EXPLORATION ET VISUALISATION DES DONNÉES

PRÉPARATION DU TERRAIN

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

QUESTIONS DE BASE

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

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

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

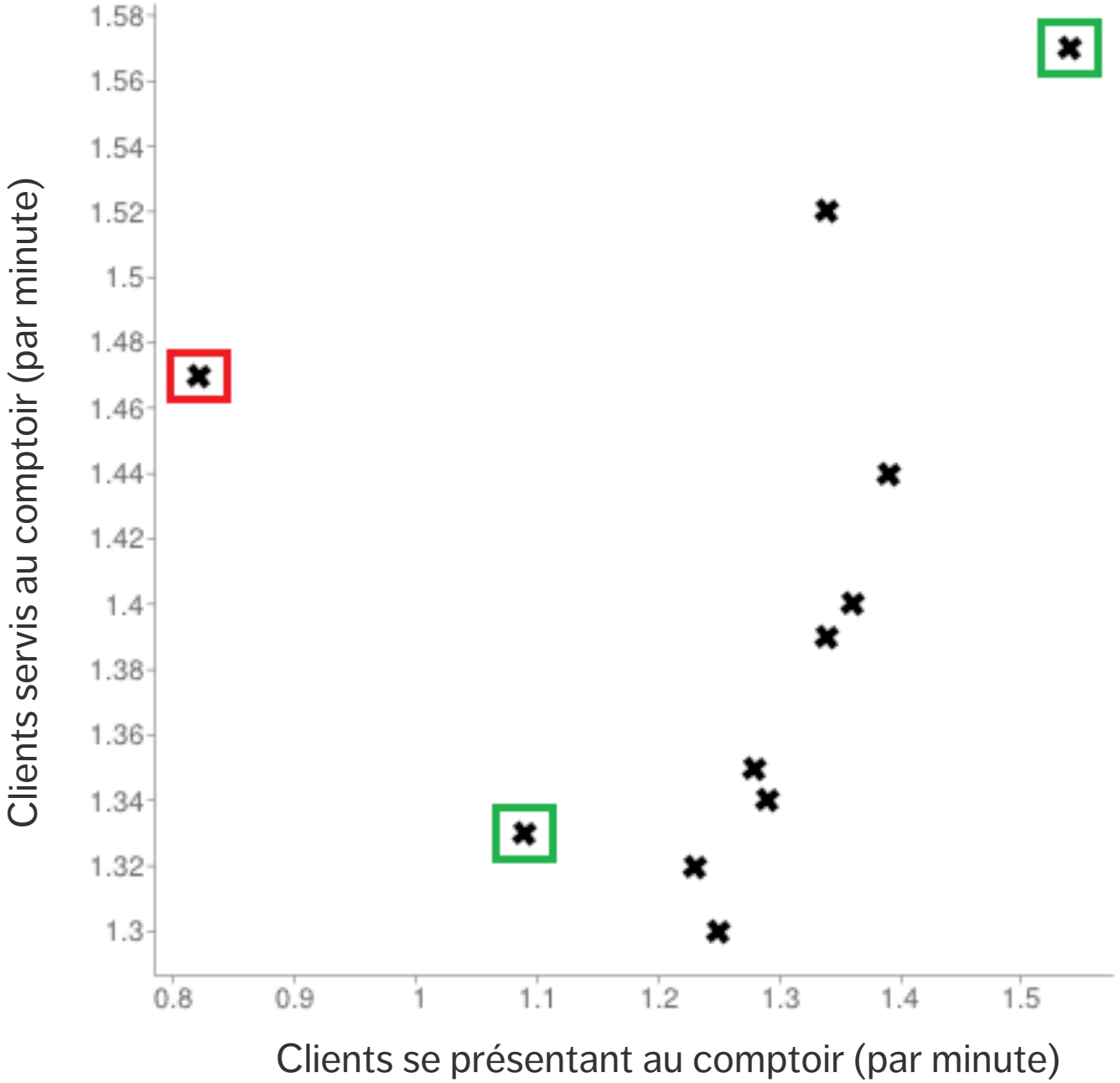
```
       Cl              NO3              NH4         
 Min.   :  0.222   Min.   : 0.000   Min.   :    5.00  
 1st Qu.: 10.994   1st Qu.: 1.147   1st Qu.:   37.86  
 Median : 32.470   Median : 2.356   Median :  107.36  
 Mean   : 42.517   Mean   : 3.121   Mean   :  471.73  
 3rd Qu.: 57.750   3rd Qu.: 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
```

