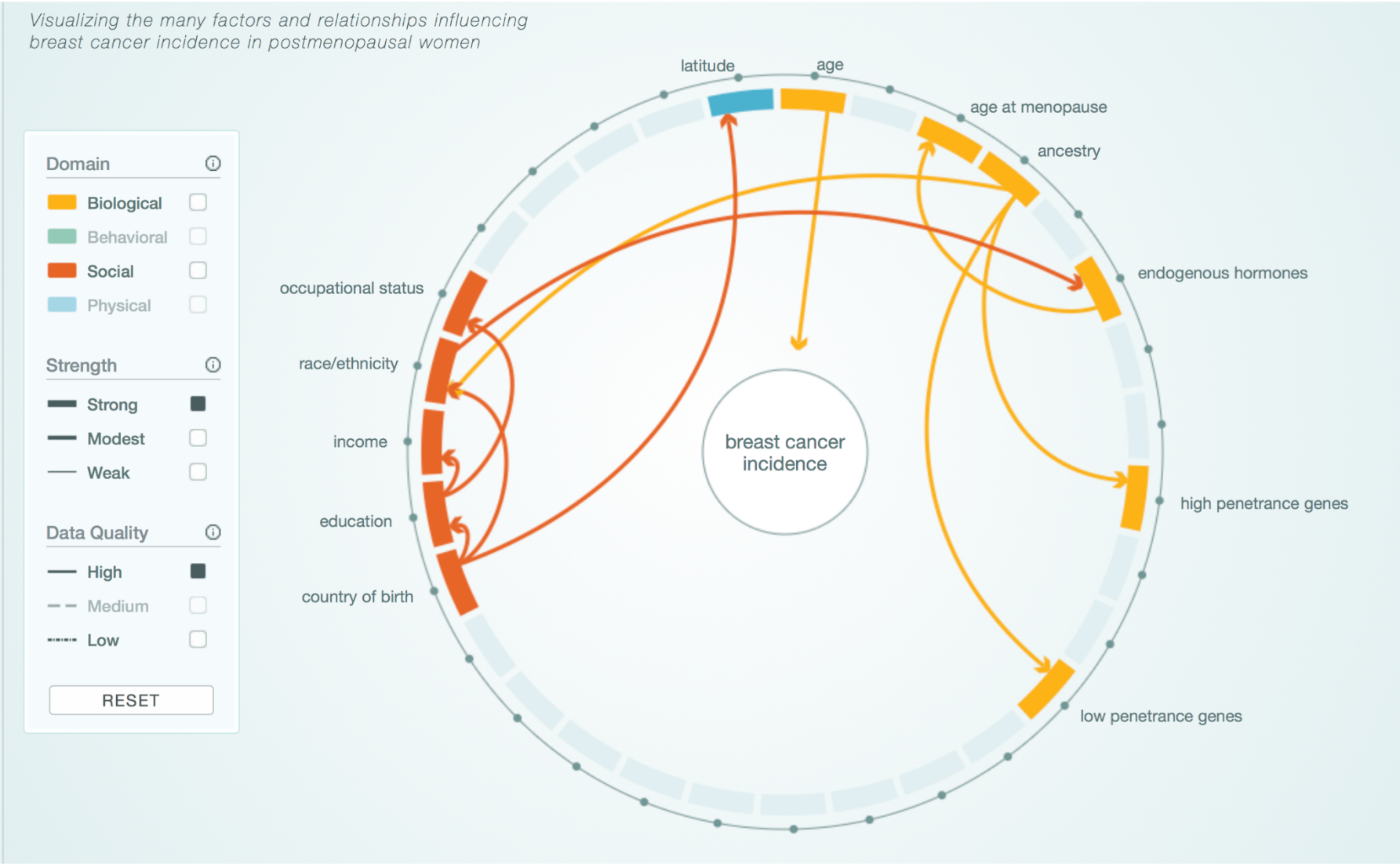
Modèle des causes du cancer du sein



Pouvez-vous inférer une relation de cause à effet de ce diagramme?

[http://www.cbcrp.org/causes/index.php]

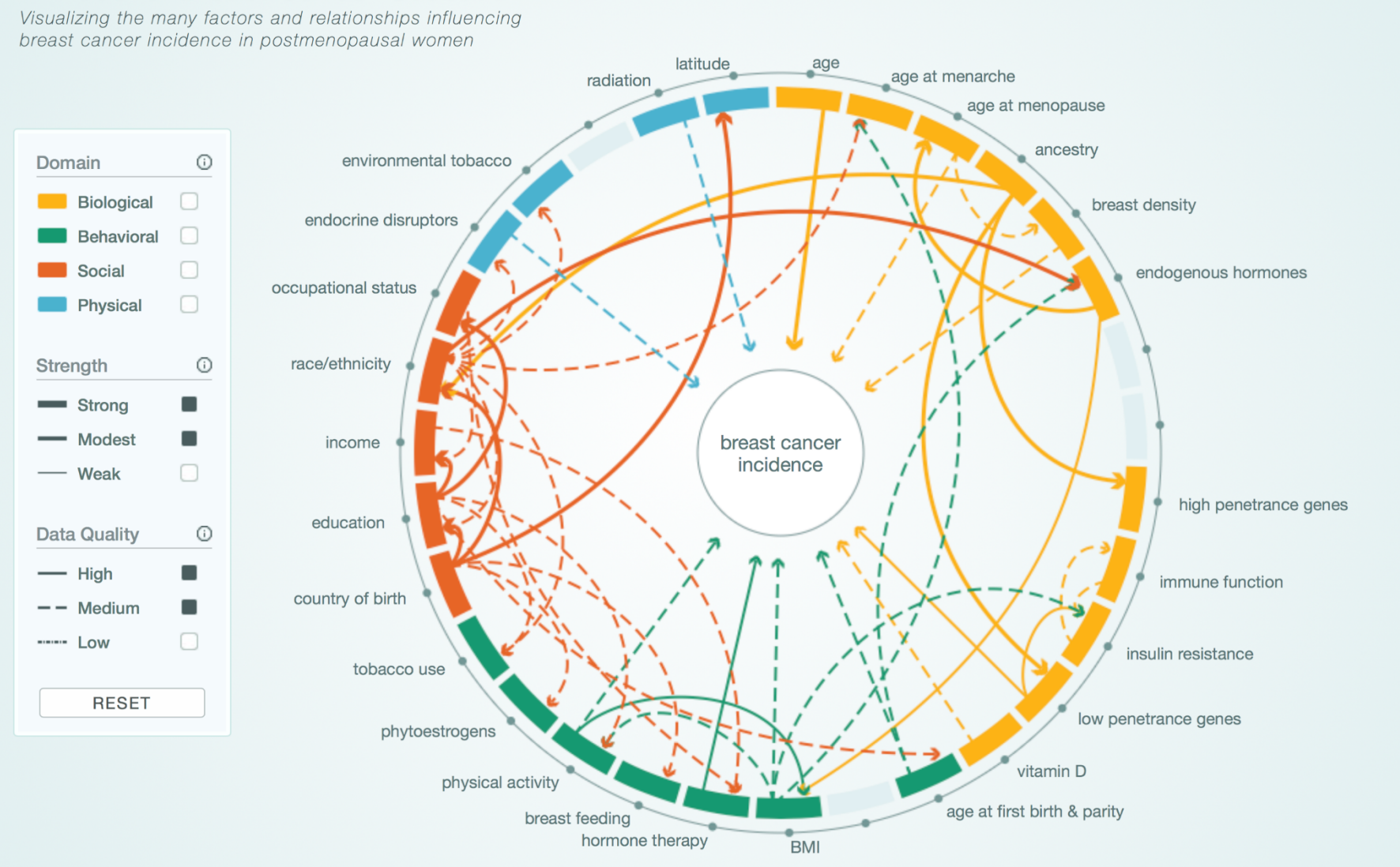
Modèle des causes du cancer du sein



Pouvez-vous inférer une relation de cause à effet de ce diagramme?

