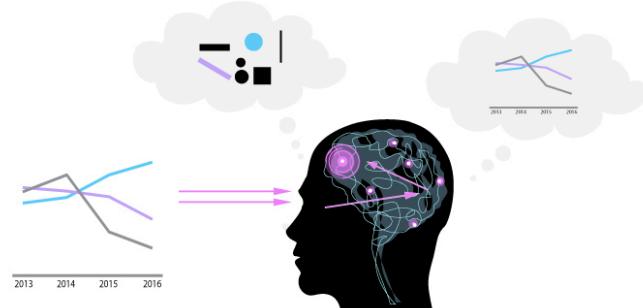
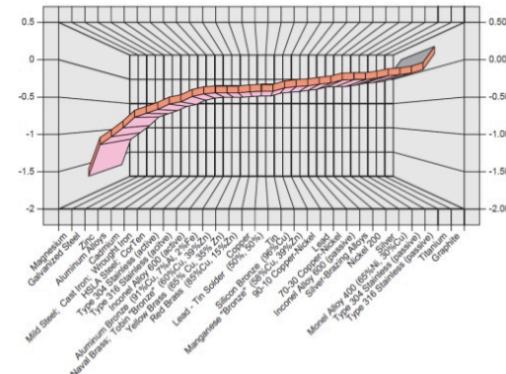
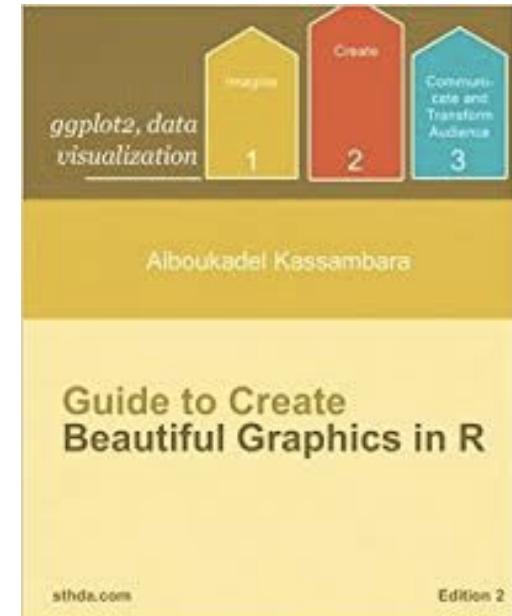
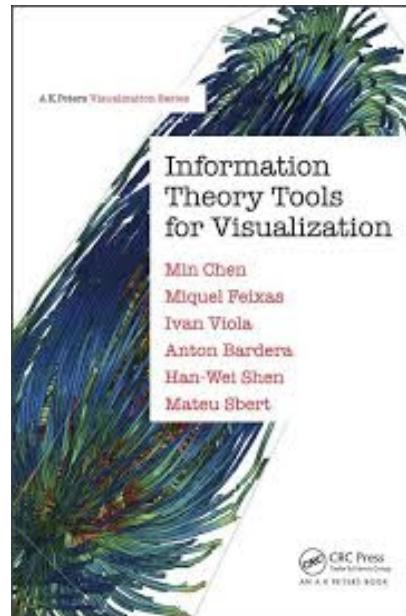
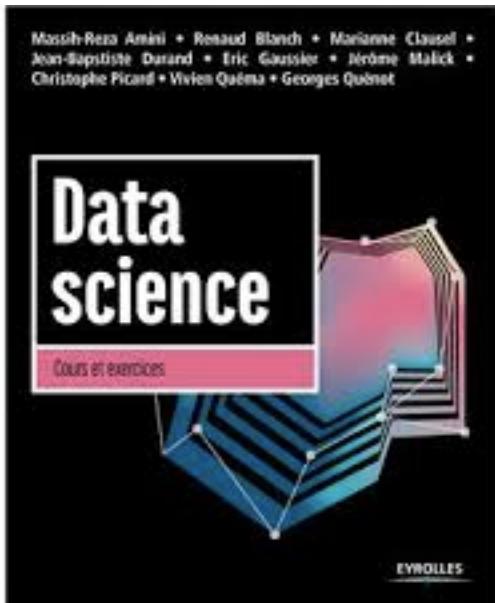


# Visualisation d'information



david.rousseau@univ-angers.fr

# Références

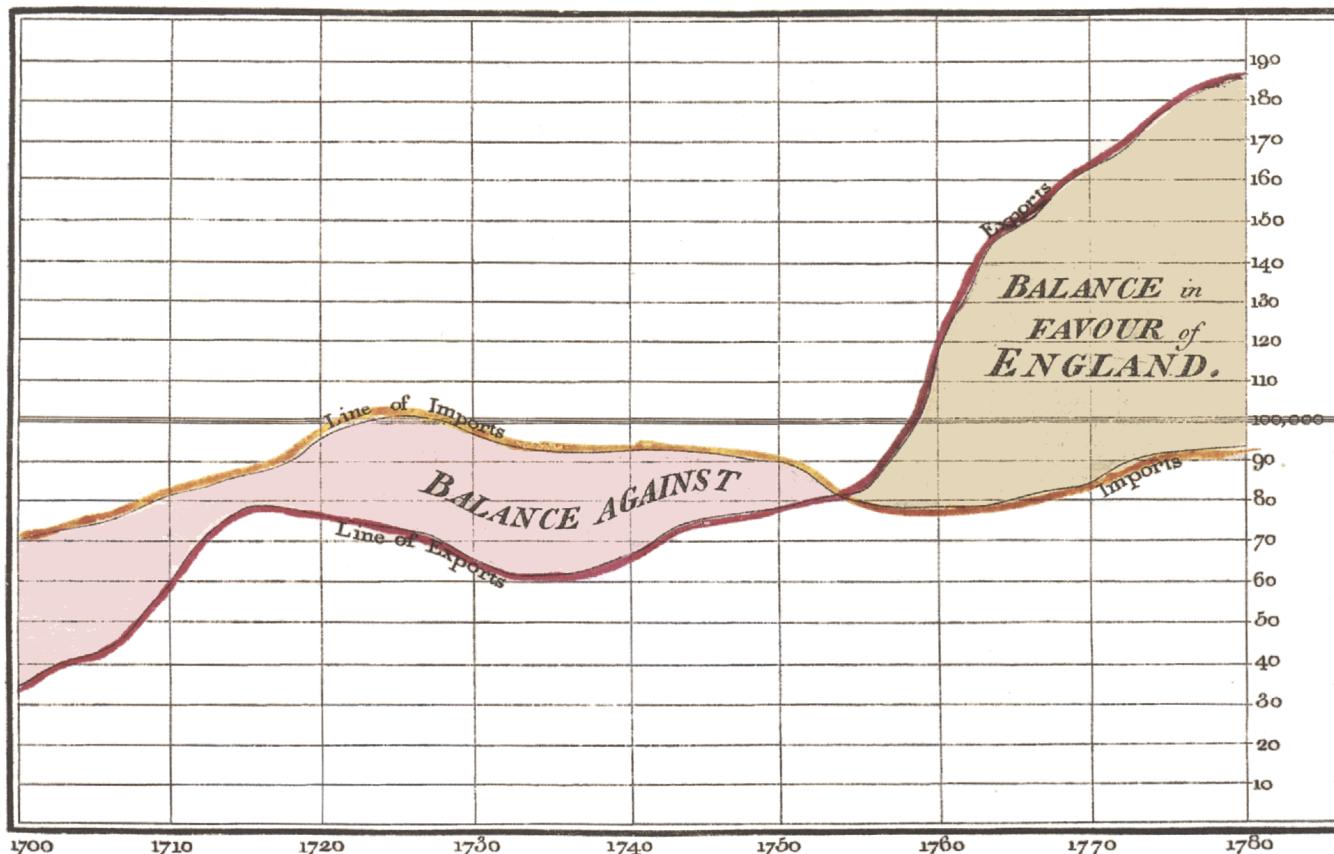


Jupiter Notebook : <https://github.com/romsson/jdev17-python-dataviz-talk>

Vidéo associées : <http://devlog.cnrs.fr/jdev2017/t7>

# Un sujet qui n'est pas nouveau

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



The Bottom line is divided into Years, the Right hand line into £10,000 each.

Published as the Act directs, 1<sup>st</sup> May 1786, by W<sup>m</sup> Playfair

Neale sculpt 352, Strand, London.

# Motivation

## Pour quoi faire ?

Proposer un cadre pour trouver des représentations graphiques à des données abstraites

Choisir des représentations utilisant au mieux les capacités du système percepteur humain

## Quand ?

Explorer les données en amont de toute analyse

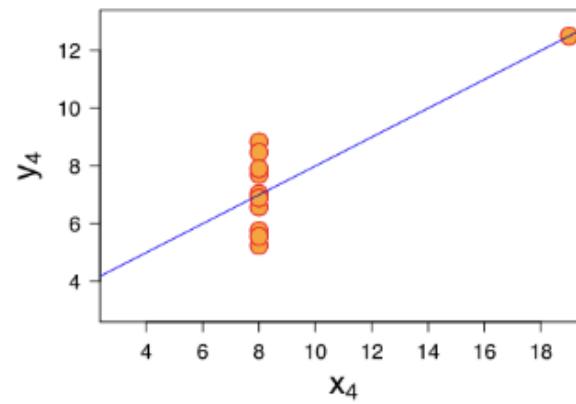
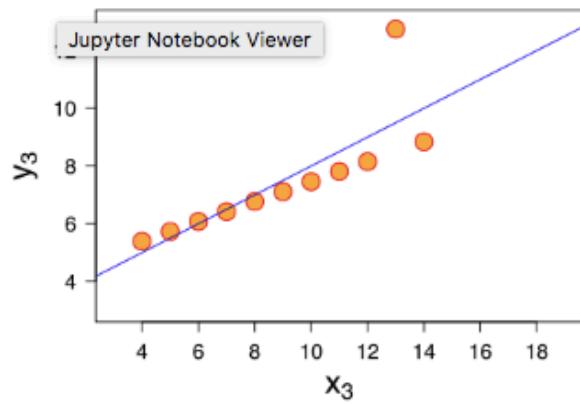
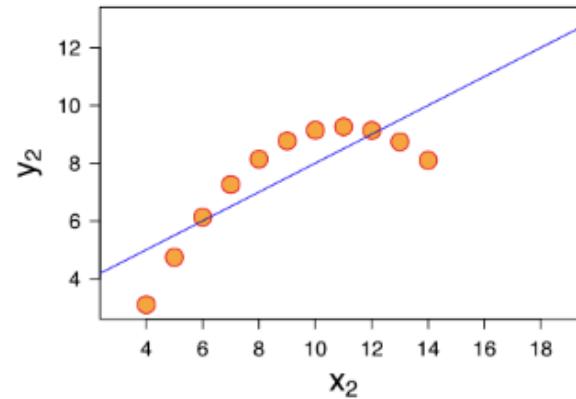
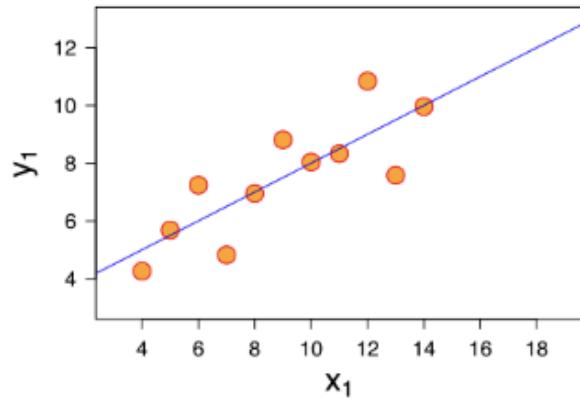
Explorer les espaces des features avant la prise de décision finale

Donner à voir les résultats d'une analyse

# Exemple |

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

# Exemple |



# Exemple 1

Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.122
Correlation	0.816
Linear regression line	$y = 3.00 + 0.500x$

## Exemple 2



Epidémie de choléra à Londres en 1854 par docteur J. Snow à droite centré sur la fontaine infectée

Batonnet noir la localisation des domiciles des victimes

# Exemple 3

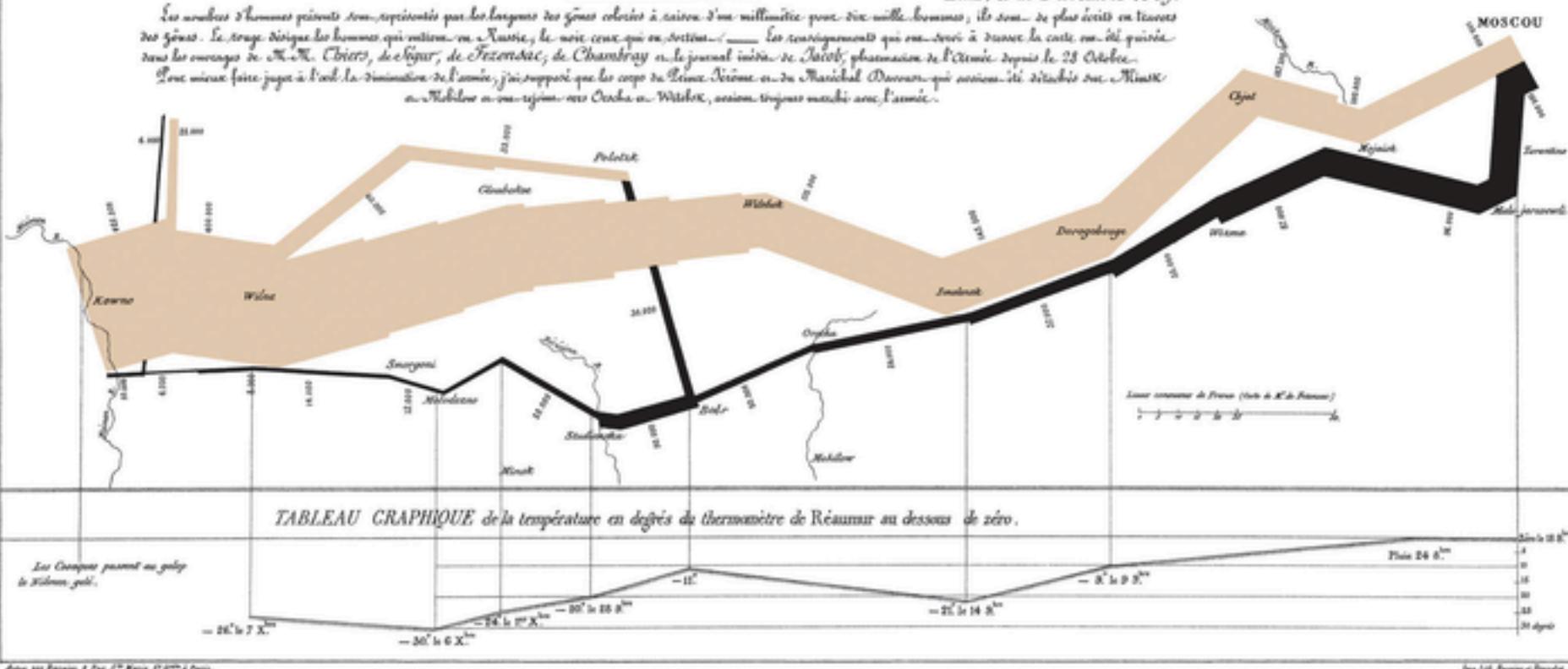
*Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.*

Dessiné par A. Minard, Ingénieur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes perdus sont représentés par les largures des bandes colorées à savoir : une millimètre pour dix mille hommes ; ils sont de plus écrits en lettres des bandes. Le rouge désigne les hommes qui rentrent en Russie ; le noir ceux qui en sortent. — Les renouvellements qui ont lieu à travers la carte sont indiqués sous les noms de Chabrol, de Ségur, de Fonscada, de Chambrey ou le journal intitulé "l'Ami" plusieurs fois par l'Armée depuis le 28 Octobre.

