Explorer le monde des **données**

par la visualisation

Antoine Cornuéjols et Christine Martin

(prenom.nom@agroparistech.fr)

AgroParisTech – INRAE MIA Paris-Saclay

EKINOCS research group

La visualisation de données

- Ensemble de techniques de communication d'un message sur des données par la visualisation

- Communiquer clairement sous forme visuelle des informations que l'on peut tirer de grandes quantités de données

https://files.ipbes.n ers_fr.pdf

Figure SPM 1 **Tendances mondiales de la capacité de la nature à maintenir ses contributions à une bonne qualité de vie, de 1970 à aujourd'hui, illustrant un déclin pour 14 des 18 catégories de contributions analysées.**

...

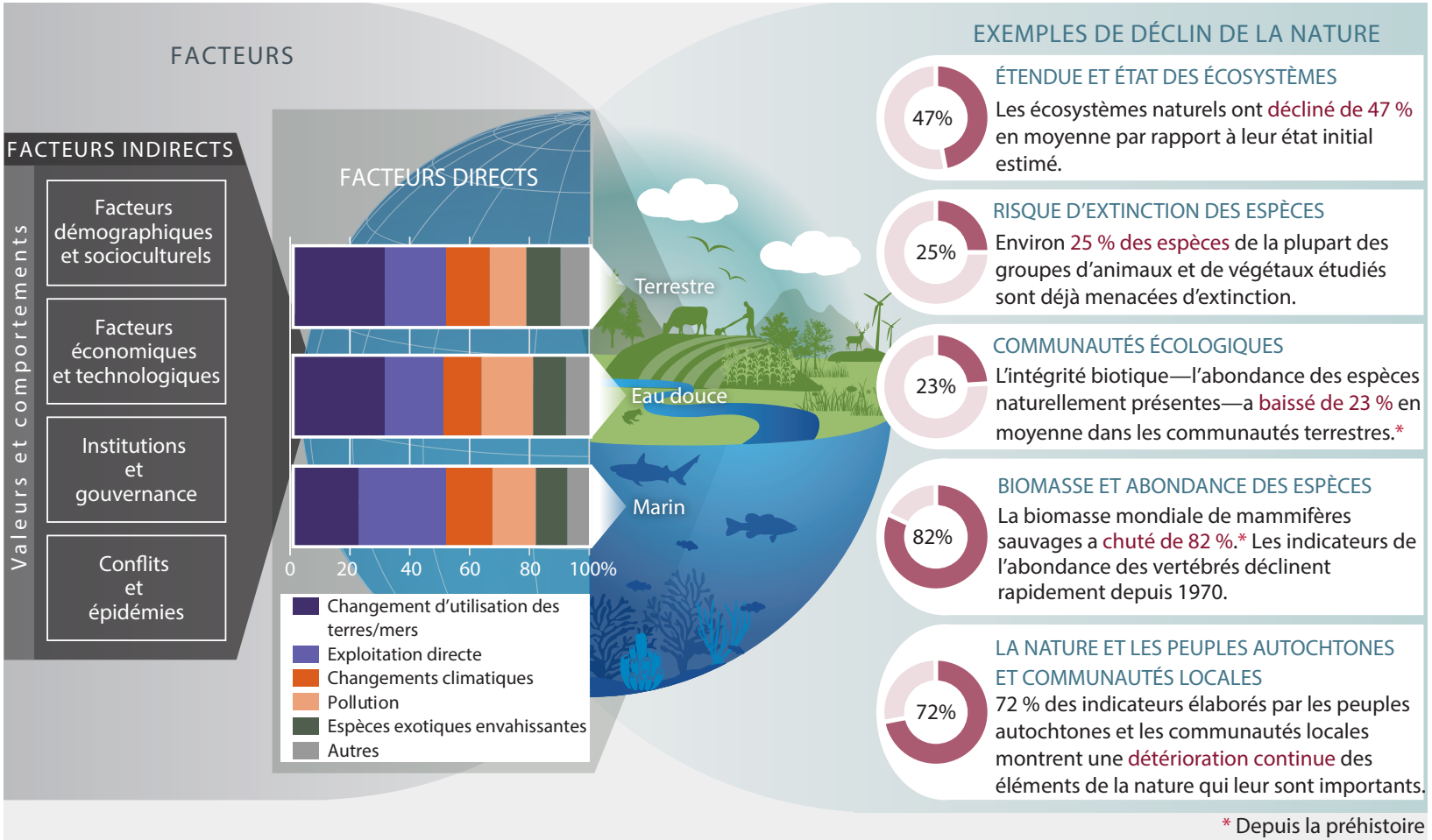


Figure SPM 2 **Exemples de déclins observés dans la nature au niveau mondial, soulignant le recul de la biodiversité provoqué par des facteurs de changement directs et indirects.**

...

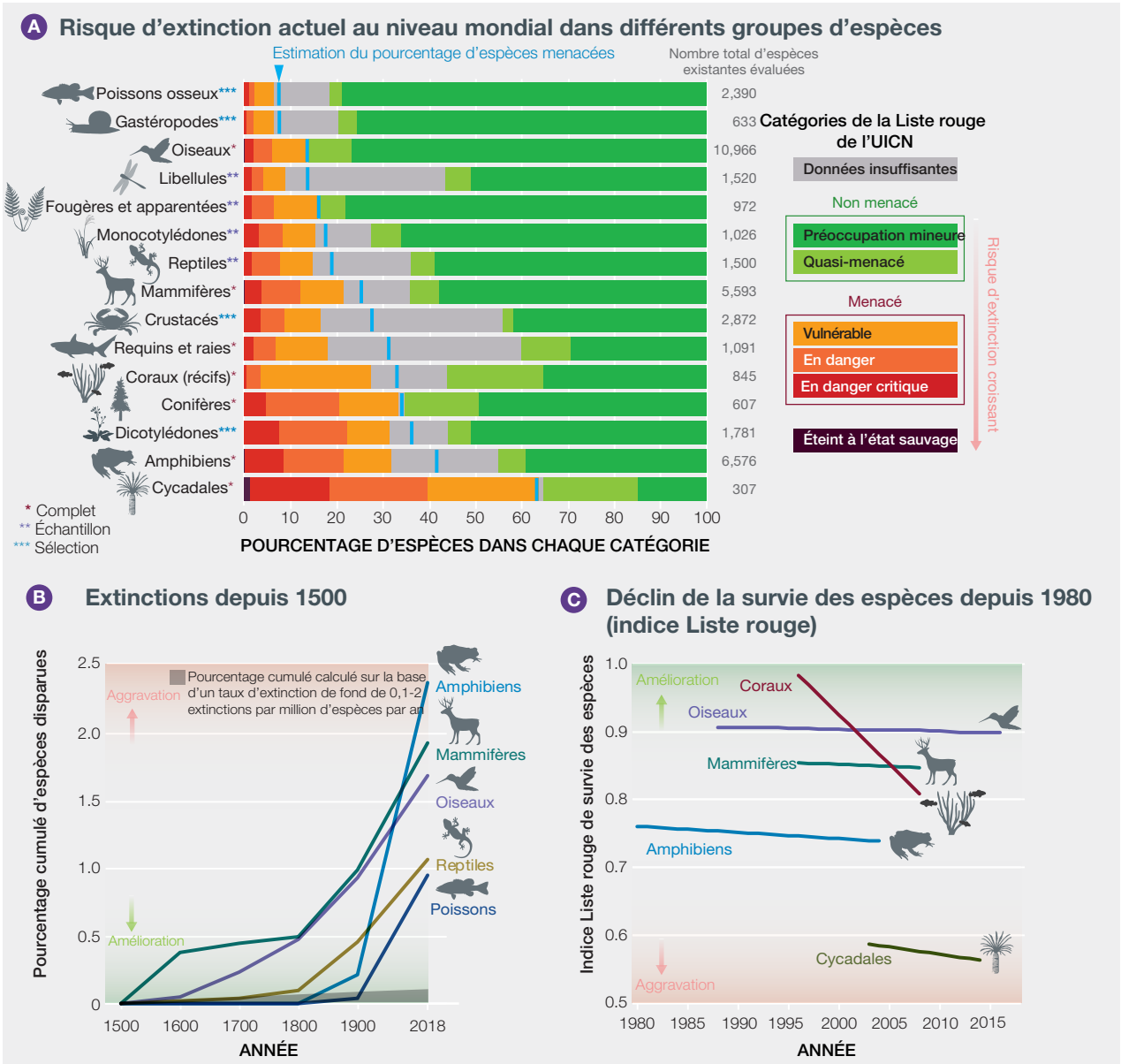


Figure SPM 3 **Une proportion importante des espèces évaluées est menacée d'extinction et les tendances générales s'aggravent, avec une forte augmentation des taux d'extinction au cours du siècle dernier.**

...