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]

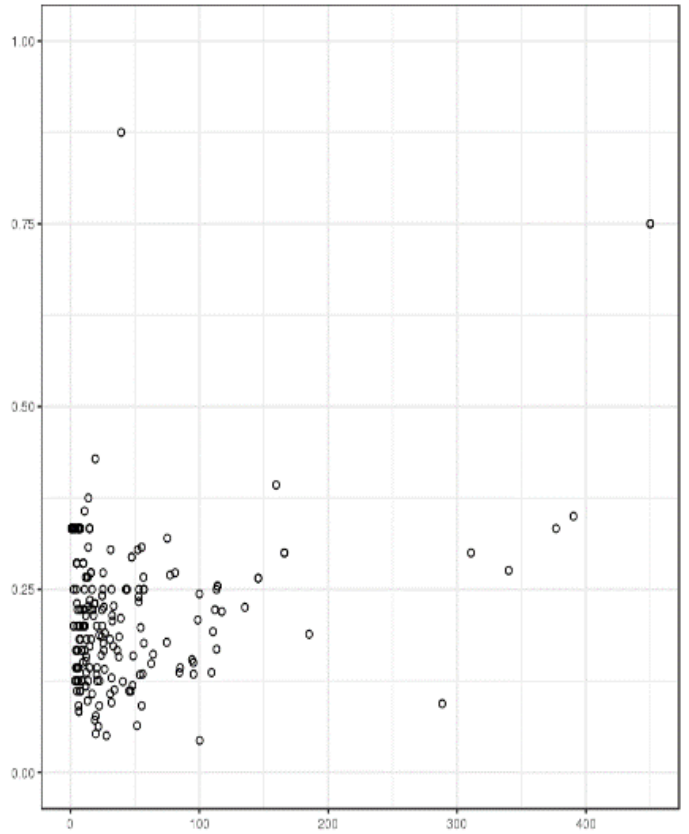


[http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff]

REPRÉSENTATION D'OBSERVATIONS À VARIABLES MULTIPLES

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

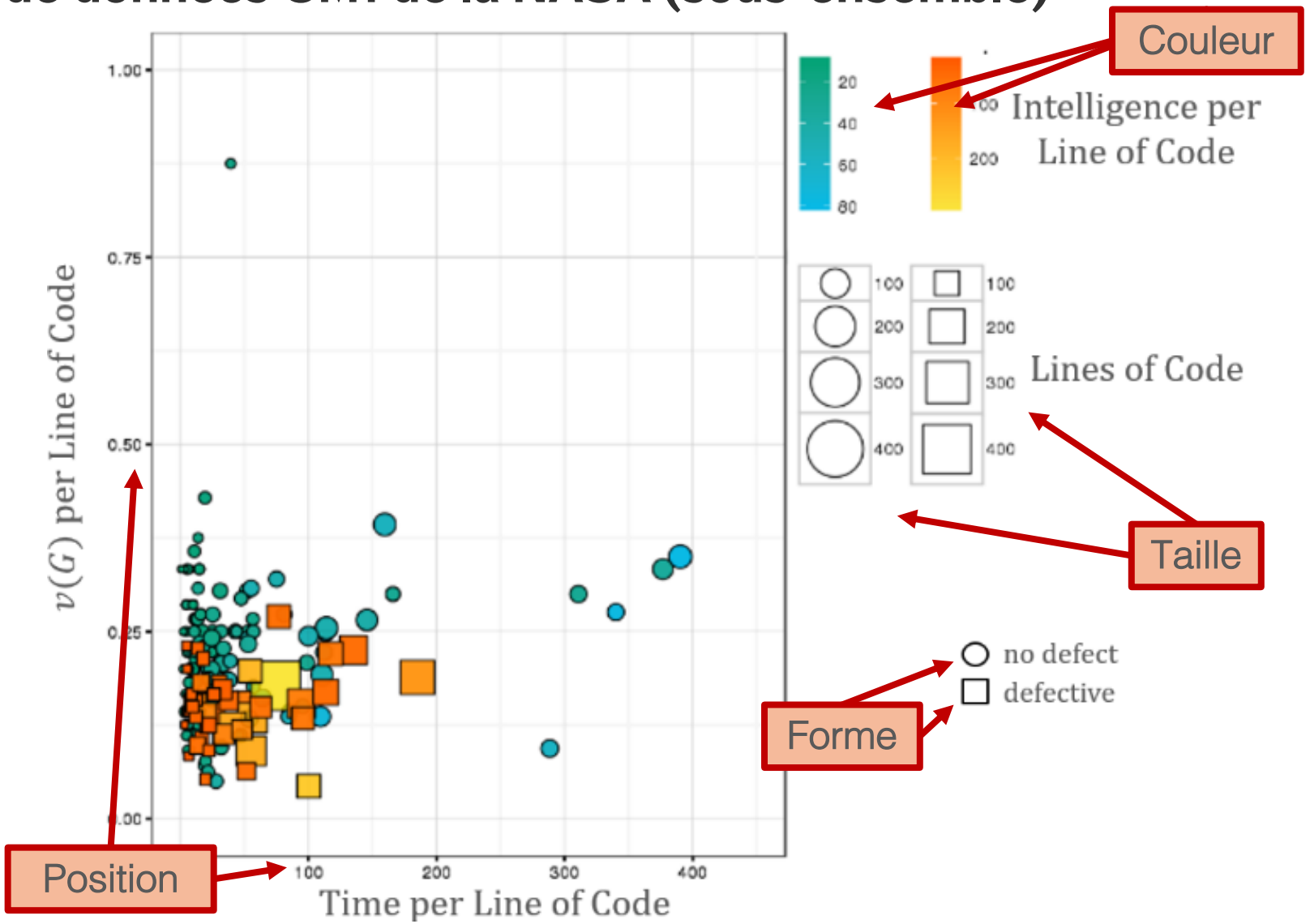
- taille
- couleur
- valeur
- texture
- orientation d'une droite
- forme
- (mouvement?)