Lors mieux faire juge à l'œil la situation de l'armée, j'ai supposé que les corps de l'Armée Nécessaire ou du Maréchal Davout qui arrivent à l'armée de Moscou au début de novembre vers Orelle ou Vitebsk, entrent toutefois avec l'armée.



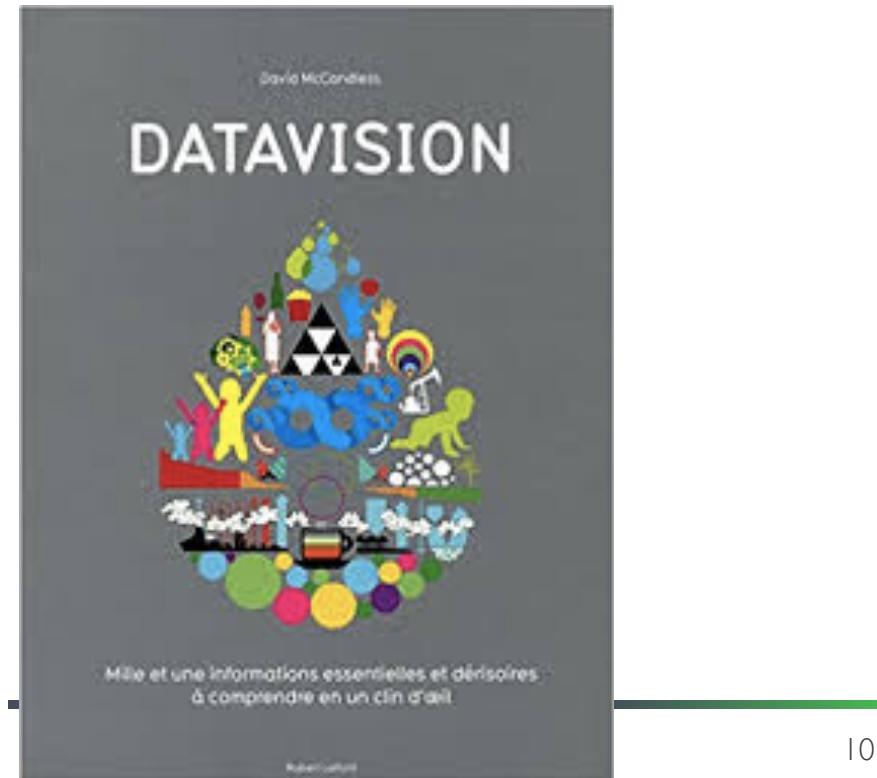
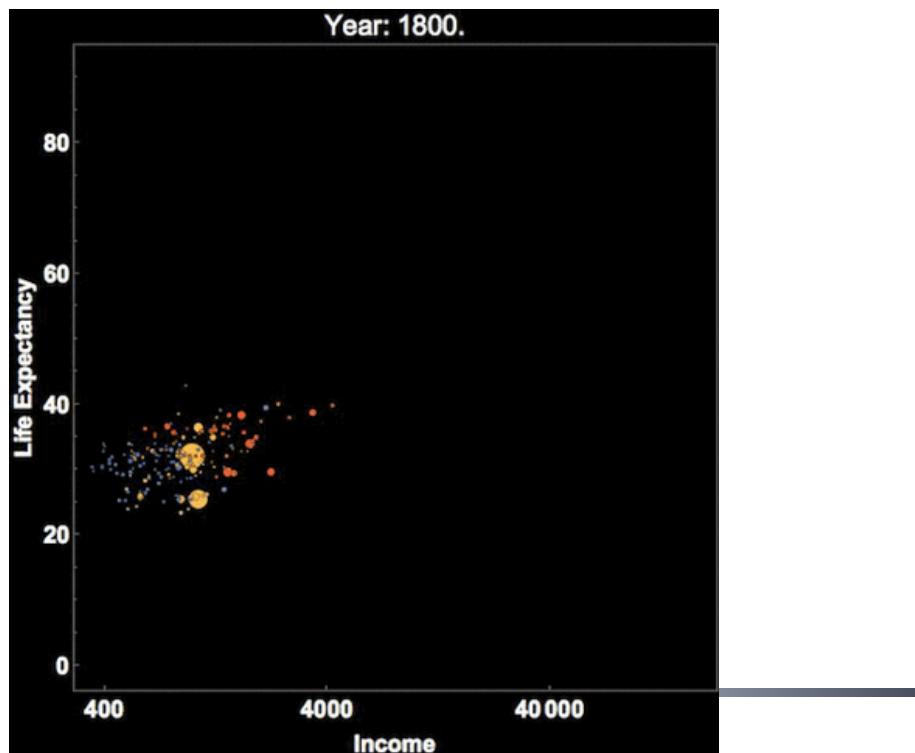
Pertes humaines lors de la campagne de Russie par C. J Minard (1869)

# Plus d'exemples ?

Histoire de l'infographie : <https://infowetrust.com/>

Logiciel de visualisation interactive développé par H. Rosling : <https://www.gapminder.org/>

Motivation de H. Rosling : [https://www.ted.com/talks/hans\\_rosling\\_shows\\_the\\_best\\_stats\\_you\\_ve\\_ever\\_seen?language=fr](https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen?language=fr)



# Offres d'emplois en data via

- <https://groups.google.com/forum/#!forum/data-vis-jobs>
- <https://twitter.com/search?q=%23dataviz>

# Des données au graphique

« Overview first, zoom and filter then details on demand » (Shneiderman 1996)

La première étape clé de la production d'une visualisation consiste à choisir l'encodage graphique

Quelle variable de l'image (position, couleur, ...) va servir à encoder quel attribut de données (âge, taille, poids, ..des individus listés dans notre jeu de données) ?

Pour cela il convient de connaître les différentes types d'attributs qu'on peut avoir à traiter

de connaître les variables de l'image qui sont disponibles pour les encoder et la manière dont elles sont perçues par l'autre humain



# Les différents attributs

**Attributs nominaux** : valeurs parmi un ensemble (fini ou non) qui ont pour seule comparaison possible l'identité.

**Exemple** : nom d'une personne (nombre infini de valeurs possibles), ou son sexe (H ou F).

**Remarque** : il est tout à fait possible de coder sous forme de nombre.

**Métriques** :

On peut donner la distributions des attributs, les k éléments les plus fréquents,  
...

Si l'ensemble est muni d'une structure (par exemple nom de ville, structuré hiérarchiquement par la géographie : département, région, état), il est possible de grouper les valeurs avant de donner une distribution

Si on doit résumer un ensemble par une unique valeur, on ne peut donner que la valeur la plus fréquente, i.e. , le mode.

# Les différents attributs

**Attributs ordonnés** : pour deux valeurs ont peut tester l'intensité mais aussi laquelle est la plus petite.

**Exemple** : échelle de Likert « pas du tout d'accord », « pas d'accord », « indifférent », « d'accord », « tout à fait d'accord », taille des vêtement

**Remarque** : aucun calcul n'est envisageable sur les valeurs de ces attributs, on ne dispose pas de structure algébrique.

**Métriques** :

Distribution, éléments les plus fréquents, bornes, minimum, maximum, quantiles

Si on doit résumer à une valeur, une alternative au mode est la médiane.

# Les différents attributs

**Attributs quantitatifs** : pour deux valeurs on peut tester l'identité, l'ordre et mesurer des différences.

**Métriques** :

Toutes (moyenne, écart type, ...)

Echelle d'intervalle (différence)

Echelle de ratio si le rapport fait sens, i.e. quand la valeur nulle n'est pas arbitraire ( mesure en mètre OK, date pas OK)



# L'image

La perception visuelle met en jeu des phénomènes complexes étudiés depuis longtemps mais non encore complètement élucidés.

Quelques modèles simplistes pour raisonner sur les visualisations

**La théorie de Gestalt** (1890 par [Christian von Ehrenfels](#)):

L'analyse des différences entre stimulus et perception

**La sémiologie de Bertin** (1967)

Un inventaire systématique des signes utilisés dans les graphiques

**Le design automatique de Mackinlay** (1986)

Un proposition de règles pour la sélection automatique de signes

**Le design selon Tufte** (1970)

Ensemble de règles pour estimer la qualité d'un visualisation

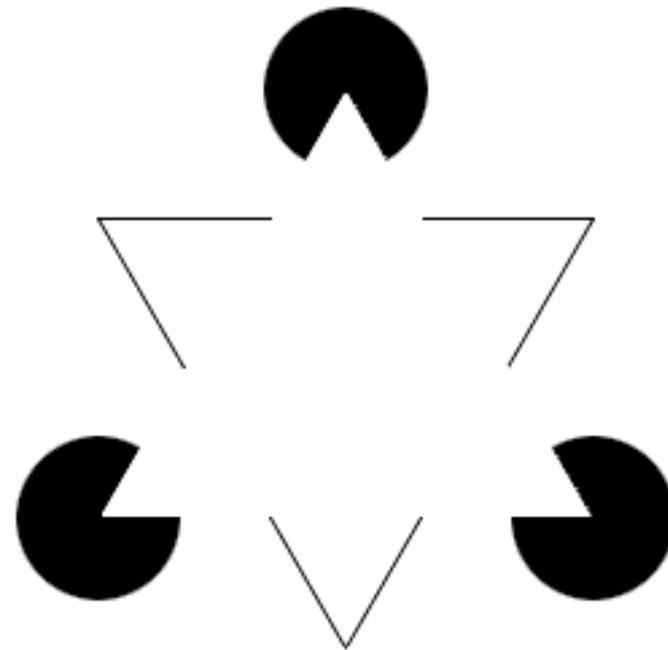
# La théorie de Gestalt

**L'émergence** où l'interprétation du cerveau comble des manques de notre système visuel en inférant à partir de diverses sources, notamment notre expérience acquise.



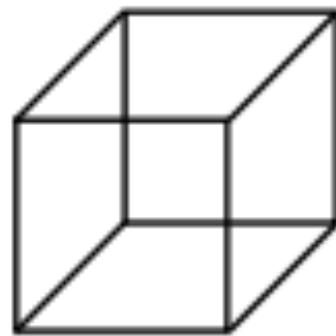
# La théorie de Gestalt

La **réification** nous fait percevoir des formes dans les interstices



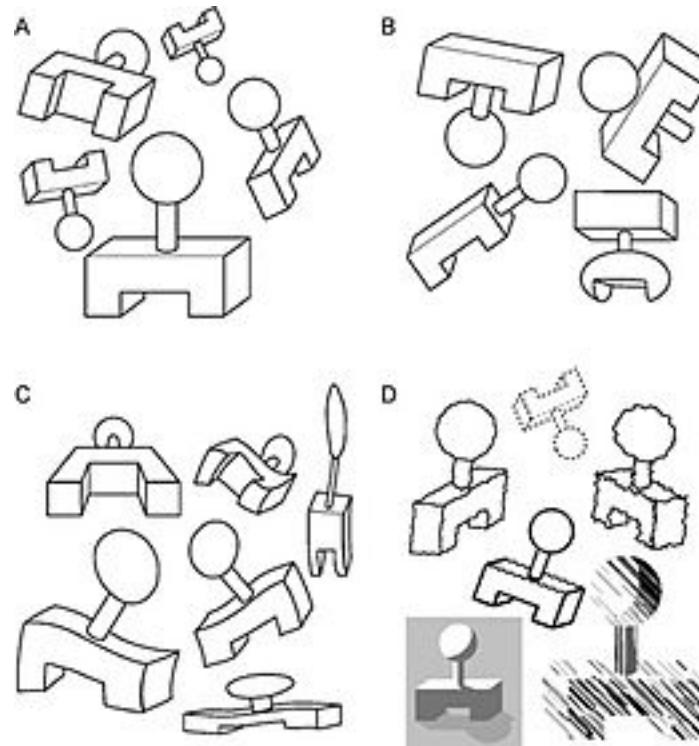
# La théorie de Gestalt

La **multistabilité**, un stimulus peut produire plusieurs perception



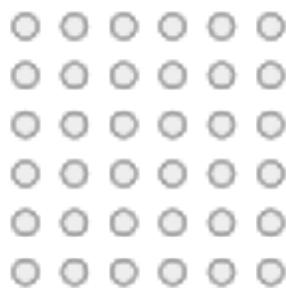
# La théorie de Gestalt

**L'invariance**, est la propriété par laquelle la perception d'objets géométriques simples est reconnue de façon robuste à des transformations géométriques simples (rotation, translation, homothétie, ...)

