Introduction à la science des données

Instructeur: Patrick Boily

uOttawa

Institut de développement professionnel
Professional Development Institute

Patrick Boily

Carrière :

Professeur [uOttawa] (~55 cours/ ~150 journées d'atelier)

Gérant ['12 – '19, CQADS/CAQAD, Carleton]

Fonctionnaire ['08 – '12, ASFC | StatCan | TC | TPSGC]

Clients :

AMC, SGDN, ACSTA, plusieurs autres (~40 projets)

Spécialités :

Visualisation des données, nettoyage des données, application d'un large éventail de méthodes quantitatives.

LES PRINCIPES DE LA VISUALISATION DES DONNÉES

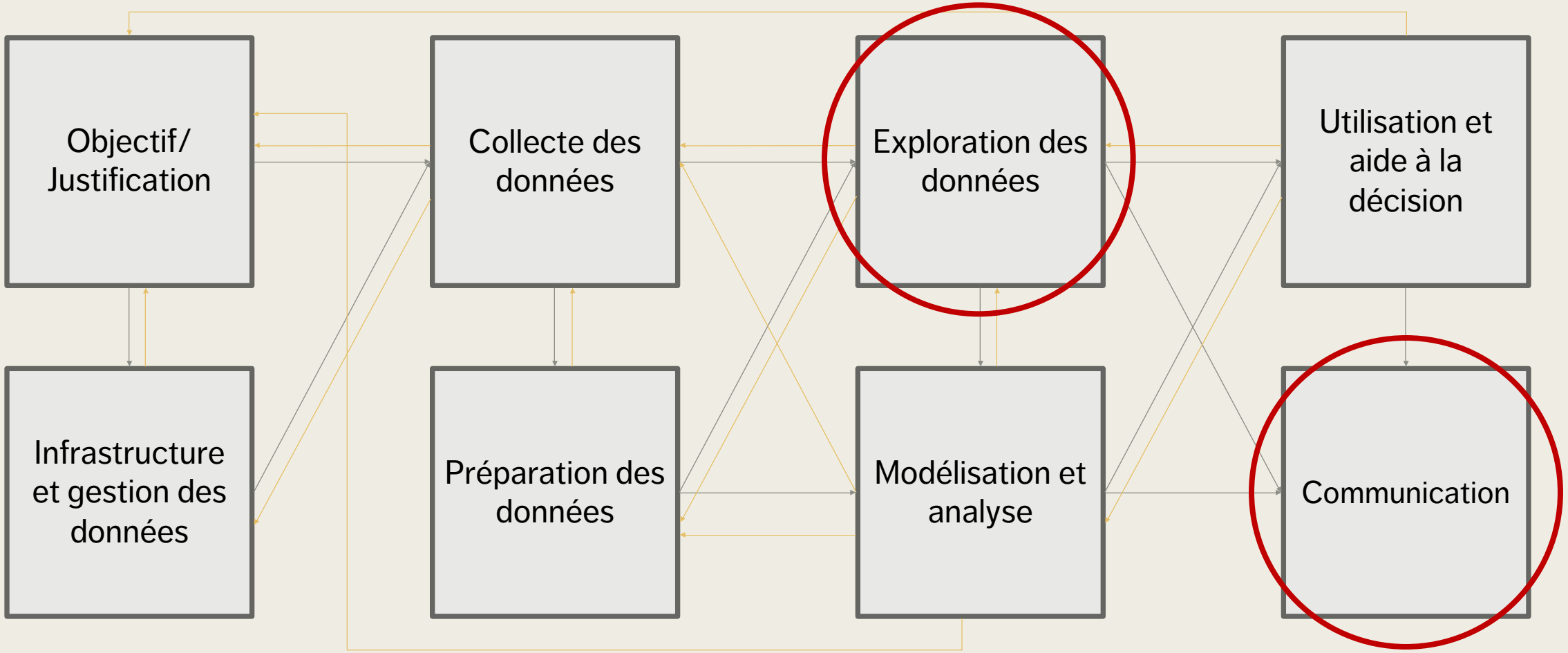
Patrick Boily
Data Action Lab | uOttawa | Idlewyld Analytics
pboily@uottawa.ca

[en collaboration avec Stephen Davies | DAVHILL]

« La recherche ne se limite plus à la collecte et au traitement des données, mais à leur gestion, leur analyse et leur visualisation. »

@DamianMingle

LE « FLUX DE TRAVAIL » DE LA SCIENCE DES DONNÉES



VISUALISATION DES DONNÉES AVANT L'ANALYSE

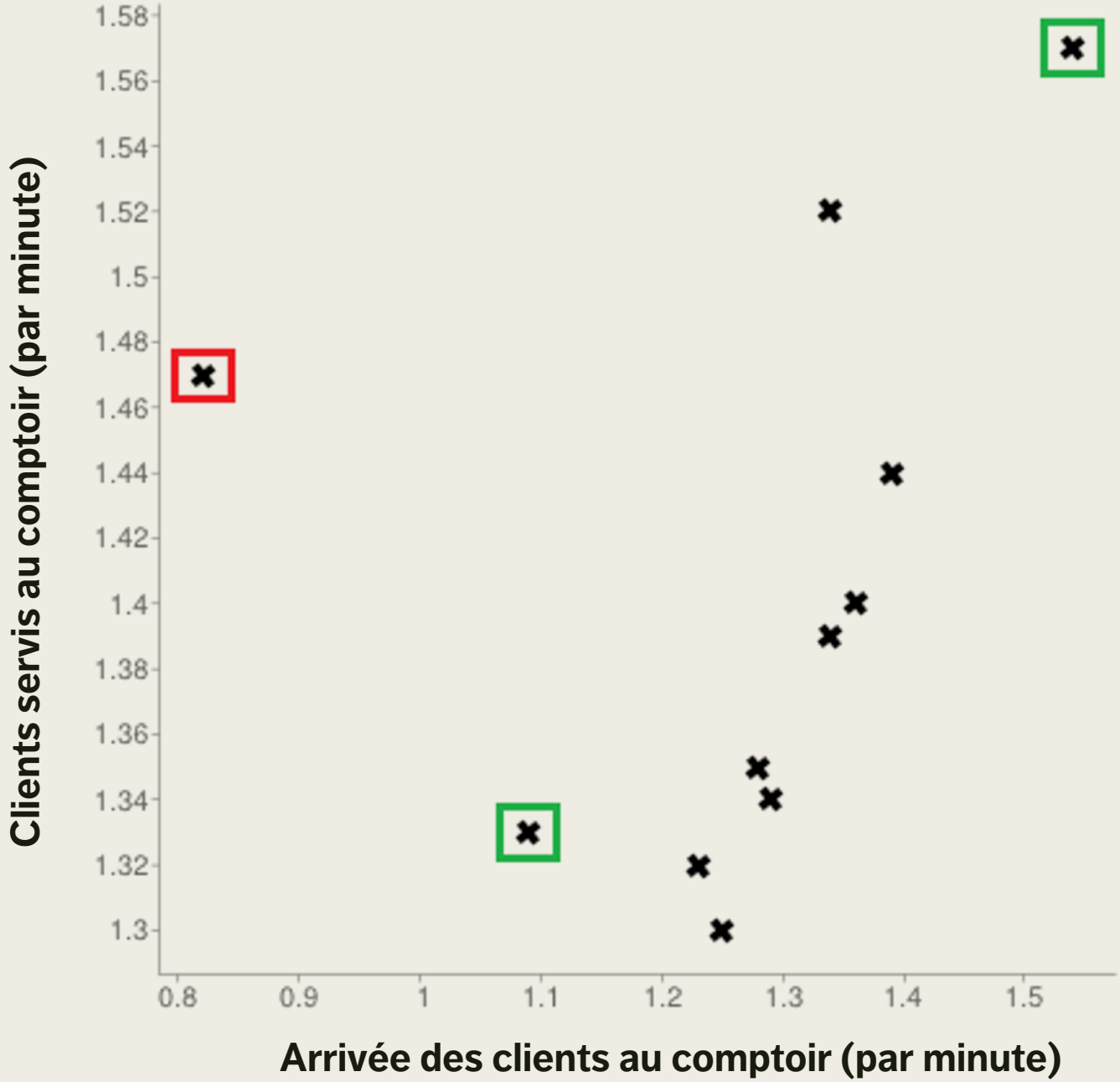
PRINCIPES DE LA VISUALISATION DES DONNÉES

VISUALISATION DES DONNÉES AVANT L'ANALYSE

La visualisation des données peut servir à préparer le terrain pour l'analyse :

- **détection des entrées irrégulières**
  entrées non valides, valeurs manquantes, valeurs aberrantes
- **donner forme aux transformations de données**
  regroupement, normalisation, transformations Box-Cox, transformations de type PCA
- **comprendre les données**
  l'analyse des données comme forme d'art, l'analyse exploratoire
- **identifier les structures de données cachées**
  les regroupements, les associations, les schémas qui informent l'étape suivante de l'analyse

[Ensemble de données personnelles]



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

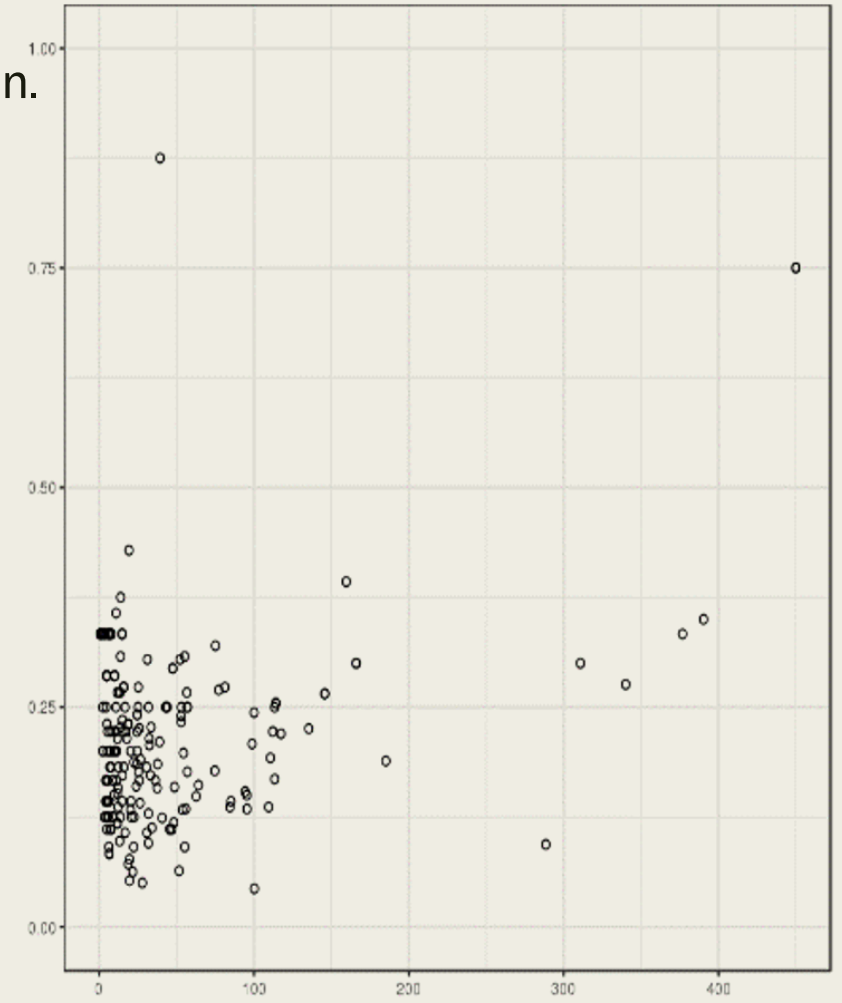
REPRÉSENTER LES OBSERVATIONS

2 variables peuvent être représentées par leur position dans le plan.

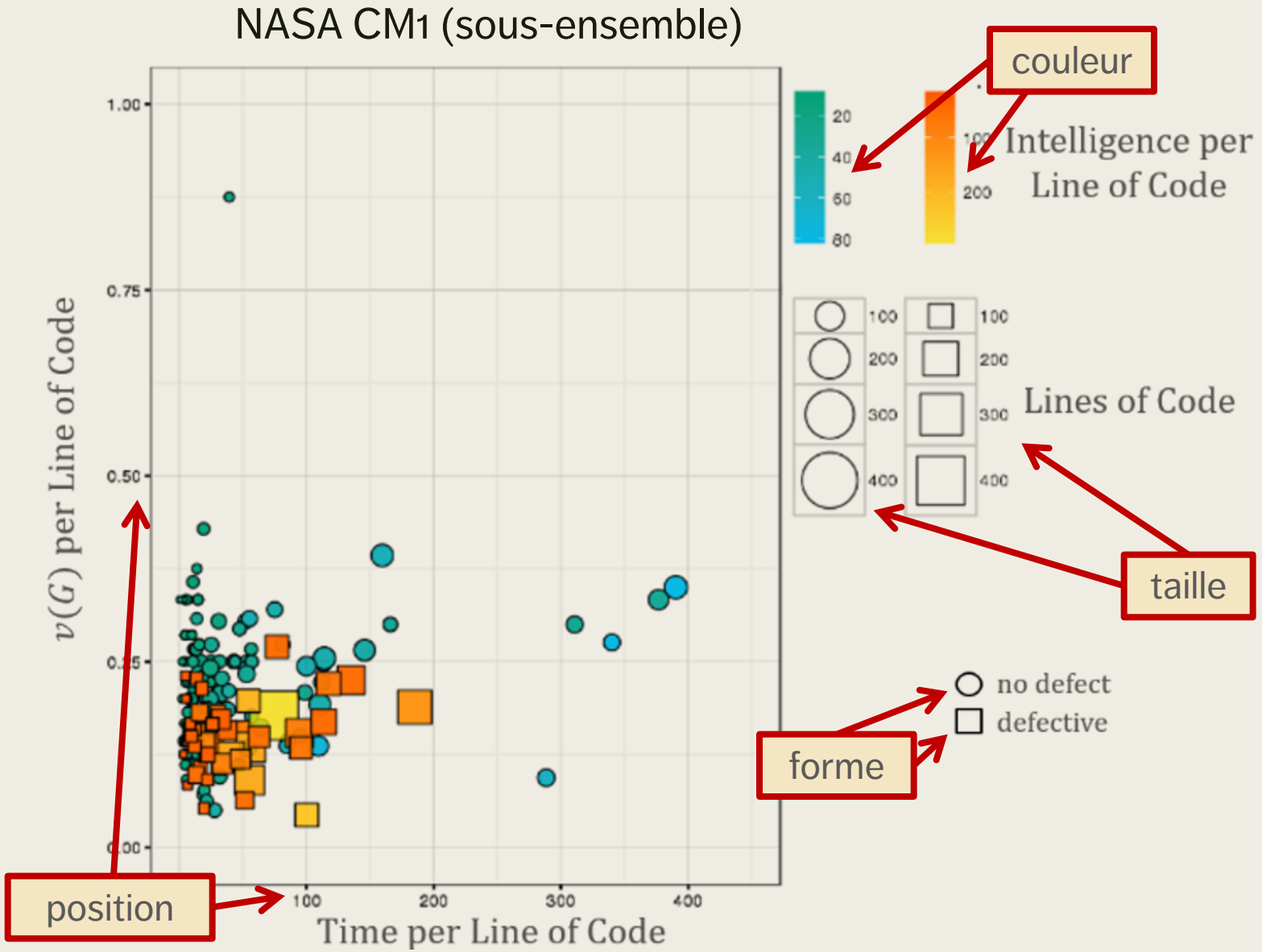
Des facteurs supplémentaires peuvent être représentés par :

- taille
- couleur
- valeur
- texture
- orientation de la ligne
- forme
- (mouvement ?)