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

[E. Tufte, *Beautiful Evidence*]

PRINCIPES FONDAMENTAUX DU DESIGN ANALYTIQUE

Le raisonnement et la communication sont interreliés 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

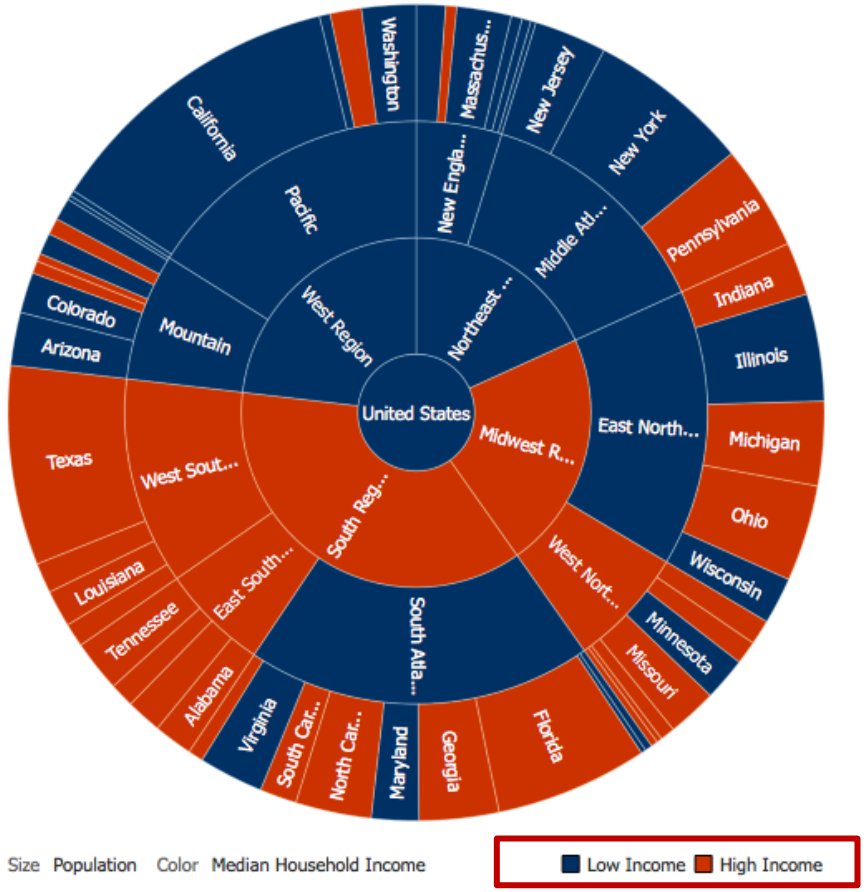
Une **méthode** et un objet à la fois (**objectivité**)