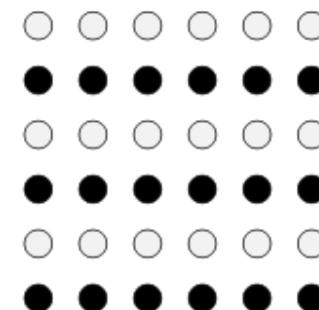


# La théorie de Gestalt

Au delà de la perception des formes Gestalt explore ce qui conduit à la perception des groupes



Proximité



Similarité



Fermeture

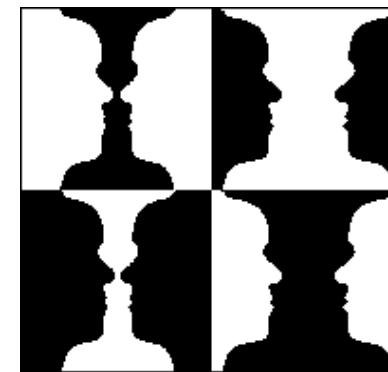
Law of Symmetry



Symétrie



Continuité



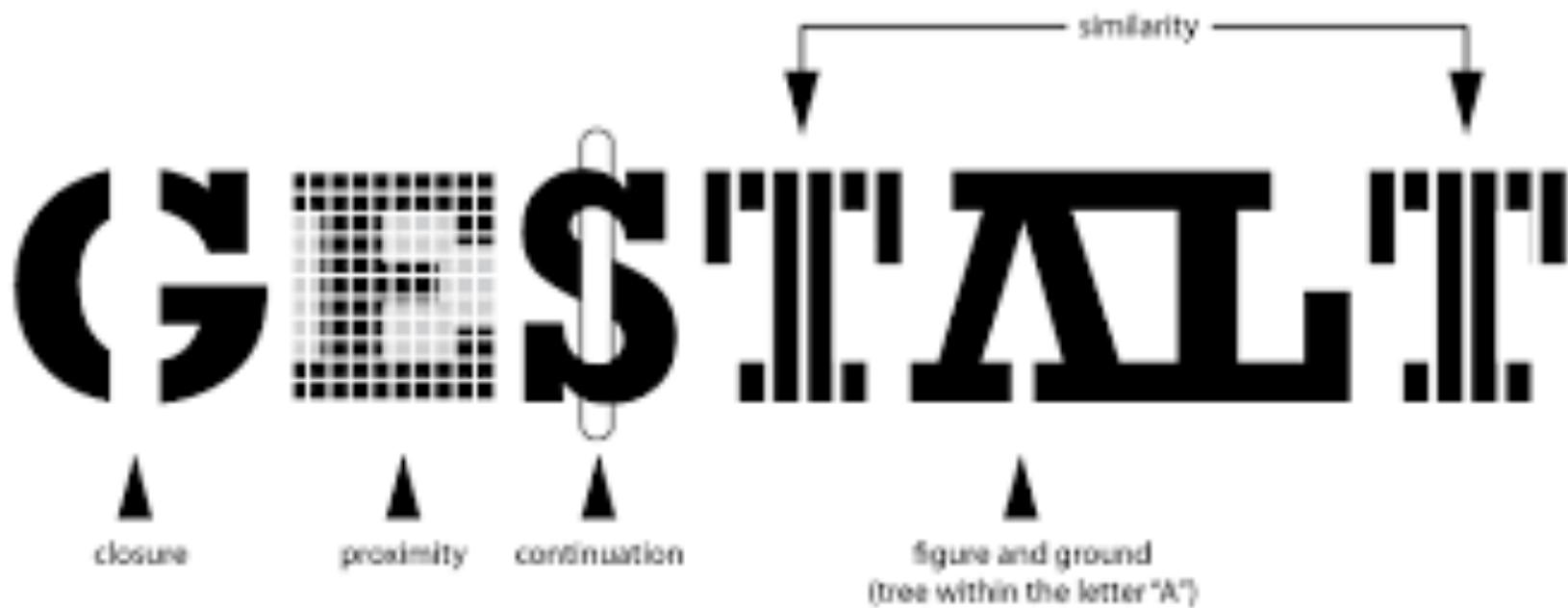
Smallness

# Exercice |



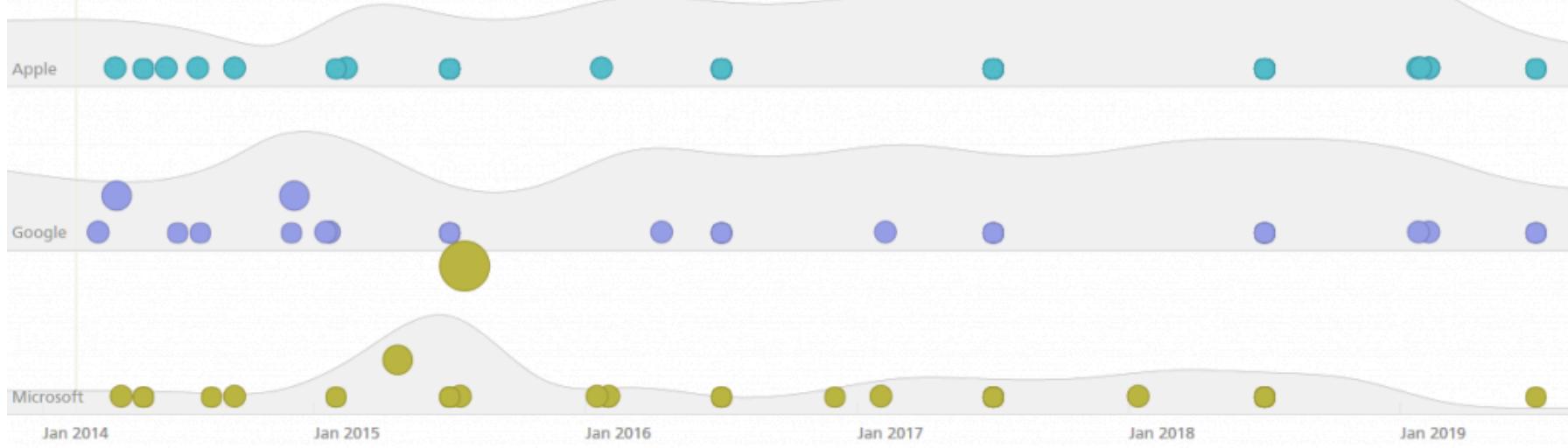
Quels sont les principes de Gestalt mobilisés dans cette data visualization ?

# Exercice 1



# Exercice 2

mentions of Apple, Google, and Microsoft across the web over the last five years.



Quels sont les principes de Gestalt mobilisés dans cette data visualization ?

<https://www.fusioncharts.com/blog/how-to-use-the-gestalt-principles-for-visual-storytelling-podv/>

## **Figure & ground**

The first thing you notice when looking at this visualization is that the bubbles stand out against the backdrop of the area charts. This is appropriate considering the designer wants the viewer to explore deeper information that's embedded within the bubbles. The area chart which is grayed out simply shows the trend over time, and isn't meant to be explored. This is a great example of the Gestalt principle of Figure & Ground.

## **Proximity**

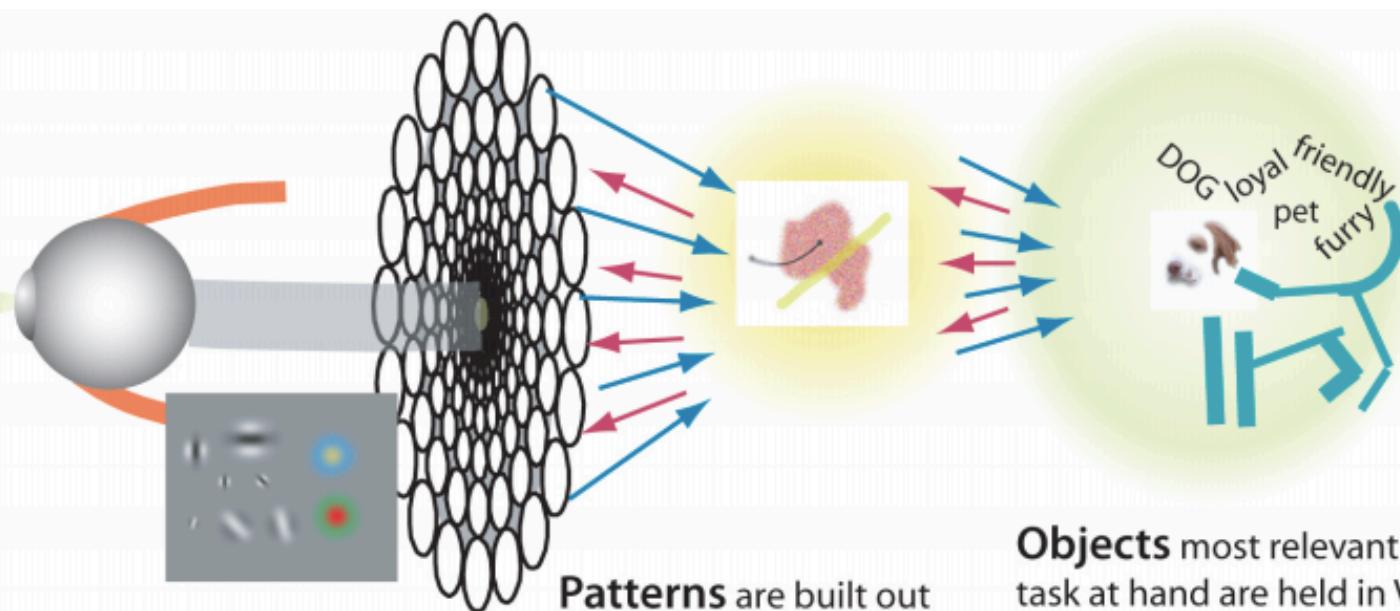
The bubbles are organized in 3 distinct groups along the horizontal line. We can identify the 3 groups of bubbles easily because of how close they are to each other. Notice that enough space is given between each group to make them distinct. This uses the principle of Proximity.

## **Similarity**

Further, we notice that the bubbles are of three colors – green, purple, and blue. This Similarity brings out the grouping even more clearly.



# Perception préattentive



**Features** are processed in parallel from every part of the visual field. Millions of features are processed simultaneously.

**Patterns** are built out of features depending on attentional demands. Attentional tuning reinforces those most relevant.

**Objects** most relevant to the task at hand are held in Visual Working Memory. Only between one and three are held at any instant. Objects have both non-visual and visual attributes.

**Bottom-up information drives pattern building**

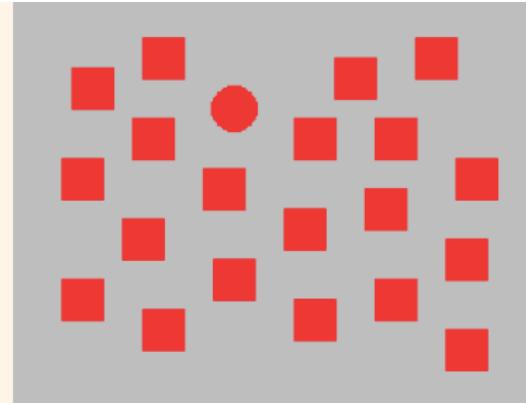
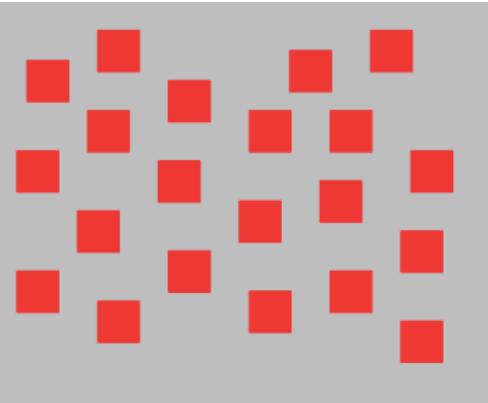
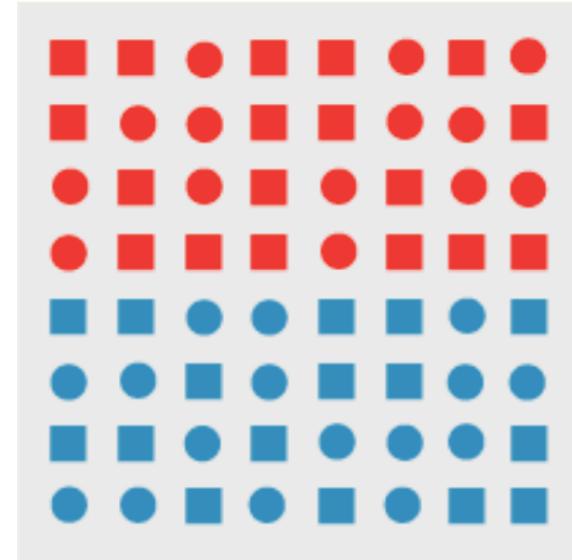
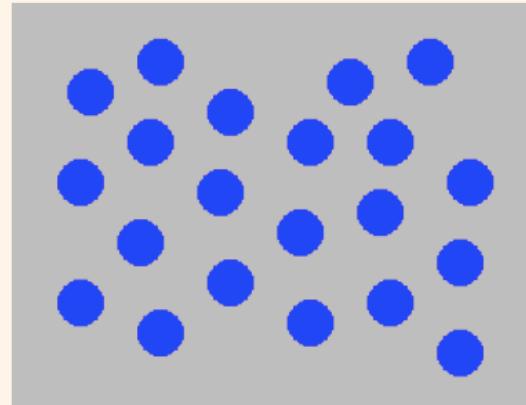
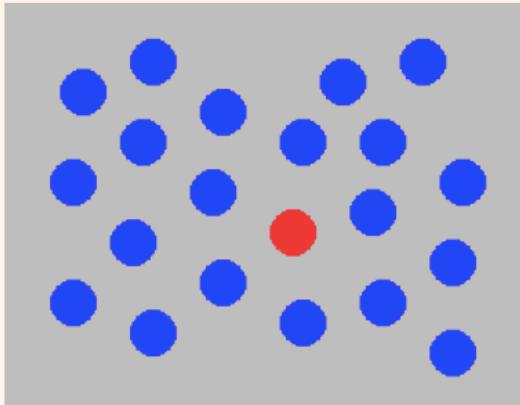
**Top-down attentional processes reinforce relevant information**

# Perception préattentive

Certains éléments attirent plus l'attention que d'autres et sont traités instantanément et en parallèle