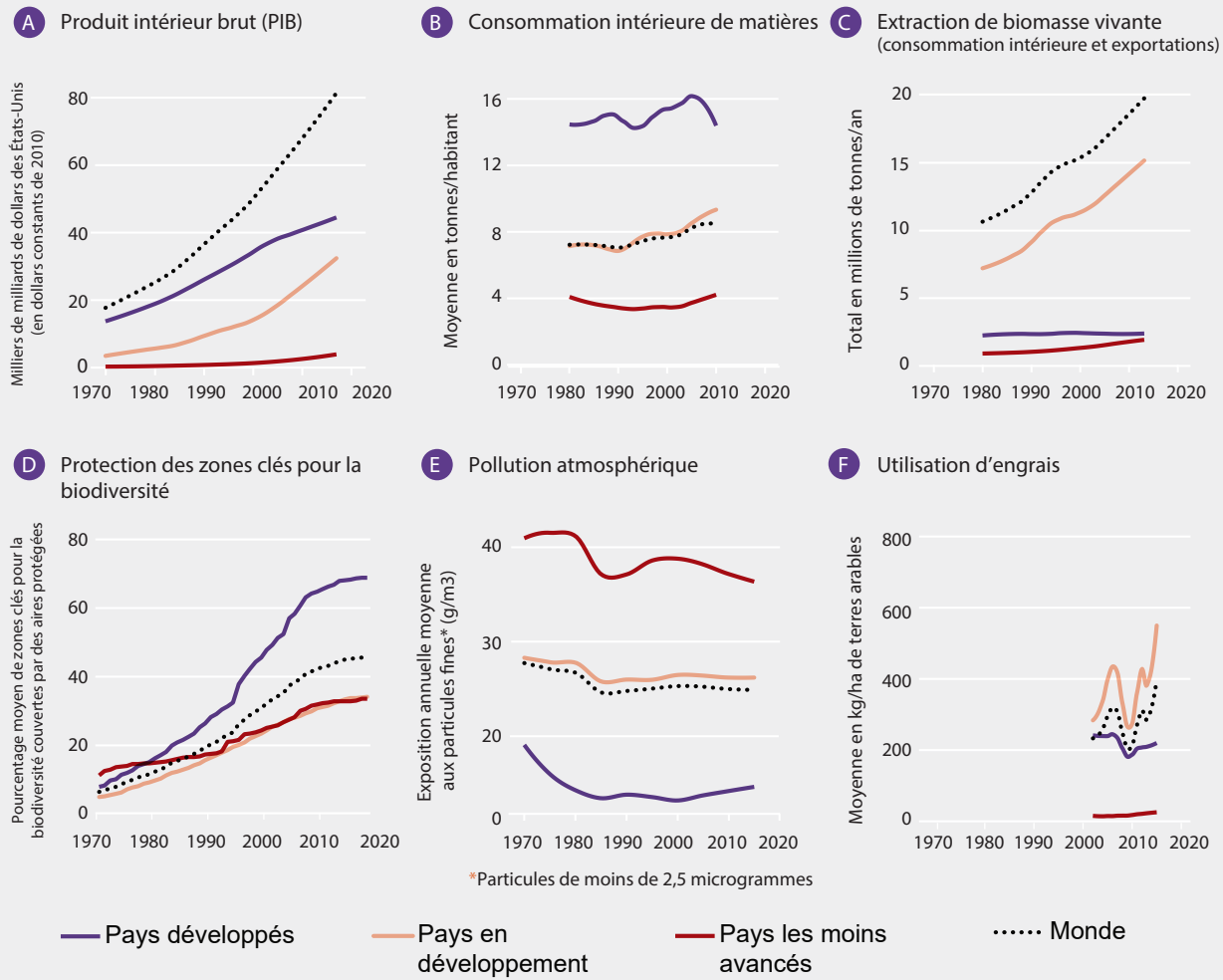


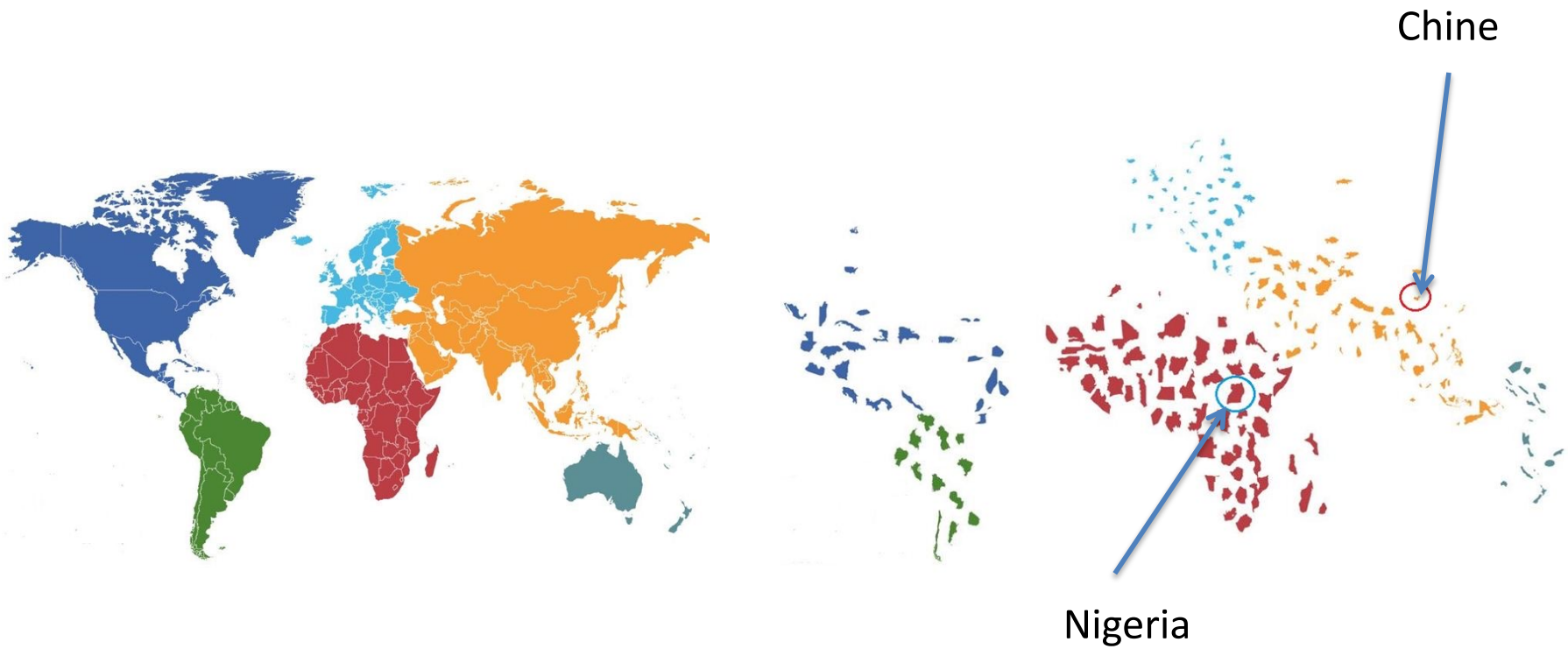
Figure SMP 4 **Trajectoires de développement depuis 1970 pour certains indicateurs clés des interactions entre l'homme et l'environnement, qui mettent en évidence une forte augmentation de l'échelle de la croissance économique mondiale et de ses impacts sur la nature, avec de grandes disparités entre les pays développés, en développement et moins avancés.**

...

- Le taux de natalité en Chine est de 8/1000
- Il est de 250/1000 au Nigéria

- Le taux de natalité en Chine est de 8/1000
- Il est de 250/1000 au Nigéria

- Le taux de natalité en Chine est de 8/1000
- Il est de 250/1000 au Nigéria



La visualisation : un outil utile

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

Raw Data from Anscombe's Quartet

La visualisation : un outil utile

Mêmes grandeurs statistiques pour les 4 paires de variables

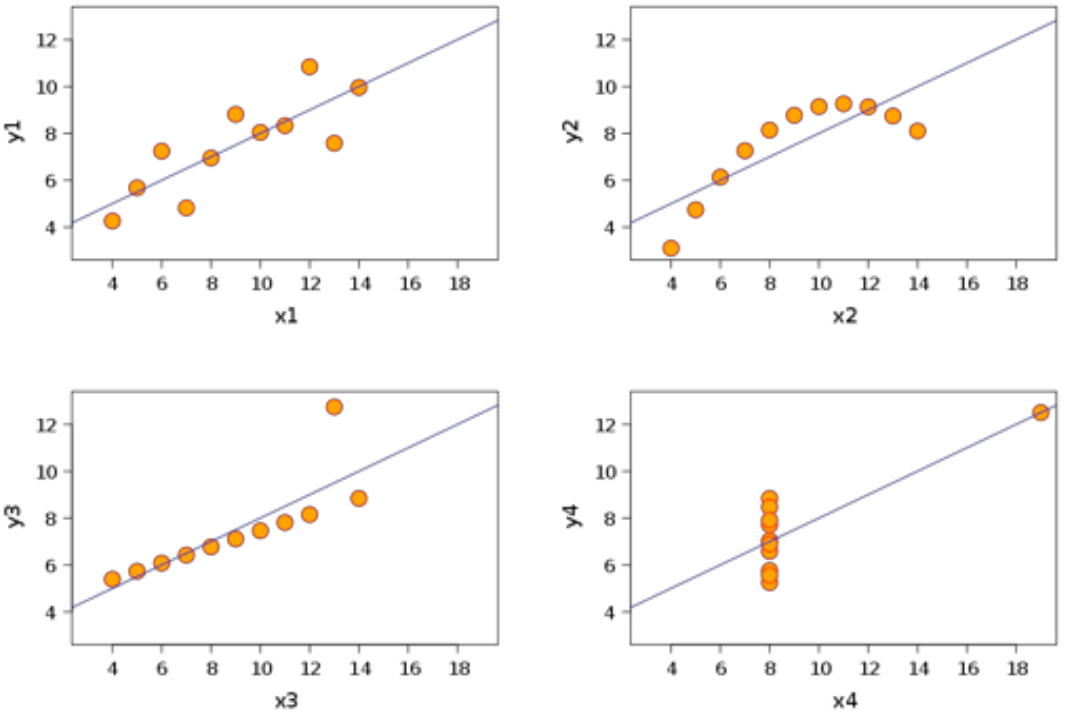
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Mean of x	9.0
Variance of x	11.0
Mean of y	7.5
Variance of y	4.12
Correlation between x and y	0.816
Linear regression line	$y = 3 + 0.5x$