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)

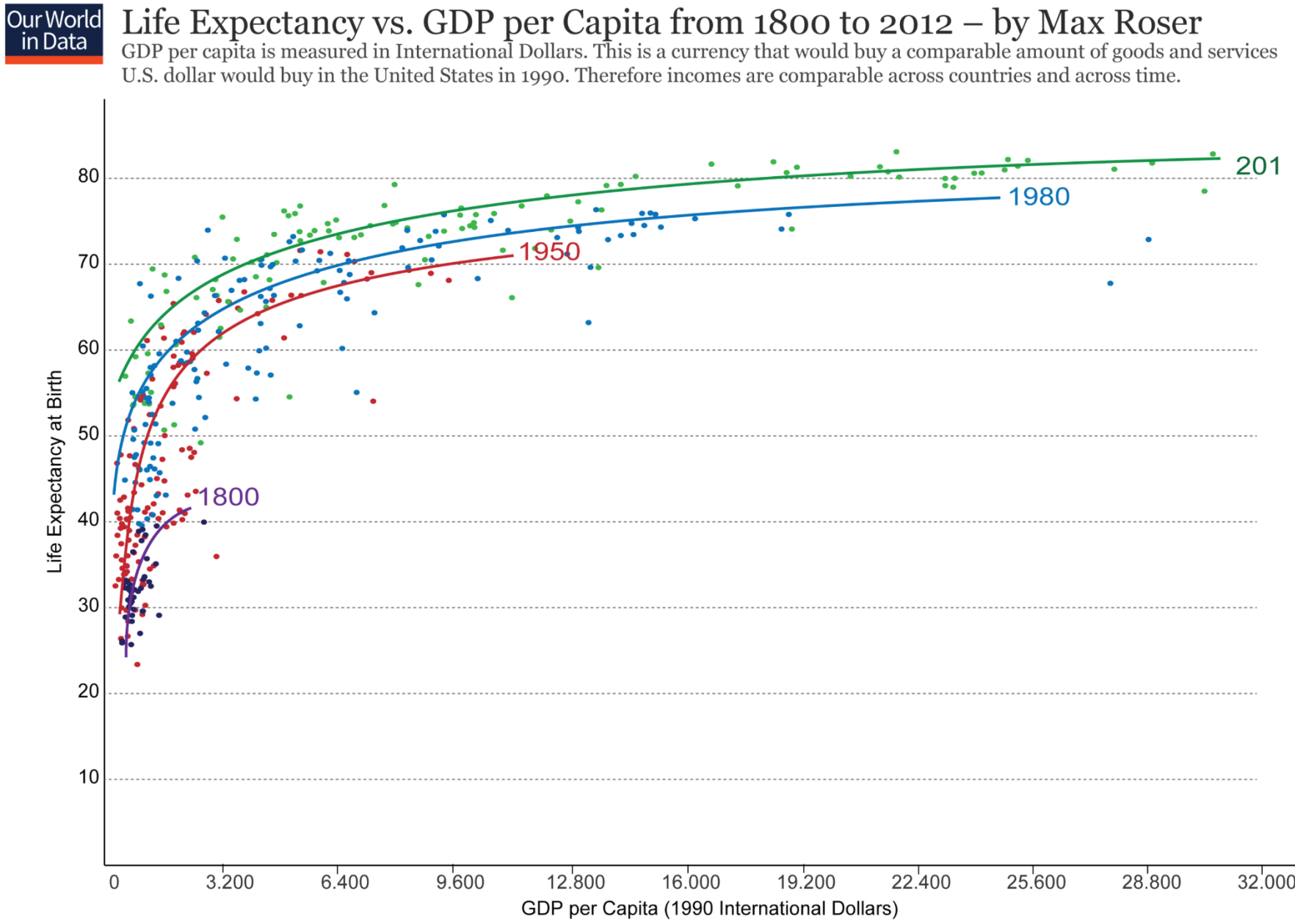
Peut être générée automatiquement

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.