[http://www.cbcrp.org/causes/index.php]

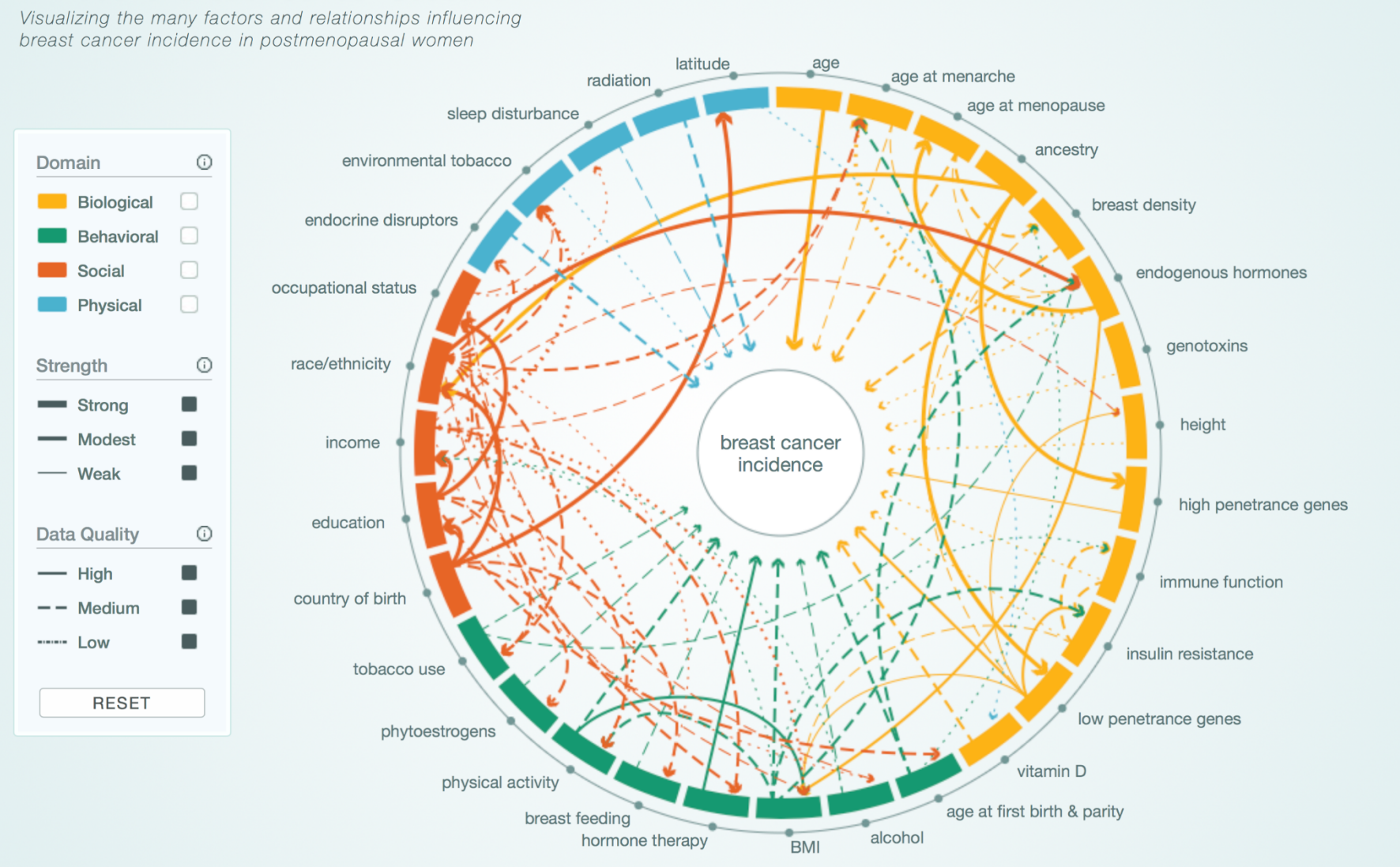
Modèle des causes du cancer du sein



Pouvez-vous inférer une relation de cause à effet de ce diagramme?

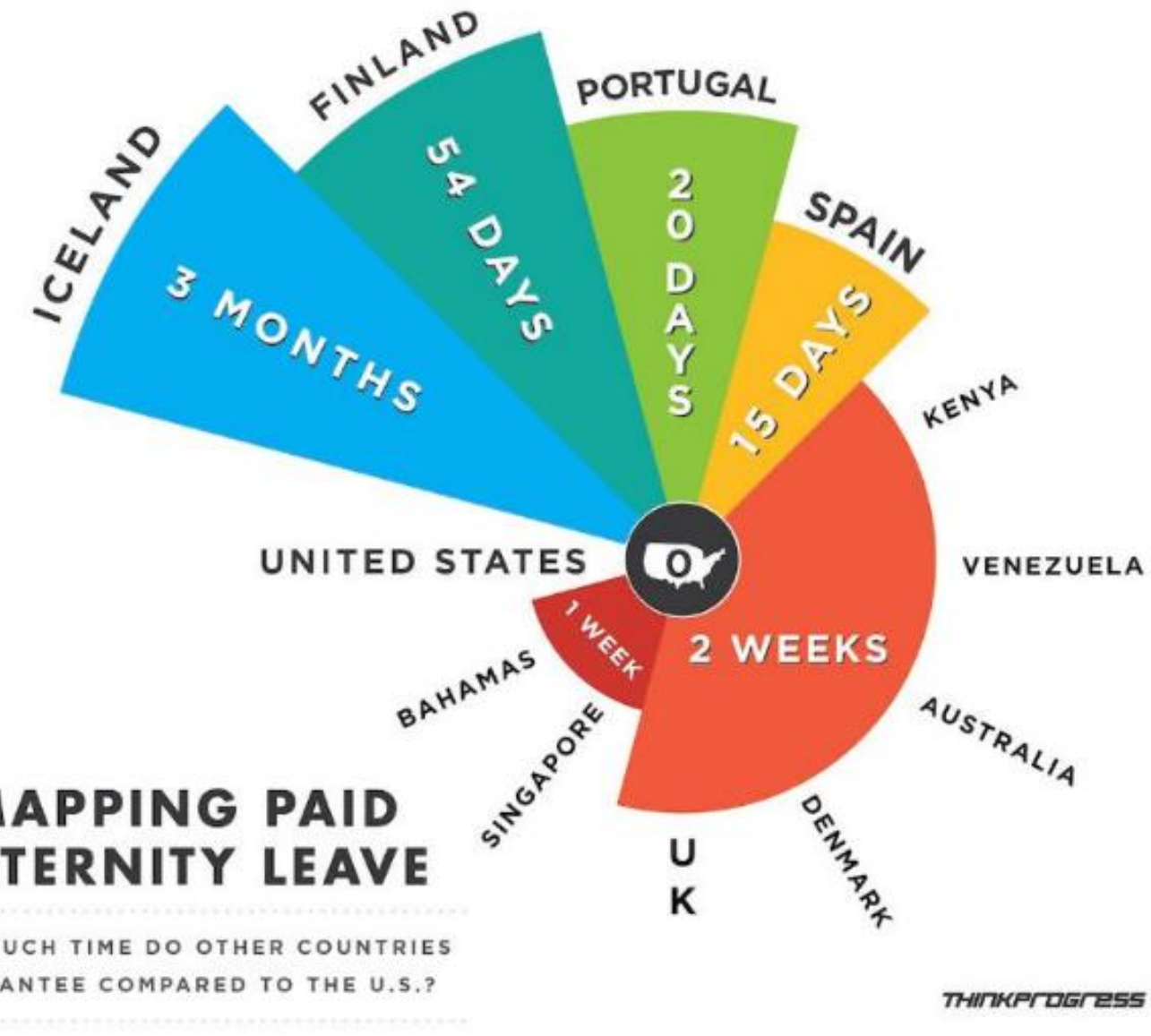
[http://www.cbcrp.org/causes/index.php]

Modèle des causes du cancer du sein



Pouvez-vous inférer une relation de cause à effet de ce diagramme?

[https://thinkprogress.org, trouvé à l'adresse http://viz.wtf]



[https://thinkprogress.org, trouvé à l'adresse http://viz.wtf]