NASA CM1 (sous-ensemble)



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



VISUALISATIONS STANDARD

Graphique en ligne/graphique en creux/ligne de chiffres

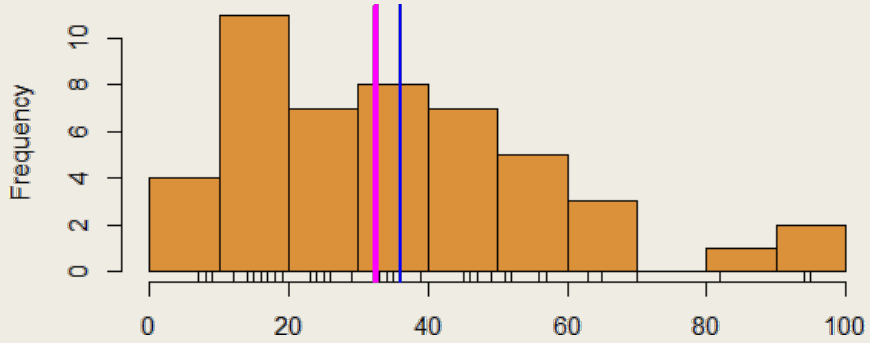
Histogramme

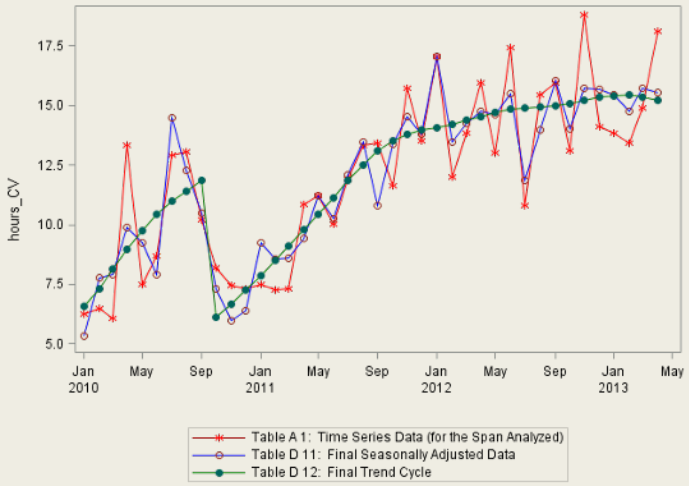
Graphique en ligne

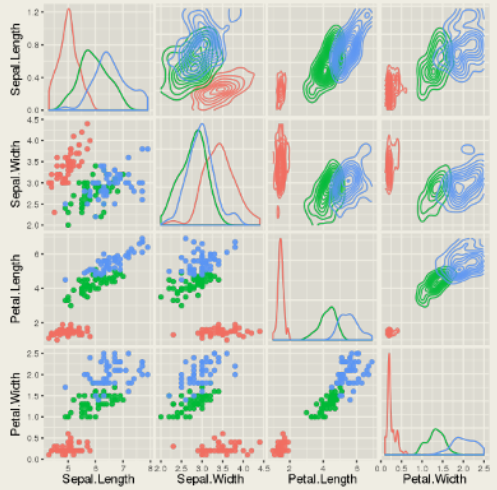
Boîte à moustache

Graphique à barres

Nuage de points







VISUALISATION DES DONNÉES POST-ANALYSE

PRINCIPES DE LA VISUALISATION DES DONNÉES

[E. Tufte, *Beautiful Evidence*]

LES PRINCIPES FONDAMENTAUX DE LA CONCEPTION ANALYTIQUE

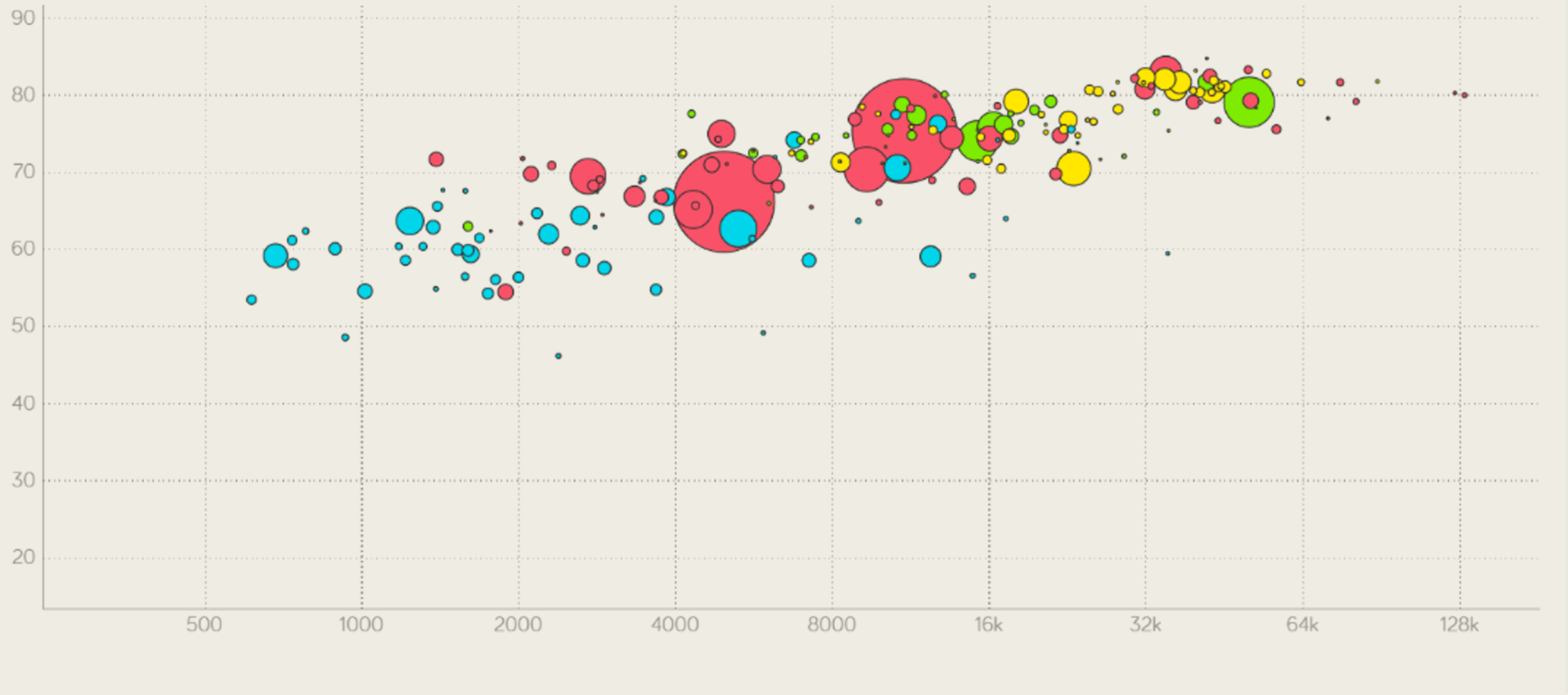
Le **raisonnement** et la **communication** de nos pensées sont intimement liés à notre vie dans un univers multivariable, causal et dynamique.

Symétrie des présentations visuelles des preuves : les consommateurs devraient rechercher exactement ce que les producteurs devraient fournir, par exemple :

- comparaisons significatives
- réseaux de causalité et structure sous-jacente
- liens multivariés
- données intégrées et pertinentes
- documentation honnête
- accent mis sur le contenu

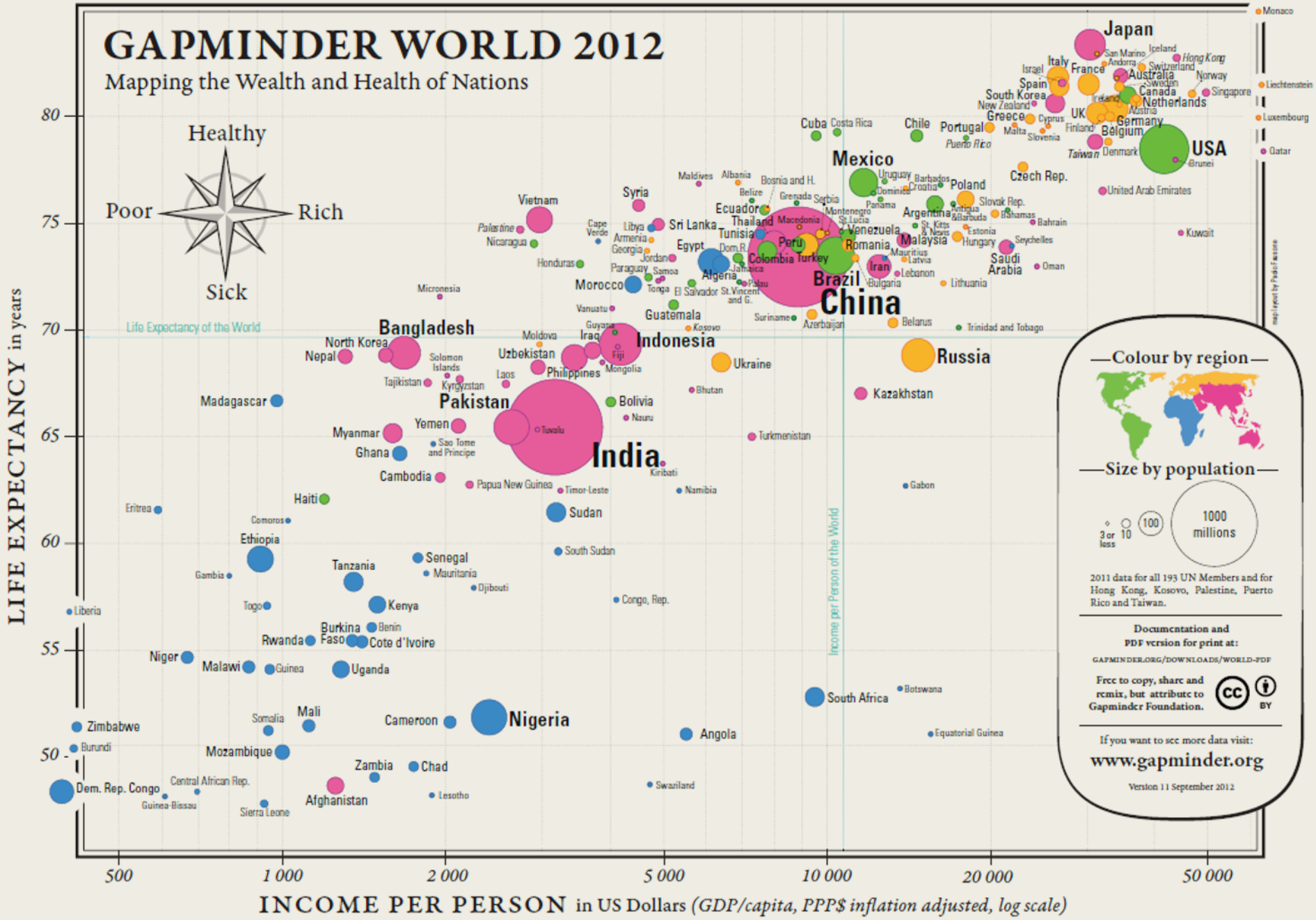
Le message est-il bien compris ?
Le message est-il transmis ?

[Gapminder Foundation]