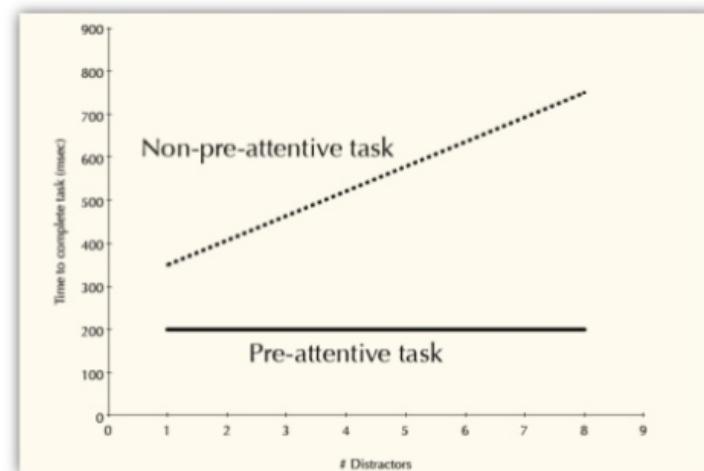
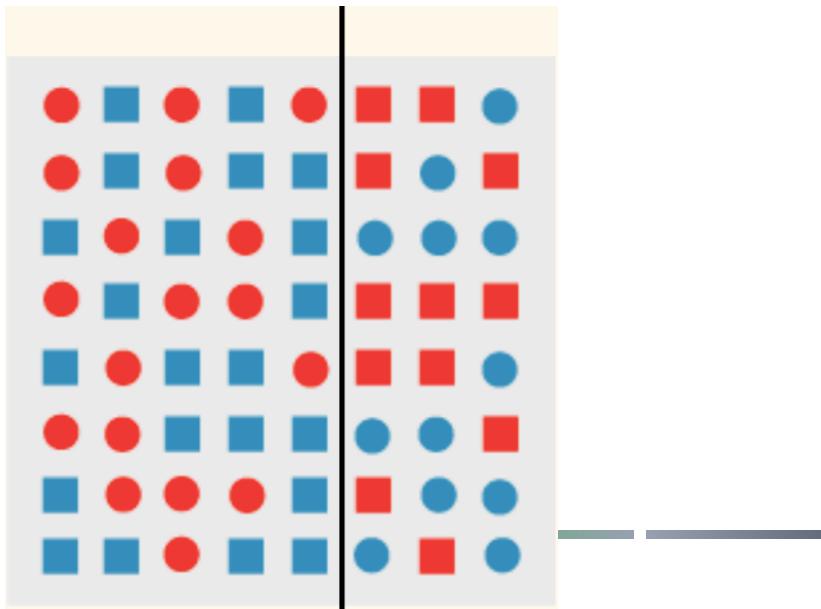
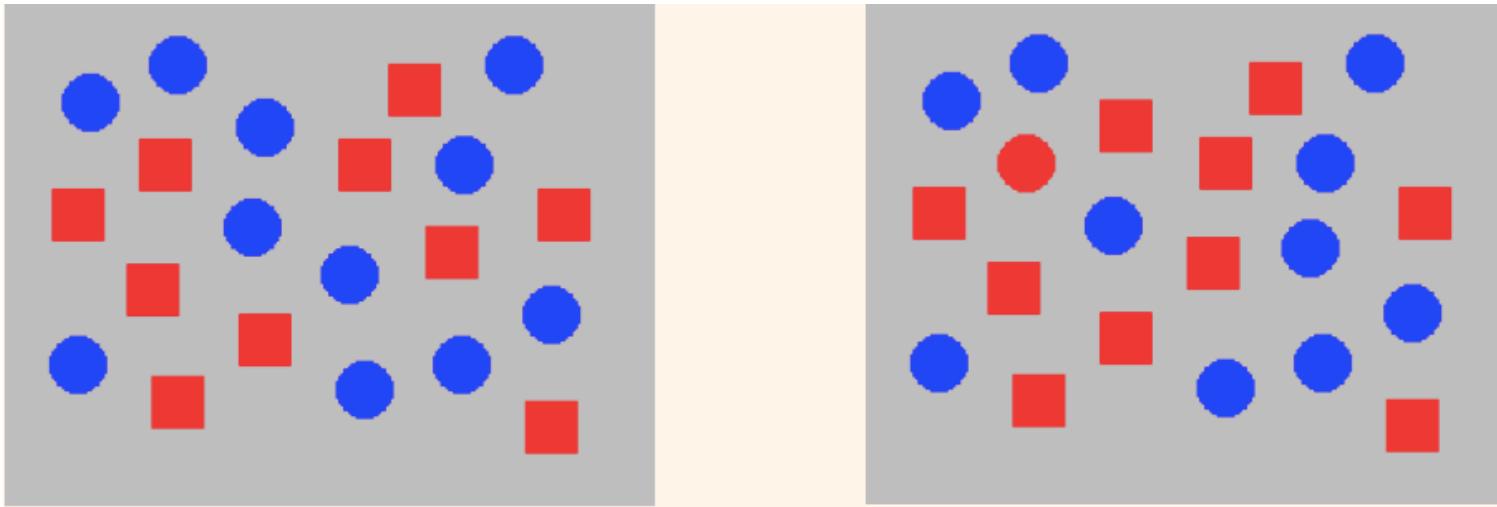
Y a-t-il un cercle rouge ?

Y a-t-il une frontière ?



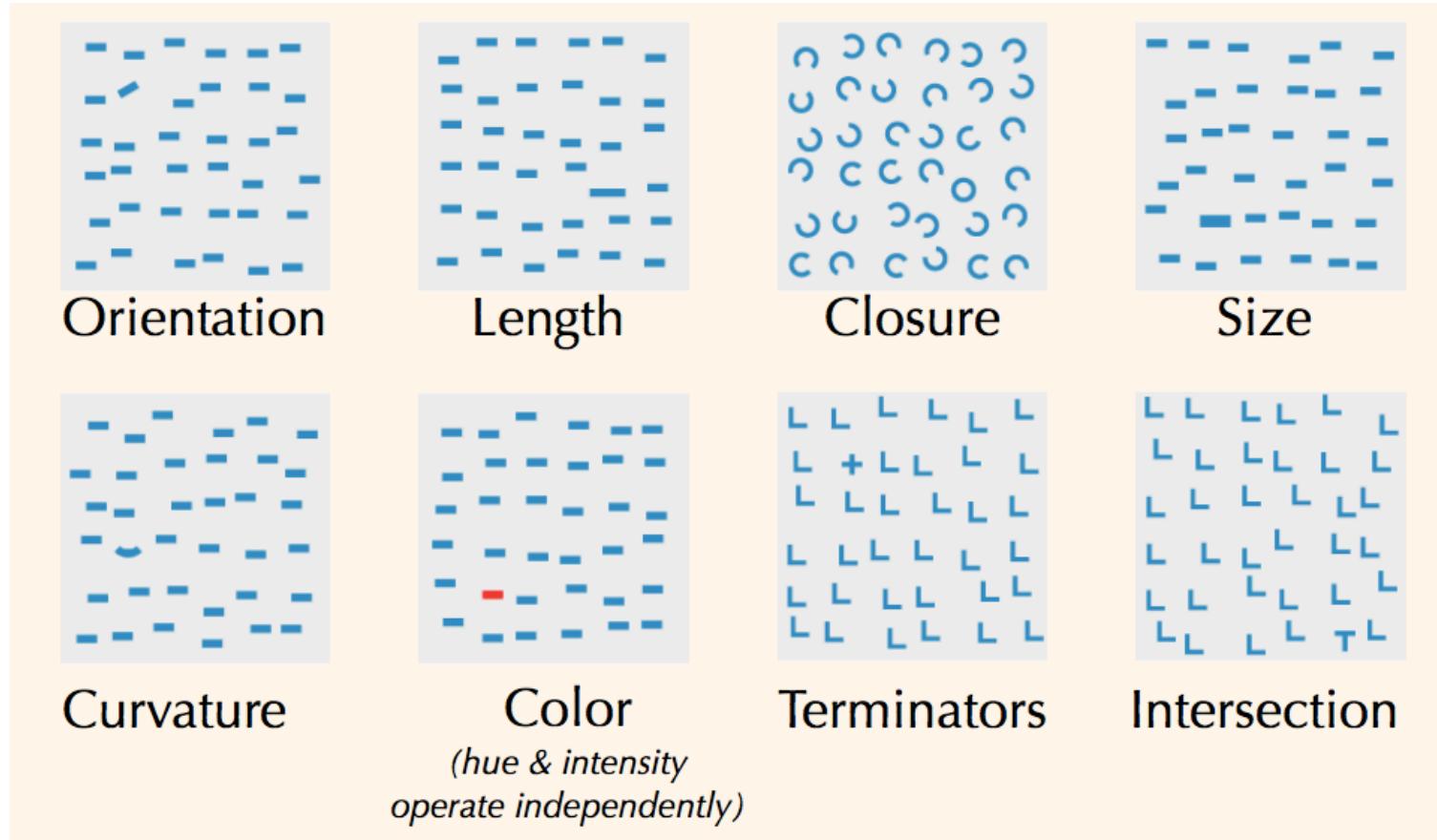
# Perception préattentive

Ceci ne fonctionne que si les distractors diffère sur le même attribut



# Perception préattentive

Liste exhaustive des attributs



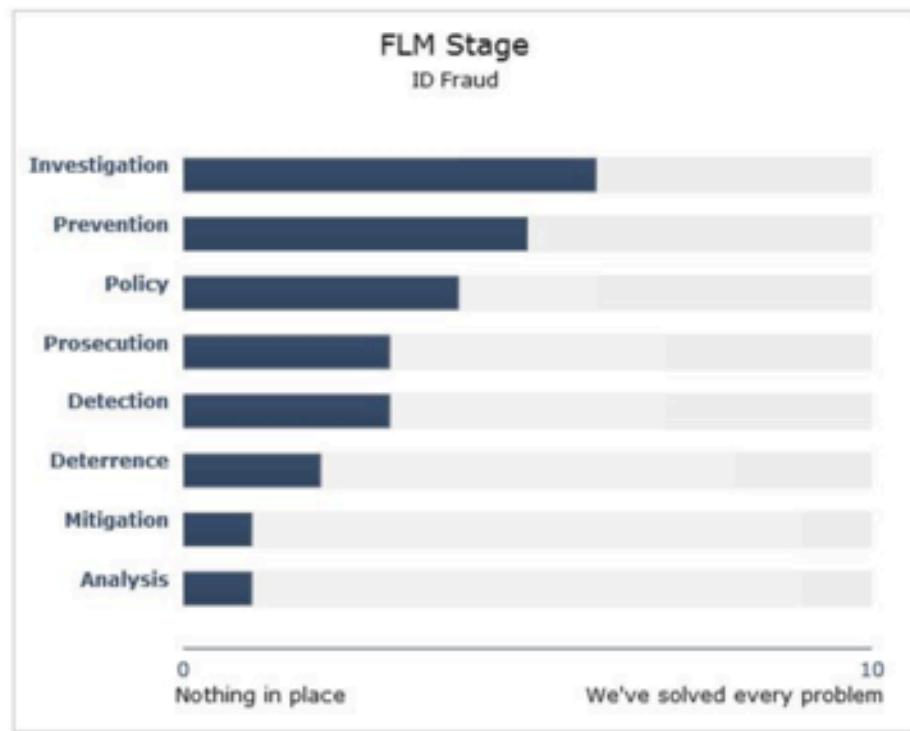
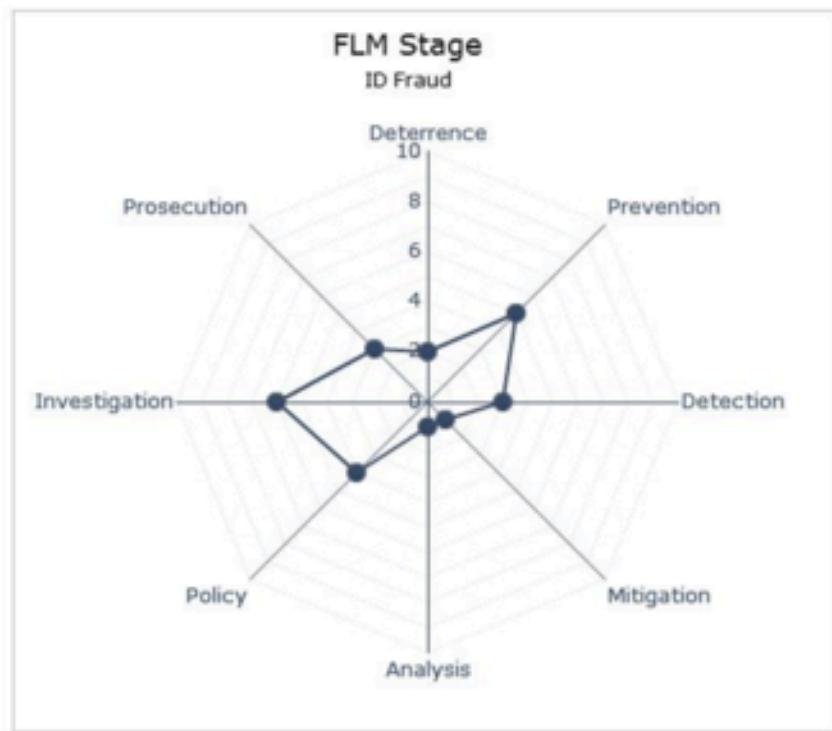
# Conséquence en data visualisation

\* Certaines tâches peuvent être réalisées « gratuitement » par le cerveau :

Target detection, Boundary detection, Region tracking, Counting (estimation)

- \* Plus l'histoire que l'on veut raconter avec les données utilise la perception préattentive plus le lecteur va comprendre l'histoire rapidement
- \* On peut facilement perdre le lecteur si on déclenche des perceptions préattentives de façon non appropriée

# Contre exemple



# La sémiologie de Bertin

La visualisation consiste à encoder les **attributs des données** à l'aide de **marques graphiques** (point, lignes, surfaces, ...).

Ces marques sont caractérisées par des **variables graphiques** (couleur, forme, ...) utilisables pour encoder les divers **attributs des données**.

Bertin a recensé de manière systématique les variables graphiques exploitables pour différencier **des marques graphiques**.

Bertin a caractérisé ces variables du point de vue des jugements visuels que ces marques autorisent :

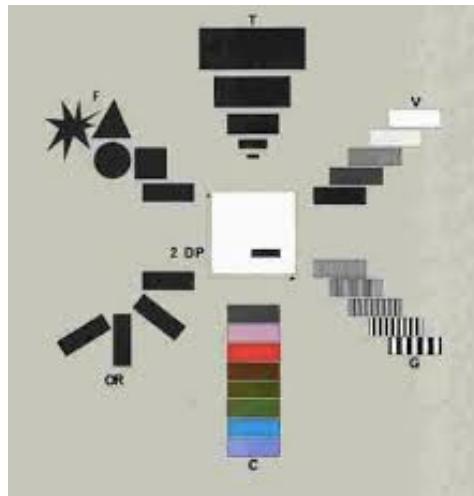
**Association** : est-ce que des marques partageant le même niveau de cette variable graphique nous apparaissent comme appartenant à un groupe ?