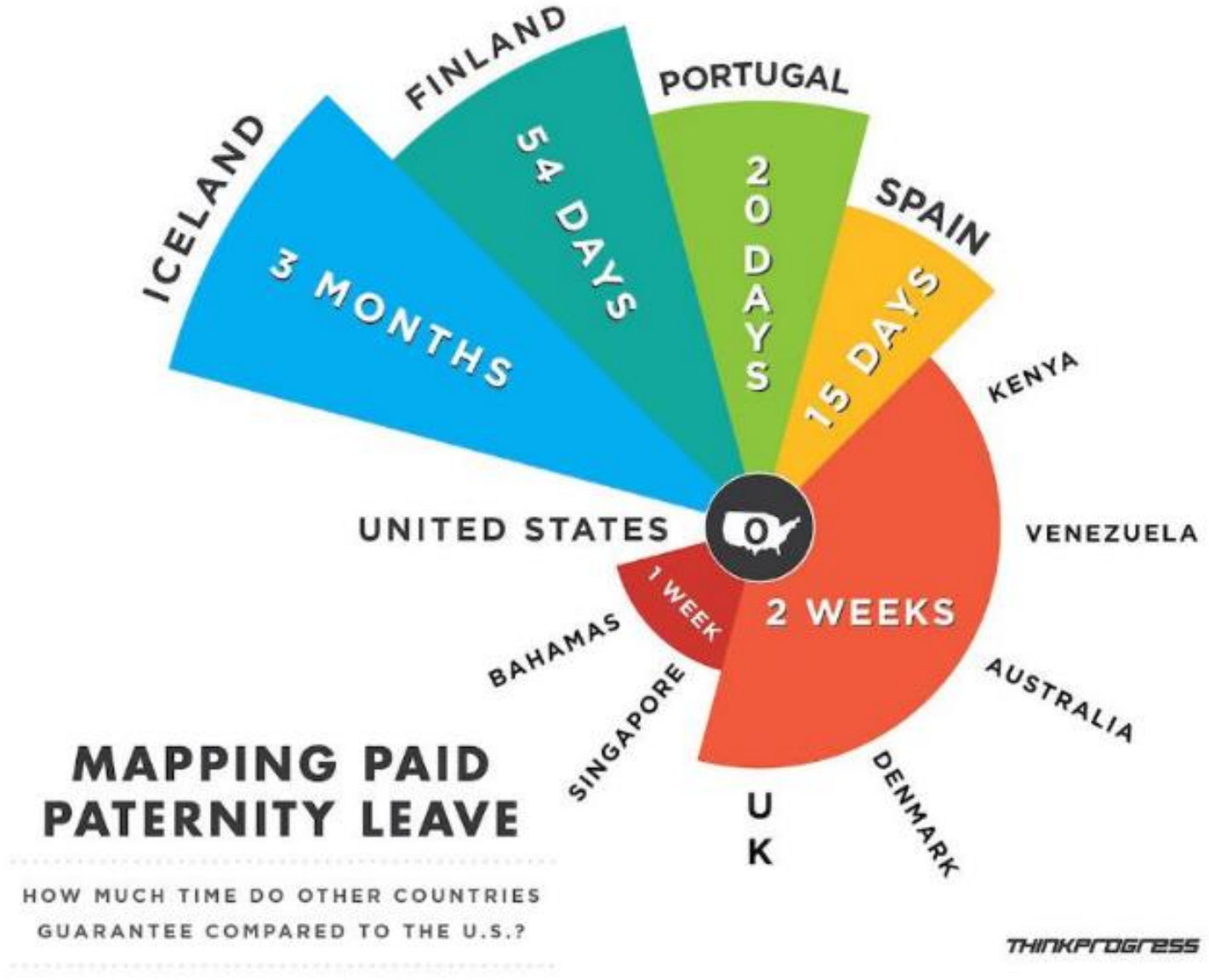
Faible densité de données

Rapport de bric-à-brac graphique élevé

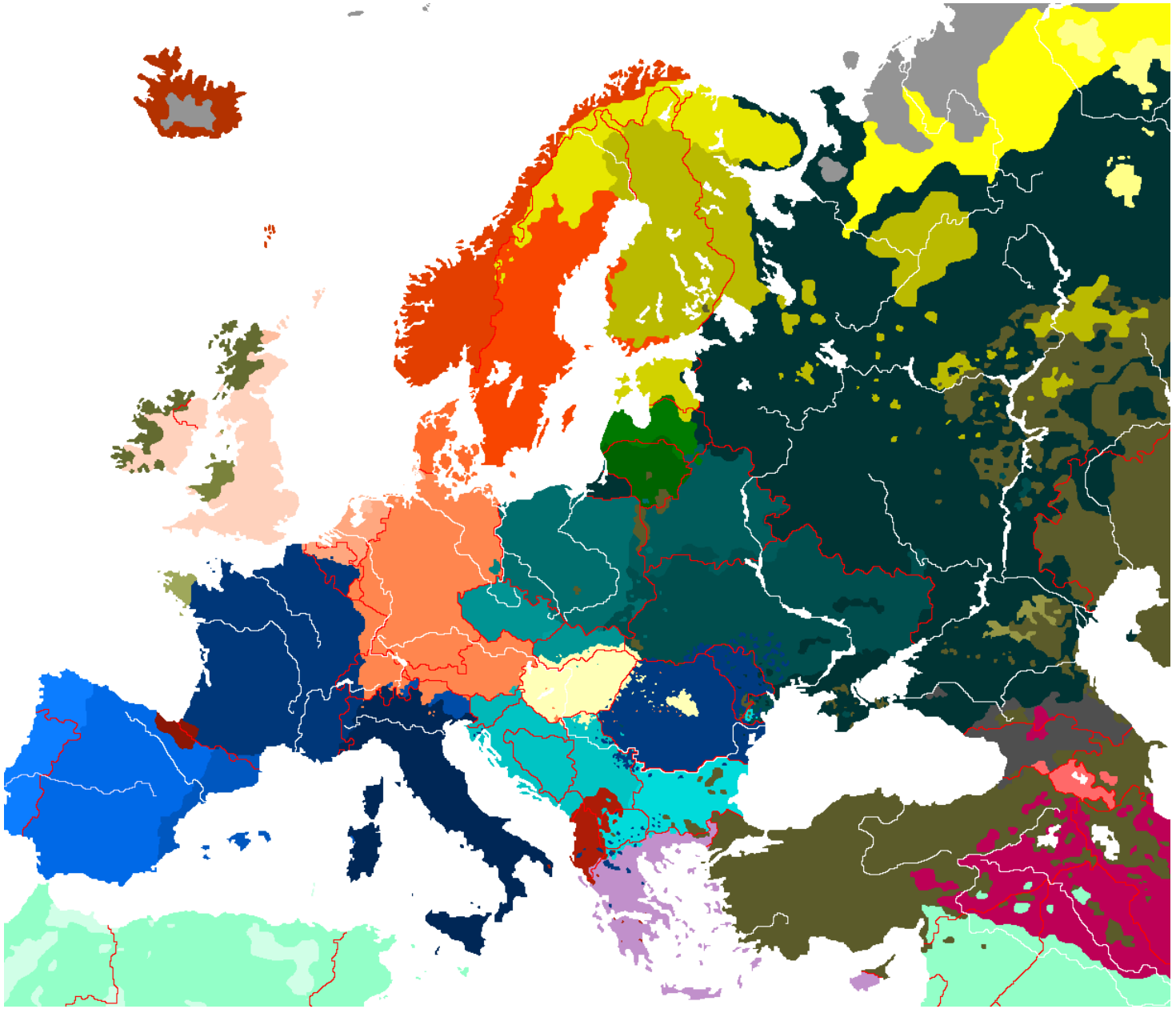
Effets d'échelle

Choix de données arbitraires

Pourquoi ne pas plutôt utiliser un **diagramme en bâtons** ou un **tableau**?



[Trouvé sur Wikipedia, auteur inconnu]



[Trouvé sur Wikipedia, auteur inconnu]