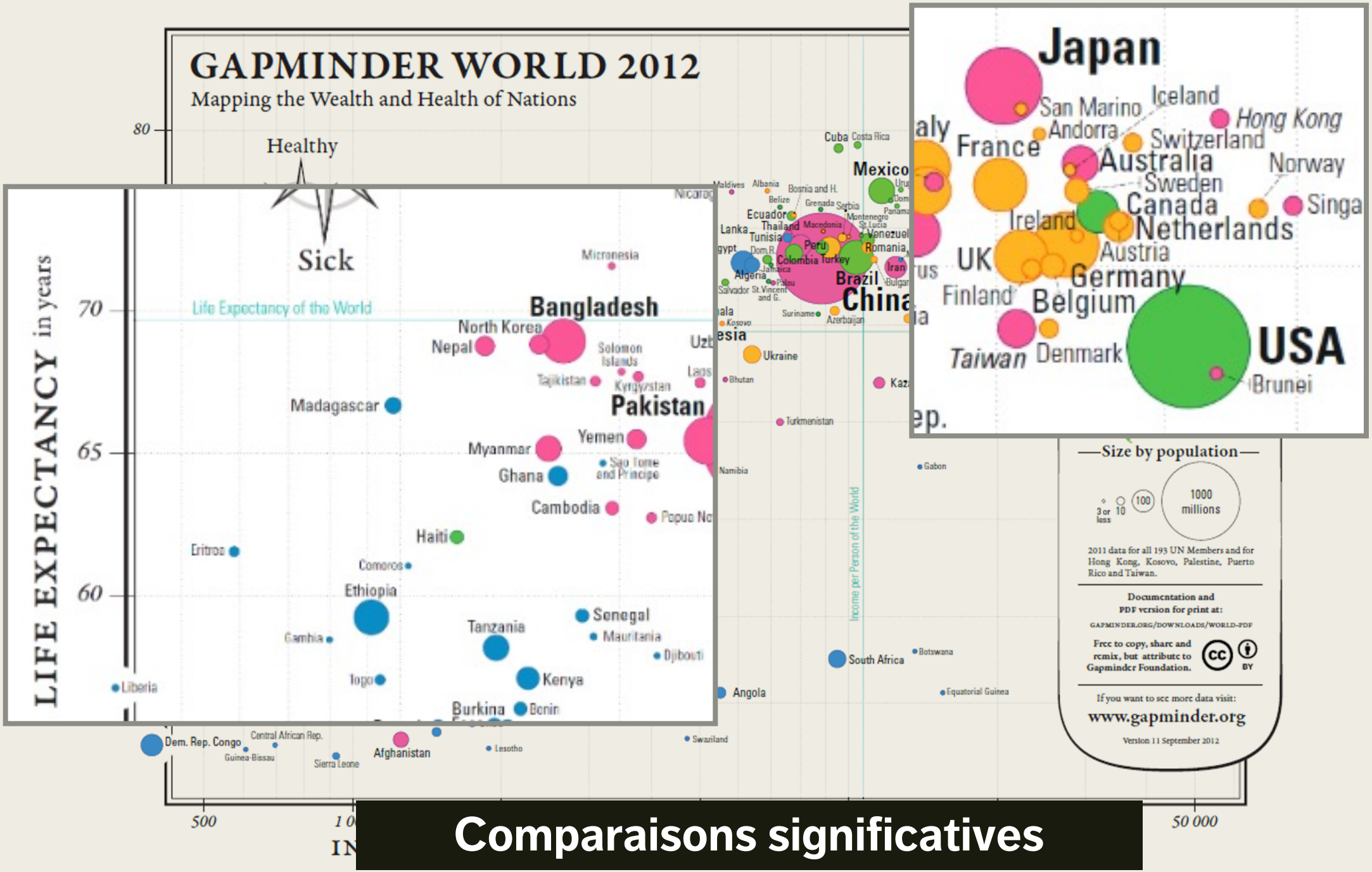


Données non intégrées

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

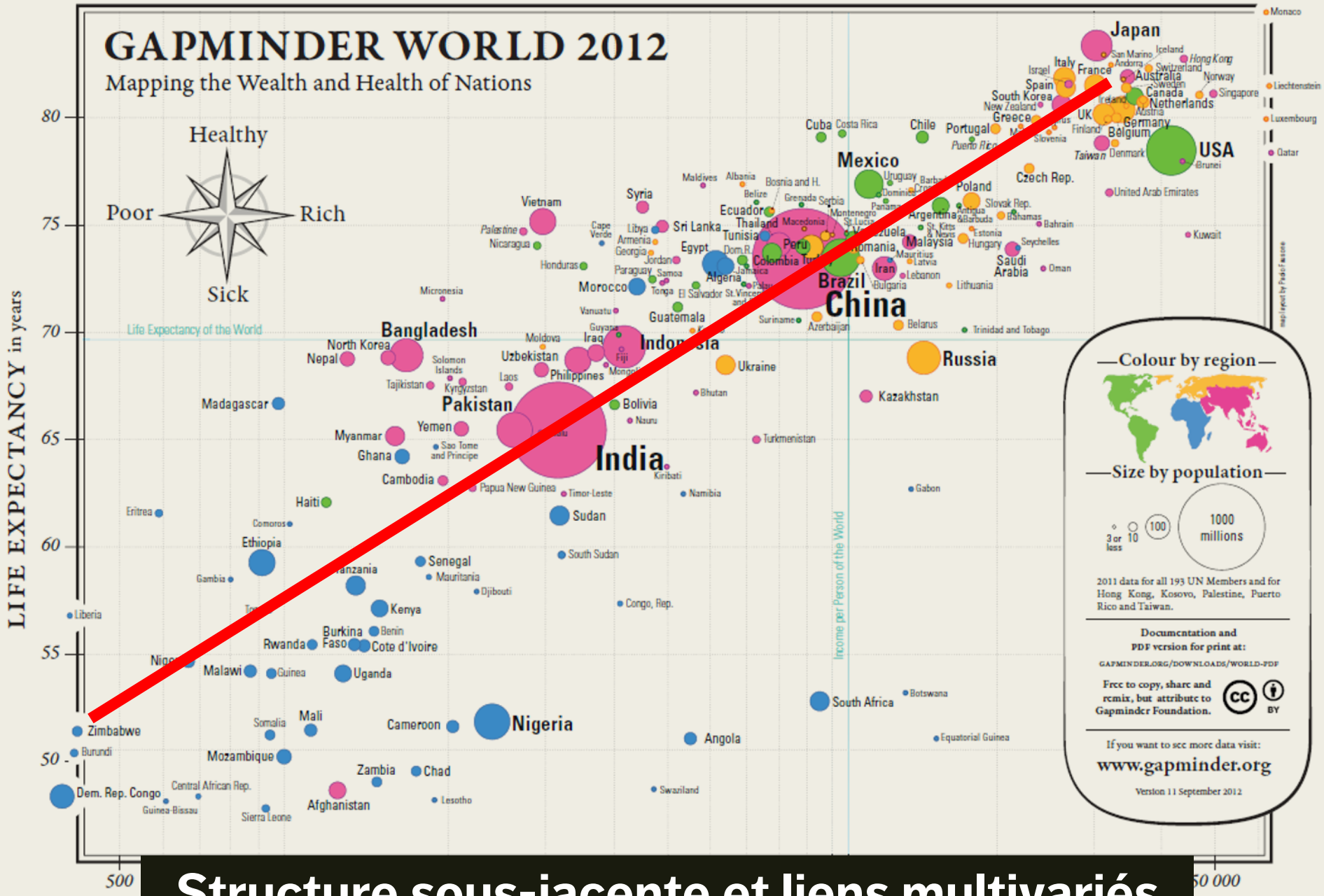


[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



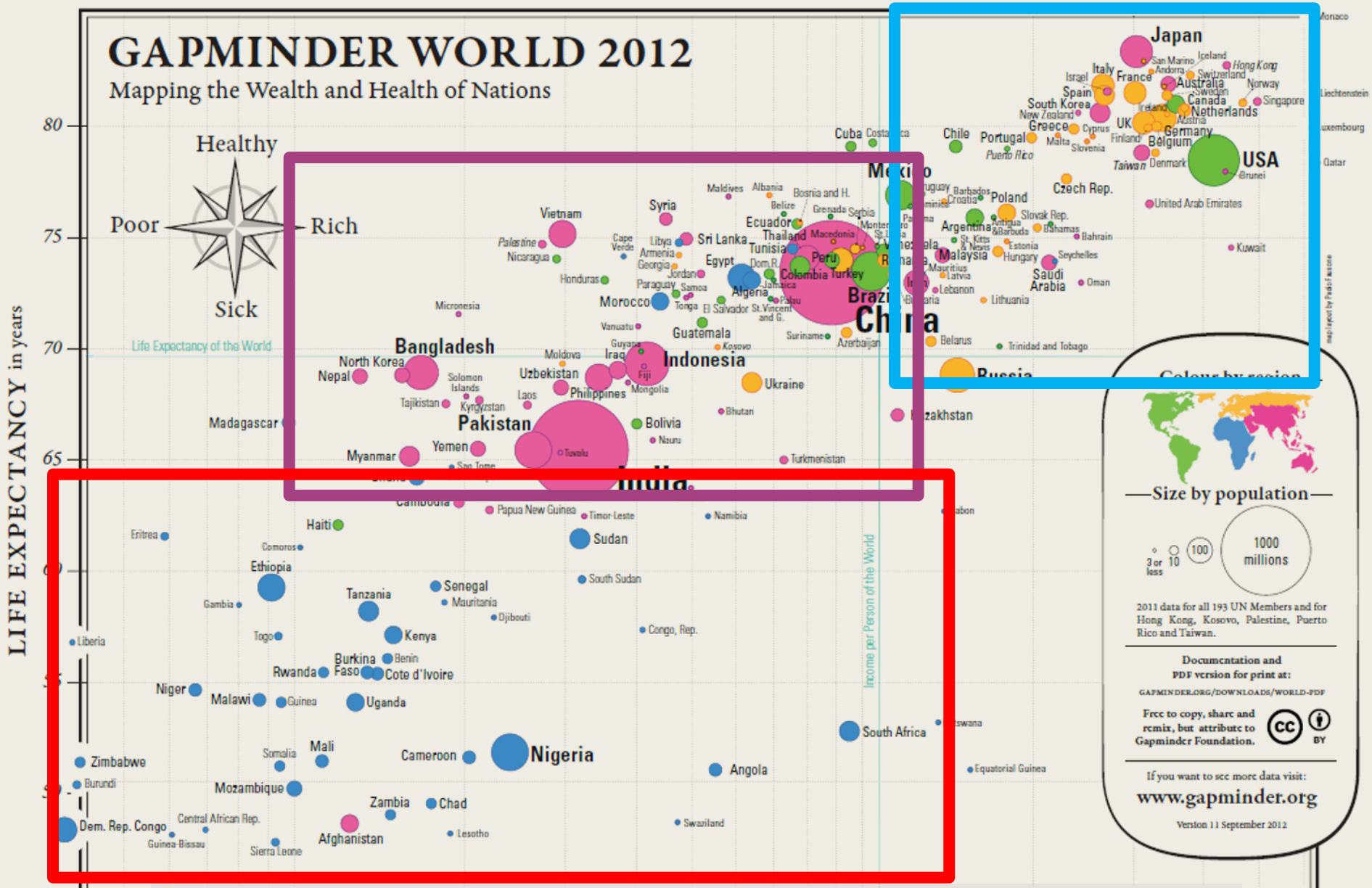
Comparaisons significatives

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



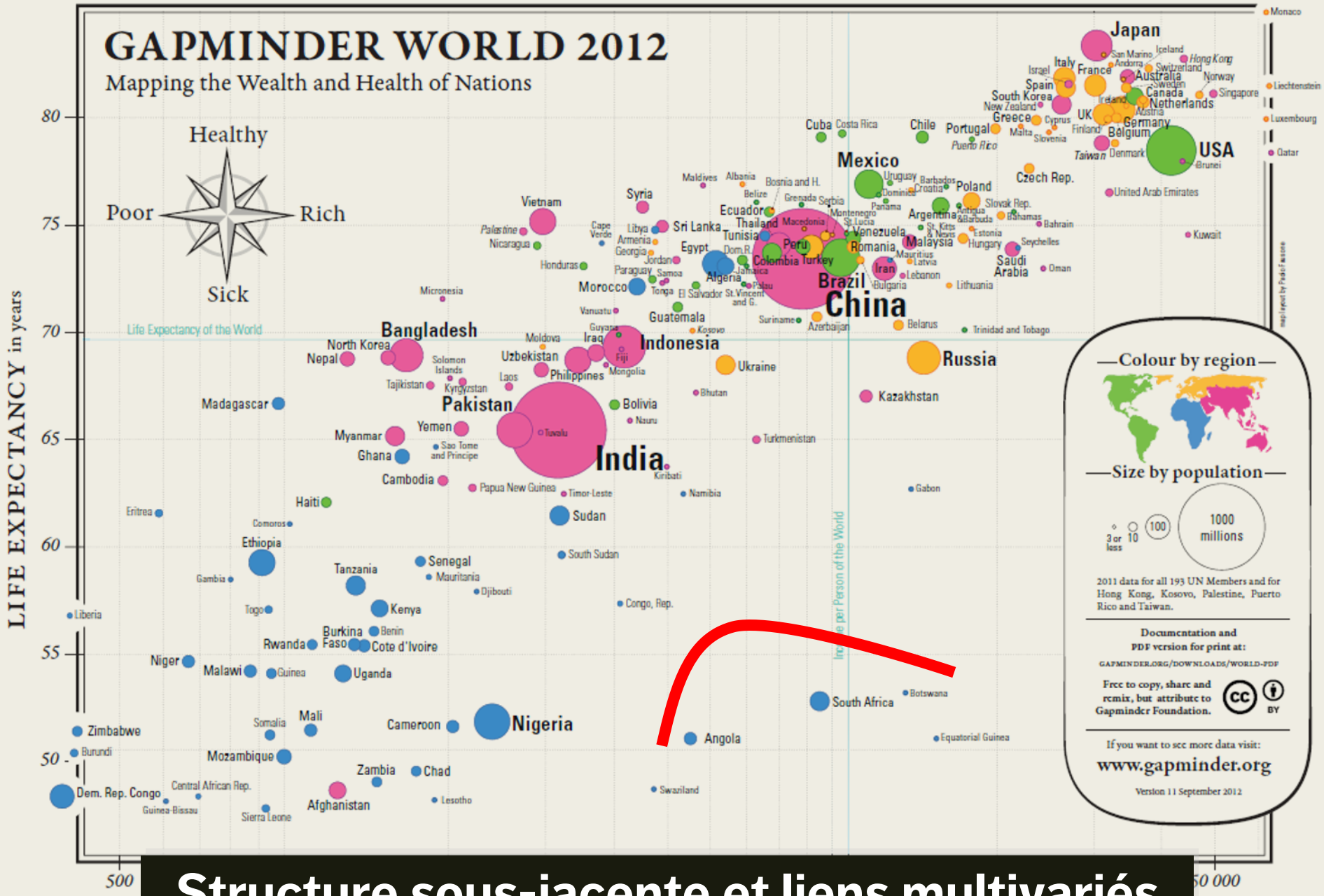
Structure sous-jacente et liens multivariés

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



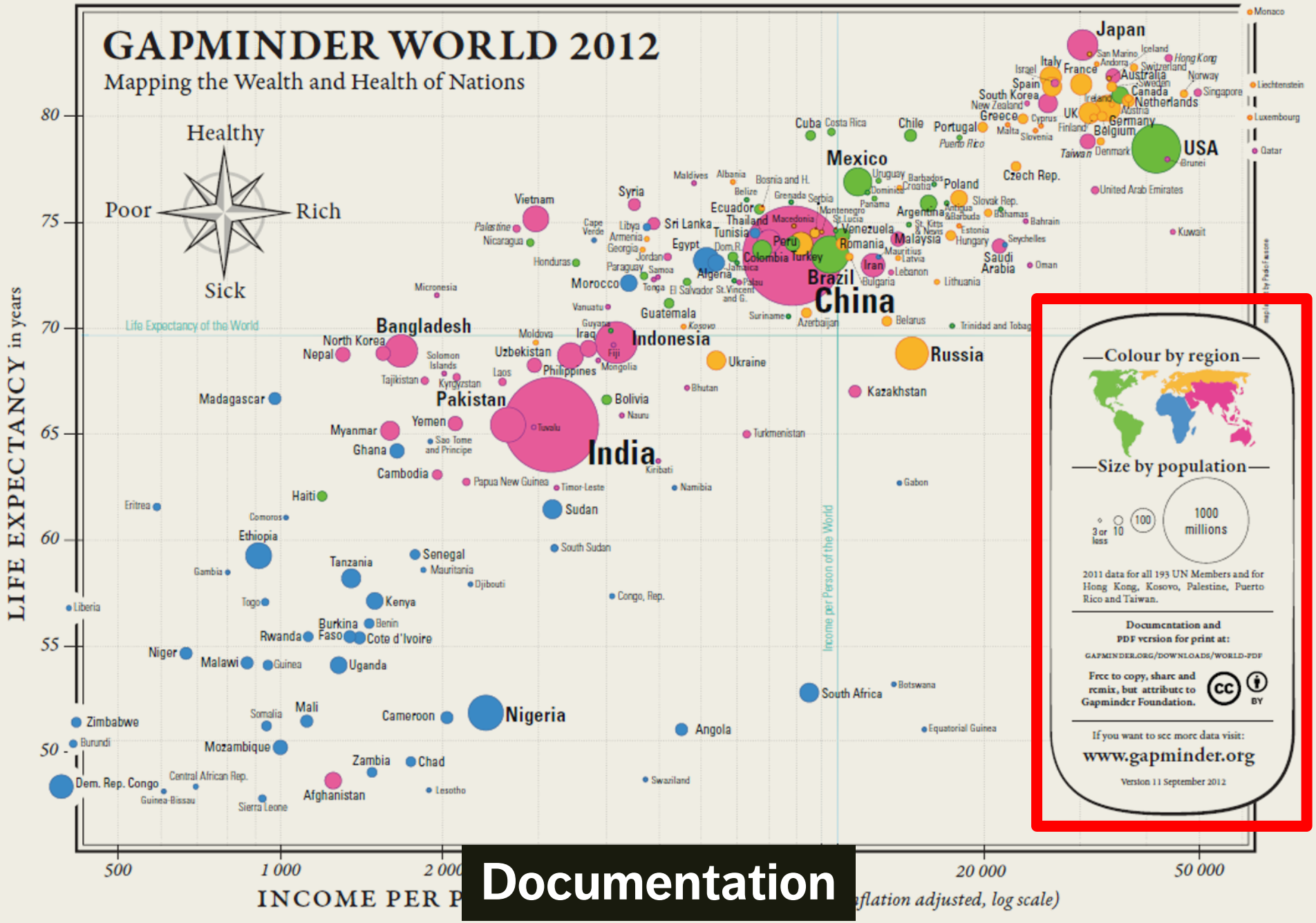
Structure sous-jacente et liens multivariés