**Sélection** : Est-ce que deux marques ayant des niveaux différents nous apparaissent différentes (et si oui combien de niveau peut on percevoir ?)

**Ordre** : peut-on ordonner des niveaux distincts ?

**Quantité** : peut-on quantifier la différence entre deux niveaux distincts

# Les marques graphiques de Bertin



Position  
Taille  
Valeur  
Texture  
Teinte  
Orientation  
Forme

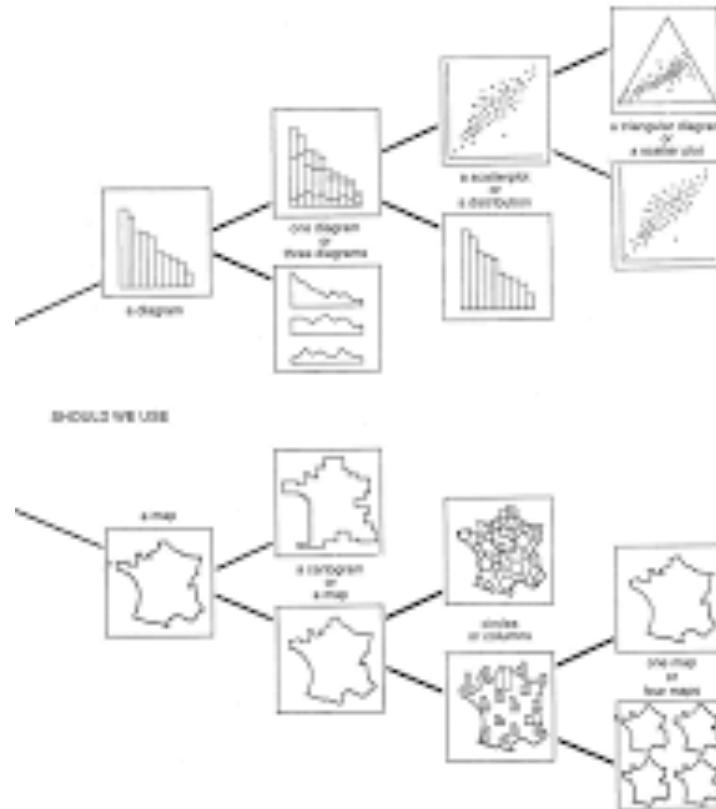
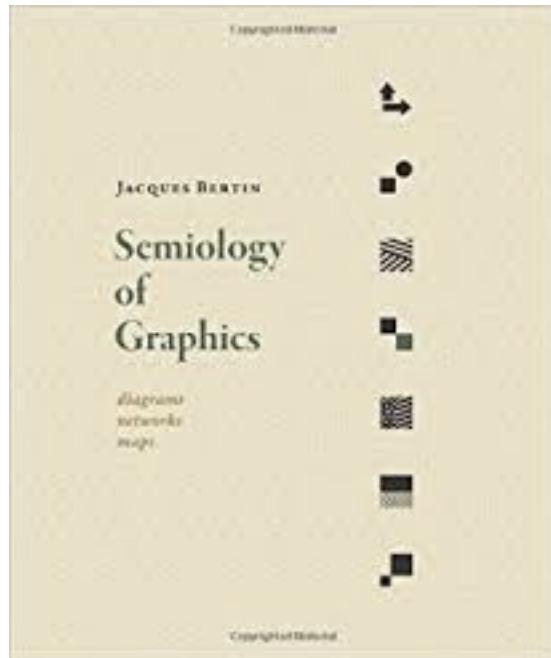


Ordre de priorité

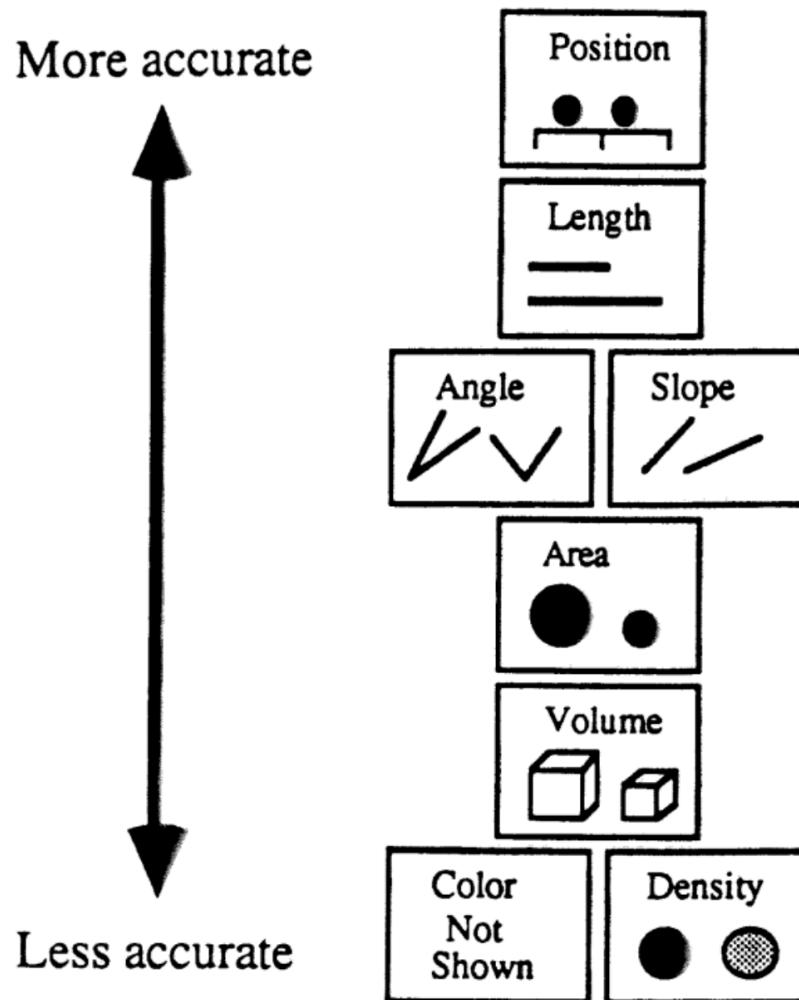
Bertin's Visual Variables

POSITION	SIZE	SHAPE	VALUE	HUE	ORIENTATION	TEXTURE
Selective Associative Ordered Quantitative	Selective Ordered Quantitative	Associative	Selective Ordered Quantitative	Selective Associative	Selective Associative (sometimes)	Selective Associative Ordered (sometimes)

# Dataviz comme un problème combinatoire

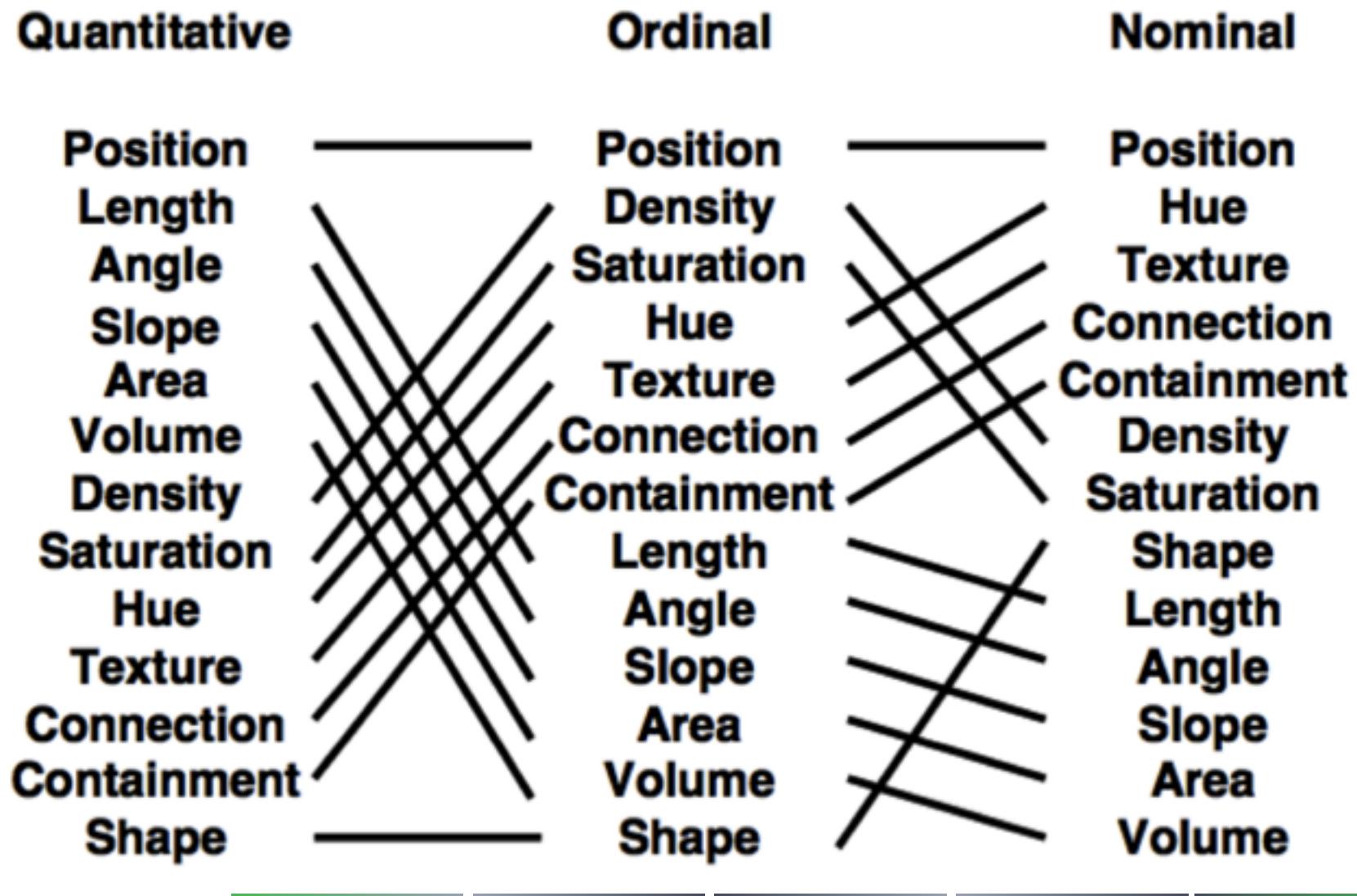


# La combinatoire de Mackinlay

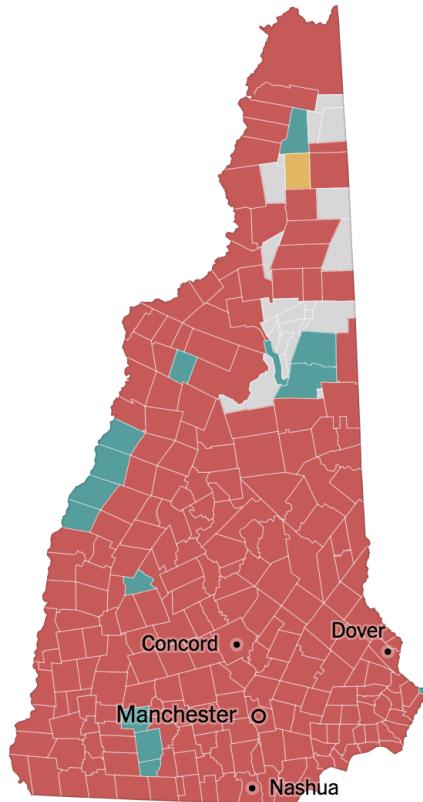


Mackinlay, Jock. ["Automating the design of graphical presentations of relational information."](#) *Acm Transactions On Graphics (Tog)* 5.2 (1986): 110-141.

# La combinatoire de Mackinlay

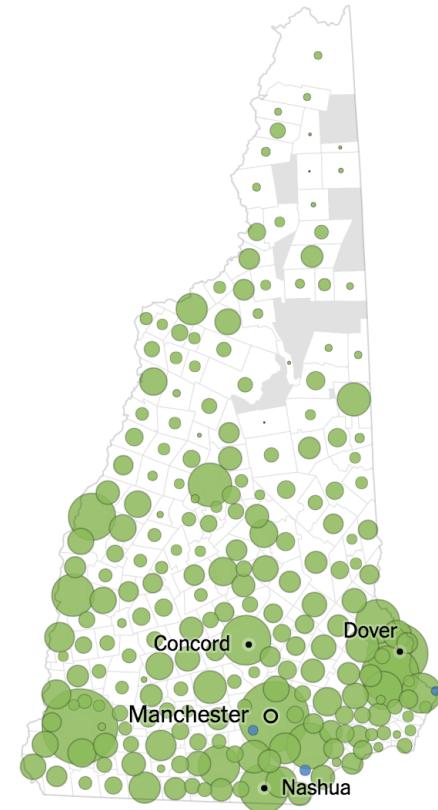


# Exemple 1



LEADER ■ Trump ■ Kasich ■ Cruz

Results Size of Lead Trump Kasich



Circle size is proportional to the size of a candidate's lead.