La visualisation : un outil utile

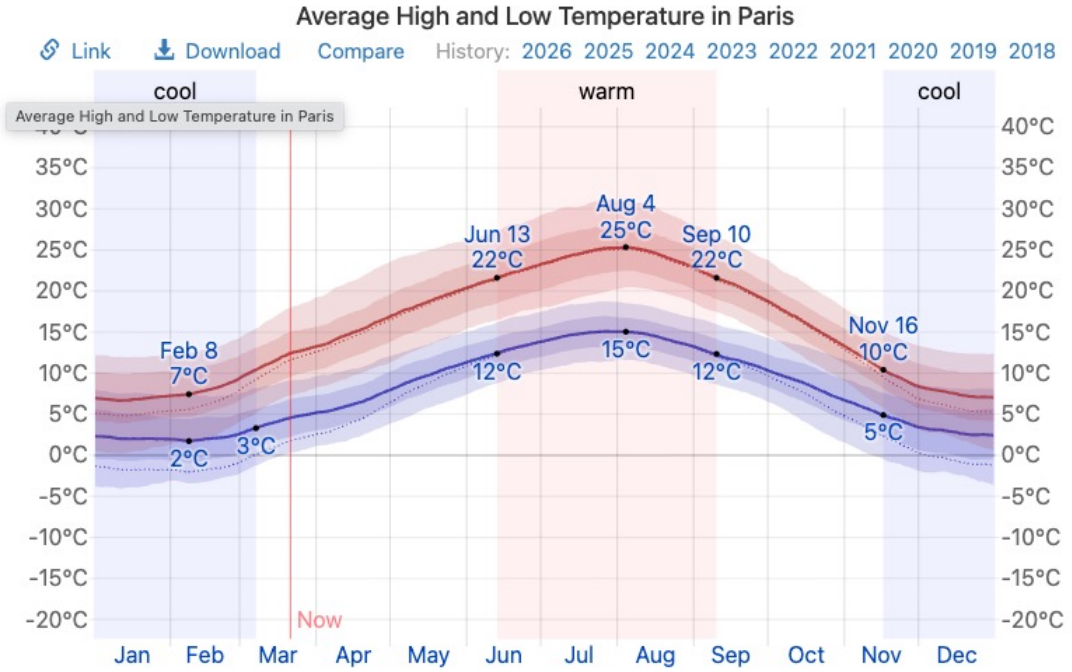
Une visualisation permet de distinguer des **comportements différents**

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



Illustration

Températures à Paris (sur la période 2018 – 2026)



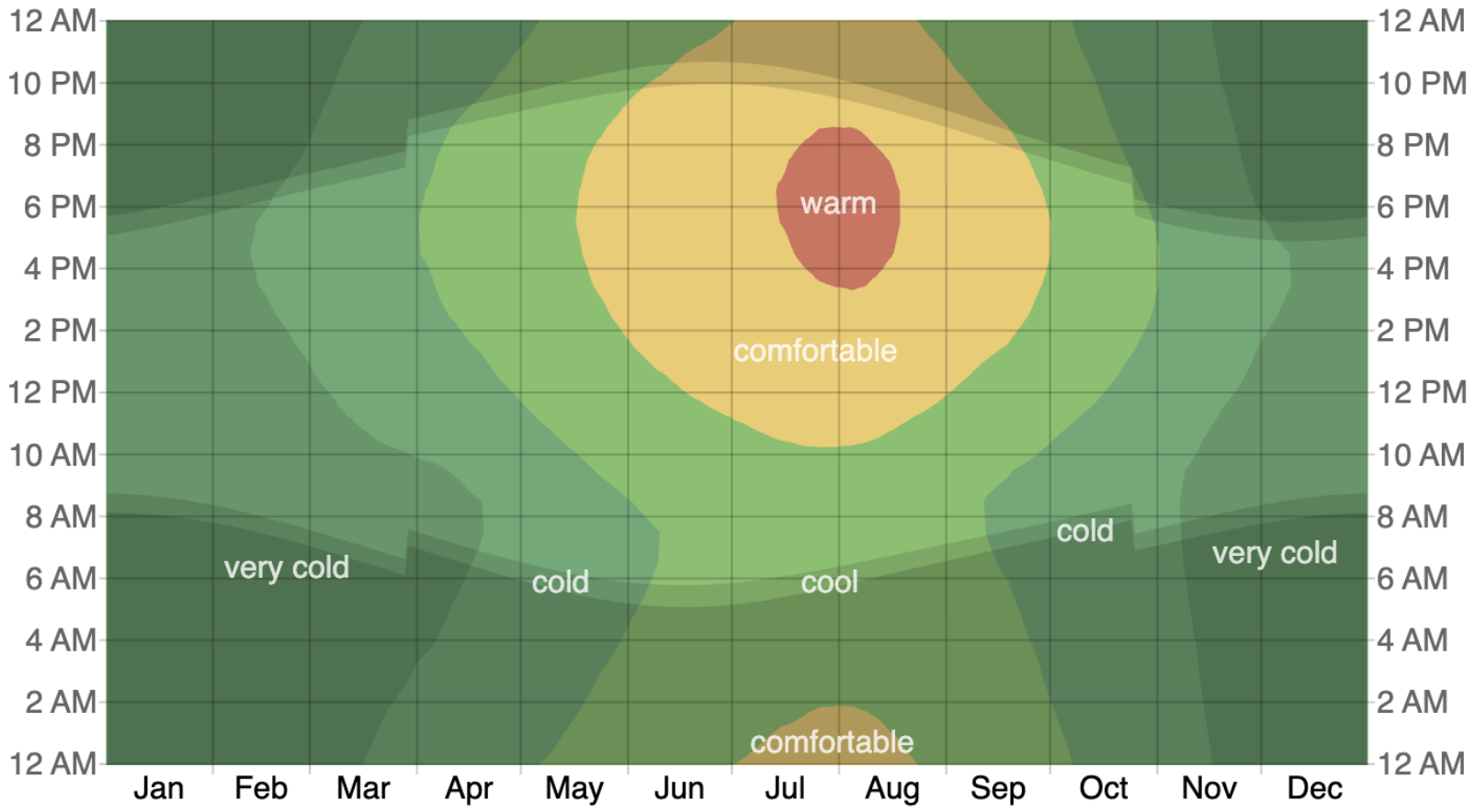
The daily average high (red line) and low (blue line) temperature, with 25th to 75th and 10th to 90th percentile bands. The thin dotted lines are the corresponding average perceived temperatures.

Average	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
High	7°C	8°C	12°C	15°C	19°C	22°C	25°C	24°C	21°C	16°C	11°C	7°C
Temp.	4°C	5°C	8°C	11°C	14°C	18°C	20°C	20°C	16°C	12°C	8°C	5°C
Low	2°C	2°C	4°C	6°C	10°C	13°C	15°C	14°C	12°C	8°C	5°C	3°C