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



Structure sous-jacente et liens multivariés

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

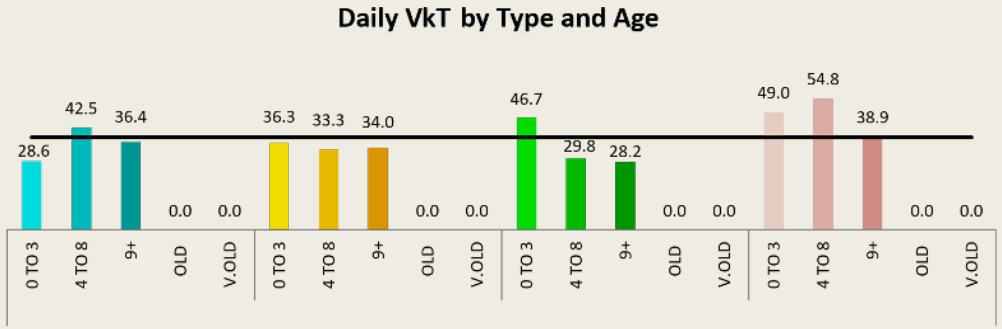


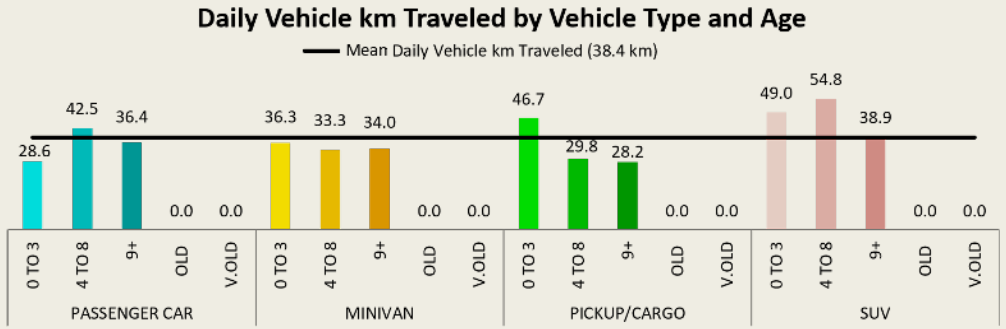
[http://flowingdata.com/tag/pitfalls]

RÈGLES FONDAMENTALES

1. Vérifiez les données
valeurs aberrantes, pics, anomalies

2. Expliquer le codage
ne supposez pas que le lecteur sait ce que tout signifie

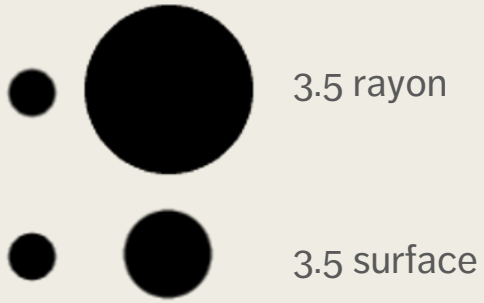




3. Étiqueter les axes
il est important de connaître l'échelle

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

RÈGLES FONDAMENTALES

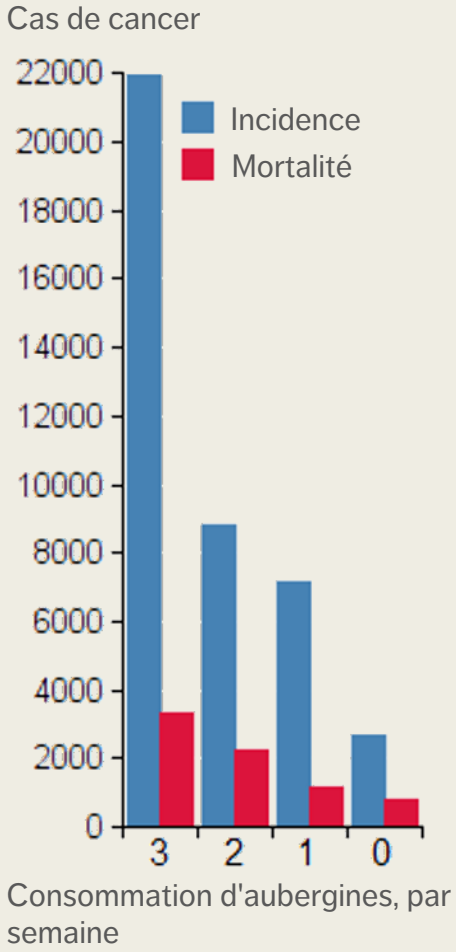


4. Inclure les unités
éliminer le besoin de deviner

5. Contrôlez votre géométrie
les cercles et les formes 2D sont dimensionnés en fonction de la surface, les diagrammes à barres en fonction de la longueur.

6. Indiquez vos sources
protégez-vous, et laissez ceux qui veulent aller plus loin le faire

7. Tenez compte de votre public
une affiche peut être verbeuse, une présentation doit être minimaliste



[http://dataphys.org/list/]

UN MOT SUR L'ACCESSIBILITÉ

Les graphiques ne peuvent généralement pas être traduits en braille. Décrire les caractéristiques et les structures émergentes d'une visualisation est une solution possible... **si on peut les repérer**.

Les analystes doivent produire des visualisations claires et significatives, mais ils doivent également les décrire, ainsi que leurs caractéristiques, de manière à ce que tout le monde puisse « voir » les informations. Pour cela, il faut que les analystes aient « vu » toutes les informations, ce qui n'est pas toujours possible.

Conditions : daltonisme, basse vision, handicap moteur, handicap cognitif, TDAH, etc.

Meilleures pratiques : texte/éléments à fort contraste, zoom/agrandissement, navigation au clavier, conception assistée, résumés courts, fonction défaire/refaire, etc. [F. Elavsky]

UN MOT SUR L'ACCESSIBILITÉ

Perception des données :

- représentations basées sur la texture
- conversion texte-parole
- son/musique
- représentations basées sur l'odeur ou le goût (?!?)

Sonifications:

- Sons TRAPPIST : le système planétaire TRAPPIST-1 traduit directement en musique
- Écoute des données du Grand collisionneur de hadrons, L. Asquith

CATALOGUE DES VISUALISATIONS

PRINCIPES DE LA VISUALISATION DES DONNÉES

A CLASSIFICATION OF CHART TYPES

Data comparison charts

Comparison

Bars

Dot plot

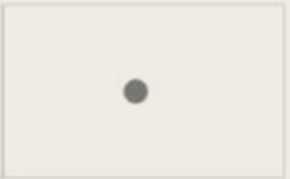
Bullet

1D Scatterplot

Heat map

Slope

Alert



Composition

Pie

Pareto

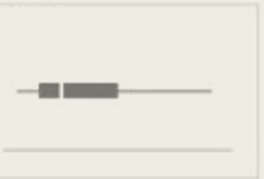
Multidimensional Pie

Distribution

Histogram

1D Scatterplot

Boxplot

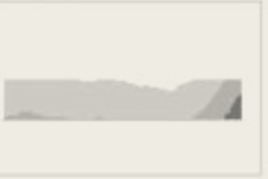


Data reduction charts

Evolution

Line

Horizon



Step

Connected Scatterplot

Relationship

Scatterplot

Connected Scatterplot

Bubble

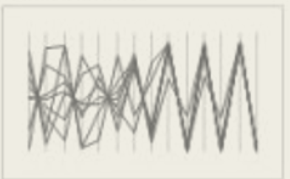
Profiling

Grouped bars

Cycle plot

Reorderable matrix

Parallel Plot



Scatterplot matrix

Horizon

Trellis

v 0.9

excelcharts.com

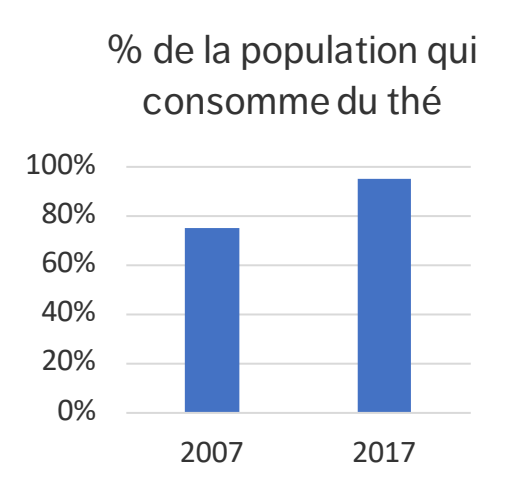
AFFICHAGE DES DONNÉES

Avec les affichages de données, nous essayons de mettre en évidence :

1. **une relation** (montrer un lien ou une corrélation entre deux ou plusieurs variables) ;
2. **une comparaison** (distinguer certaines variables des autres et montrer comment ces deux variables interagissent) ;
3. **une composition** (rassembler différents types d'informations qui forment un tout et les afficher ensemble), et
4. **une distribution** (présenter une collection d'informations liées ou non pour voir comment elles sont corrélées, le cas échéant, et pour comprendre s'il y a une interaction entre les variables).

TEXTE SIMPLE ET TABLEAUX

Un ou deux chiffres sur lesquels se concentrer peuvent aider à « planter le décor » et à attirer l'attention sur une partie du rapport.



95% de la population boit du thé aujourd'hui, contre 75% en 2007

Les tableaux interagissent avec notre système **verbal** (nous les **lisons**) :

- utilisé pour **comparer** des valeurs
- le lecteur recherchera **leurs** lignes

La conception de la table doit être discrète :

- les **données** doivent ressortir, pas les bordures
- tableau dense : utilisez une couleur de ligne **alternée**

Utilisez la couleur pour exprimer l'intensité/magnitude :

- utiliser la **saturation d'une seule couleur**
- utiliser une légende pour éliminer les valeurs

TABLEAUX ET CARTES THERMIQUES DE TABLEAUX

Nom	L'année dernière	Cette année
Ron	20	30
Fred	30	40
George	10	15

Nom	L'année dernière	Cette année
Ron	20	30
Fred	30	40
George	10	15

	Last Year	This Year	Next Year	Optimum
George	20	20	20	20
Peter	40	35	30	25
John	10	10	5	5
Sandra	25	30	35	40

	Last Year	This Year	Next Year	Optimum
George	20	20	20	20
Peter	40	35	30	25
John	10	10	5	5
Sandra	25	30	35	40

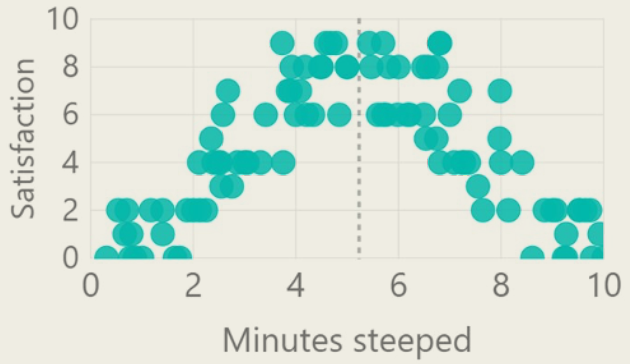
	Last Year	This Year	Next Year	Optimum
George				
Peter				
John				
Sandra				

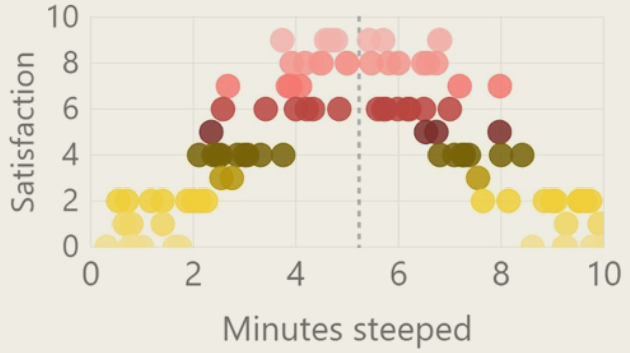
DIAGRAMMES DE DISPERSION (« SCATTERPLOTS »)

Montrer la relation entre 2 variables (diagramme de dispersion) ou 3 variables (diagramme à bulles) :

- utiliser des lignes moyennes (lignes en pointillés) pour fournir un contexte
- beaucoup moins d'options dans Power BI que dans R ou Excel
- envisagez d'utiliser des regroupements pour plus de clarté (par exemple, des **gradients** de couleur).

Combien de temps faut-il laisser infuser la tasse de thé parfaite ?





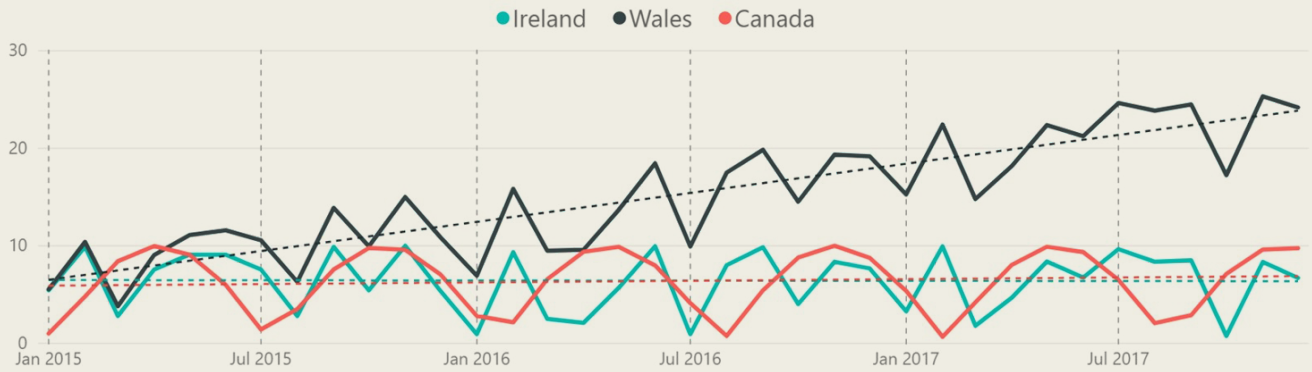
GRAPHIQUE EN LIGNE

Le graphique en ligne peut montrer une seule série de données (particulièrement utile pour les **séries chronologiques**).

L'échelle de l'axe doit être **claire** et **pertinente**.

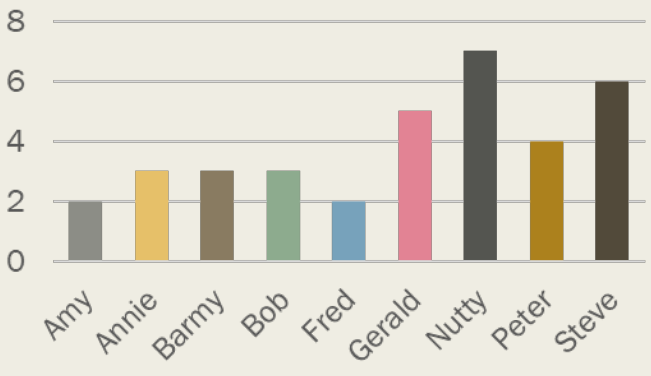
Vous pouvez souhaiter « **ancrer** » l'axe des y si vous utilisez des filtres dynamiques.

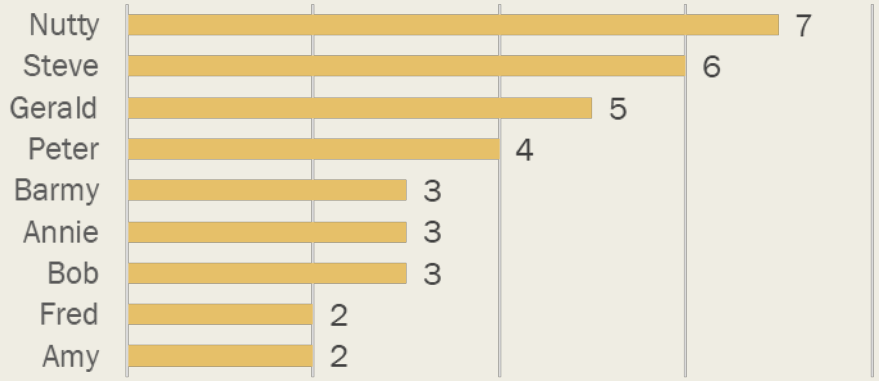
- sinon, le graphique peut se déplacer au fil des interactions.



Comparaison des pays - tasses de thé bues par semaine et par personne

GRAPHIQUE À BARRES





Polyvalent et utile.

Utilisez TOUJOURS (?) une ligne de base zéro.

Utilisez l'axe du graphique OU les étiquettes de données : l'axe pour les déclarations générales, les étiquettes de données pour les détails.

Les graphiques horizontaux sont apparemment **plus faciles à lire** (selon de nombreuses études).

Pensez à l'ordre des catégories.