Results Size of Lead Sanders Clinton

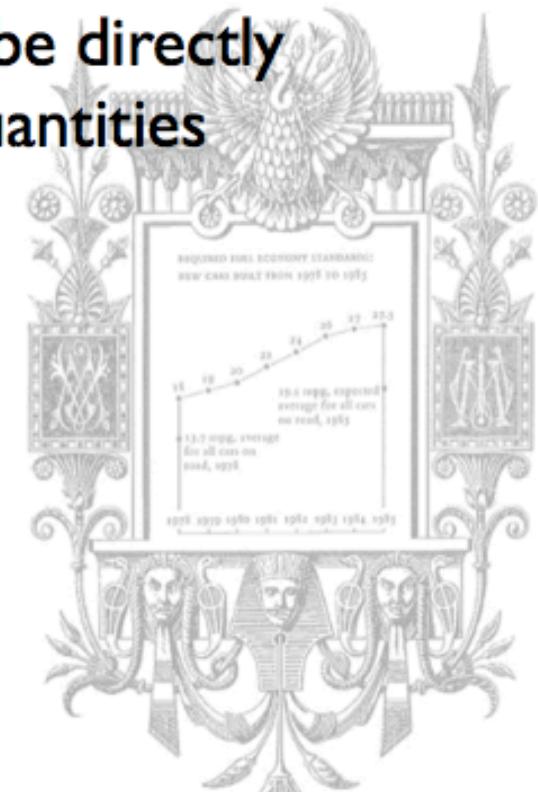
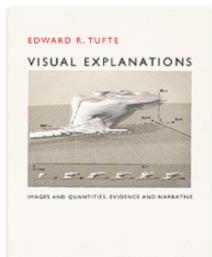
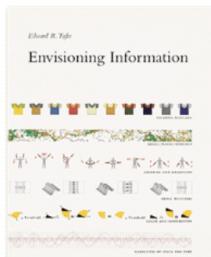
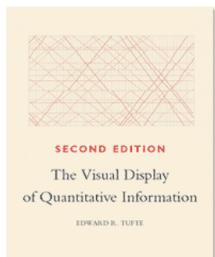
# Les principes de design de Tufte

**Size of the graphic effect should be directly proportional to the numerical quantities (“lie factor”)**

**Maximize data-ink ratio**

**Avoid chart junk**

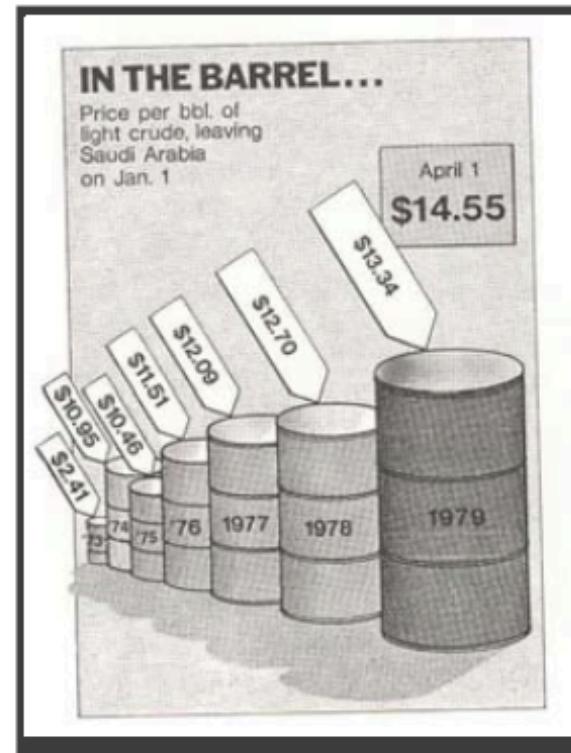
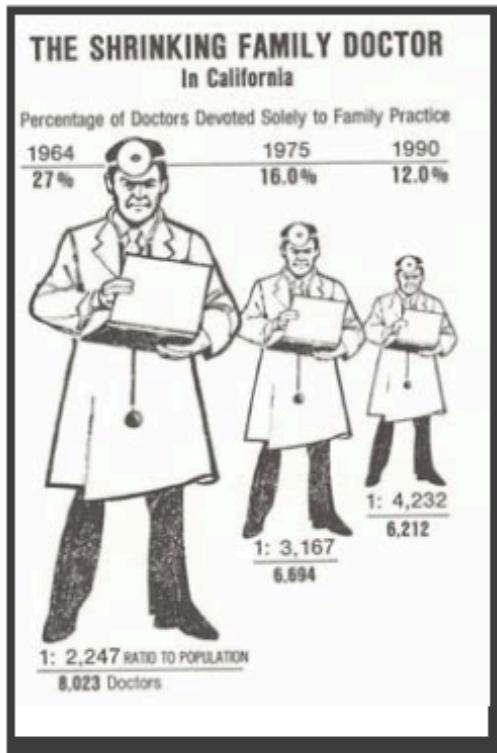
Edward Tufte



# The Lie Factor

Size of effect shown in graphic

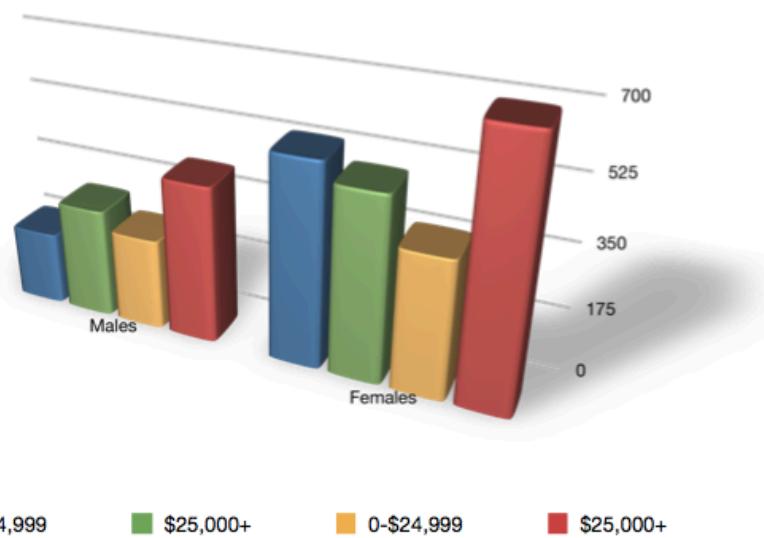
Size of effect in data



Tufte, VDQI

## Maximize Data-Ink Ratio

$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$

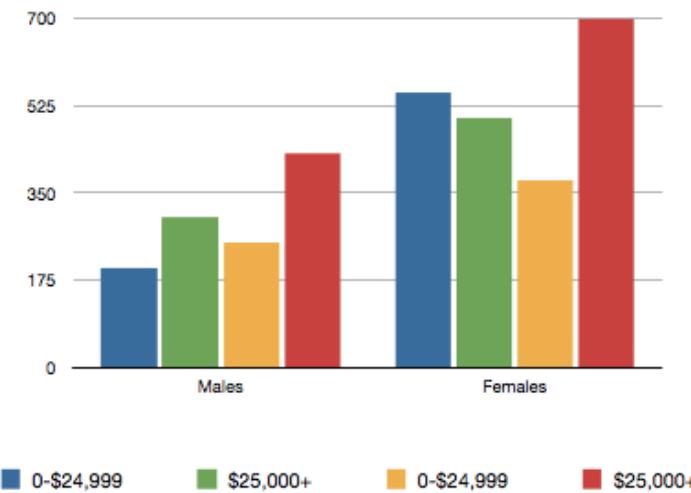


Edward Tufte

Et choisir plutôt

# Maximize Data-Ink Ratio

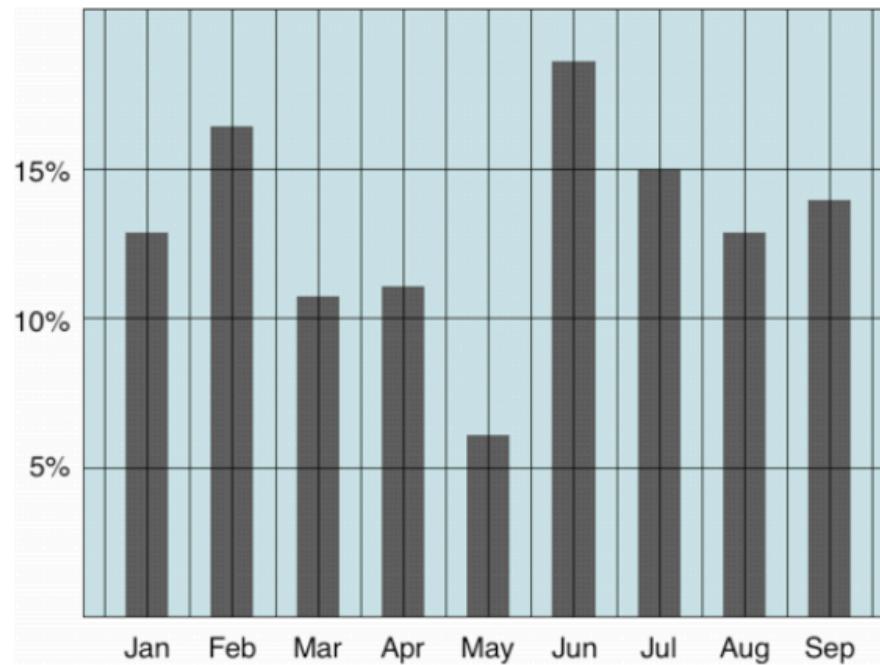
$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$



Edward Tufte

# Avoid Chart Junk

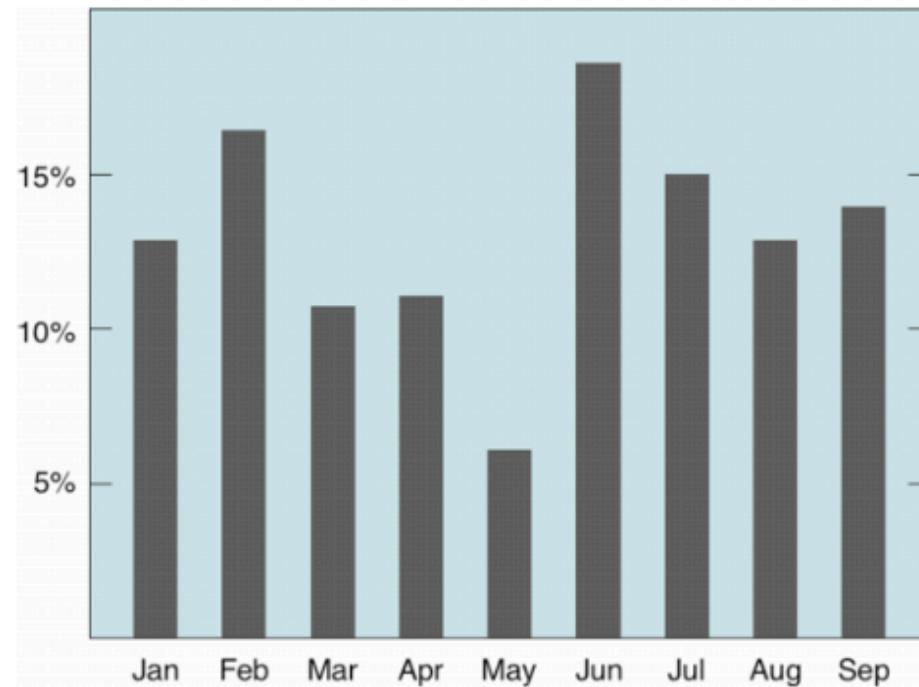
Extraneous visual elements that distract from the message



ongoing, Tim Brey

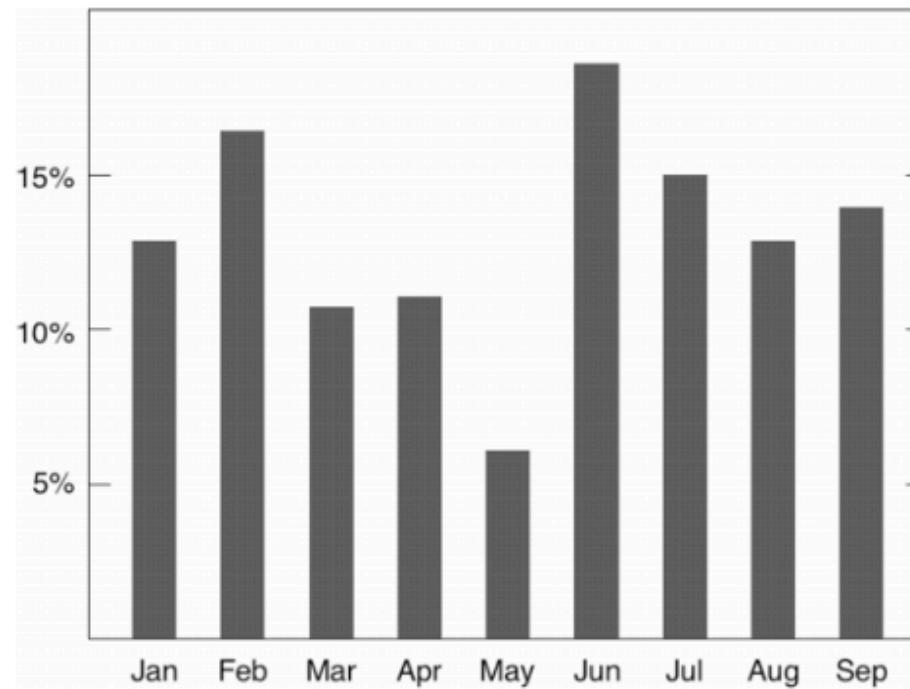
# Avoid Chart Junk

Extraneous visual elements that distract from the message



# Avoid Chart Junk

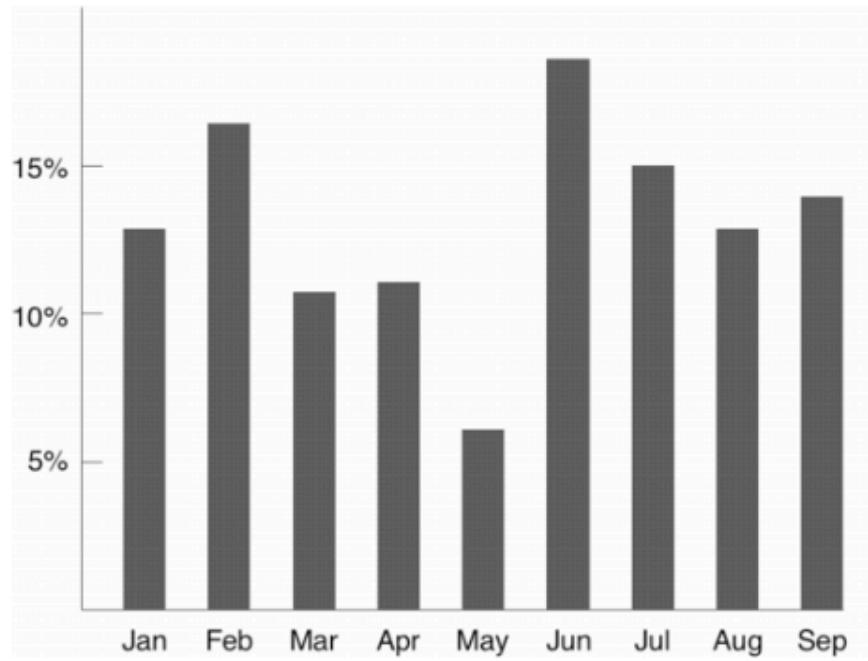
Extraneous visual elements that distract from the message