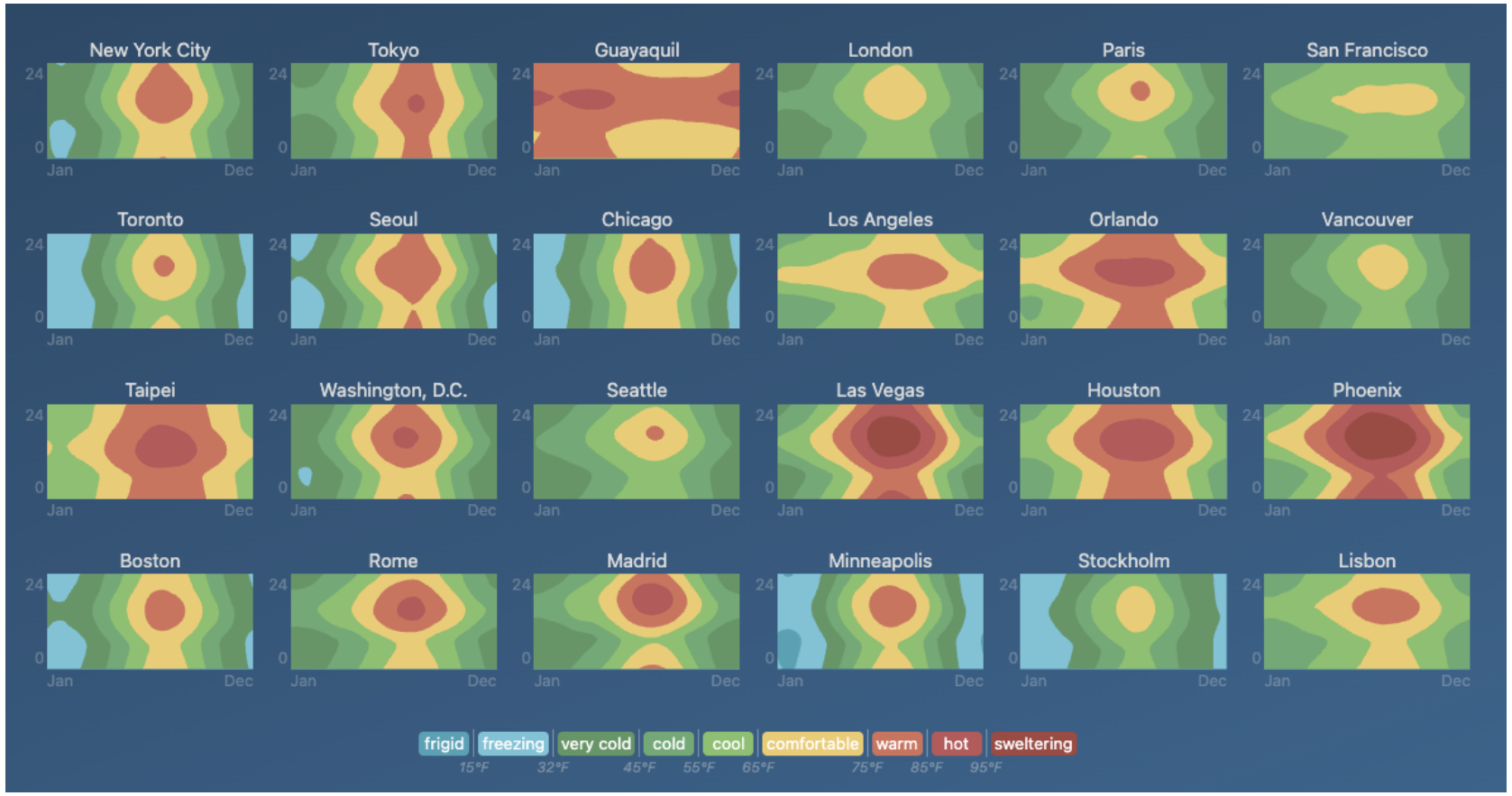
Weatherspark (https://weatherspark.com/y/47913/Average-Weather-in-Paris-France-Year-Round)

Illustration (une autre représentation)

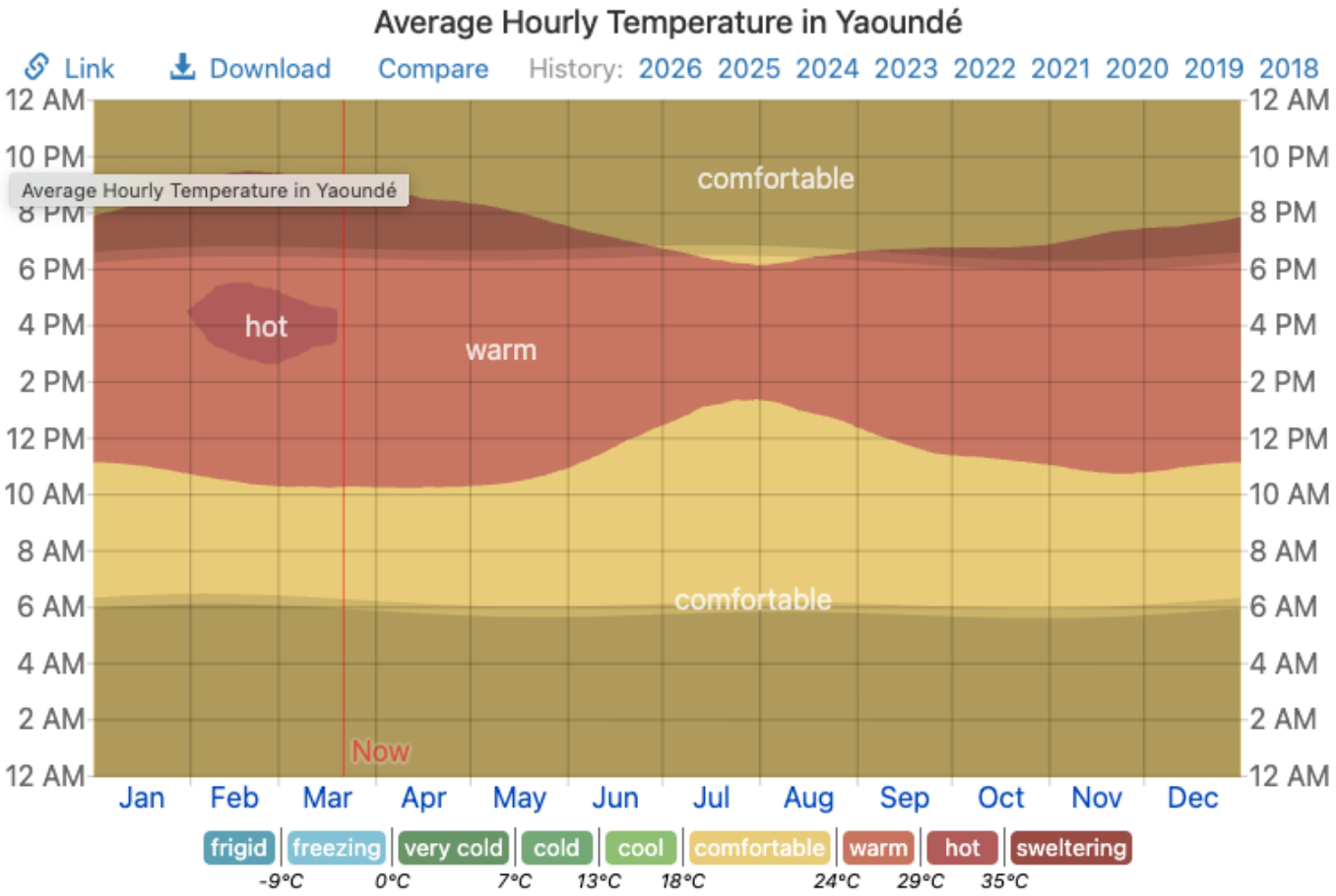
Températures à Paris (sur la période 2018 – 2026)



Weatherspark (https://weatherspark.com/y/47913/Average-Weather-in-Paris-France-Year-Round)



Weatherspark (https://weatherspark.com/y/47913/Average-Weather-in-Paris-France-Year-Round)



The average hourly temperature, color coded into bands. The shaded overlays indicate night and civil twilight.

Weatherspark (https://weatherspark.com/y/47913/Average-Weather-in-Paris-France-Year-Round)

Cartes

FIGURE 2-5 Distance to McDonald's *(2010) by Stephen Von Worley, http://datafl.ws/24y*

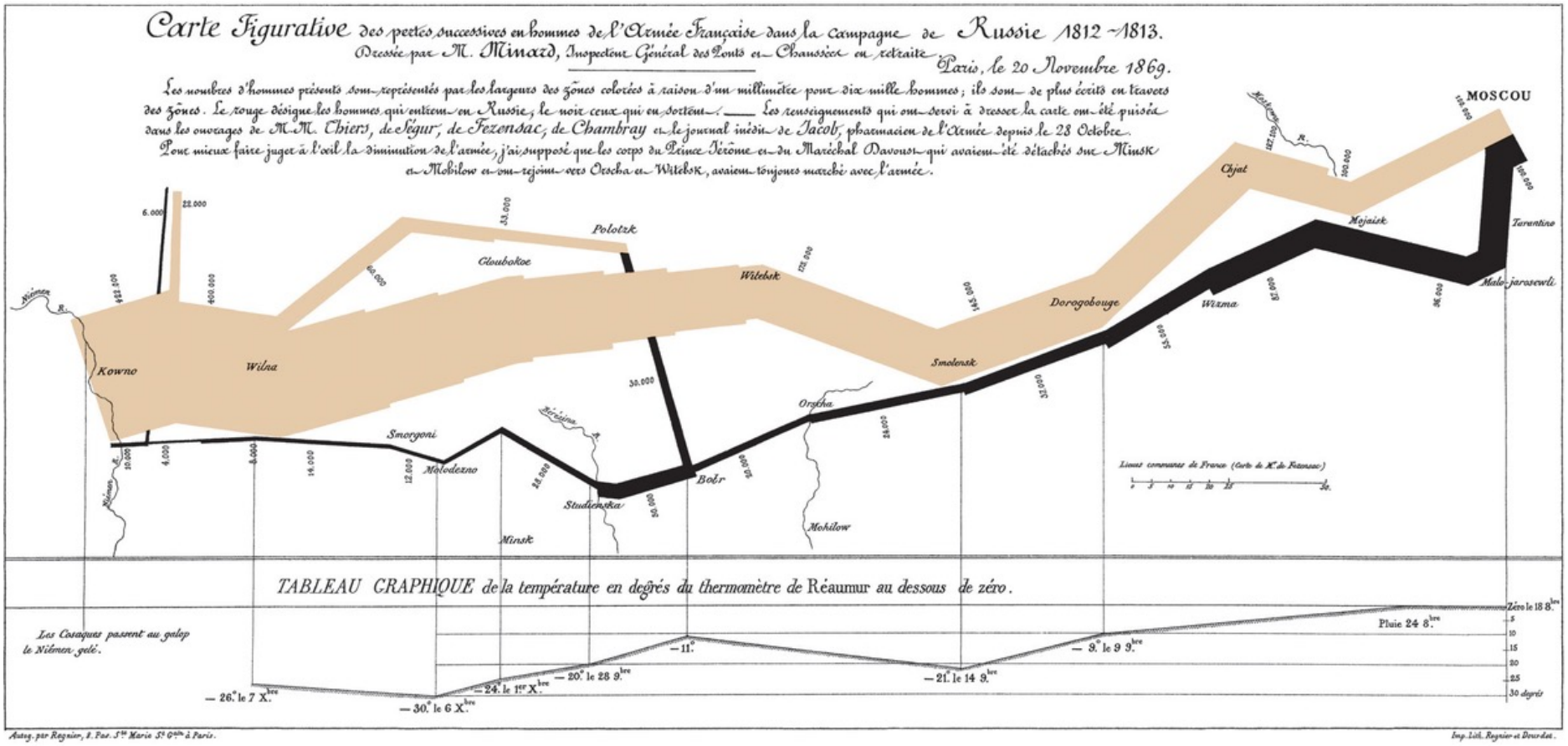
Distance au plus proche McDonald.