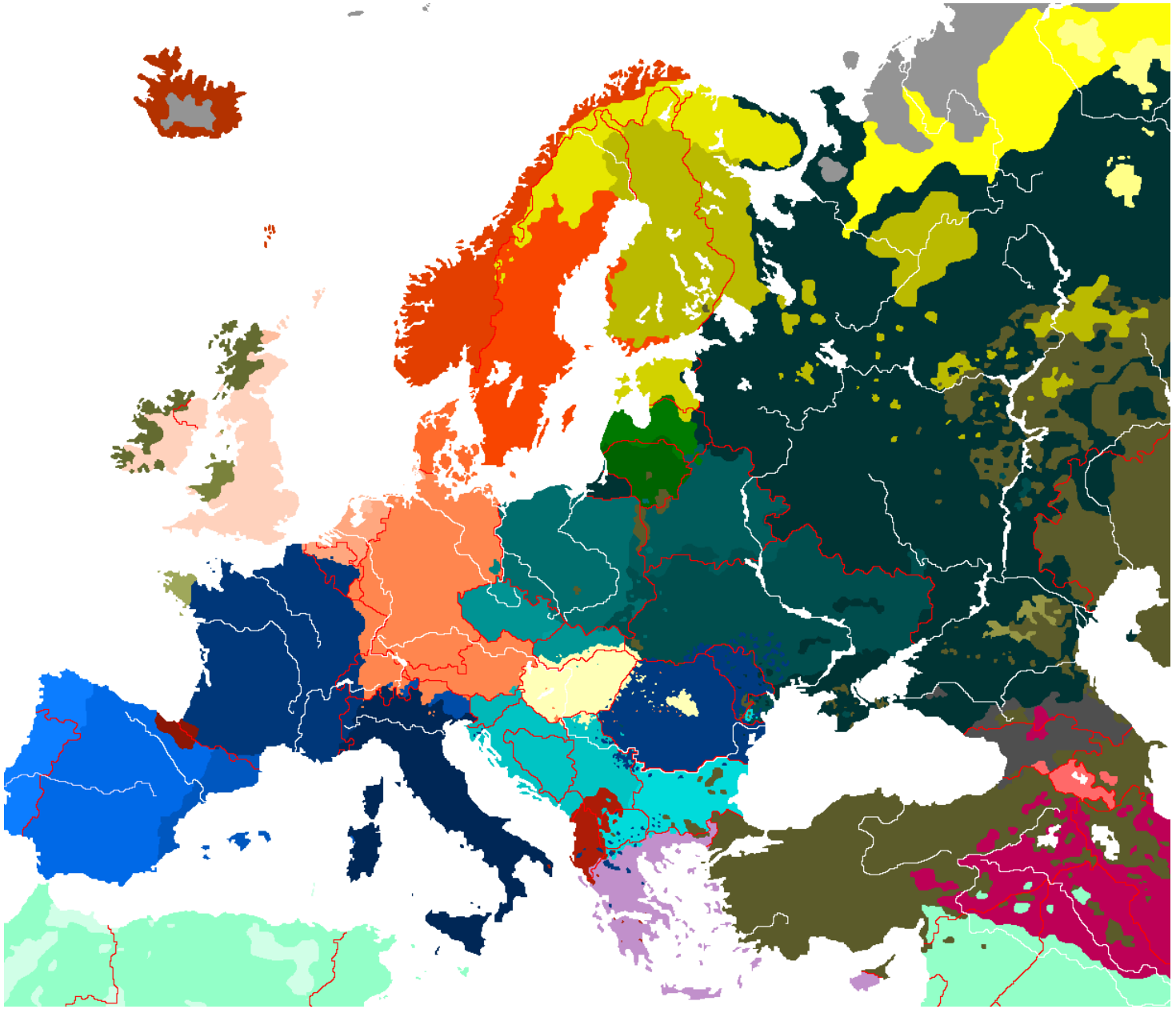
Encodage?

Densité de population?

Langues secondes?

Rivières?

Aucune source



STRUCTURE LOGIQUE DES GRAPHIQUES

EXPLORATION ET VISUALISATION DES DONNÉES

*non nécessaire

INTRODUCTION À GGPLOT2

La fonction *ggplot2* est en fait un jeu d'outils permettant de transformer des données en éléments d'affichage visuels. Elle permet à l'utilisateur de commander les détails de l'affichage graphique.

Aspect le plus important, la fonction *ggplot2* peut servir à établir la **structure logique** du graphique.

Un graphique *ggplot2* comporte deux éléments principaux (et des termes optionnels) :

- une fonction esthétique (**aes** – liens entre les données et les éléments graphiques)
- une fonction de géométrie (**geom** – type de graphique)
- *facets, *coordinates, *scales, *labels, *guides, etc.

[Healey, K., *Data Visualization: A Practical Introduction*]

GRAMMAIRE DE GGPLOT2

1. Tidy Data

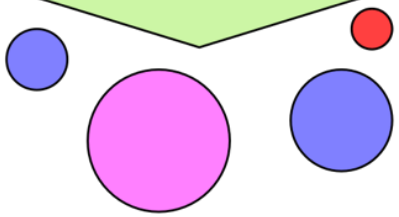
```
p ← ggplot(data = gapminder, ...
```

gdp	lifexp	pop	continent
340	65	31	Euro
227	51	200	Amer
909	81	80	Euro
126	40	20	Asia

2. Mapping

```
p ← ggplot(data = gapminder, mapping =
      aes(x = gdp, y = lifexp, size = pop,
        color = continent))
```

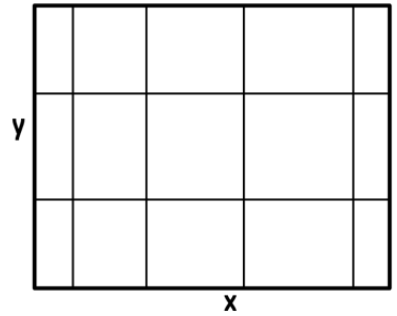
3. Geom



```
p + geom_point()
```

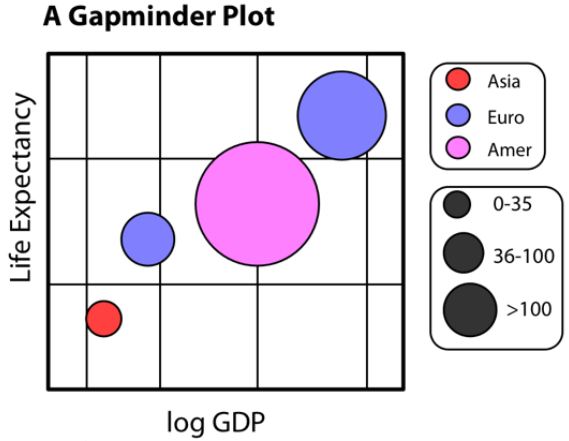
4. Co-Ordinates & Scales

```
p + coord_cartesian() + scale_x_log10()
```



5. Labels & Guides

```
p + labs(x = "log GDP", y = "Life
Expectancy", title = "A Gapminder Plot")
```



GRAMMAIRE DE GGPLOT2 – GEOM

La source de données et les variables sont précisées au moyen de `ggplot()`.

Les diverses fonctions *geom* précisent la **manière** dont ces variables seront représentées visuellement :

- au moyen de points, de barres, de lignes, de zones ombragées, etc.

Il existe présentement 37 géométries.

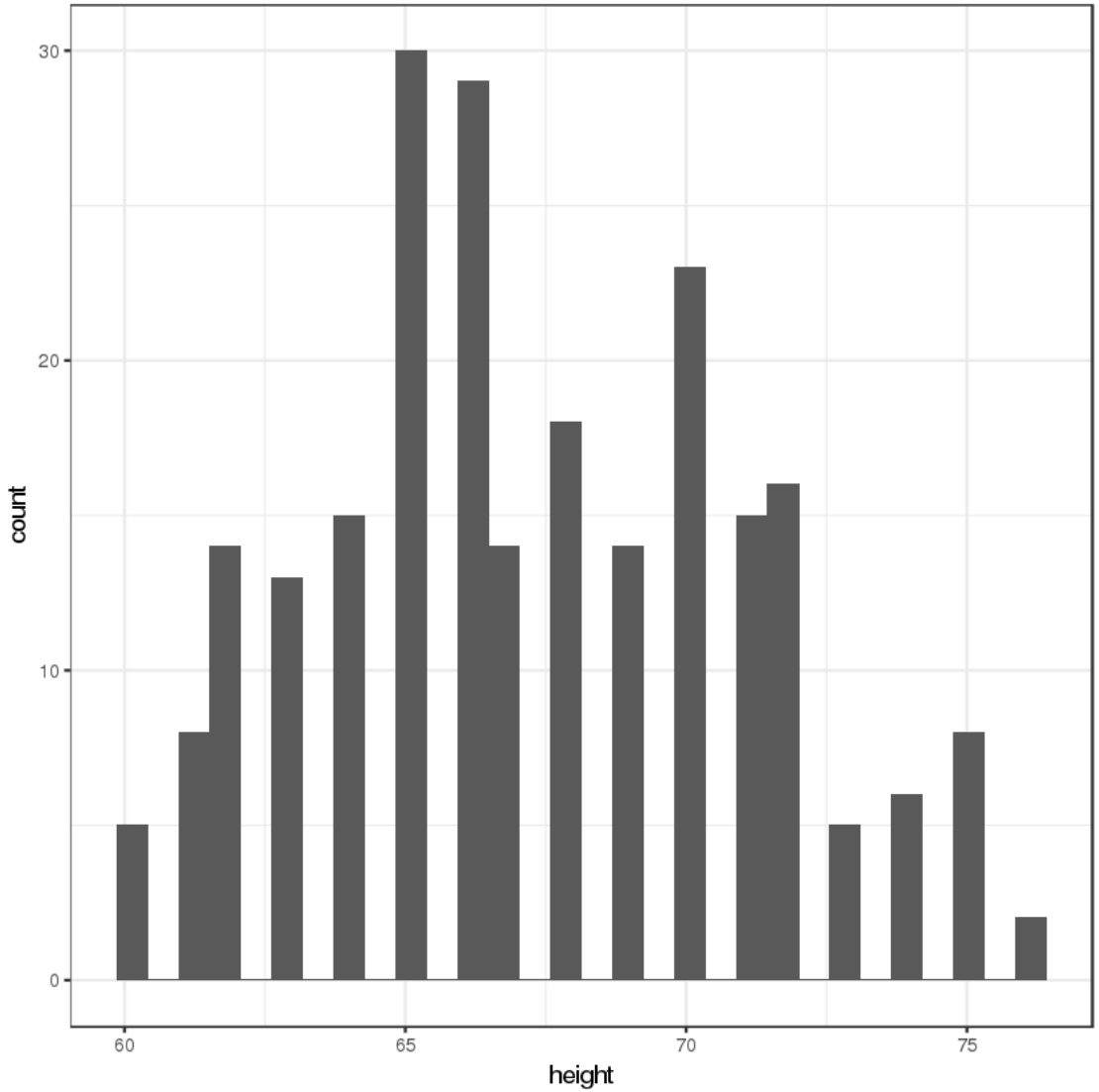
GRAMMAIRE DE GGPLOT2 – GEOM()

```
library("ggplot2")
data(singer, package="lattice")
# Using data from the 1979 ed. of the
# New York Choral Society

# Histogram of heights
ggplot(singer, aes(x=height)) +
  geom_histogram()

# Boxplot of heights by voice part
ggplot(singer, aes(x=voice.part, y=height)) +
  geom_boxplot()
```

Selon vous, quelle sera la forme du graphique?