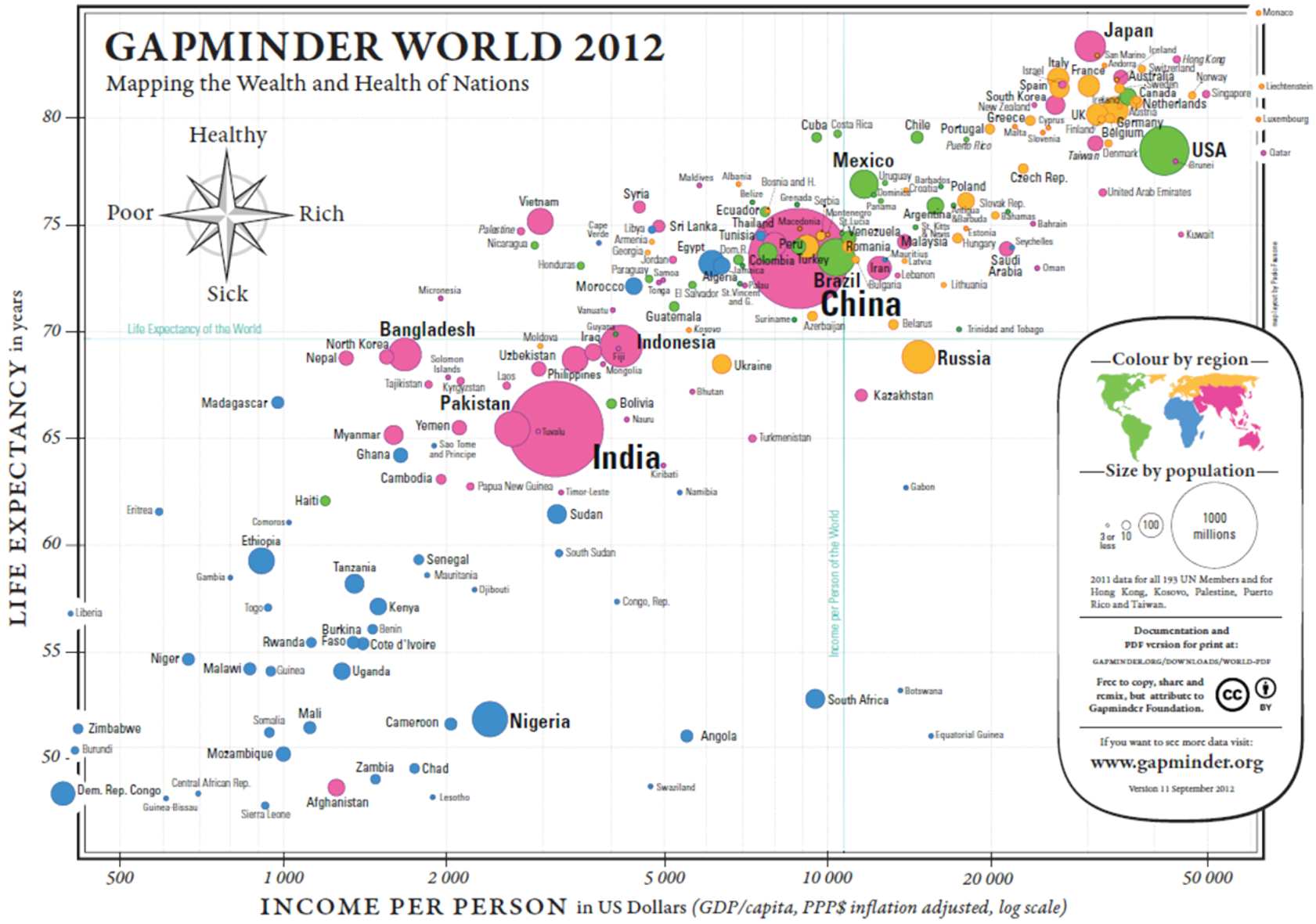
En général, plus le PIB d'un pays élevé, meilleure est son espérance de vie.

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.

data-action-lab.com

[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.