Plus c'est lumineux, moins il faut de temps pour accéder à un big mac.

Intensité lumineuse la nuit : USA

[Nathan YAU (2013) « Data points. Visualization that means something », p. 49-50]

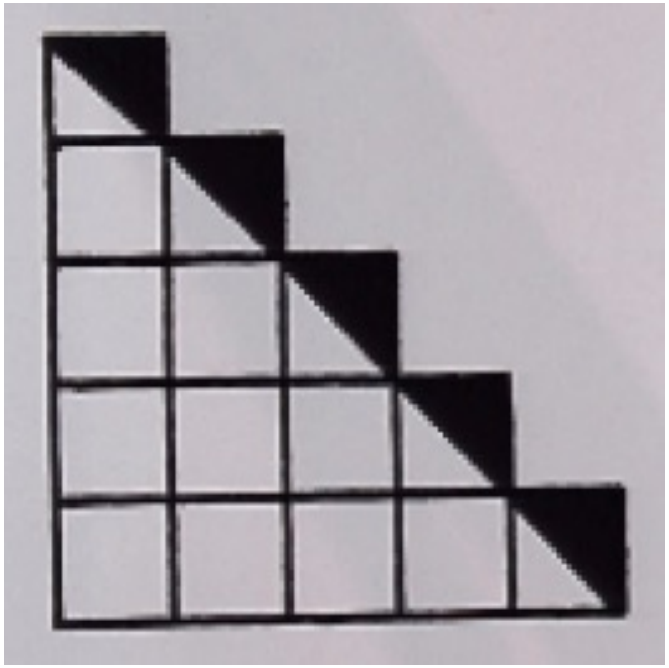
La campagne de Russie (1812-1813)



Spatial – temporel – température – taille armée

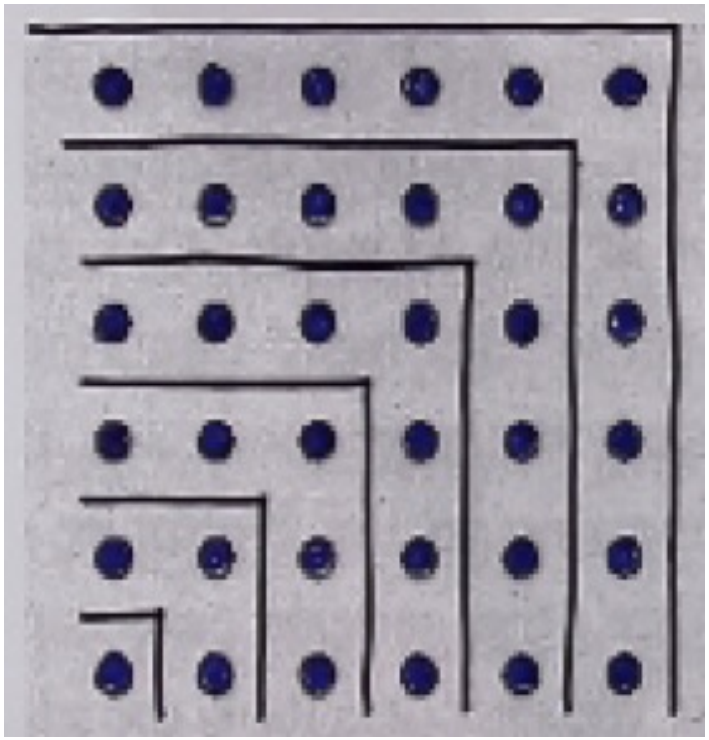
« Raisonnement » visuel

$$1 + 2 + 3 + \ldots + n \overset{?}{=} \frac{n^2}{2} + \frac{n}{2}$$



« Raisonnement » visuel

$$1 + 3 + 5 + \ldots + (2n - 1) \overset{?}{=} n^2$$



Plan

1. Puissance (et limites) de l'appareil visuel
2. Qu'est-ce qu'une représentation
3. Les primitives visuelles
4. Types de données et types de visualisation
   - Démarche
   - Données catégorielles
   - Séries temporelles
   - Données multi-variées
5. Outils et librairies

Formidable puissance et limites aussi

de notre **appareil visuel**

Pourquoi la visualisation

- Notre **principal canal sensoriel**
- Nous sommes **très bons pour reconnaître** (certains) **patterns visuels**
- **Très efficace** pour comprendre, expliquer, raisonner et convaincre

Processus préattentifs

Processus préattentifs

- Certaines inférences semblent être réalisées sans que l'**attention** ne soit mobilisée
- Généralement en moins de 200 à 250ms (une saccade visuelle prend 200ms)
- Semble se réaliser en parallèle par un système visuel bas-niveau

Combien de 3 ?

128176875613897654698450698560498282676298098584582245098564589450984509809435859091030209905959595772564675050678904567884578980982167765487636490856091294968 6

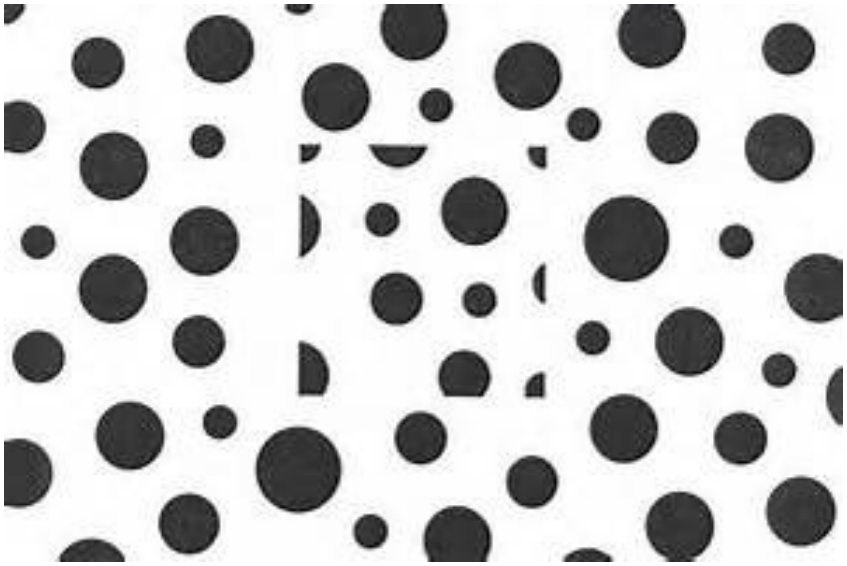
Nécessite l'attention

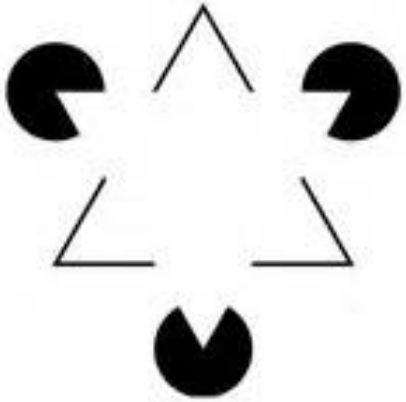
Combien de 3 ?