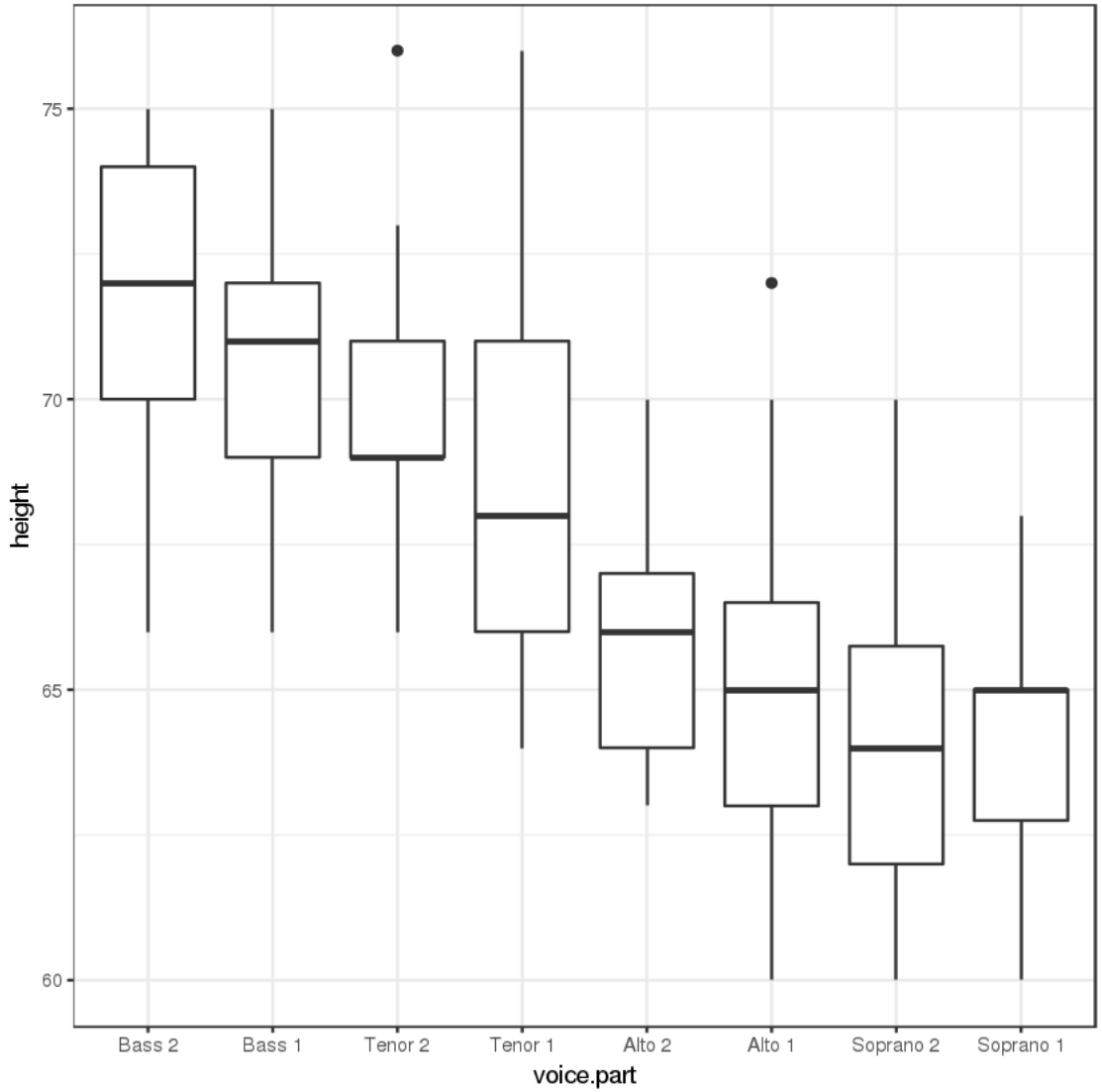
GRAMMAIRE DE GGPLOT2 – GEOM()

```
library("ggplot2")
data(singer, package="lattice")
# Using data from the 1979 ed. of the
# New York Choral Society

# Histogram of heights
ggplot(singer, aes(x=height)) +
   geom_histogram()

# Boxplot of heights by voice part
ggplot(singer, aes(x=voice.part, y=height)) +
   geom_boxplot()
```



GRAMMAIRE DE GGPLOT2 – GEOM()

```
library("ggplot2")
data(singer, package="lattice")
# Using data from the 1979 ed. of the
# New York Choral Society

# Histogram of heights
ggplot(singer, aes(x=height)) +
   geom_histogram()

# Boxplot of heights by voice part
ggplot(singer, aes(x=voice.part, y=height)) +
   geom_boxplot()
```



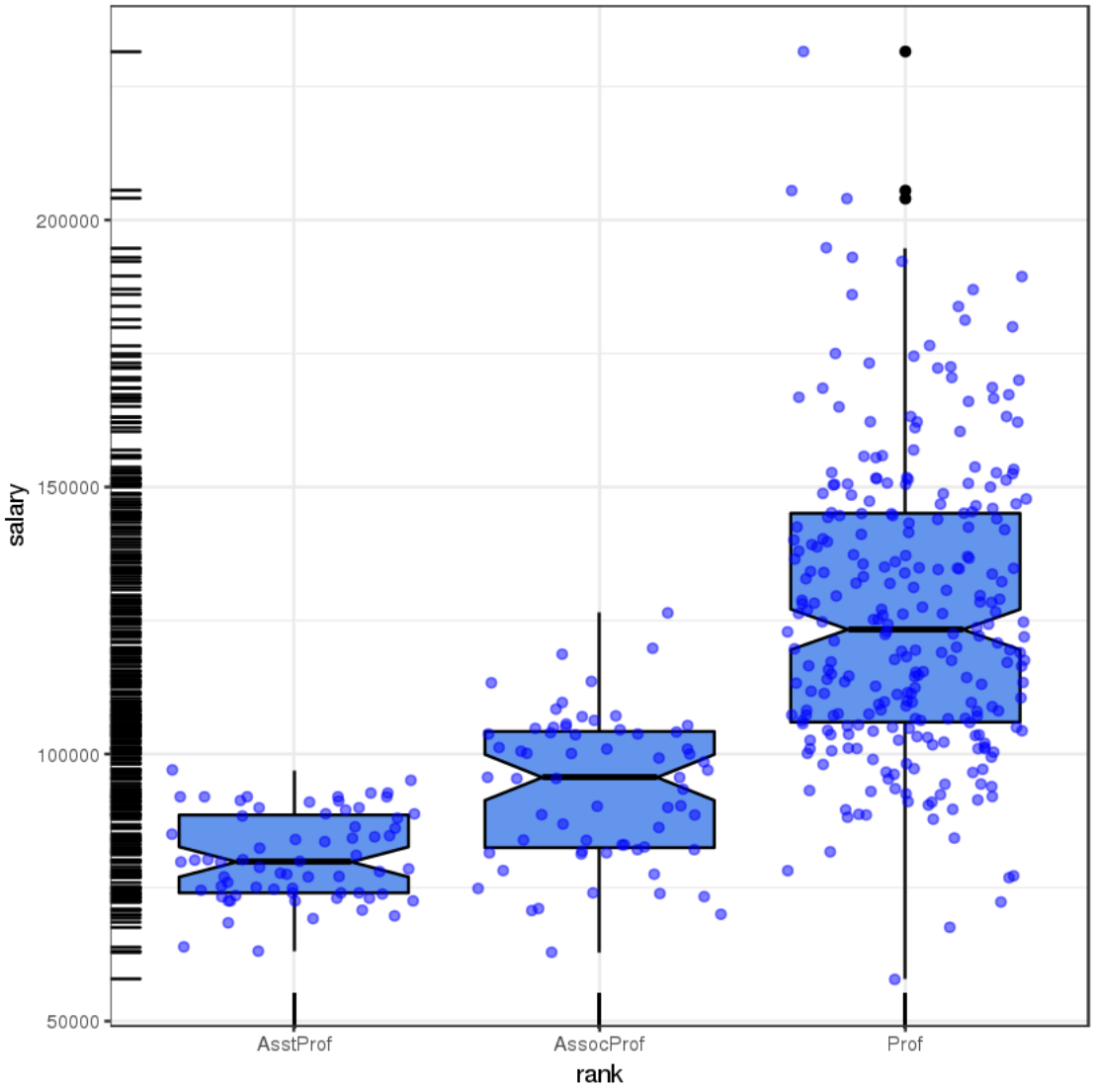
GRAMMAIRE DE GGPLOT2 – GEOM()

```
library(ggplot2)
data(Salaries, package="car")
# Using data on salaries of a sample of
# US university professors (2018-2019)
# var: rank, sex, yrs.since.phd, yrs.service, salary

ggplot(Salaries, aes(x=rank, y=salary)) +
  geom_boxplot(fill="cornflowerblue",color="black", notch=TRUE) +
  geom_point(position="jitter", color="blue", alpha=.5) +
  geom_rug(side="l", color="black")
```