ongoing, Tim Brey

# Avoid Chart Junk

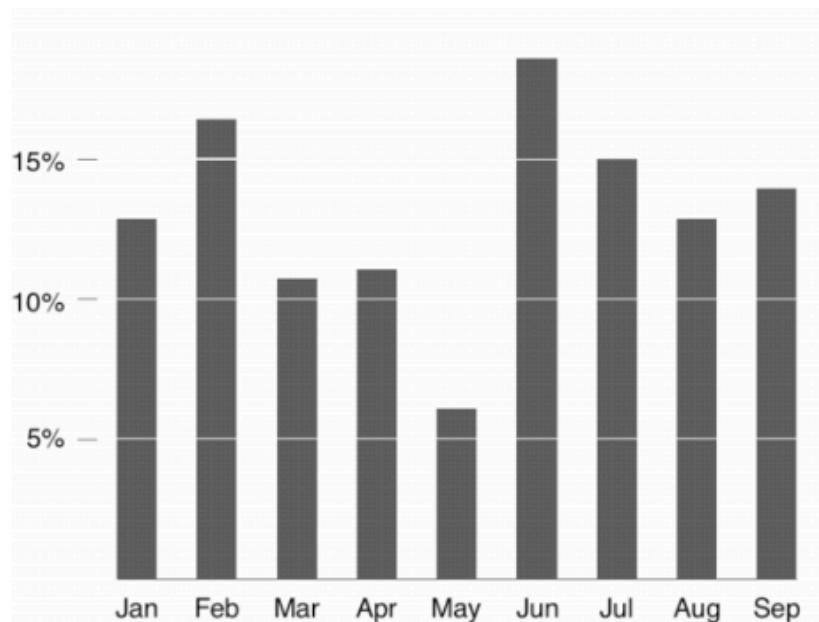
Extraneous visual elements that distract from the message



ongoing, Tim Brey

# Avoid Chart Junk

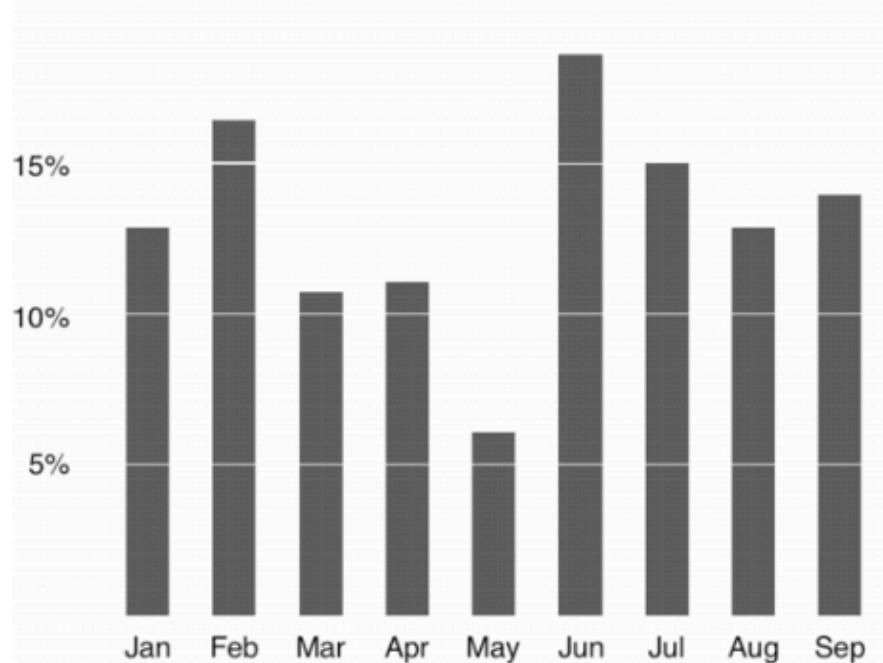
Extraneous visual elements that distract from the message



ongoing, Tim Brey

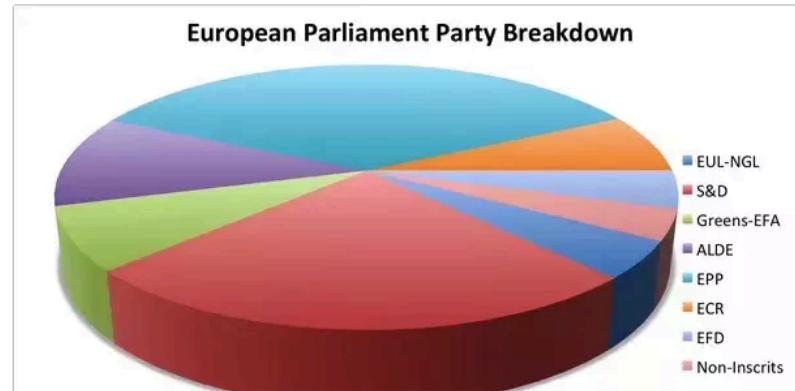
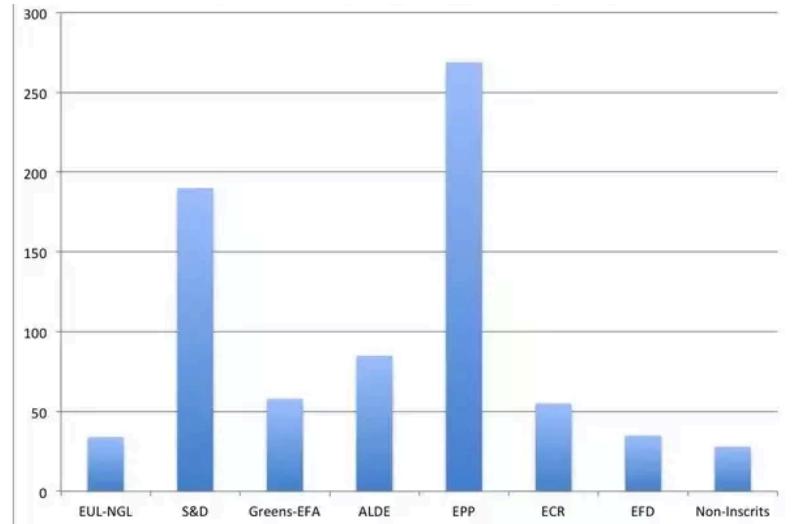
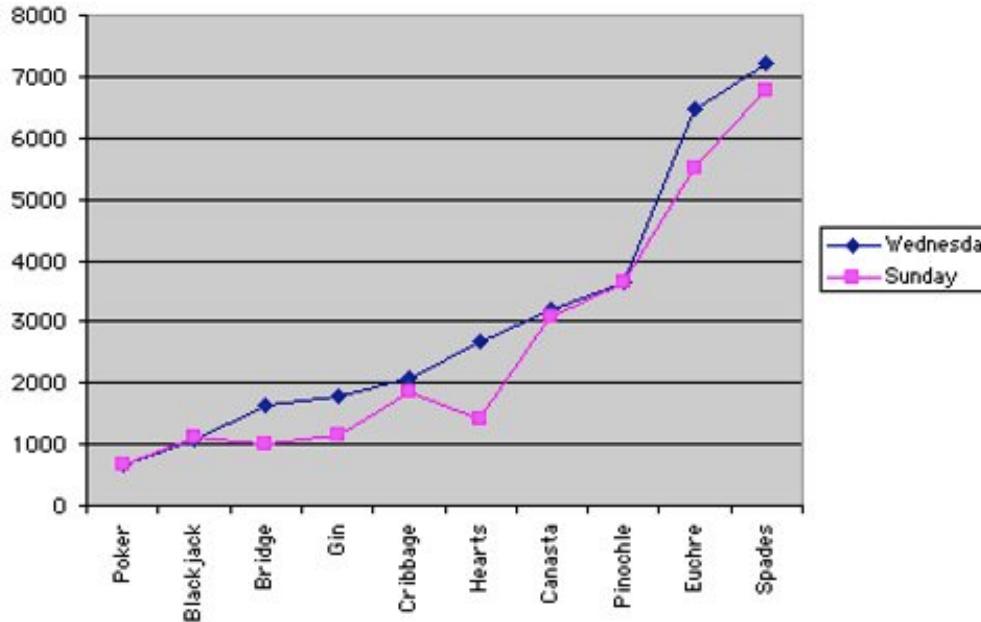
# Avoid Chart Junk

Extraneous visual elements that distract from the message



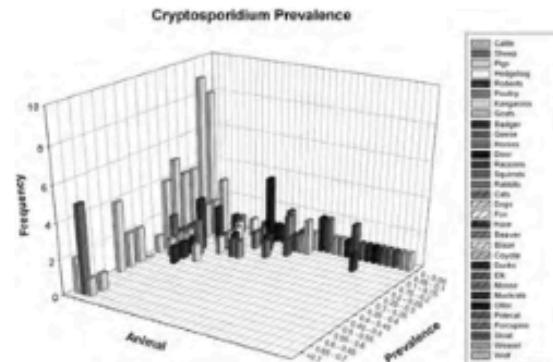
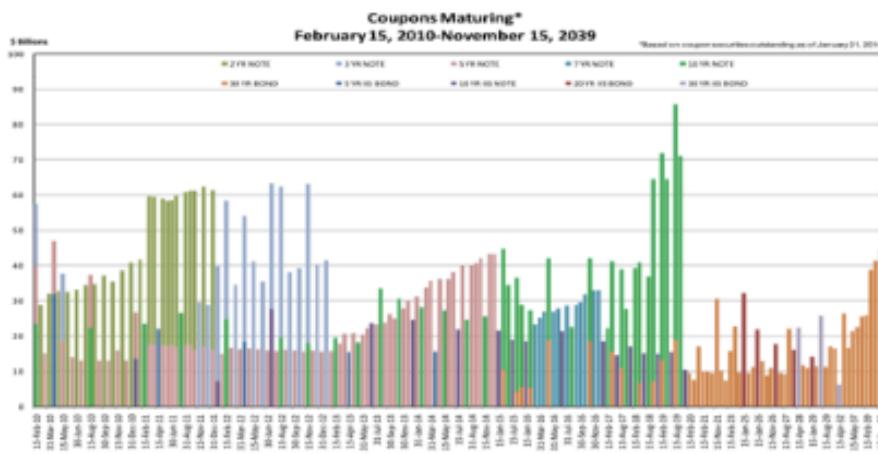
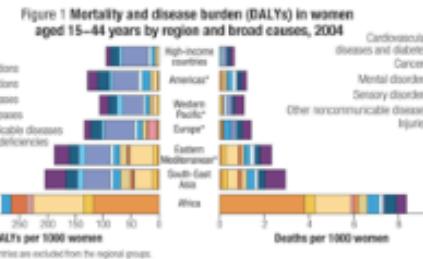
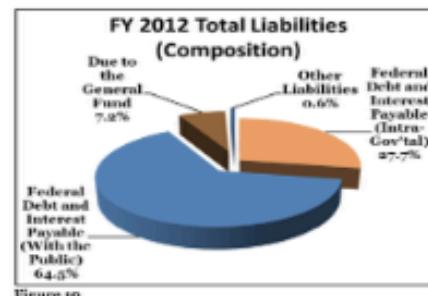
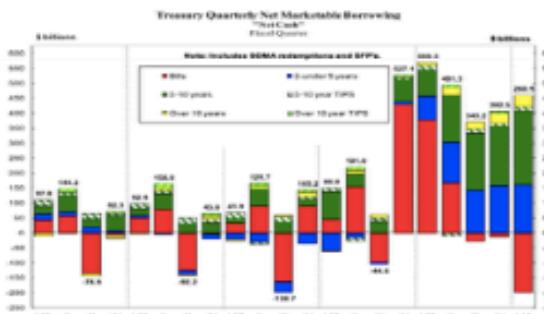
ongoing, Tim Brey

# Contre exemples |



# Contres exemples 2

## Not Effective...

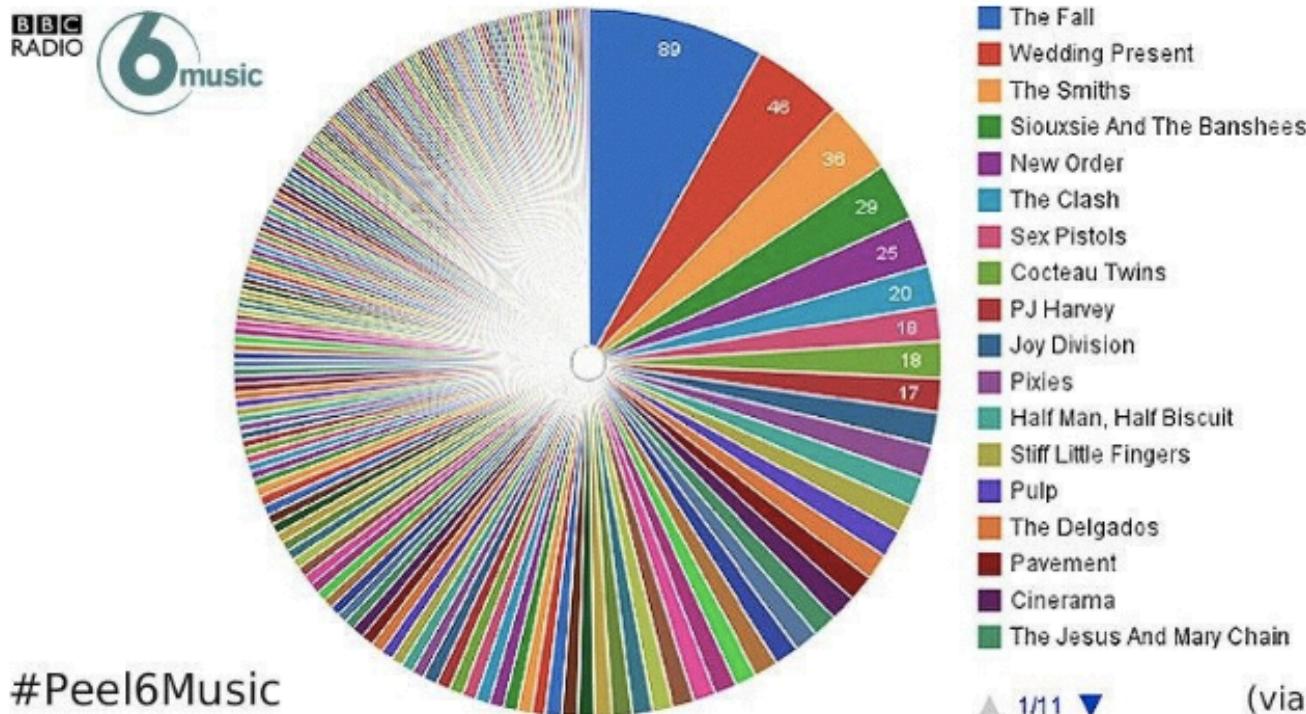


Sources: US Treasury and WHO reports

# WTF Visualizations

<http://wtfviz.net>

John Peel's most played artists in his Festive 50s



# En pratique

## Visualization Templates (or Charts)

### A CLASSIFICATION OF CHART TYPES



<http://www.datavizcatalogue.com/search.html>.