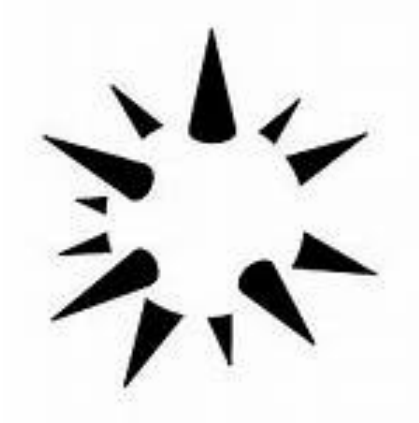
1281768756138976546984506985604982826762
9809858458224509856458945098450980943585
9091030209905959595772564675050678904567
8845789809821677654876364908560912949686

Ne nécessite **pas** l'attention

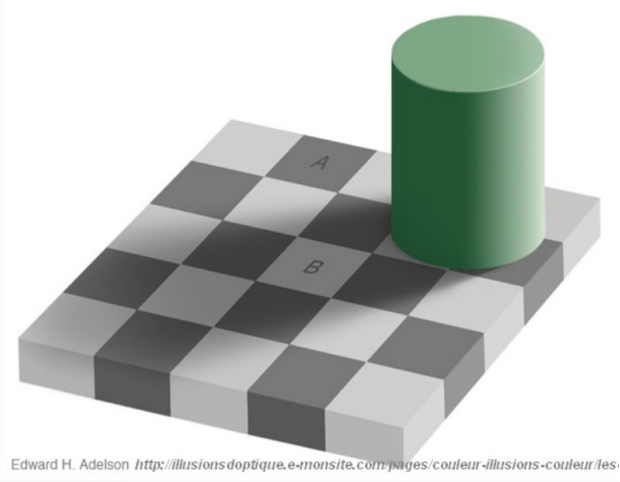
Voir = compléter / interpréter







Les « primitives visuelles » [Bertin, 1967]

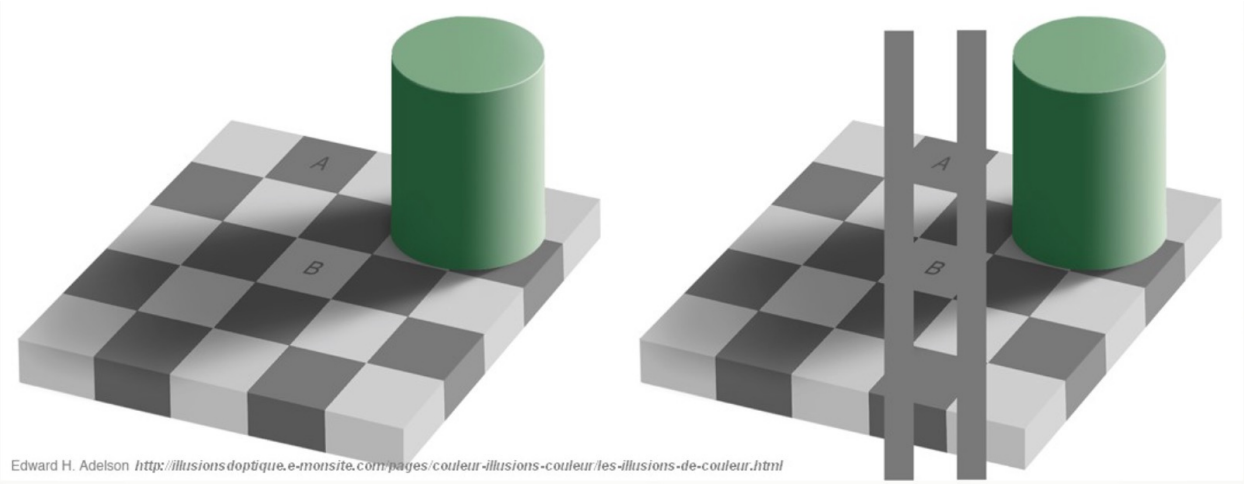


[Stephen FEW (2009)
« Now you see it », p. 48]

La case **B** paraît nettement
plus claire que la case **A**

Nous percevons la couleur en **contexte**

Les « primitives visuelles » [Bertin, 1967]



[Stephen FEW (2009)
« Now you see it », p. 48]

La case B paraît nettement plus claire que la case A

Pourtant …

Nous percevons la couleur en **contexte**

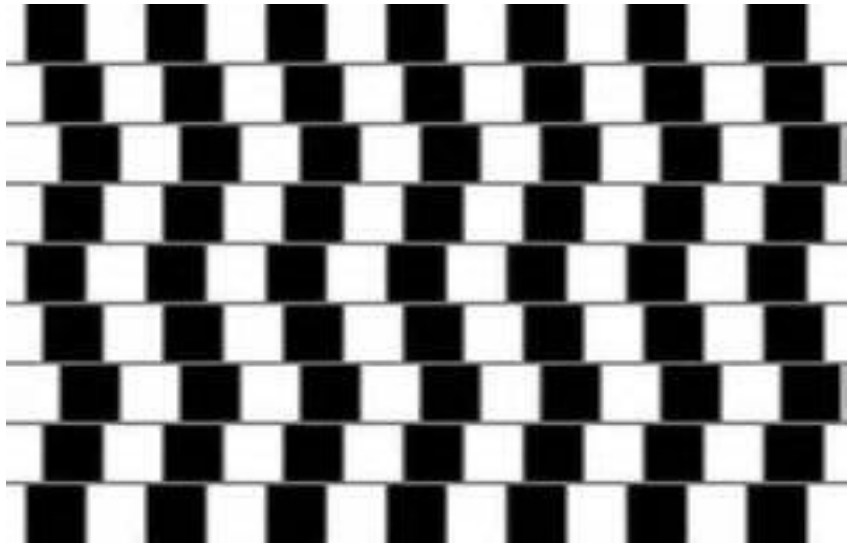
Nos sens sont limités

[Stephen FEW (2009)
« Now you see it », p. 35]

Quand on montre alternativement les deux photos, la plupart des gens ne voient pas de différence

http://www.bing.com/images/search?view=detailV2&ccid=duaEBXN%2b&id=E2EDFDB4D9270FD2A804EC8BE45AA554A6BB62DF&thid=OIP.duaEBXN-kIftUuyJw5krygHaFj&q=blindness+to+change+tourists&simid=608047137953351566&selectedIndex=0&ajaxhist=0

Interpretative vision and its illusions



Interprétation(S)

Chacun perçoit selon sa perspective

Plan

1. Puissance (et limites) de l'appareil visuel
2. Qu'est-ce qu'une représentation
3. Les primitives visuelles
4. Types de données et types de visualisation
   - Démarche
   - Données catégorielles
   - Séries temporelles
   - Données multi-variées
5. Outils et librairies

Qu’est-ce qu’une représentation ?

Représentation

- Représentation = un système de signes
  qui tient lieu d'autre chose

- Exemple : le nombre trente-quatre

34	100010	XXXIV
Décimal	Binaire	Romain

Représentation et raisonnement

- Décimal : puissances de 10

  (arbitraire, vient de nos dix doigts)

- Binaire : puissances de 2

- Romain : très difficile de faire des additions

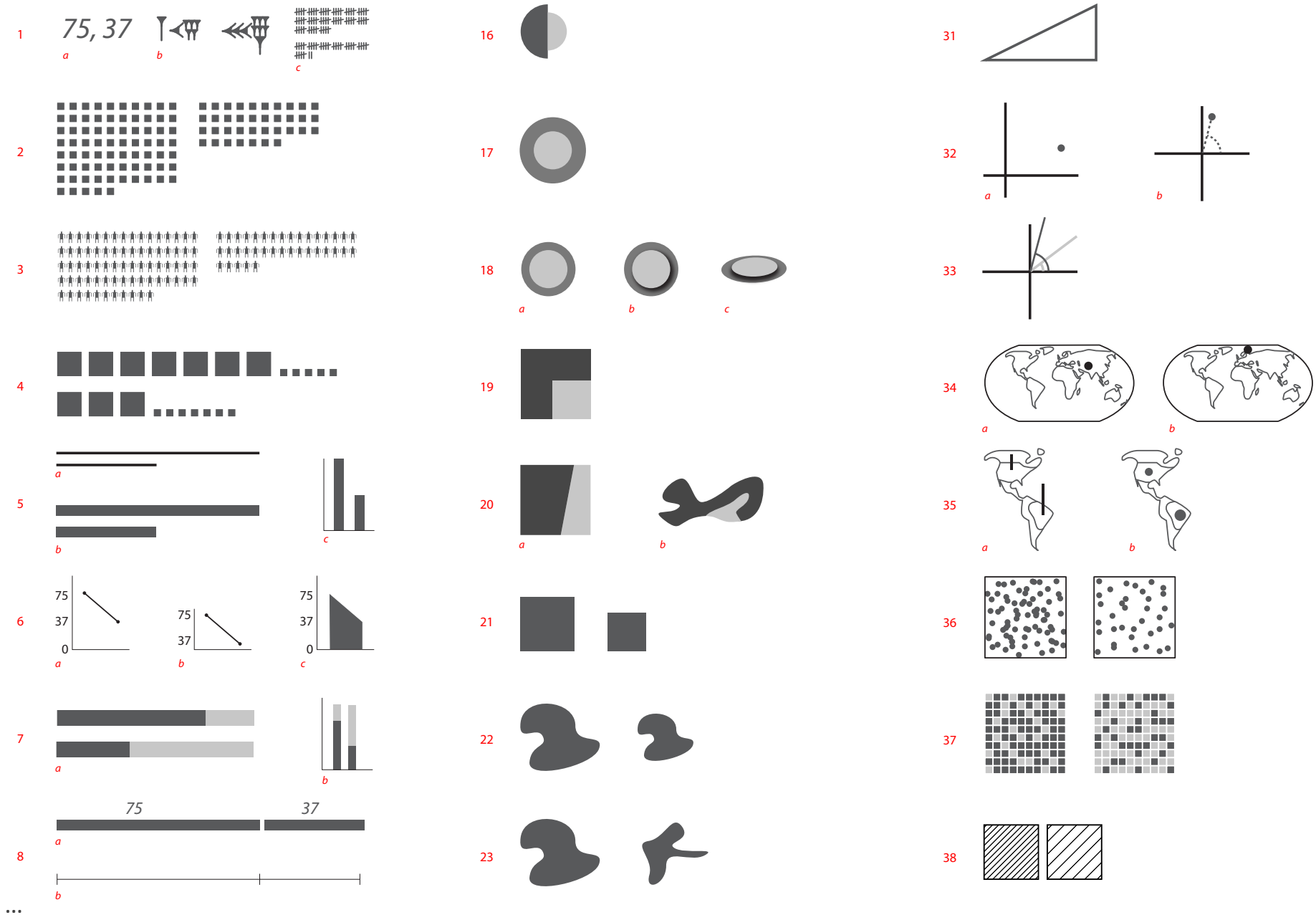
  et des multiplications

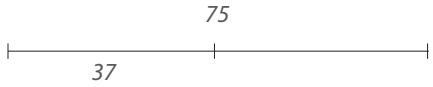
- Chaque type de données est mieux représenté par certains types de représentations