TYPES DE GRAPHIQUES

Diagrammes à barres empilées

Diagrammes à barres à 100

Graphiques de zone

Treemaps

Graphiques à jauges

Cartes thermiques et cartes de choroplèthe

Cartes géographiques

Coordonnées parallèles

Visages de Chernoff

Nuages de mots

Diagrammes de réseau

Dendrogrammes et arbres

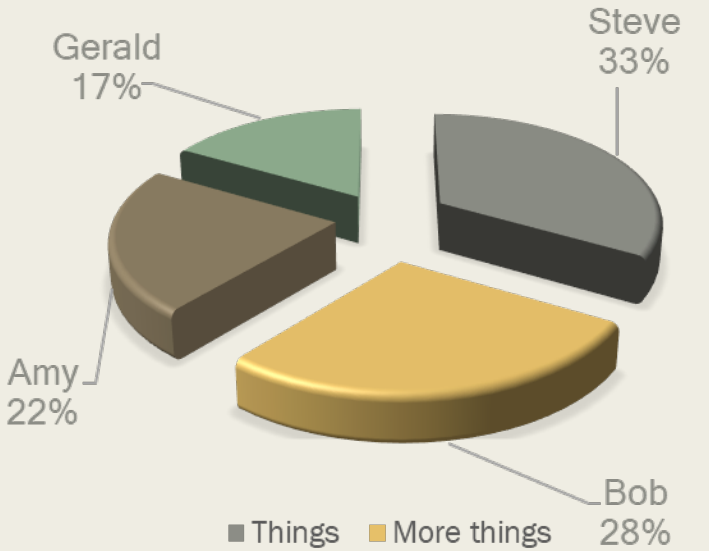
Sparklines

Graphiques interactifs

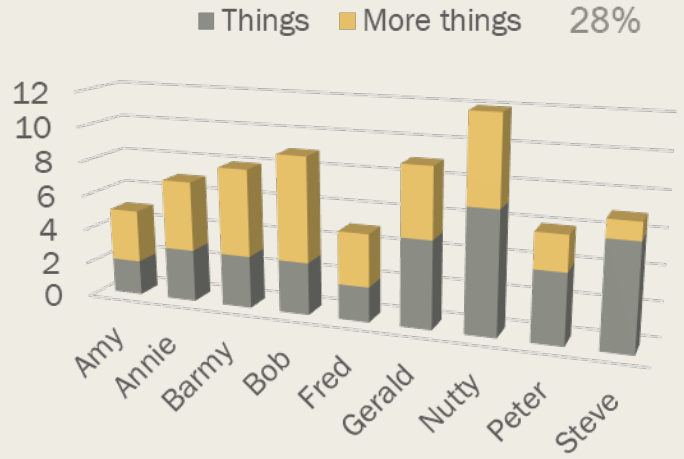
Petits multiples, etc.

GRAPHIQUES À ÉVITER

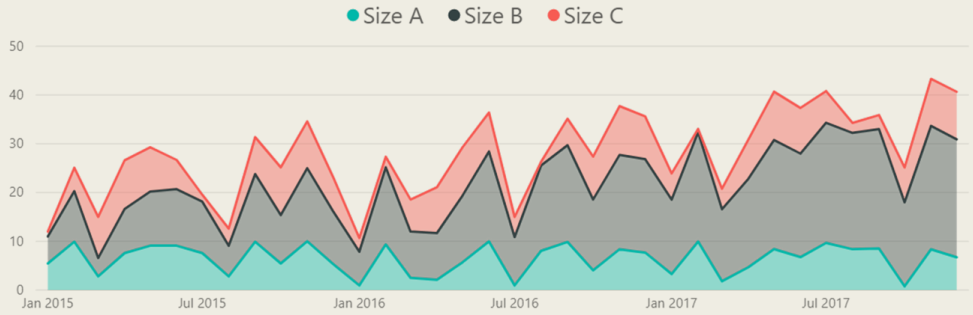
ÉVITEZ (?) tout ce qui comporte un arc (à l'exception de graphiques à jauge) : tarte, beignet, etc. : le cerveau humain a du mal à **comparer les arcs** – sans étiquette, quelle est la différence entre Steve et Bob ?



ÉVITEZ les graphiques en 3D : il est difficile de les comparer visuellement (et ils sont trop **encombrés**).



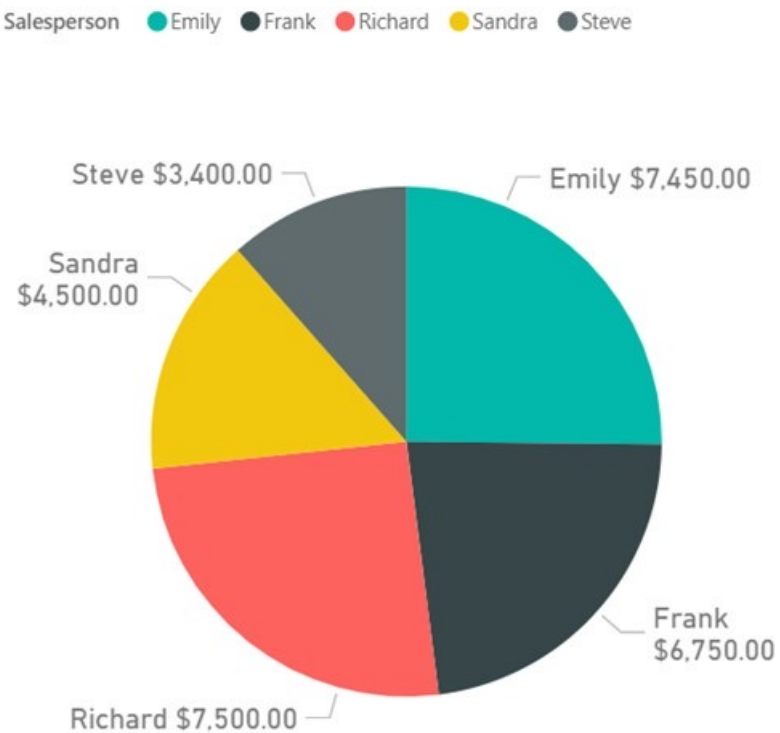
ÉVITEZ les graphiques à zones empilées : ils sont bcp trop déroutants.



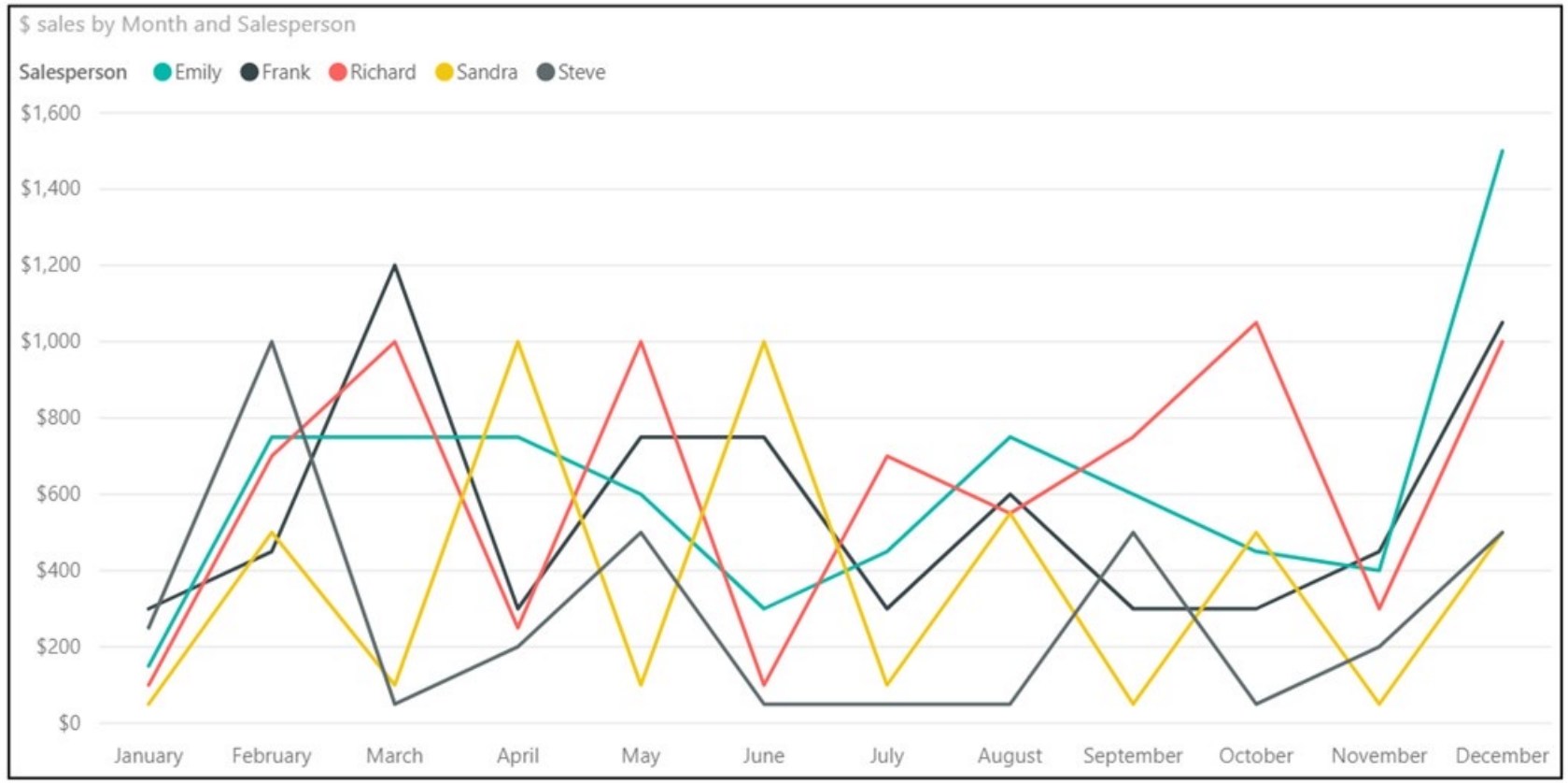
Sales Dashboard

$ sales

$ sales by Salesperson



$ sales by Month and Salesperson



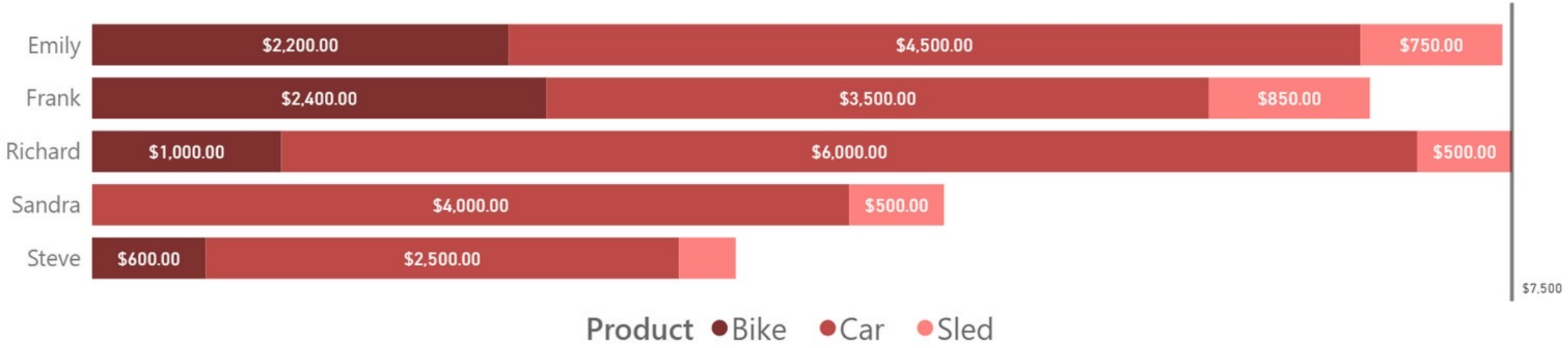
$ sales by Product and Salesperson

Sales Dashboard

Annual Sales for 2017

Total Sales

$29.6K



POINTS À RETENIR

Effective data visualizations **provide insights** and **facilitate understanding**.

The basic principles can guide your visualization design and consumption.

Be **creative** but keep your data and your representations **honest**.

Be mindful of attempts to distort trends and conclusions with flashy visuals.

Data and code should be made available along with the displays.

RÈGLES DE BASE CONCEPTION ET MISE EN PAGE

PRINCIPES DE LA VISUALISATION DES DONNÉES

TRAITEMENT VISUEL

La perception est **fragmentée** – les yeux scrutent constamment.

La pensée visuelle recherche des modèles

- **Processus pré-attentifs :** rapides, instinctifs, efficaces, multitâches
  recueillir des informations et construire des modèles :

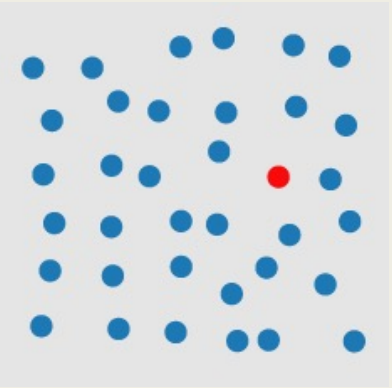
  caractéristiques → modèles → objets

- **Processus attentif :** lent, délibéré, cibléd
  découvrir des caractéristiques dans les modèles :

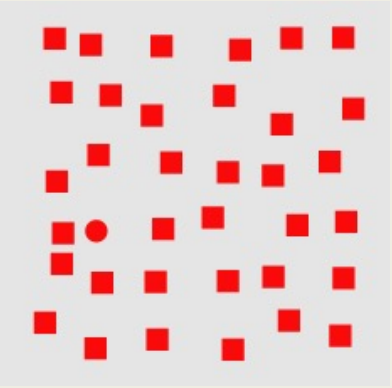
  objets → modèles → caractéristiques

Défi : mettre en évidence un aspect d'un graphique peut rendre les autres aspects plus difficiles à voir.

pré-attentifs

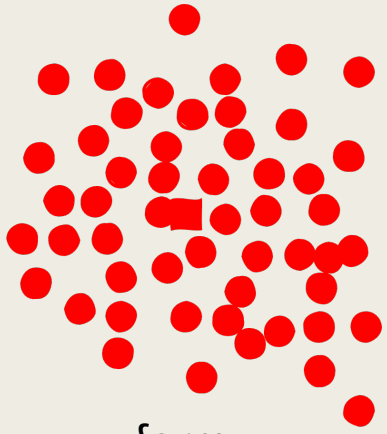


attentif

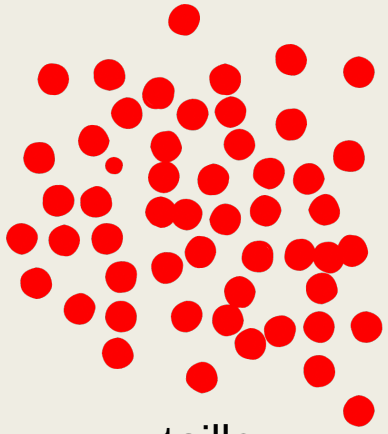


[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]