Selon vous, quelle sera la forme du graphique?

GRAMMAIRE DE GGPLOT2 – GEOM()

```
library(ggplot2)
data(Salaries, package="car")
# Using data on salaries of a sample of
# US university professors (2018-2019)
# var: rank, sex, yrs.since.phd, yrs.service, salary

ggplot(Salaries, aes(x=rank, y=salary)) +
   geom_boxplot(fill="cornflowerblue",color="black", notch=TRUE) +
   geom_point(position="jitter", color="blue", alpha=.5) +
   geom_rug(side="l", color="black")
```



GRAMMAIRE DE GGPLOT2 – ESTHÉTIQUE

L'**esthétique** désigne les attributs affichés des données.

Vous devez faire correspondre une donnée à un attribut (comme la taille ou la forme d'un repère) et créer la légende appropriée.

Vous précisez l'esthétique au moyen de la fonction `aes()`.

Vous pouvez préciser l'esthétique dans la fonction *data* ou *geom*. Si vous précisez l'esthétique dans la fonction *data*, l'esthétique vise alors toutes les fonctions *geom* précisées.

GRAMMAIRE DE GGPLOT2 – ESTHÉTIQUE

Les caractéristiques esthétiques offertes avec `geom_point()` (scatterplot), p. ex., sont les suivantes :

- `x`, `y`, `alpha`, `color`, `fill`, `shape`, `size`

Il existe une **différence importante** entre les caractéristiques (comme la couleur et la forme) selon qu'elles sont précisées à l'intérieur et à l'extérieur de la fonction `aes()` :

- à l'intérieur : la couleur ou la forme choisie repose automatiquement sur les données
- à l'extérieur : la caractéristique ne vise pas les données.

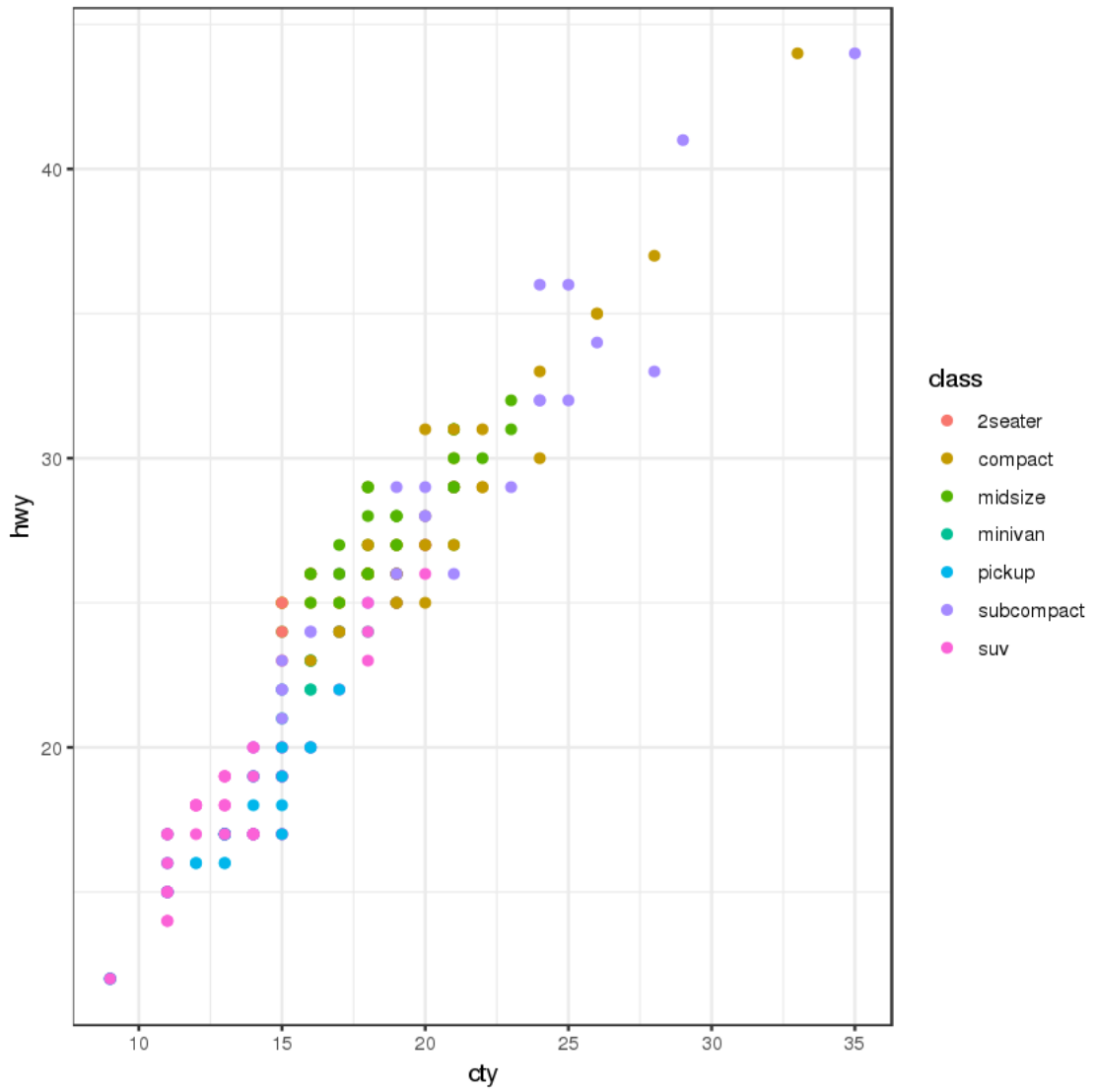
GRAMMAIRE DE GGPLOT2 – AES()

```
library(ggplot2)
# Using the mpg dataset

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(aes(colour = class))

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(colour = "red")
```

Selon vous, quelle sera la forme du graphique?