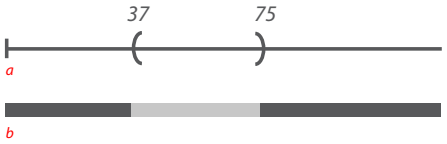
- Mais il y a toujours plusieurs manières de représenter chaque type de données

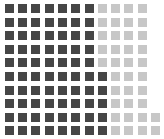
Énormément de manières de représenter quelque chose

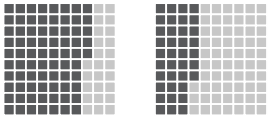
- E.g. comment représenter les quantités 75 et 37 ?


1
75, 37
a
b
c
2
3
4
5
a
b
c
6
75
37
0
a
75
37
b
75
37
0
c
7
a
b
8
75
37
a
b
...
16
17
18
a
b
c
19
20
a
b
21
22
23
31
32
a
b
33
34
a
b
35
a
b
36
37
38

9 

10 

11 

12 

13 

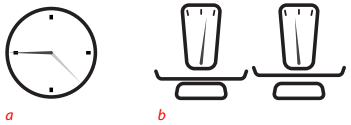
14 

15 

24 

25 

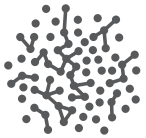
26 

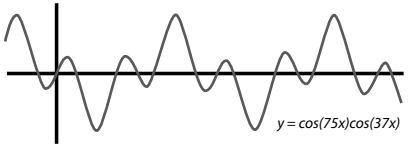
27 

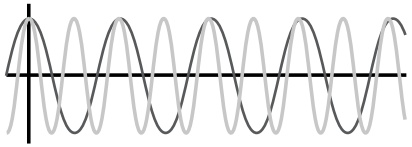
28 

29 

30 

39 

40 

41 

42 *animation: two pulses with 75 and 37 beats per minute*

43 *animation: two points rotating with 75 and 37 revolutions per minute*

44 *two sounds, 75hz and 37hz*

45 

...

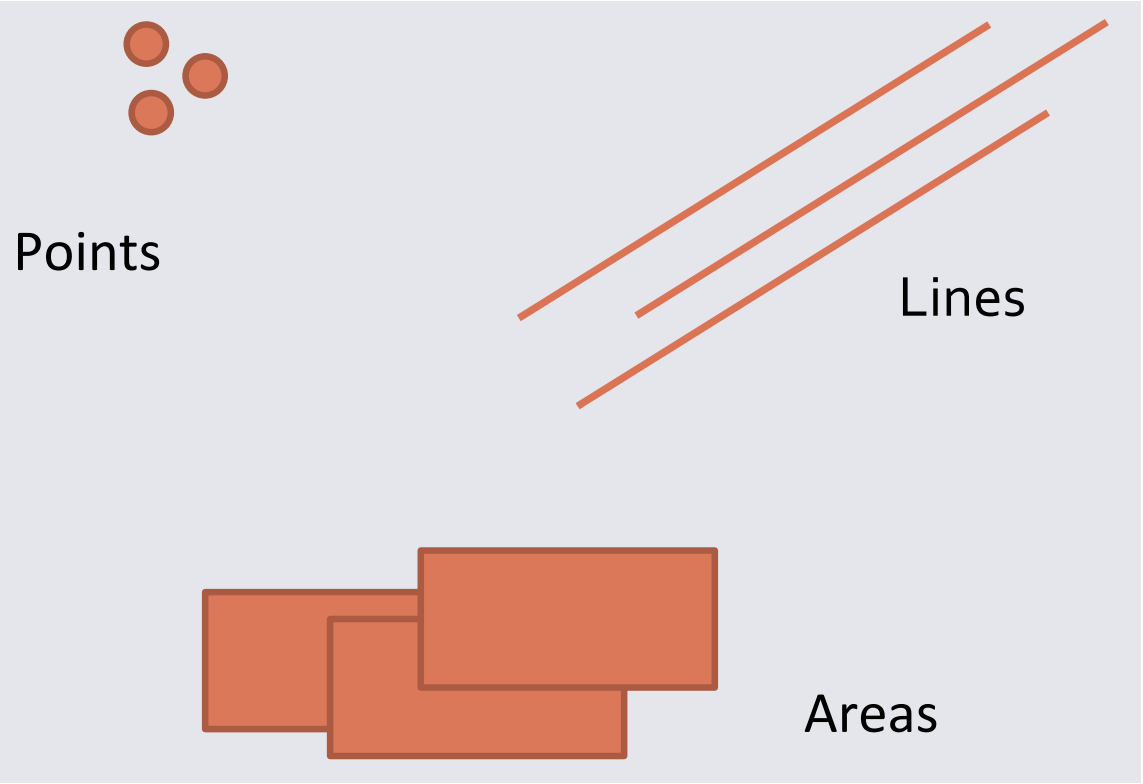
Plan

1. Puissance (et limites) de l'appareil visuel
2. Qu'est-ce qu'une représentation
3. Les primitives visuelles
4. Types de données et types de visualisation
   - Démarche
   - Données catégorielles
   - Séries temporelles
   - Données multi-variées
5. Outils et librairies

Les primitives de visualisation

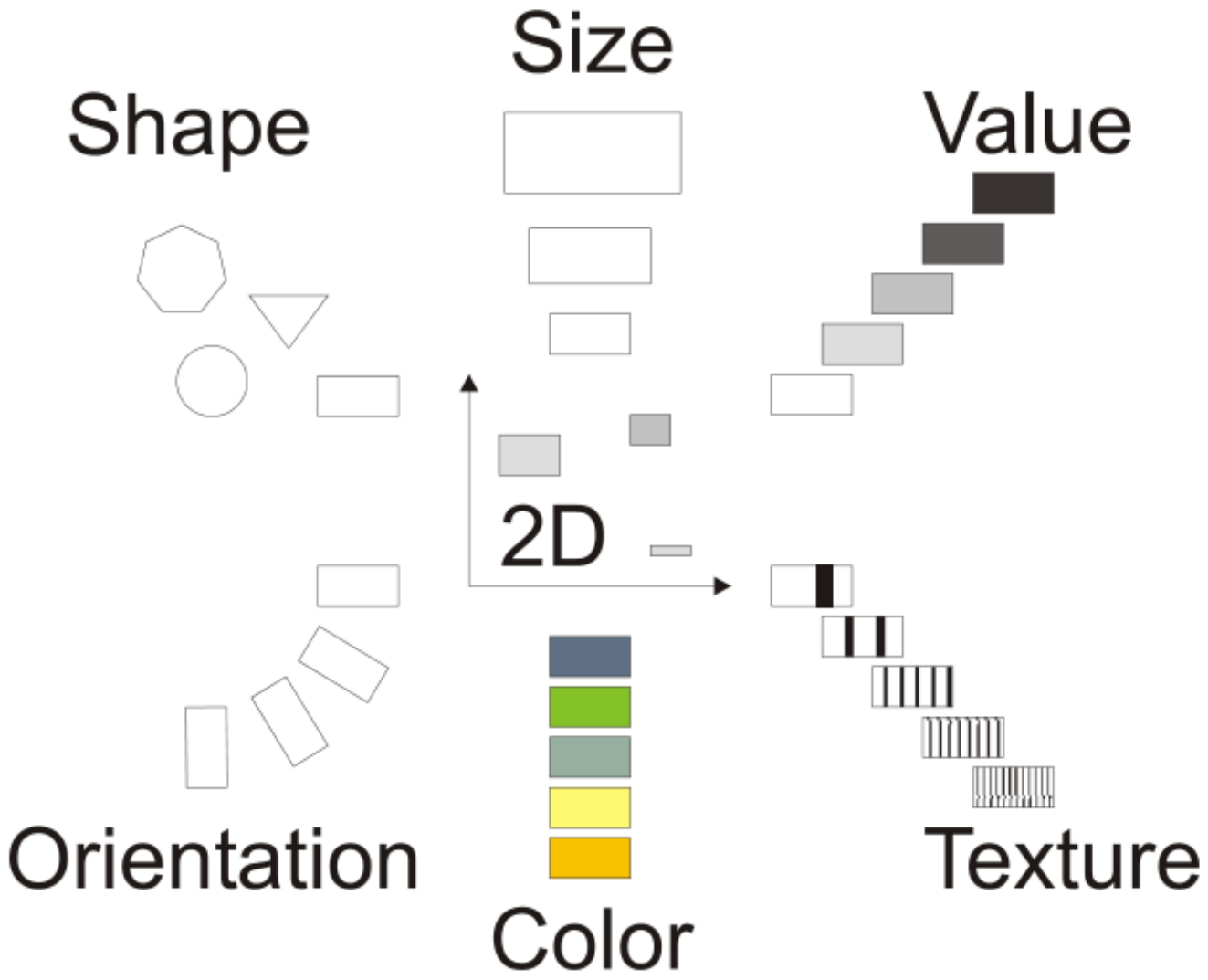
Les « marques »