ATTRIBUTS PRÉ-ATTENTIFS

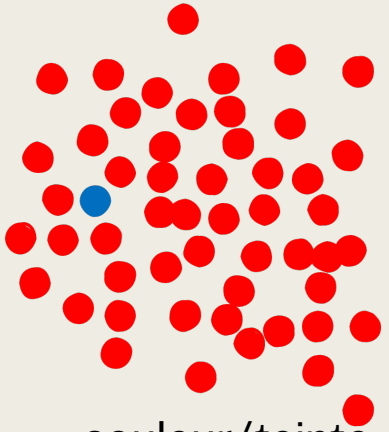


form

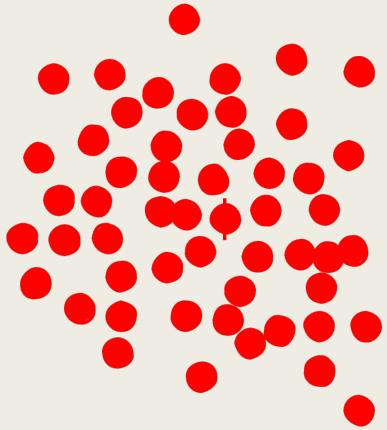


taille

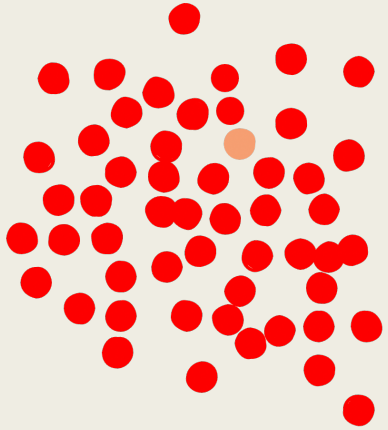
netteté



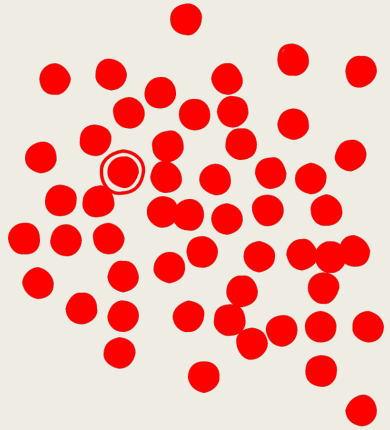
couleur/teinte



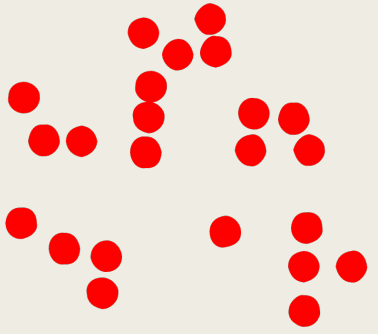
marquages



intensité/valeur



clôture



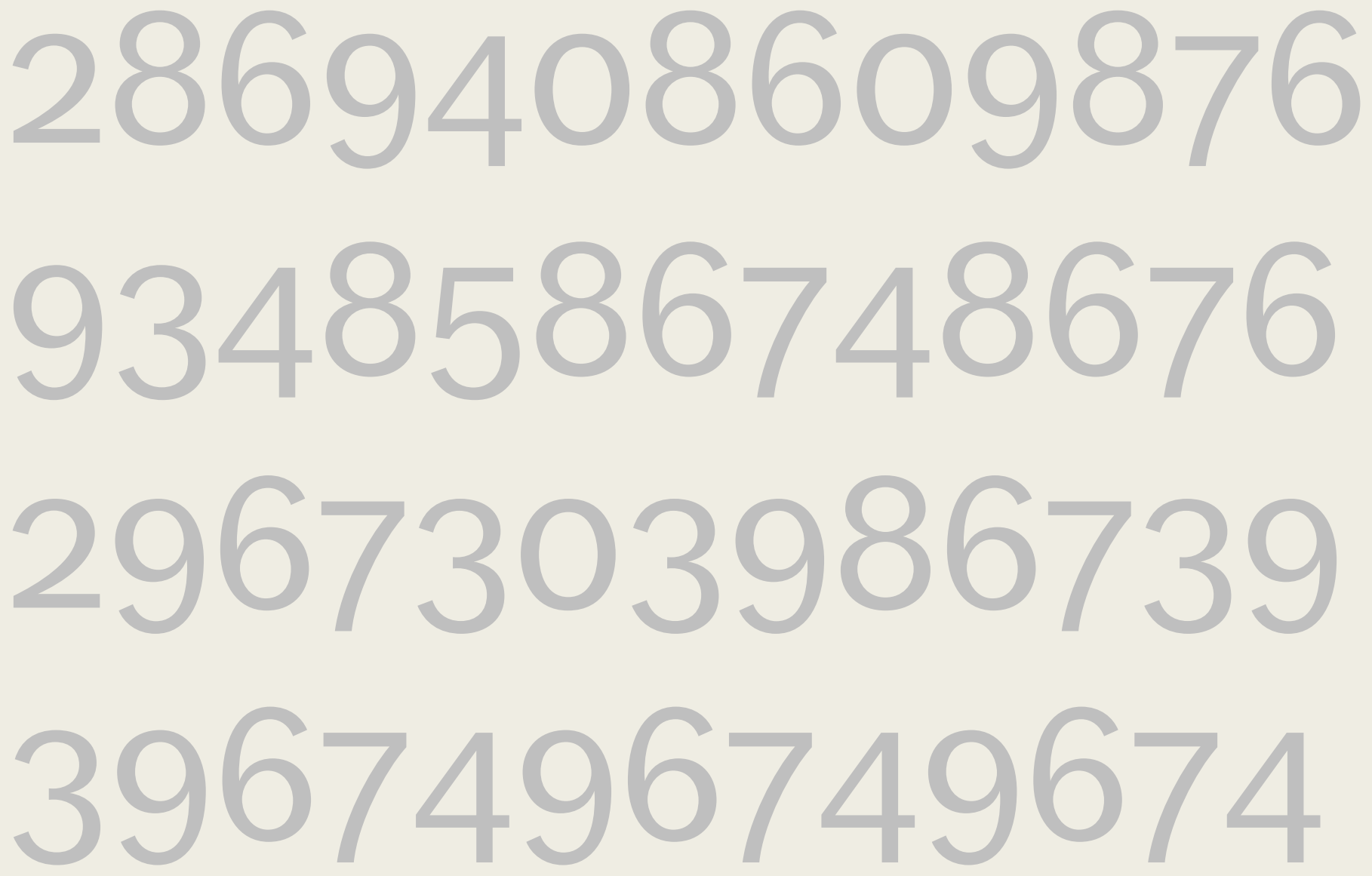
numérosité

[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]

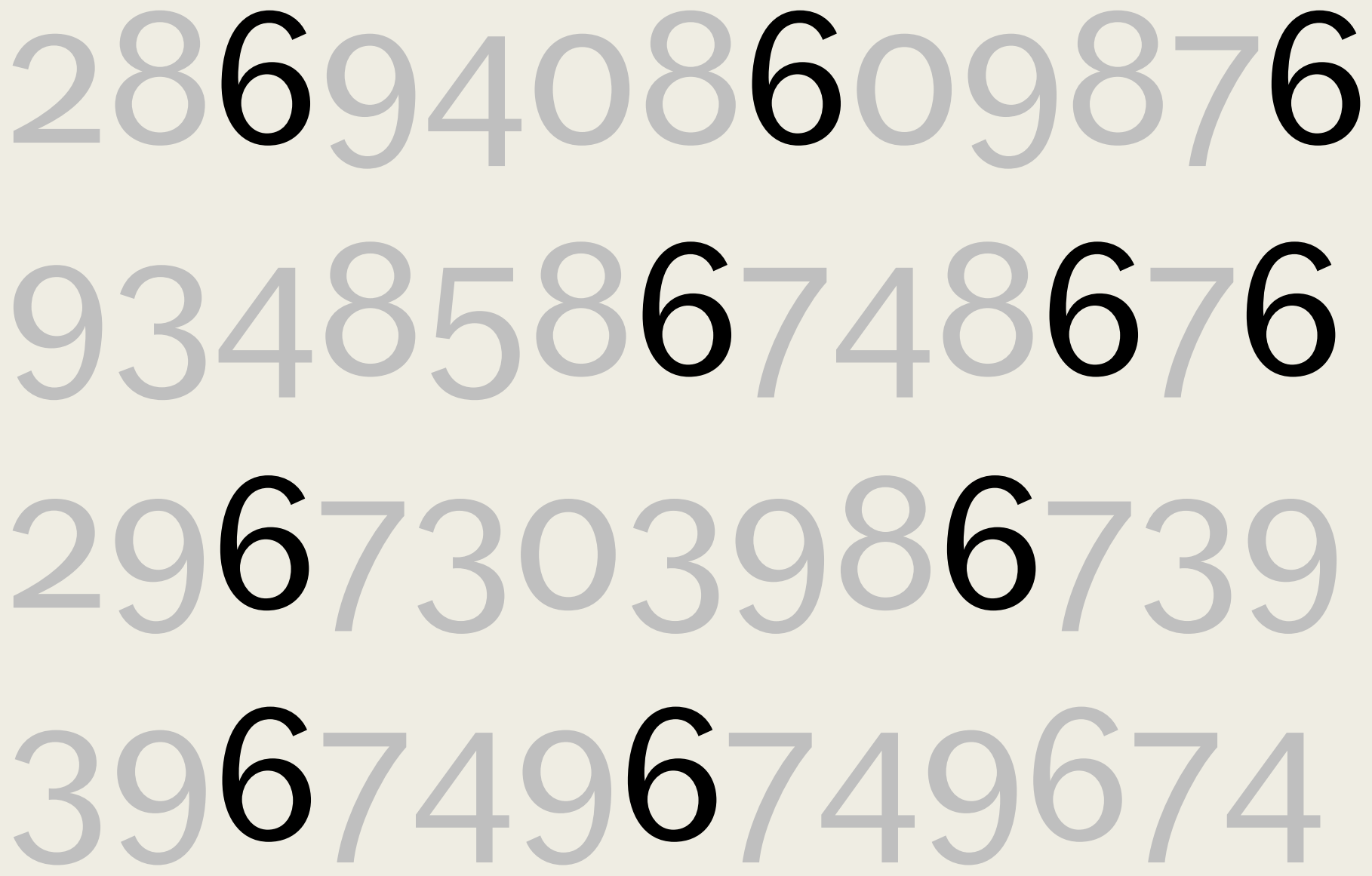
ATTRIBUTS PRÉ-ATTENTIFS

Combien de 6 y a-t-il
sur la prochaine diapositive ?

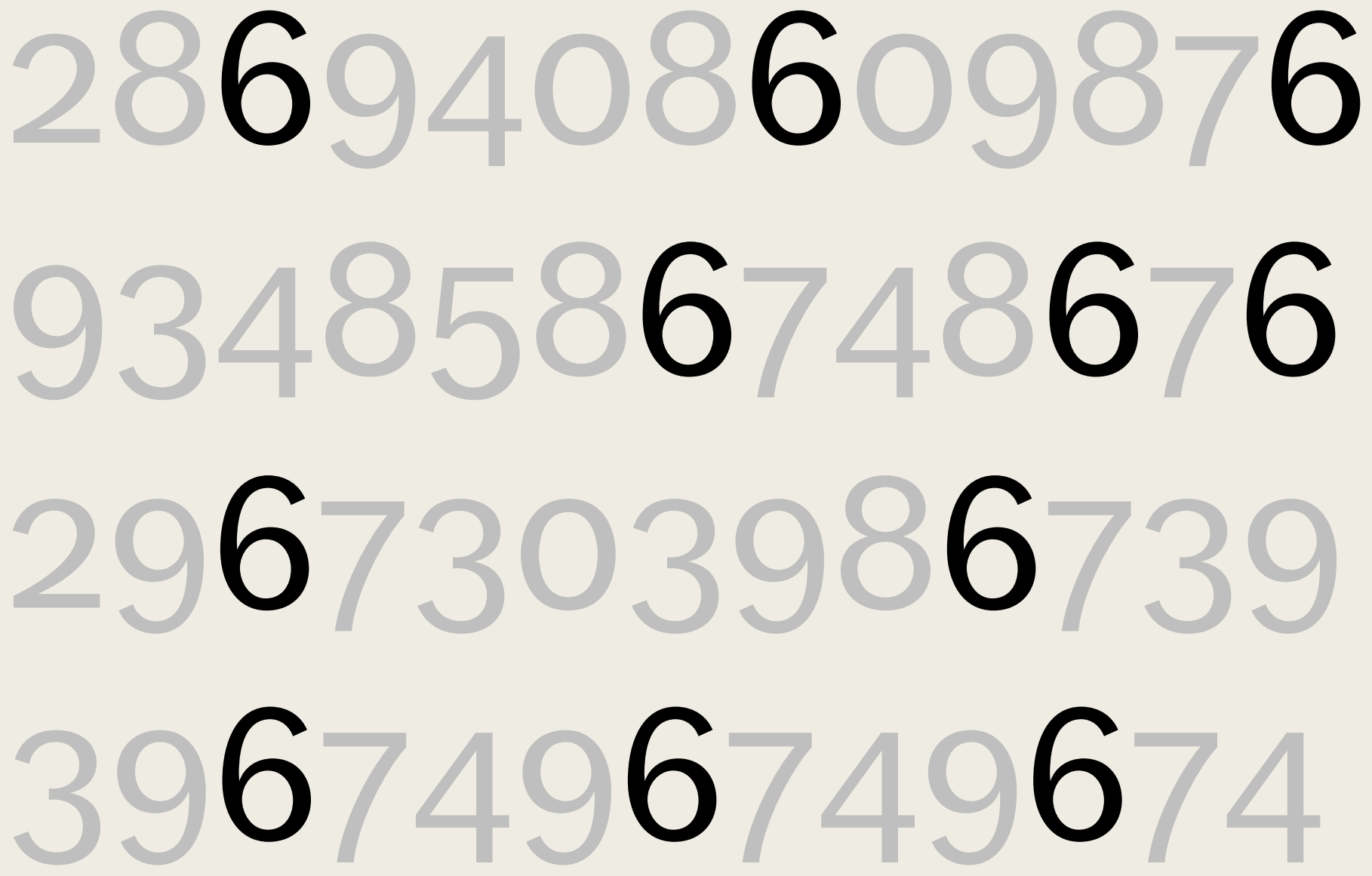
[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]



[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]

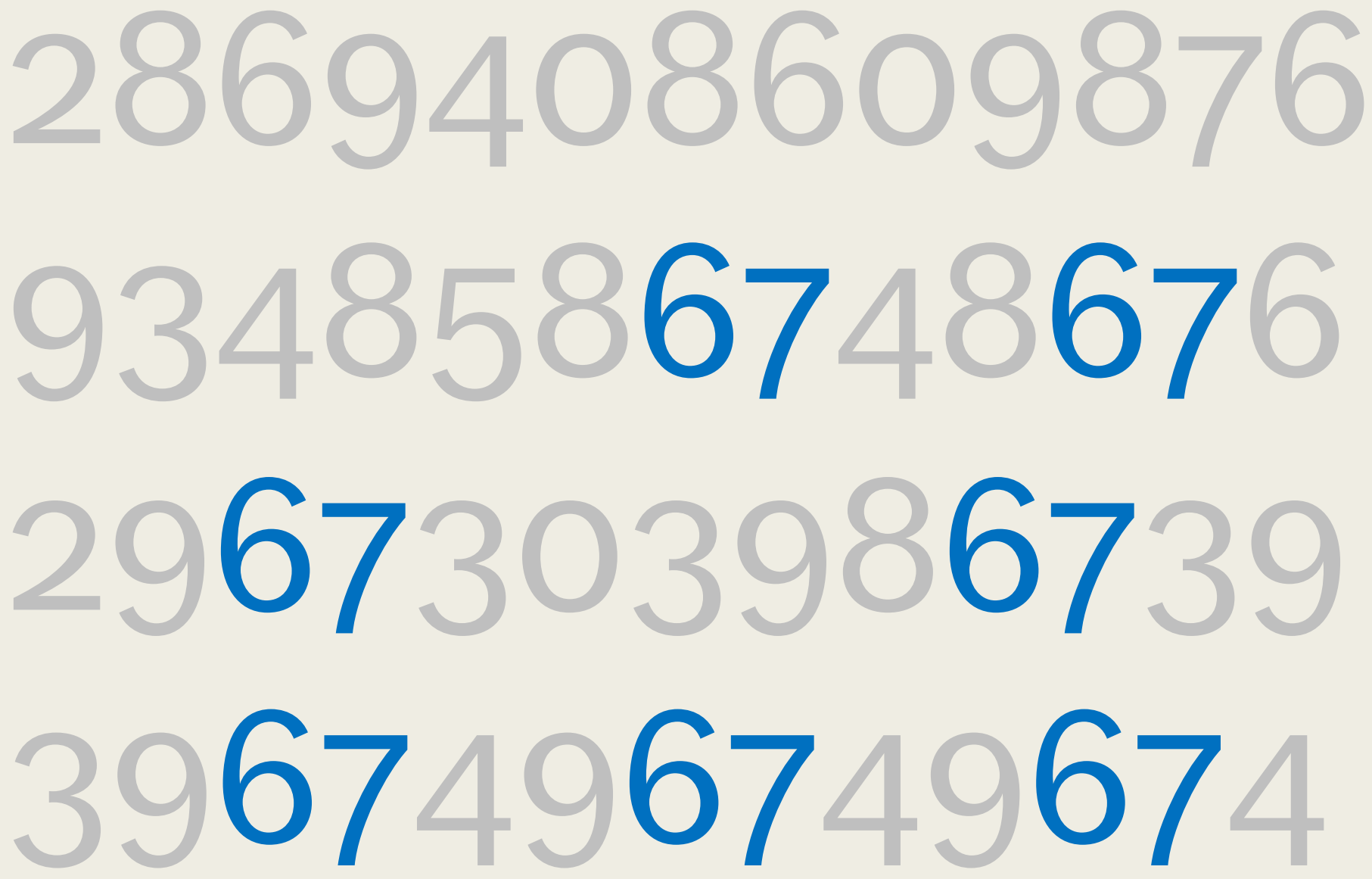


[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]



[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]

[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]



[P. Boily, S. Davies, J. Schellinck, *The Practice of Data Visualization*]

DÉSENCOMBREMENT

L'ENNEMI, C’EST LE DÉSORDRE !