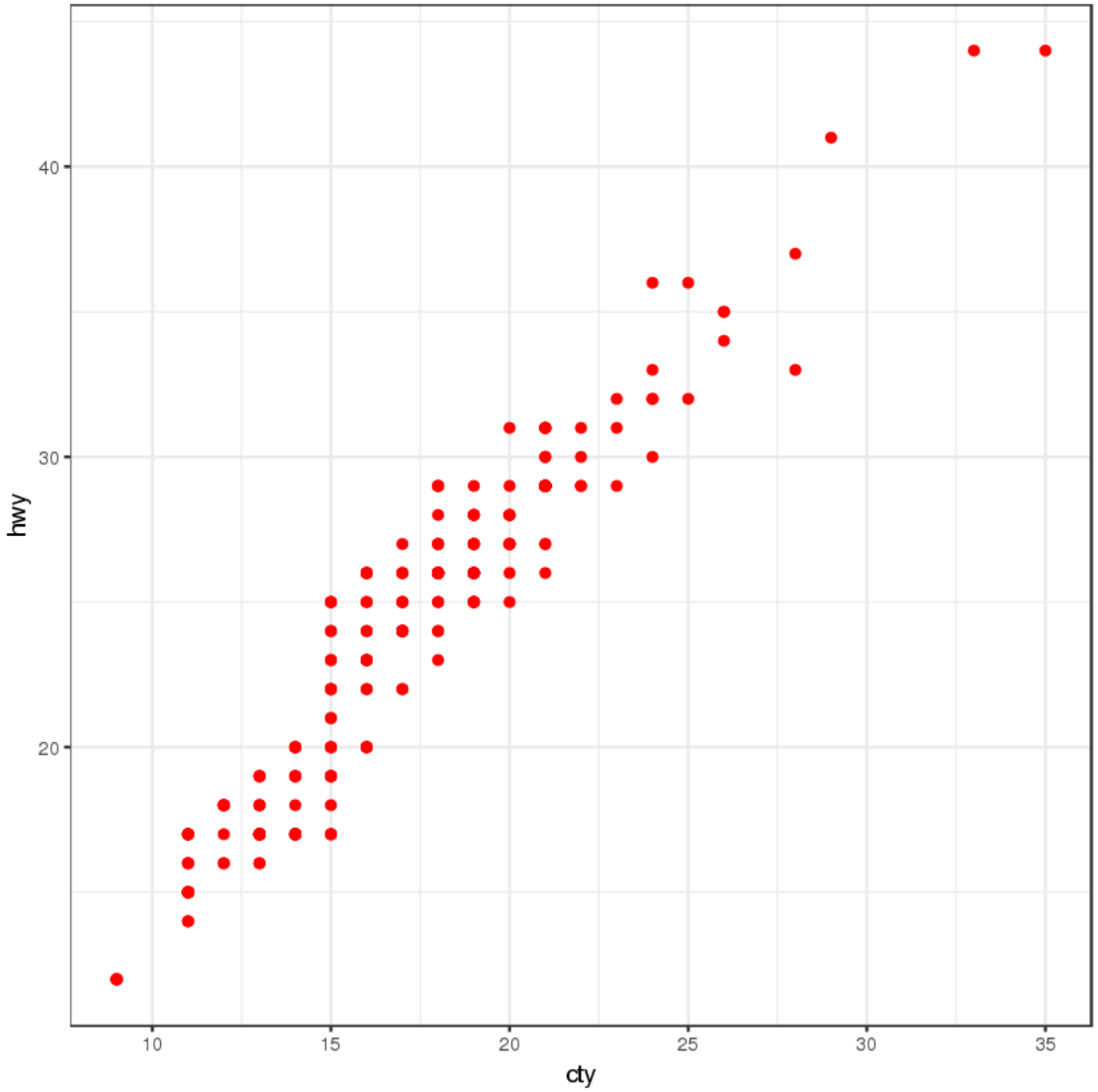
GRAMMAIRE DE GGPLOT2 – AES()

```
library(ggplot2)
# Using the mpg dataset

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(aes(colour = class))

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(colour = "red")
```



GRAMMAIRE DE GGPLOT2 – AES()

```
library(ggplot2)
# Using the mpg dataset

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(aes(colour = class))

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(colour = "red")
```



INTRODUCTION AUX TABLEAUX DE BORD

CONCEPTS AVANCÉS DE LA VISUALISATION DES DONNÉES ET DE LA CRÉATION DE RAPPORTS

« Si un arbre tombe dans la forêt et si personne ne l'entend tomber, est-ce qu'il fait du bruit? »

(vieille énigme)

RAPPORT ET DÉPLOIEMENT

Une analyse est bonne seulement si on la **transmet** ou si on la **déploie**.

Questions essentielles :

- Qui reçoit le rapport?
- Quels sont les flux de travail menant à sa création?
- Est-ce que les données peuvent donner lieu à des politiques utiles?

La création automatique de rapports devrait **régulièrement** faire l'objet d'une vérification et d'une validation.

RAPPORT ET DÉPLOIEMENT

La **communication** devrait avoir lieu à diverses étapes du projet, et non seulement à son achèvement :

- vous devez maintenir les commanditaires et les clients au fait des principaux points;
- vous pouvez délaisser les détails techniques, mais vous devez tout de même les documenter.

Le **scénario idéal** consiste à utiliser un logiciel d'analyse qui permet aussi de créer des rapports :

- minimise l'erreur humaine liée à la fonction copier-coller;
- supprime le besoin de maintenir la séparation entre l'analyse et la création de rapports;
- facilite le partage du travail avec les autres membres du projet.

Vous pouvez simplifier davantage le processus en procédant à un déploiement directement sur le web.

DISCUSSION

Quels sont vos outils de création de rapport favoris?

Jusqu'à quel point devez-vous tester un produit avant de le déployer?

Quel est le coût du déploiement d'un produit défectueux?

TABLEAUX DE BORD

Un **tableau de bord** est un affichage visuel des données qui sert à surveiller les états et à faciliter la compréhension.

Exemples :

- affichage interactif qui permet à l'utilisateur d'explorer les réclamations d'assurance automobile selon la ville, la province, l'âge du conducteur, etc.;
- fichier PDF qui montre les principaux paramètres de vérification et qui est envoyé par courriel chaque semaine au DG d'un ministère;
- écran monté au mur qui montre en temps réel les statistiques d'un centre d'appel;
- application mobile qui permet aux administrateurs d'un hôpital de voir les délais d'attente chaque heure et chaque jour pour l'année courante et l'année précédente.

QUELQUES QUESTIONS À PRENDRE EN COMPTE

Sur le tableau de bord d'une automobile, l'automobiliste doit comprendre **d'un coup d'œil** un petit nombre d'**indicateurs importants** (vitesse, niveau d'essence, phares, etc.). Un tableau de bord qui ne tient pas compte de ces deux caractéristiques peut donner lieu à des conséquences catastrophiques.

Vous devez répondre aux questions suivantes avant de concevoir un tableau de bord :

- Qui est l'**utilisateur** du tableau de bord?
- Quels renseignements doit **transmettre** le tableau de bord?
- Quelles données (catégories) seront utilisées?
- Qu'est-ce qui **figurera** dans le tableau de bord?
- Comment le tableau de bord va-t-il **aider** l'utilisateur?

[https://www.geckoboard.com/blog/building-great-dashboards-6-golden-rules-to-successful-dashboard-design]

LIGNES DIRECTRICES LIÉES À LA CONCEPTION D'UN TABLEAU DE BORD

Nick Smith propose les six règles d'or suivantes :

- **Tenez compte du public** (Qui voulez-vous informer? Est-ce que le DG a réellement besoin de savoir que les serveurs fonctionnent à 88 p. cent de leur capacité?)
- **Sélectionnez le bon type de tableau de bord** (opérationnel, stratégique, analytique)
- **Groupez les données logiquement, utilisez intelligemment l'espace** (séparez les secteurs fonctionnels : produit, ventes et marketing, finances, personnes, etc.)
- **Utilisez des données adaptées au public** (portée et étendue des données, différents tableaux de bord pour différents services, etc.)
- **Évitez d'encombrer le tableau de bord** (présentez seulement les paramètres les plus importants)
- **Actualisez les données à la fréquence appropriée** (en temps réel, chaque jour, chaque semaine, chaque mois, etc.)

[https://dashboard.edmonton.ca]

TRANSPORTATION | LIVABILITY | ENVIRONMENT | URBAN FORM | ECONOMY | FINANCE

Meets or Exceeds Target | Near Target | Needs Improvement | Measuring | Collecting Data

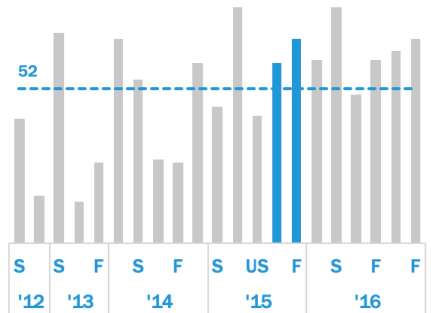
IDLEWYLD Sysabee DAVHILL

data-action-lab.com

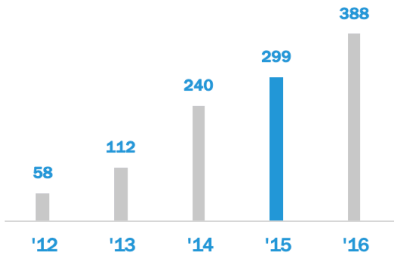
Course Metrics

[https://bigbookofdashboards.com/dashboards.html]