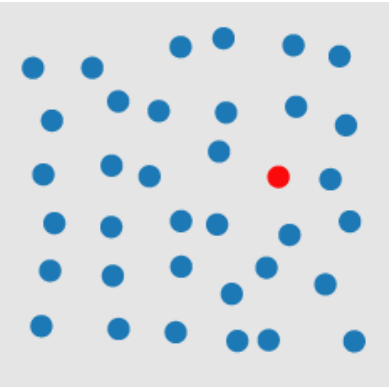
- **Traitement préattentif :** rapide , instinctif, efficace, superficiel, collecte d'information et détection de motifs.

  caractéristiques → motifs → objets

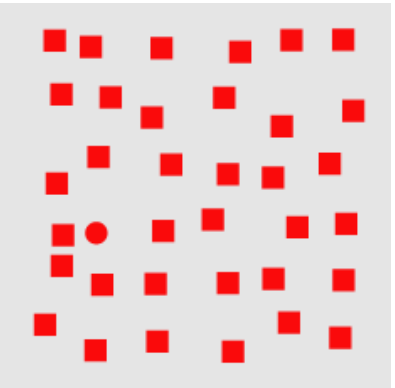
- **Traitement attentif :** lent, délibéré, focalisé, découverte de caractéristiques à l'intérieur des motifs.