<https://developers.google.com/chart/interactive/docs/gallery>

# Encodage avancé

## ■ Réutilisation de variables graphiques

Si jeu de données avec de nombreux attributs quantitatifs, d'égale importance a priori,

On aimeraient pouvoir utiliser les variables de positions les plus expressives, pour tous ces attributs.

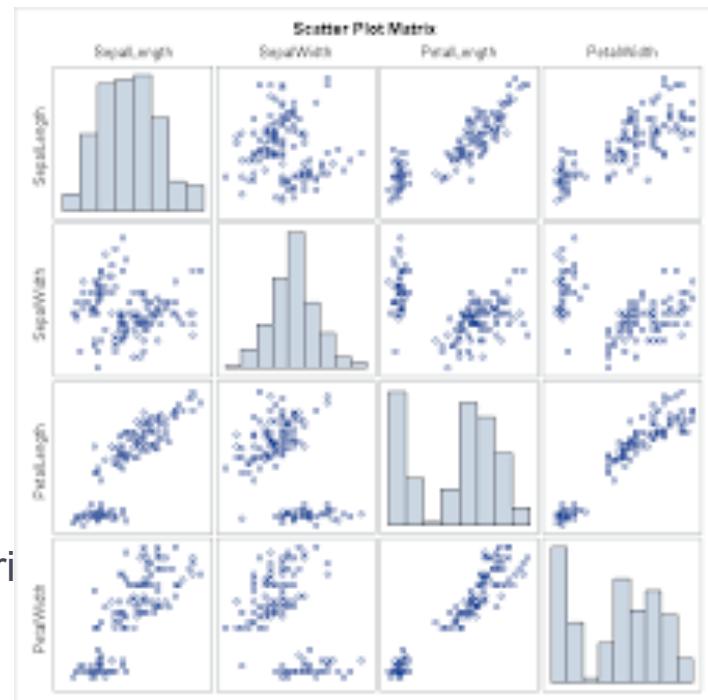
Or on ne dispose que de deux variables de position.

Idée : construire pour chaque couple d'attributs un nuage de points

Les relations (corrélations ou autre) entre attributs Apparaissent du fait des groupements ou alignements

Souvent on représente sur la diagonal les histogrammes de Chaque variable

Ceci permet d'avoir en une vue les statistiques mono et bivari du premier et du deuxième ordre



# Encodage avancé

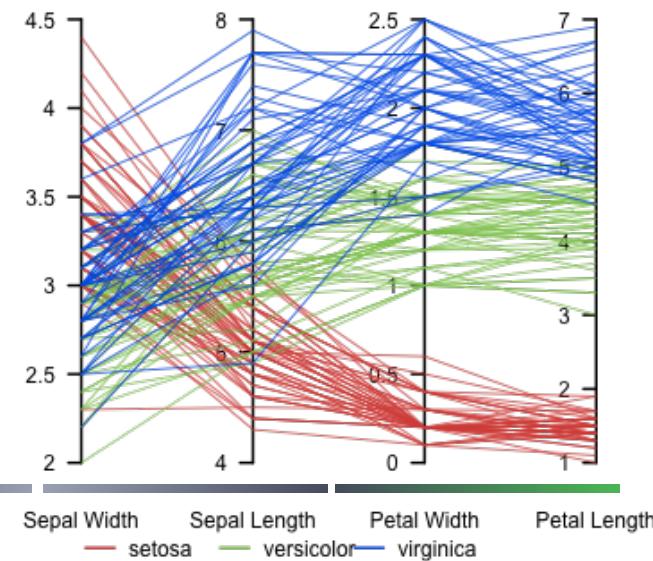
## ■ Réutilisation de variables graphiques

Une autre idée : utiliser autant d'axes verticaux que d'attributs quantitatifs

Les attributs ne sont plus représentés par des points mais par des lignes qui relient les différentes coordonnées le long de ces axes

Cette visualisation appelée coordonnées parallèles permet de repérer grâce aux motifs des lignes les corrélations positives entre deux variables successives

Parallel coordinate plot, Fisher's Iris data



# Encodage avancé

## ■ Composition de variables graphiques

Dans les deux exemples précédents, la réutilisation d'une variable graphique concerne les variables de position

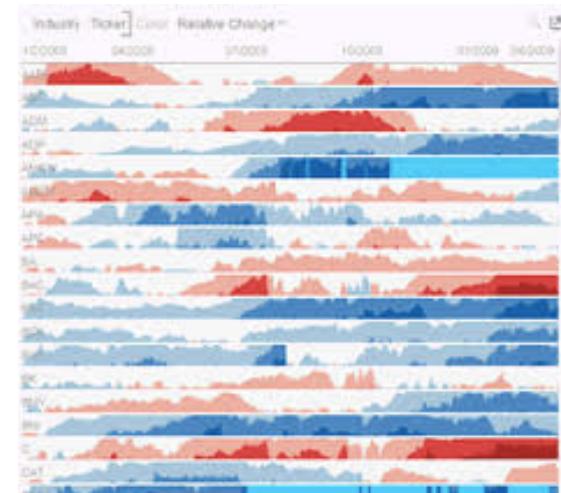
Cela permet la juxtaposition de visualisations mais inconvénient espace alloué à chaque sous-partie est réduit.

Ici utilisation conjointe de plusieurs variables graphiques pour représenter un seul attribut de données

Exemple : les Horizon graphs

On code l'écart à une moyenne avec trois attributs (par exemple signe, quotient, et reste de la valeur absolue divisée par une constante

Par trois variables graphiques (couleur, valeur et hauteur)



# Encodage avancé

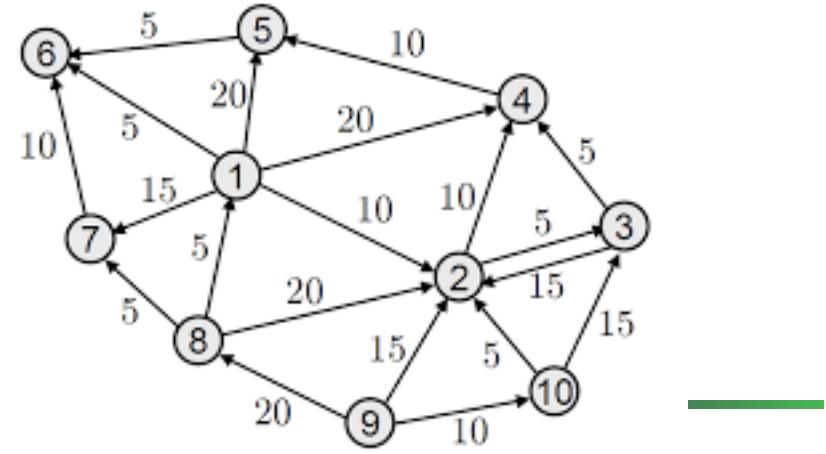
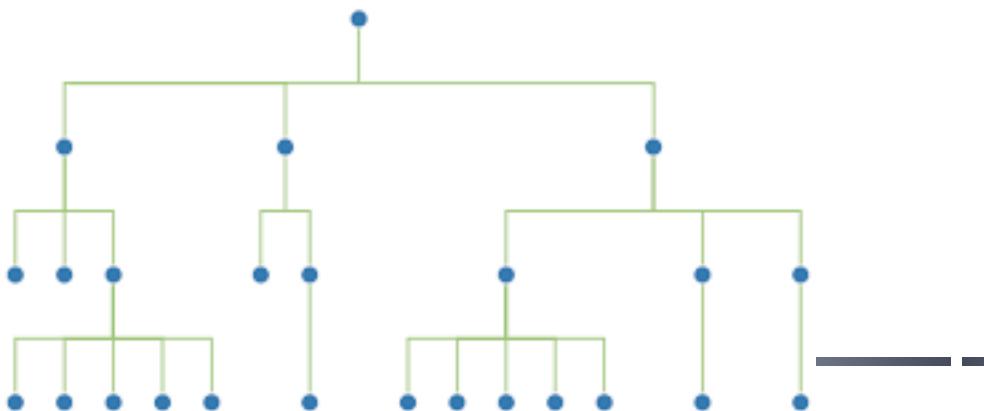
## ■ Encodage des liens entre individus

Beaucoup de jeux de données comportent des informations de relation entre individu

Si pour chaque individu on dispose d'un lien vers un parent, on parle de hiérarchie

Alors que si plusieurs liens sont possibles, on parle de réseaux.

Parfois les liens eux-mêmes peuvent comporter des attributs(nature et longueur d'un axe routier reliant deux villes, ...)

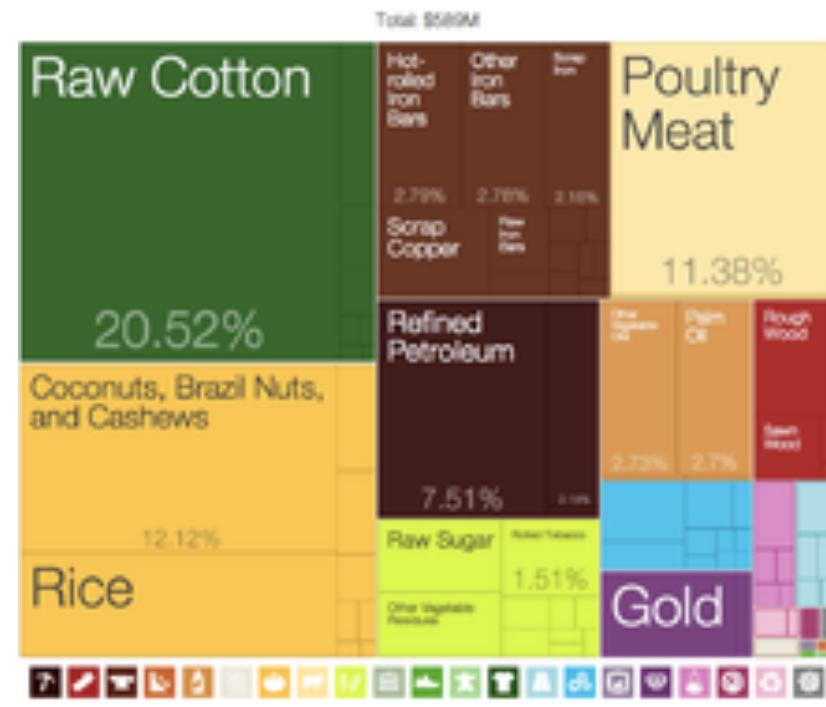
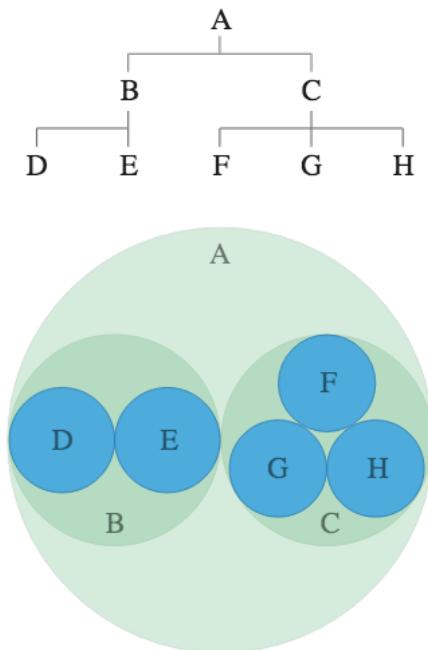


# Encodage avancé

#### ■ Encodage des liens entre individus

Des alternatives existent à ces représentations

# Les Tree-maps

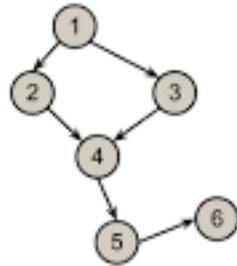


# Encodage avancé

## ■ Les réseaux

Observer la matrice d'adjacente peut être une bonne alternative la visualisation du graphe pour de grands réseaux

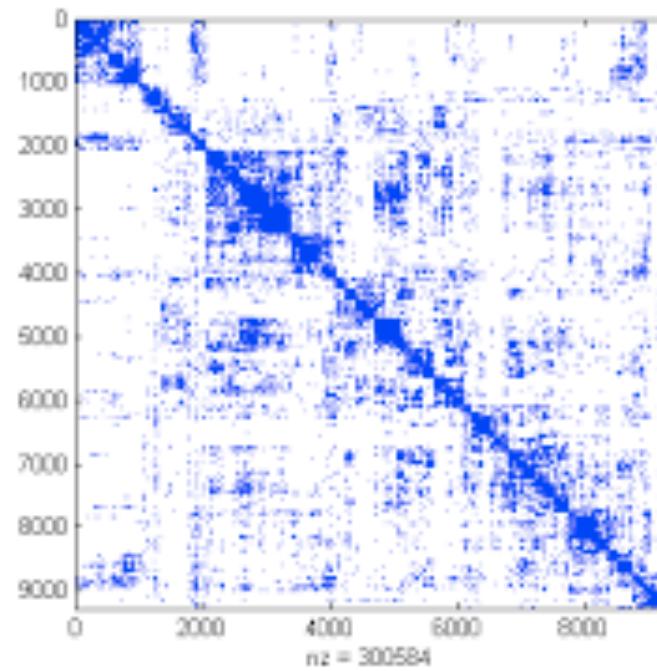
Directed Graph & Adjacency Matrix



Undirected Graph

	1	2	3	4	5	6
1	0	1	1	0	0	0
2	-1	0	0	1	0	0
3	-1	0	0	1	0	0
4	0	-1	-1	0	1	0
5	0	0	0	-1	0	1
6	0	0	0	0	-1	0

Adjacency Matrix



<https://gephi.org/>

# Take home message

Our brains take lots of perceptual “shortcuts”... ... which can either help or harm our visualizations!

Quelques bonnes pratiques à connaître.

- <https://lyondataviz.github.io/teaching/lyon1-m2/2017/tp1.html>
- <https://lyondataviz.github.io/teaching/lyon1-m2/2017/projets.html>
- <http://www.cs.ubc.ca/group/infovis/resources.shtml>
- <http://cs.colby.edu/courses/S14/cs251/goodbad.php>

# Thanks for your attention