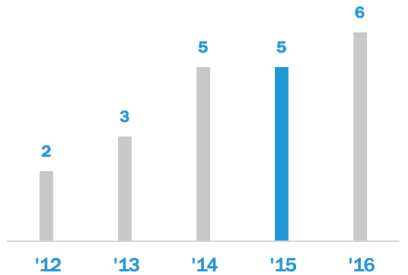
Students



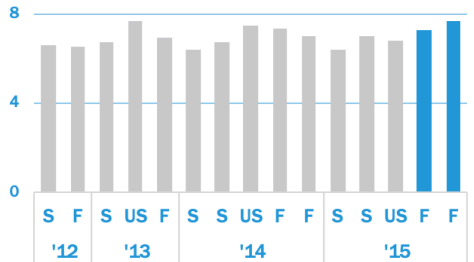
Enrollments



Classes



Ratings



1097

Total Students in five years

687

Total Students in 2015-2016

21

Total Classes in five years

7.7 of **8**

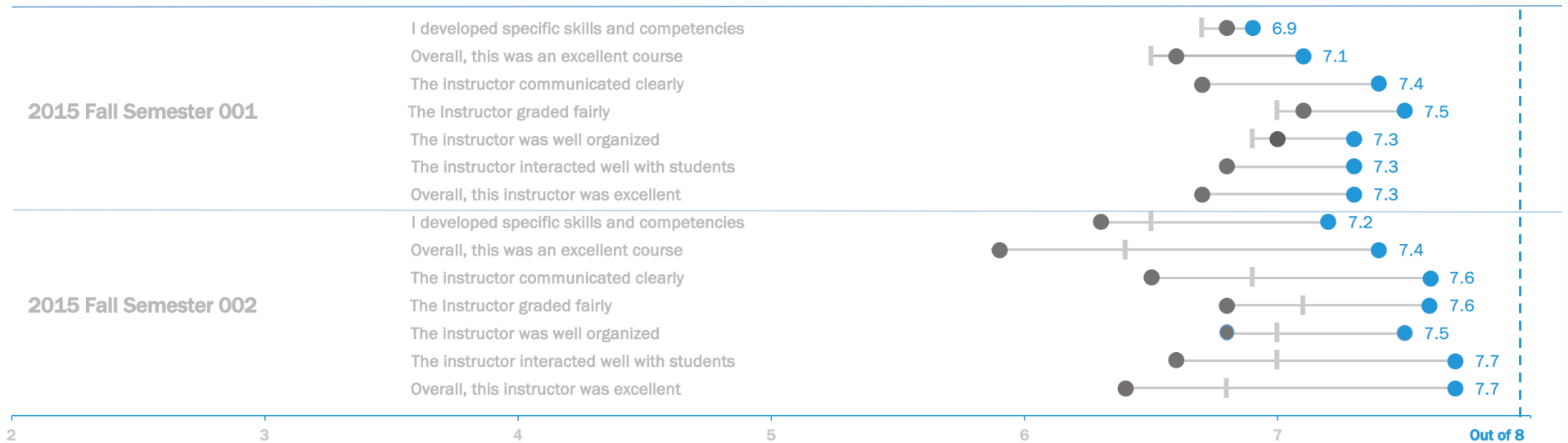
Most recent instructor rating (out of 8.0)

Semesters

Questions

BANA College Shaffer

Ratings



Course Metrics Dashboard created by Jeffrey A. Shaffer. Data from University of Cincinnati Course Evaluations. Blue indicates the 2 most recent rating periods.

TABLEAU DE BORD – POINTS FORTS

Principaux paramètres faciles à voir

Palette de couleurs simplifiée

Possibilité d'un tableau statique ou interactif

Clarté du sommaire et des détails

DISCUSSION

Aucun tableau de bord n'est parfait – aucun ensemble de graphiques ne conviendra à toutes les personnes qui les utiliseront.

Un tableau de bord doit transmettre la **vérité** et être **fonctionnel**; un tableau de bord **élégant** (agréable, plaisant) convient davantage.

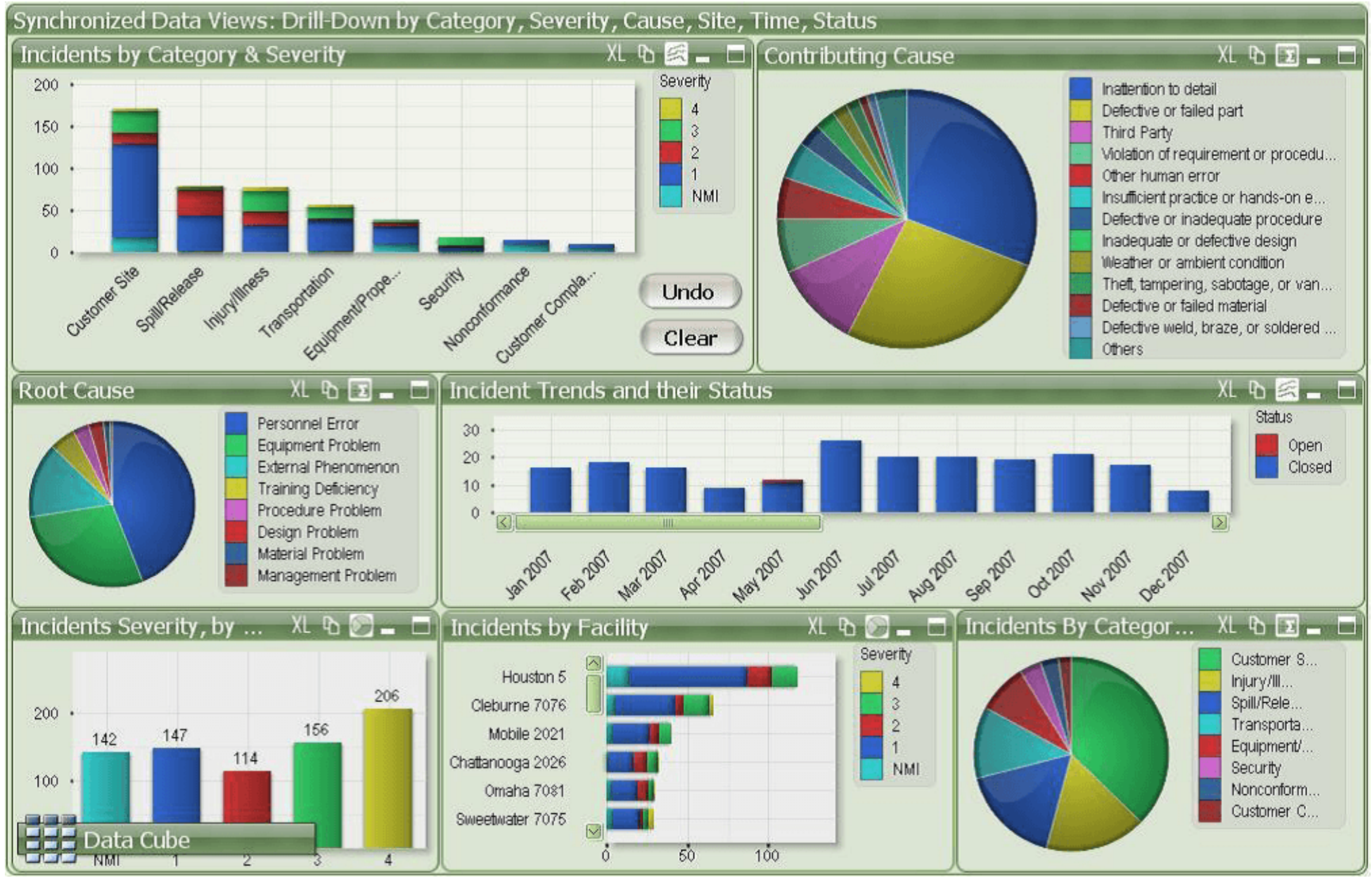
Aucun tableau de bord **n'est complet**. Un bon tableau de bord mène tout de même à un cul-de-sac, mais il devrait permettre à l'utilisateur de se demander « Pourquoi? Quelle est la cause fondamentale du problème? ».

Outils : Excel, Power BI, Tableau, R + Shiny, Geckoboard, Matillion, etc.

EXERCICE

Examinez les prochains tableaux de bord. D'un coup d'œil, pouvez-vous en déterminer le public? Quels sont leurs points forts? Quelles sont leurs limites? Comment pourriez-vous les améliorer?

[https://www.matillion.com/wp-content/uploads/2014/11/qlikview-poor-use-of-dashboard-software.png]



[https://www.geckoboard.com/assets/2-terrible-dashboard-example-min.png]

QUELS EN SONT LES DÉFAUTS?

Tableau de bord 1 : compréhension impossible en un coup d'œil, surutilisation des couleurs, diagrammes à secteurs??

Tableau de bord 2 : visualisations 3D, délimitations et arrière-plan distrayants, absence de données filtrées, nombre insuffisant d'étiquettes et contexte insuffisant.

Tableaux de bord 3 et 4 : ...