  objets → motifs → caractéristiques

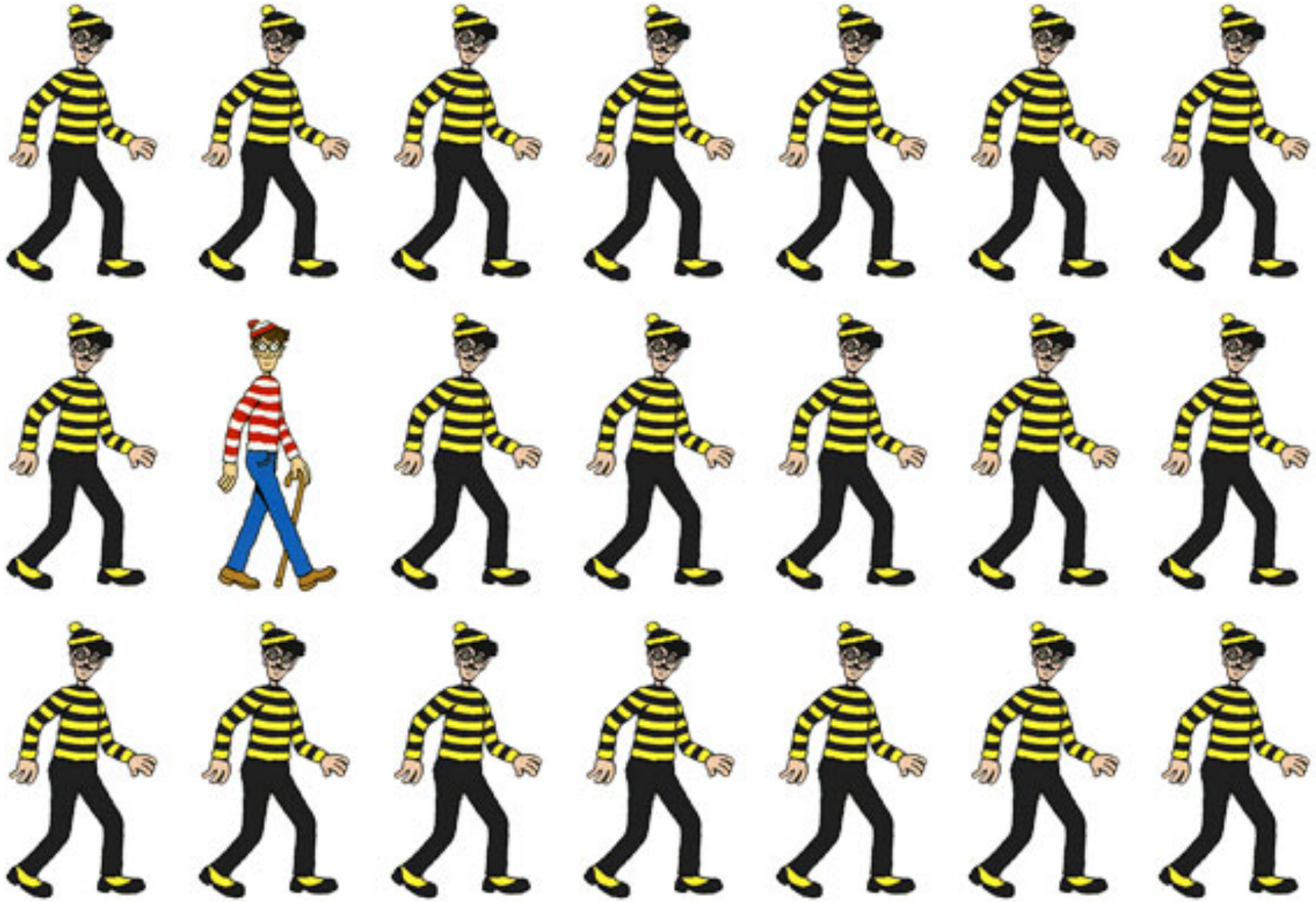
Vision préattentive



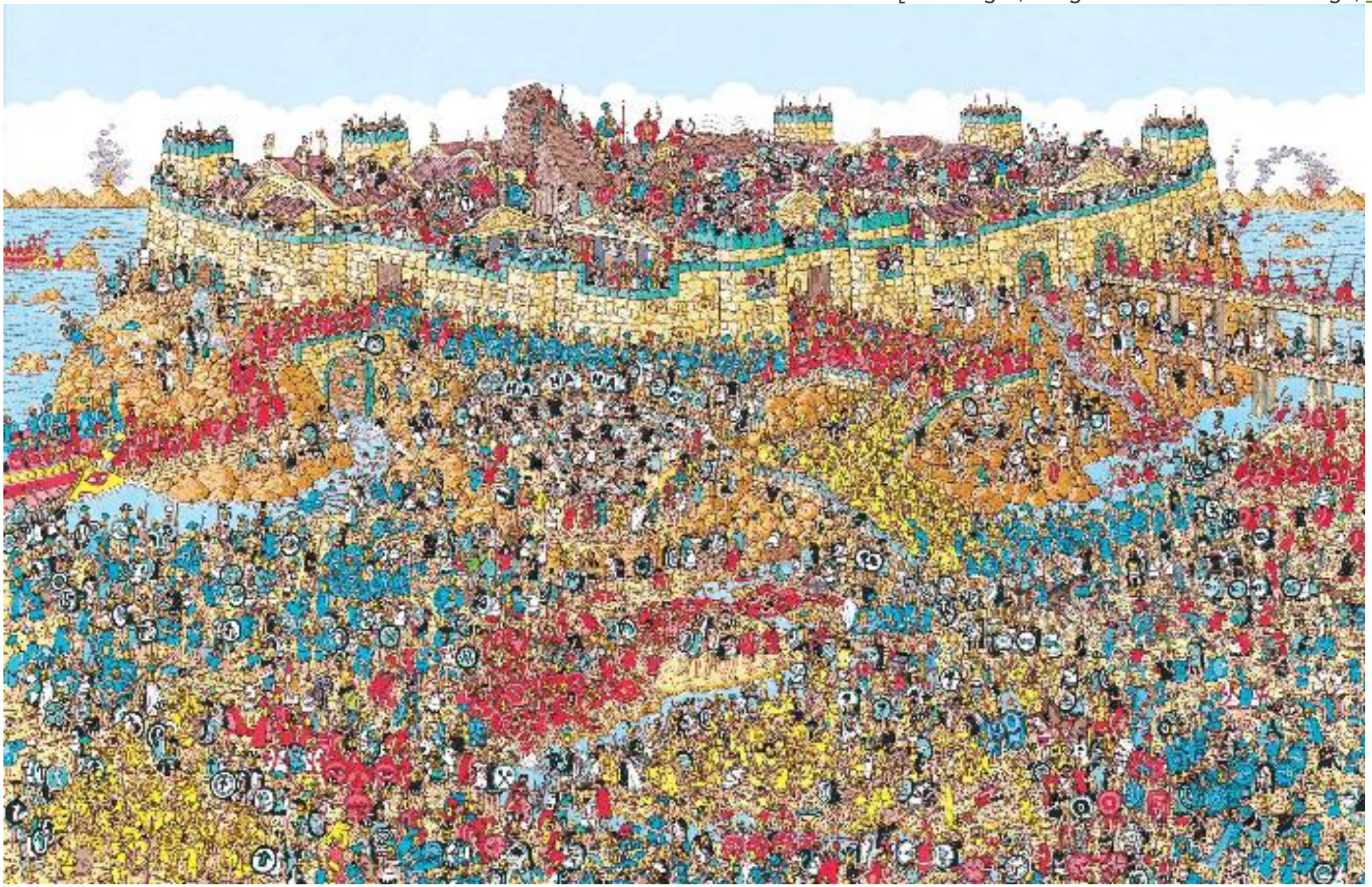
Vision attentive



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



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



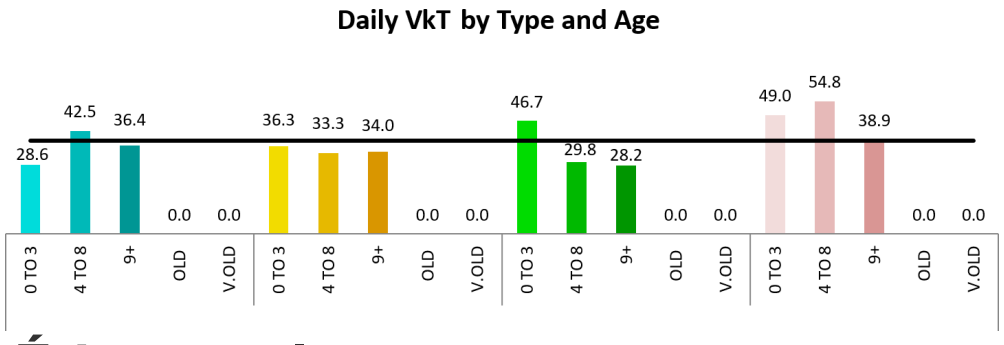
RÈGLES DE BASE

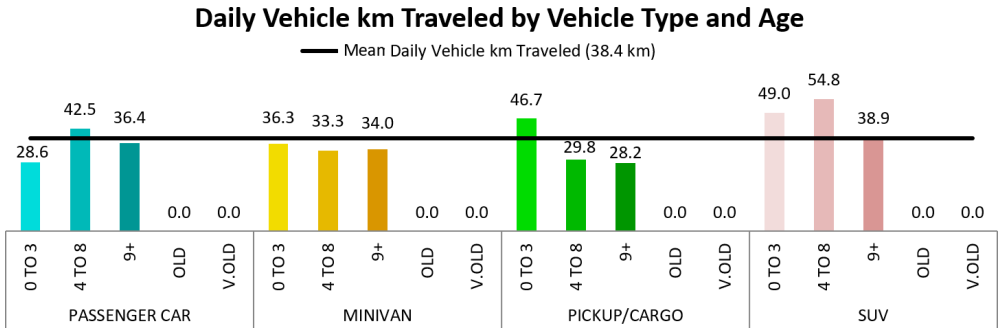
1. Examiner les données

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

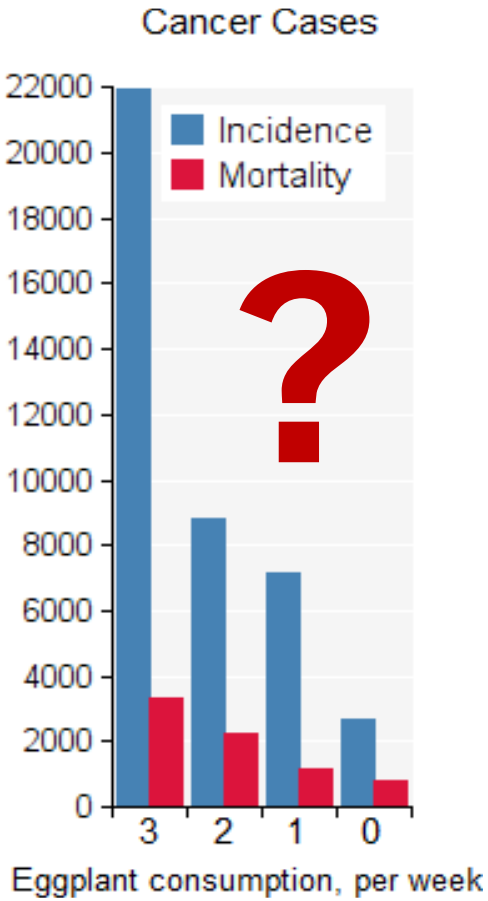
Ne pas forcer le lecteur à faire des suppositions.

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.

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 MULTIVARIÉES

Cartes de densité/Choroplètes

Coordonnées parallèles

Cartes géographiques/Distortions

Graphiques à bulles

Nuages de mots et visualisations de texte

Diagrammes de réseau

“Sparklines”

Miniatures (“Small Multiples”)

Visualisations interactives

Visualisations animées, etc.

Il y a plusieurs exemples dans la presentation longue.

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).

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

GRAPHIQUES TROMPEURS

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

GRAPHIQUES TROMPEURS