- chaque élément d'une page ajoute une **charge cognitive**
- Identifiez tout ce qui n'apporte pas de valeur ajoutée et supprimez-le.
- Considérez la charge cognitive comme l'effort mental nécessaire pour traiter l'information (plus il est faible, mieux c'est).
- Tufte fait référence au **ratio données/encre** : « plus la part d'encre d'un graphique consacrée aux données est importante, mieux c'est ».
- Dans Resonate, Duarte parle de « maximiser le rapport signal/bruit », le signal étant l'information ou l'histoire que nous voulons communiquer.

DÉSENCOMBREMENT

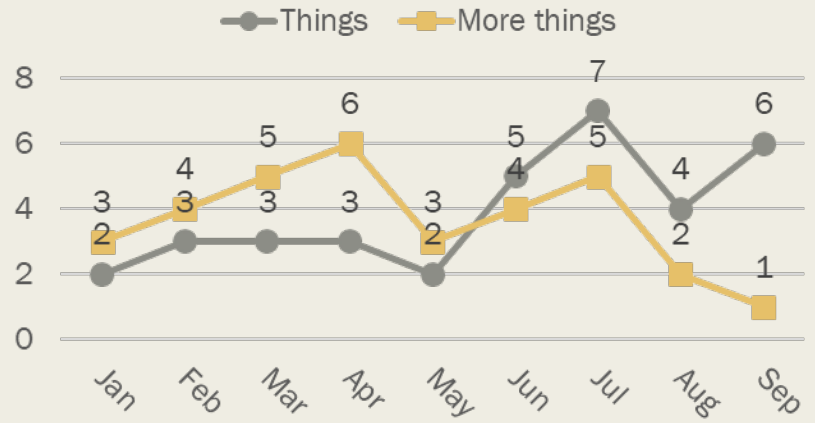
Utilisez les **principes de la Gestalt** pour organiser/ souligner les données dans un graphique.

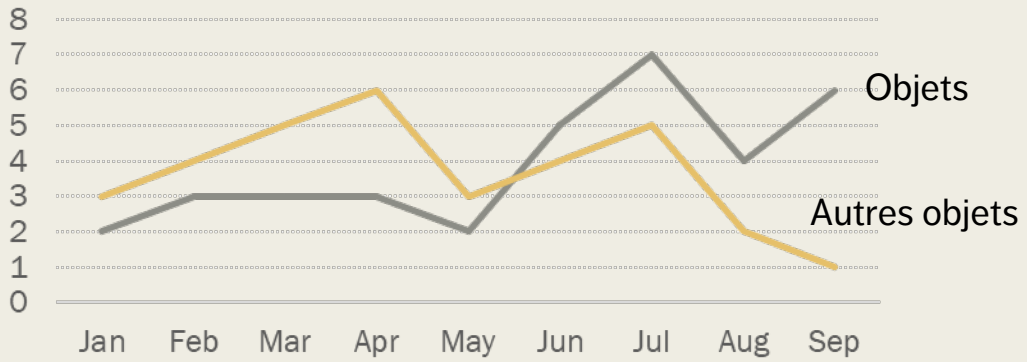
Alignez tous les éléments (graphiques, texte, lignes, titres, etc.).

- **Ne vous fiez pas** à l'œil, utilisez des cases de position et des valeurs.

Graphiques :

- supprimez les bordures, les lignes de la grille, les marqueurs de données
- clarifiez les étiquettes des axes
- étiqueter les données directement





DÉSENCOMBREMENT

Utilisez une police, une taille de police, une couleur et un alignement **uniformes**.

Ne faites pas pivoter le texte à un angle autre que 0 ou 90 degrés.

Utilisez des **espaces blancs** :

- les marges doivent rester libres de texte et de visuels
- n'étirez pas les visuels jusqu'au bord de la page ou trop près d'autres visuels
- considérez l'espace blanc comme une bordure

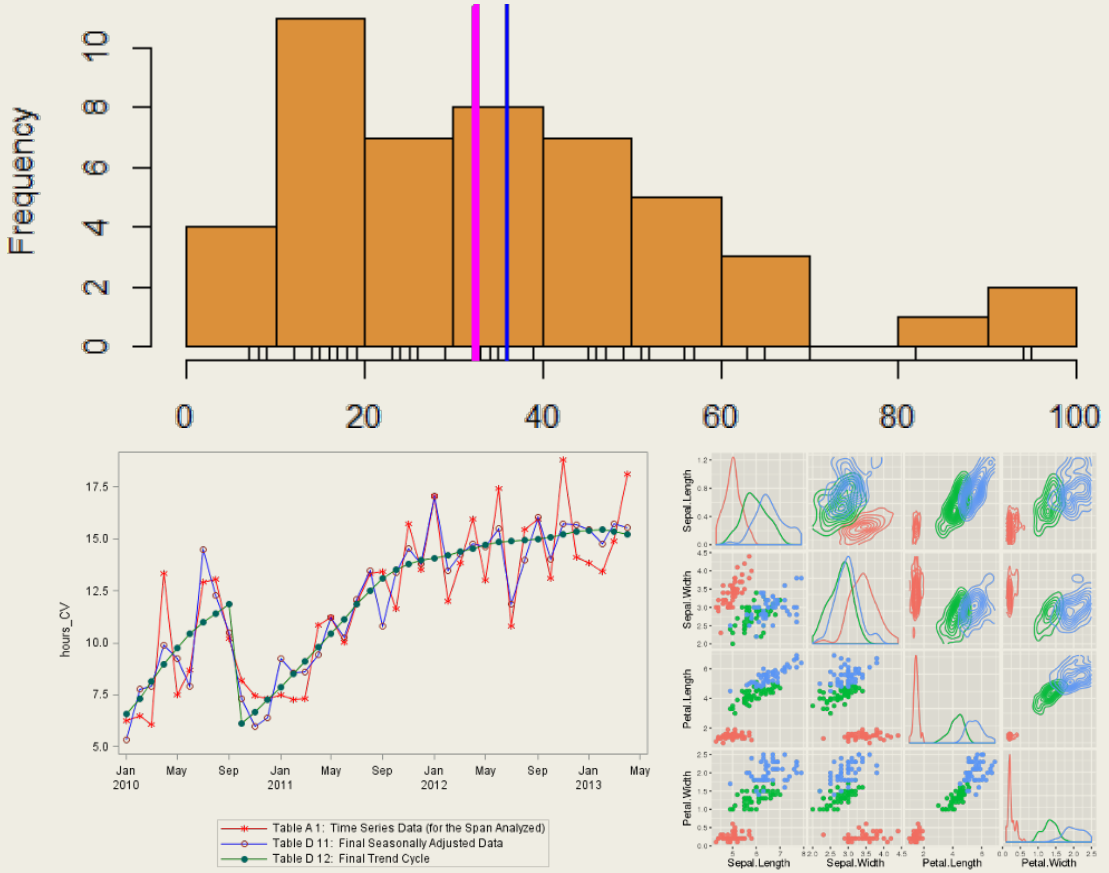
TAILLES DES GRAPHIQUES

En supposant que le tableau ait été désencombré :

- les choses d'importance égale ont la **même taille** ;
- la taille des autres éléments est proportionnelle à leur **importance**.

Comme il est rare que l'on mette plus de 3 ou 4 tableaux sur une page, les options de taille sont limitées.

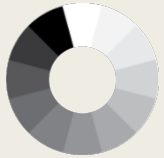
Exception perpétuelle : les **cartes géographiques** peuvent nécessiter plus d'espace.



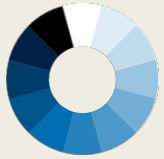
[http://www.deanenettles.com/webexamples/colorexamples]

PALETTES DE COULEURS

Achromatique



Monochromatique



Complémentaire

Complémentaire divisé

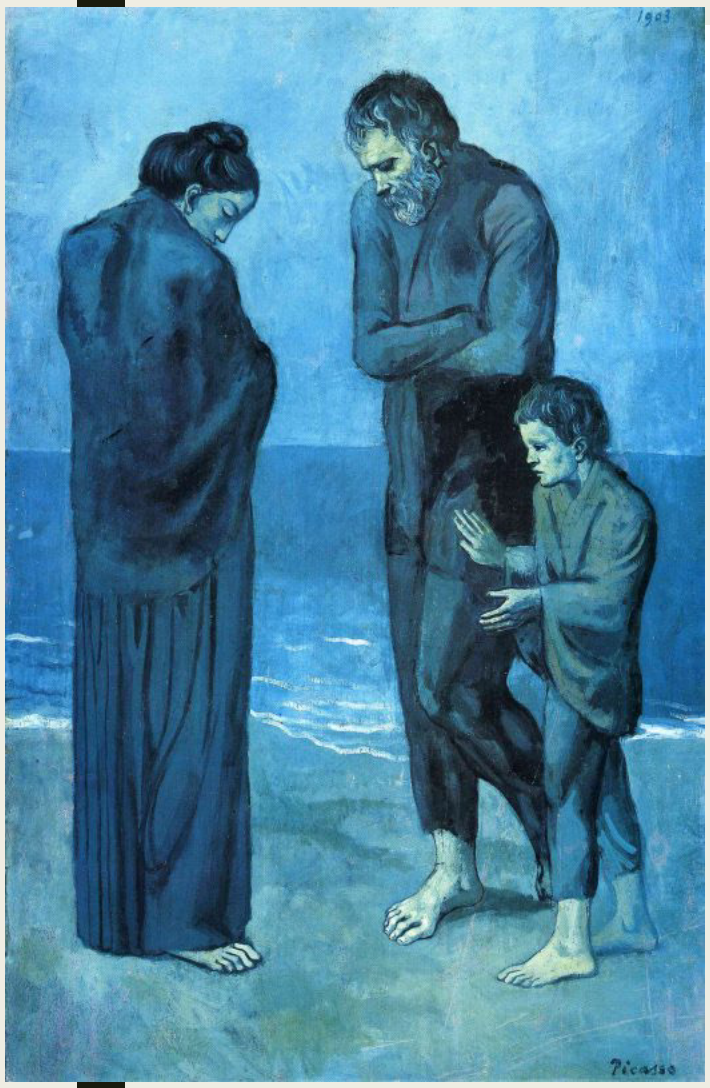
Complémentaire divisé à gauche (ou à droite)

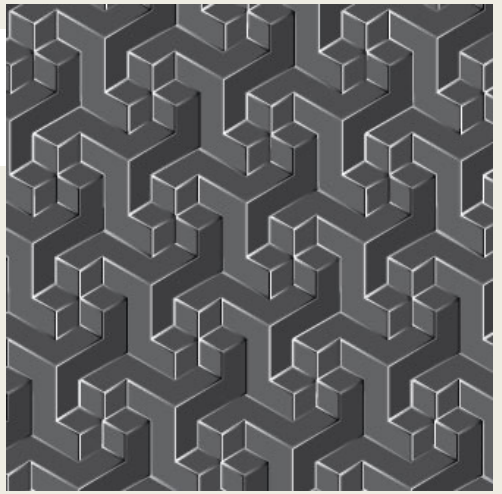
Analogique

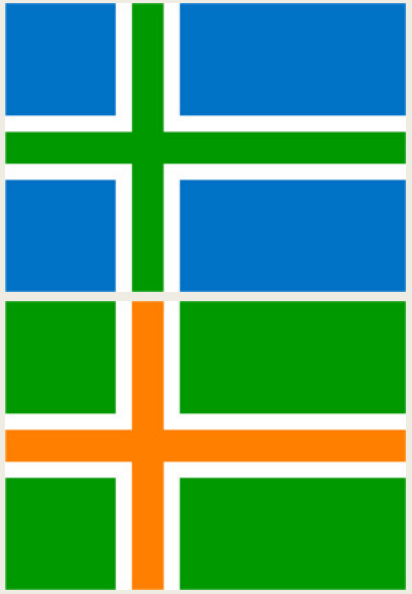
Diade de couleurs

Triade de couleurs

Tétrade de couleurs

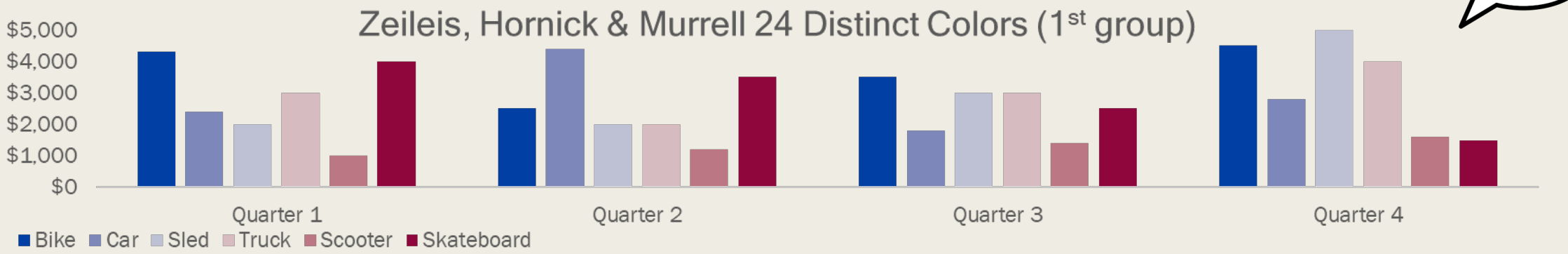


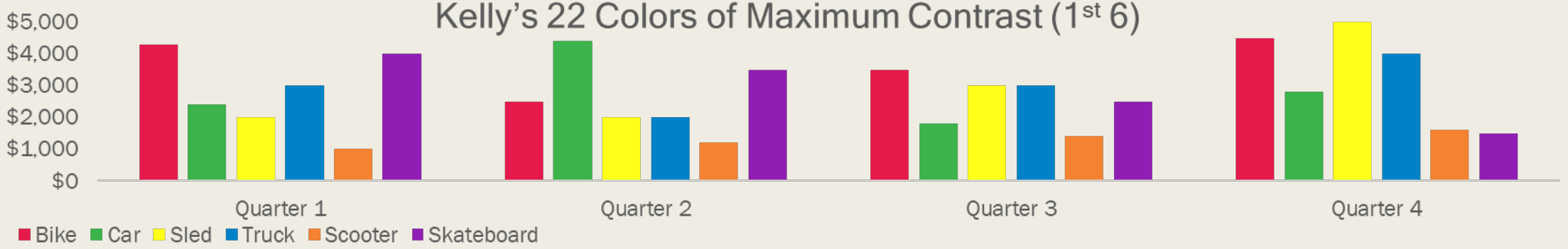


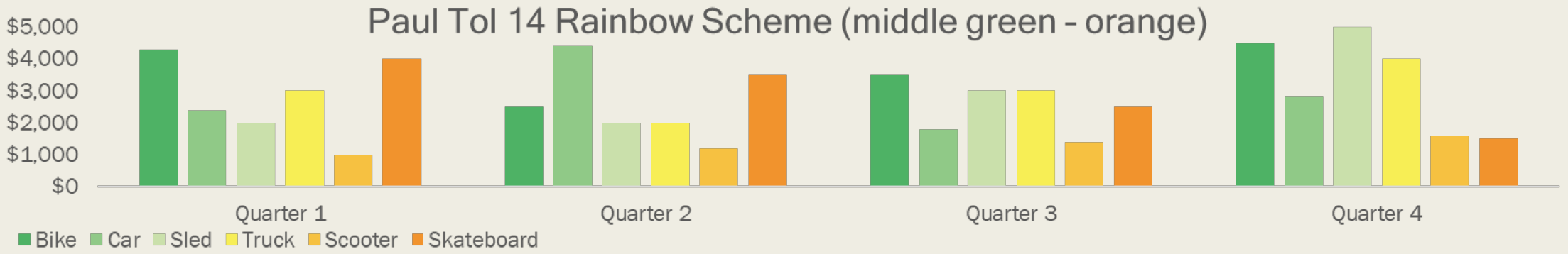


Pouvez-vous identifier les schémas de couleurs sous-jacents de ces images ?

Pouvez-vous identifier les schémas de couleurs sous-jacents de ces images ?


Zeileis, Hornick & Murrell 24 Distinct Colors (1st group)
$5,000
$4,000
$3,000
$2,000
$1,000
$0
Quarter 1
Quarter 2
Quarter 3
Quarter 4
Bike
Car
Sled
Truck
Scooter
Skateboard


Kelly's 22 Colors of Maximum Contrast (1st 6)
$5,000
$4,000
$3,000
$2,000
$1,000
$0
Quarter 1
Quarter 2
Quarter 3
Quarter 4
Bike
Car
Sled
Truck
Scooter
Skateboard


Paul Tol 14 Rainbow Scheme (middle green - orange)
$5,000
$4,000
$3,000
$2,000
$1,000
$0
Quarter 1
Quarter 2
Quarter 3
Quarter 4
Bike
Car
Sled
Truck
Scooter
Skateboard

PALETTES DE COULEURS

En ce qui concerne la couleur, **il faut en faire moins** : utilisez-la avec parcimonie (on apprend aux graphistes à "bien faire les choses, en noir et blanc").

Sur la base des principes de la Gestalt, les schémas **monochromes** peuvent être particulièrement efficaces.

Le cas échéant, choisissez un schéma basé sur l'identité de l'entreprise (cela maximise l'adhésion).

Créez un modèle (et respectez-le).

Téléchargez des images pour voir à quoi ressemblent les graphiques en fonction des différents degrés de daltonisme :

- https://www.color-blindness.com/coblis-color-blindness-simulator *(il existe d'autres outils)*

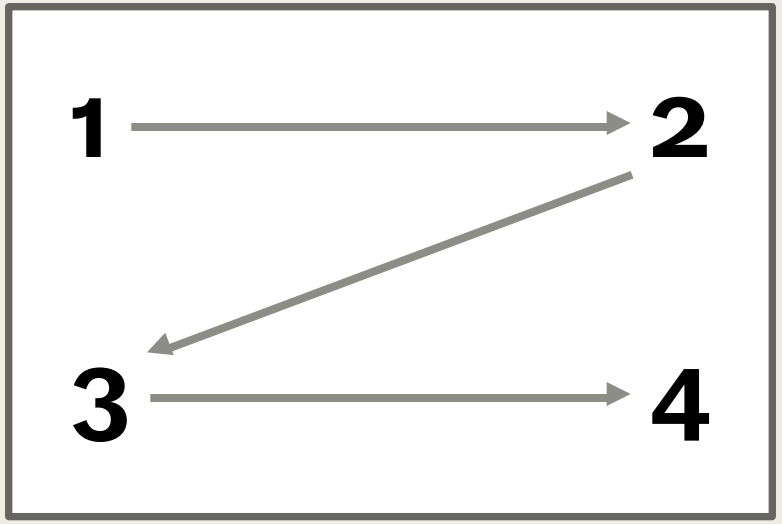
POSITION

Comment placer les éléments dans un graphique ou un tableau de bord ?

En Occident, la plupart des gens commencent **en haut à gauche** et zigzaguent jusqu'en bas à droite.

Une règle simple : ne pas faire travailler les gens trop fort.

- message principal : en haut à gauche/en haut à droite
- informations par ordre de préférence
- les gens se concentrent moins lorsqu'ils scannent, il faut donc que les informations soient moins complexes lorsque vous vous déplacez vers le coin inférieur.



TABLEAUX DE BORD (*DASHBOARDS*)

PRINCIPES DE LA VISUALISATION DES DONNÉES

[Wexler, Shaffer, Cotgreave, *The Big Book of Dashboards*]

TABLEAUX DE BORD

Un **tableau de bord** (*dashboard*) est un affichage visuel des données utilisé pour surveiller les conditions et/ou faciliter la compréhension.

Exemples:

- écran interactif qui permet aux gens d'explorer les demandes d'assurance automobile par ville, province, âge du conducteur, etc.
- PDF montrant les principaux paramètres d'audit qui est envoyé par courriel à la DG d'un département sur une base hebdomadaire.
- écran mural qui affiche les statistiques du centre d'appels en temps réel.
- une application mobile qui permet aux administrateurs d'hôpitaux d'examiner les temps d'attente sur une base horaire et quotidienne pour l'année en cours et l'année précédente.

POINTS DE RÉFLEXION

Dans le tableau de bord d'une voiture, un petit nombre **d'indicateurs clés** (vitesse, niveau d'essence, feux, etc.) doivent être compris en **un coup d'œil**. Un design qui ne tient pas compte de ces deux caractéristiques peut avoir des conséquences catastrophiques.

Il faut répondre aux questions suivantes avant de concevoir le tableau de bord :

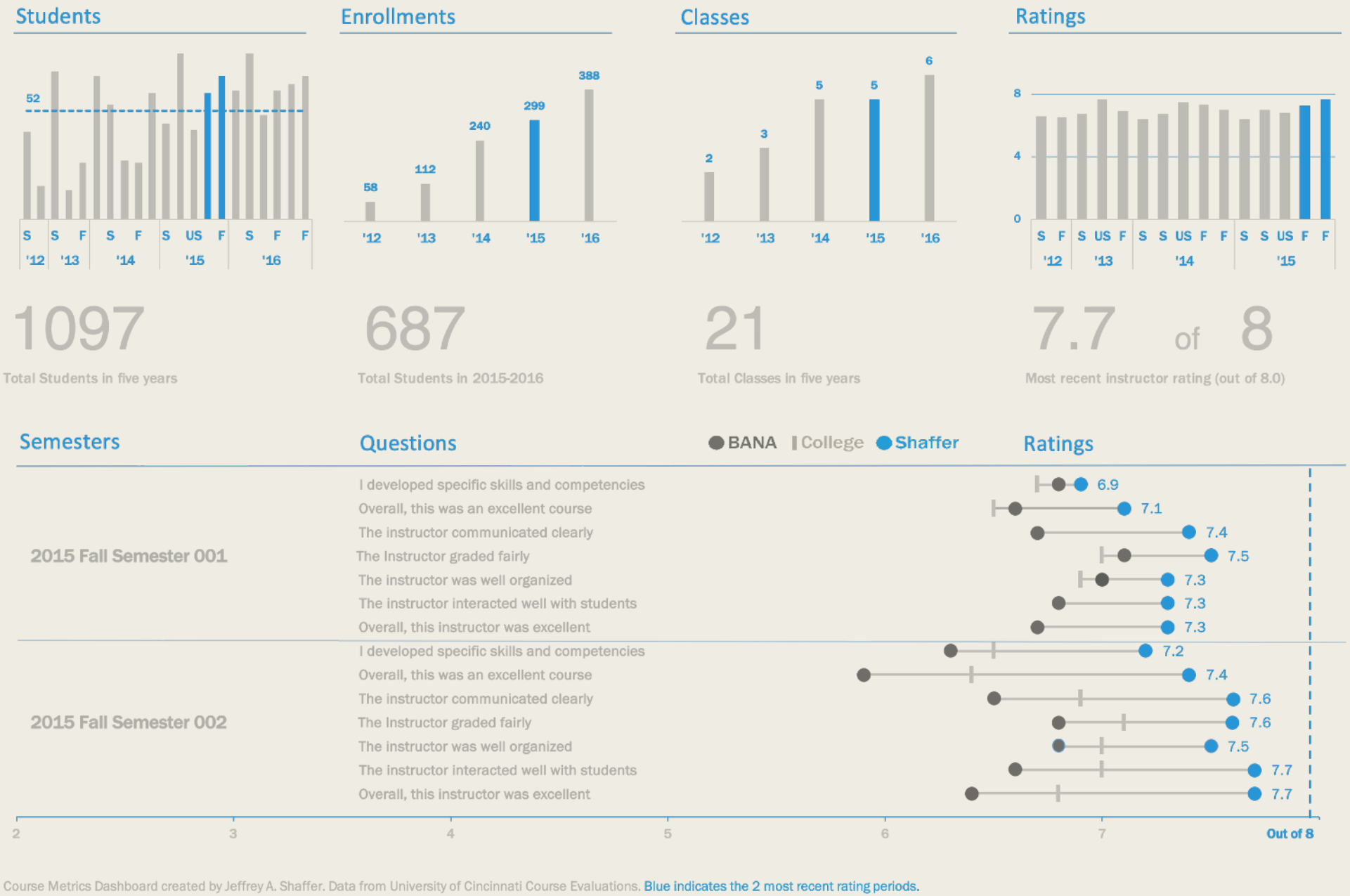
- Qui est le **client** du tableau de bord ?
- Quelle **histoire** le tableau de bord raconte-t-il ?
- Quelles données (catégories) seront utilisées ?
- Qu'est-ce qui **apparaîtra** sur le tableau de bord ?
- Comment le tableau de bord peut-il aider le client ?

Course Metrics

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

Points forts :

- Des indicateurs clés faciles à lire.
- Schéma de couleurs simple
- Possibilité d'être statique ou interactif
- La vue d'ensemble et les détails sont clairs



ÉVALUATION DU TABLEAU DE BORD

Il n'existe pas de tableaux de bord parfaits - aucune collection de graphiques ne conviendra jamais à tous ceux qui la rencontrent.

Tous les tableaux de bord doivent être **véridiques** et **fonctionnels**, mais les tableaux de bord qui sont également **élégants** (agréables, plaisants) vous mèneront plus loin.

Tous les tableaux de bord sont **incomplets**. Les bons tableaux de bord mènent toujours à des impasses, mais ils doivent permettre aux utilisateurs de se demander : « Pourquoi ? Quelle est la cause fondamentale de ce problème ? ».

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

EXERCISE

Considérez les tableaux de bord suivants.

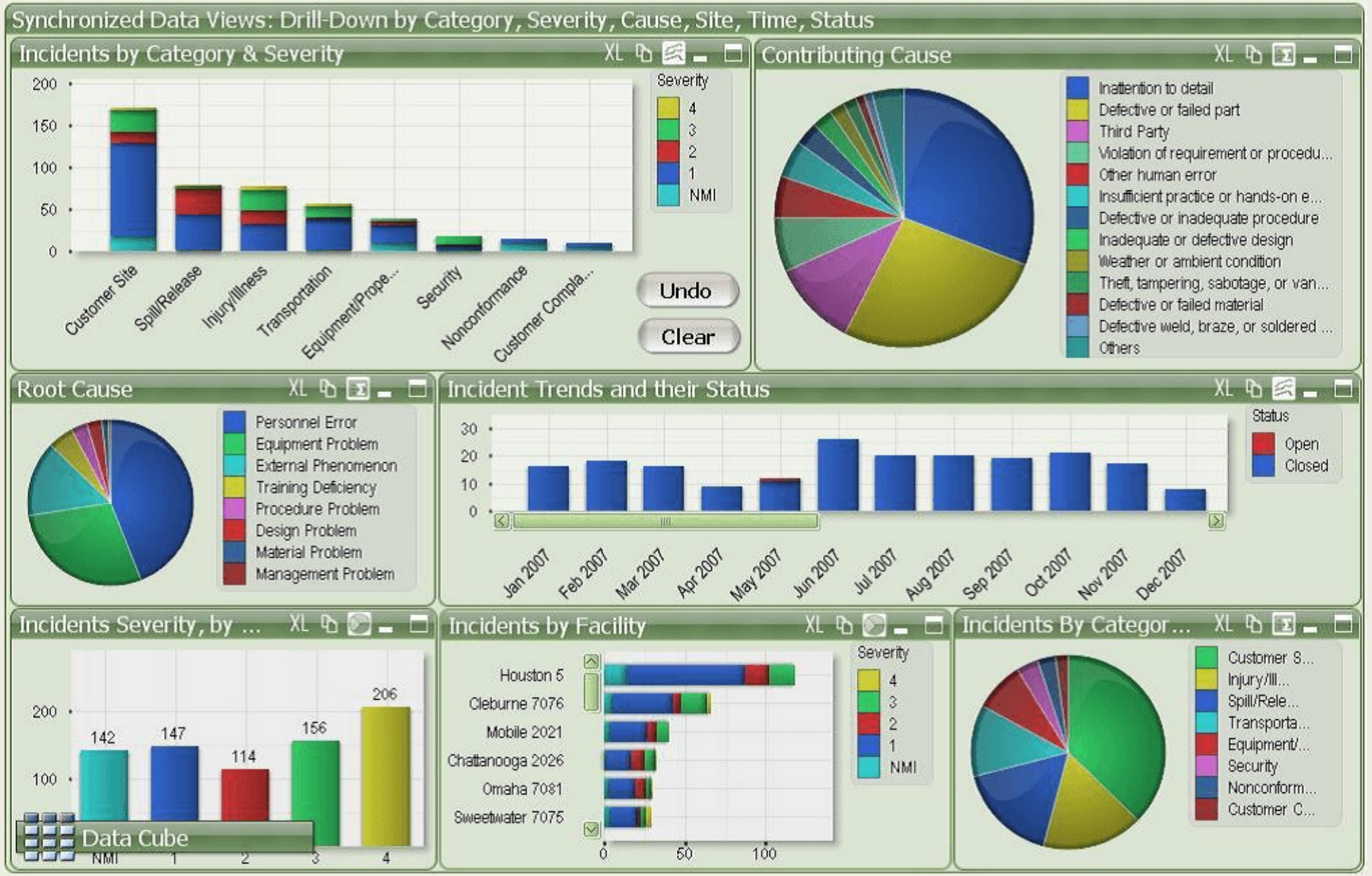
Pouvez-vous déterminer, en un coup d'œil, qui est leur public ?

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]