Marques



Les variables visuelles applicables aux « marques »

Variables visuelles



Indices pour la vision

[Nathan YAU (2013)
« Data points. Visualization that means something », p. 95]

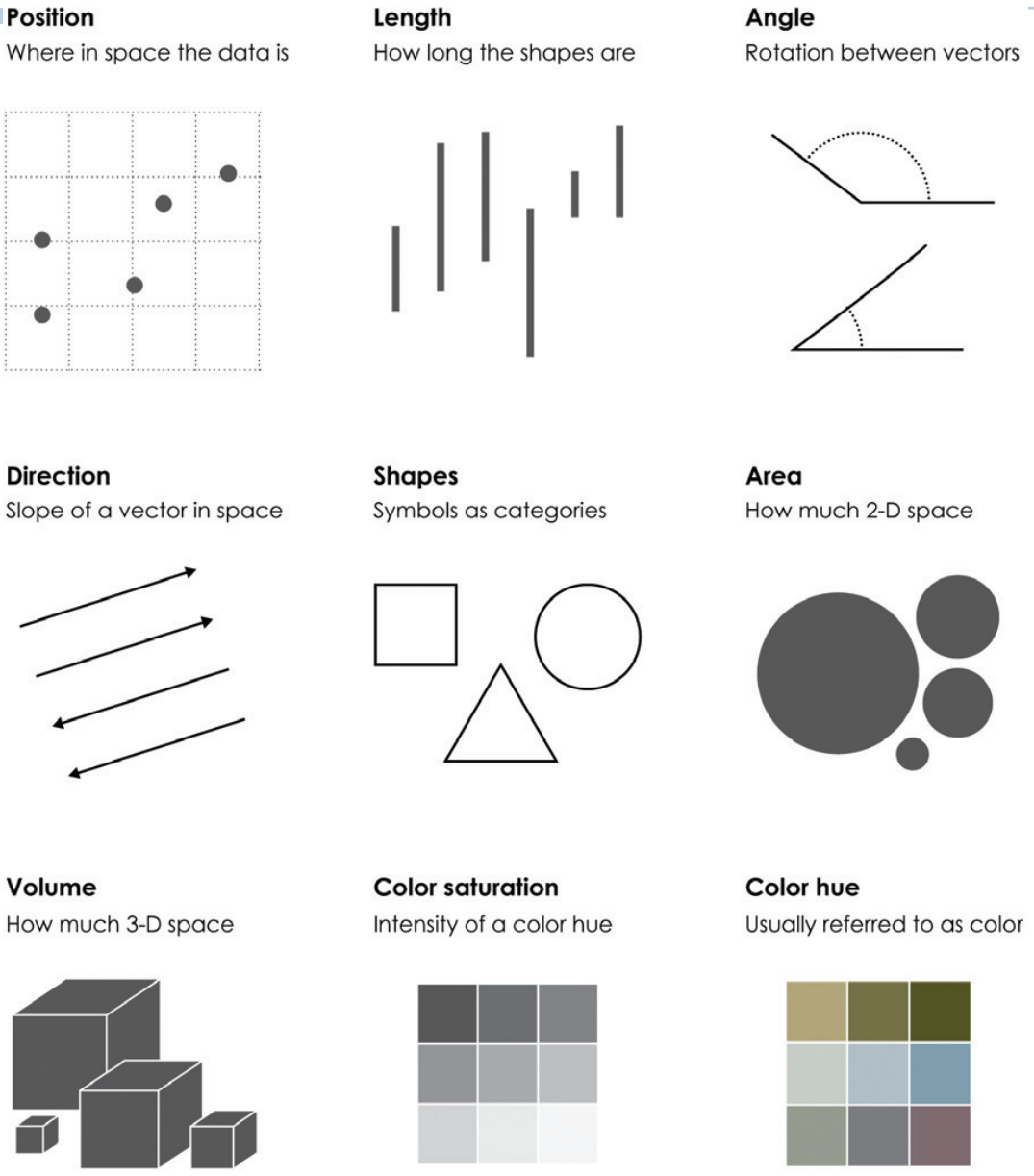
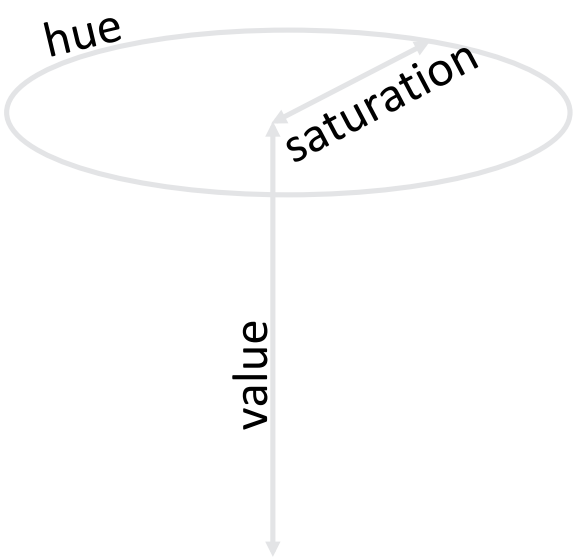


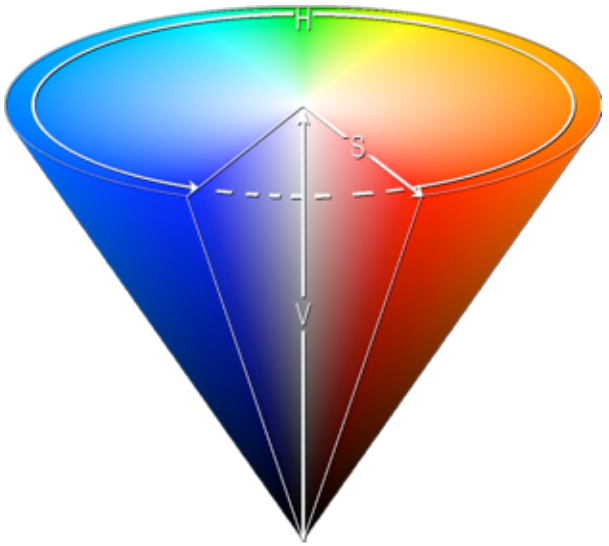
FIGURE 3-3 *Visual cues*

Variable visuelle : la couleur

Modèle HSV

- **Hue** : ce qu'évoque usuellement la « couleur »
- **Saturation** : intensité
- **Valeur** : clair / sombre





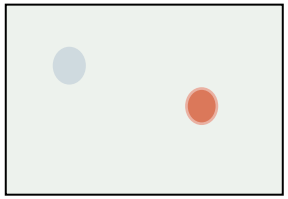
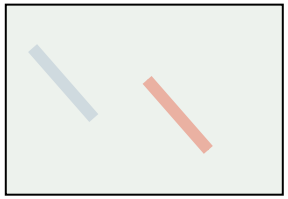
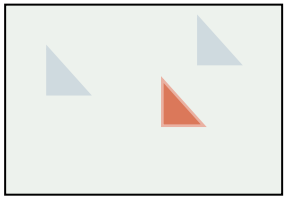
Caractéristiques des variables visuelles

- Sélective
  - Est-ce qu'un changement dans cette variable permet de **sélectionner un individu** dans un groupe ?
- Associative
  - Est-ce qu'un changement dans cette variable permet de **percevoir un groupe** ?
- Quantitative
  - **Peut-on lire un nombre** à partir d'un changement dans cette variable ?
- Ordre
  - Est-ce que des changements dans cette variable peuvent être **perçus comme ordonnés** ?
- Longueur
  - **Combien de distinctions sont possibles** si l'on change cette variable ?

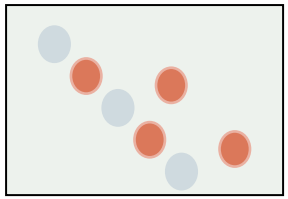
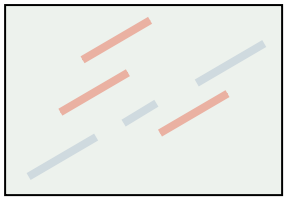
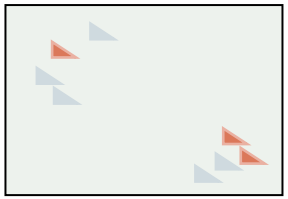
Variable visuelle : la couleur

- Illustration

✓ selective

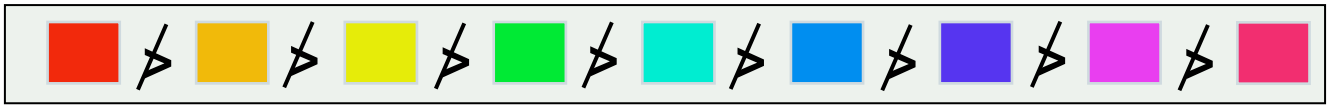




✓ associative

≠ quantitative

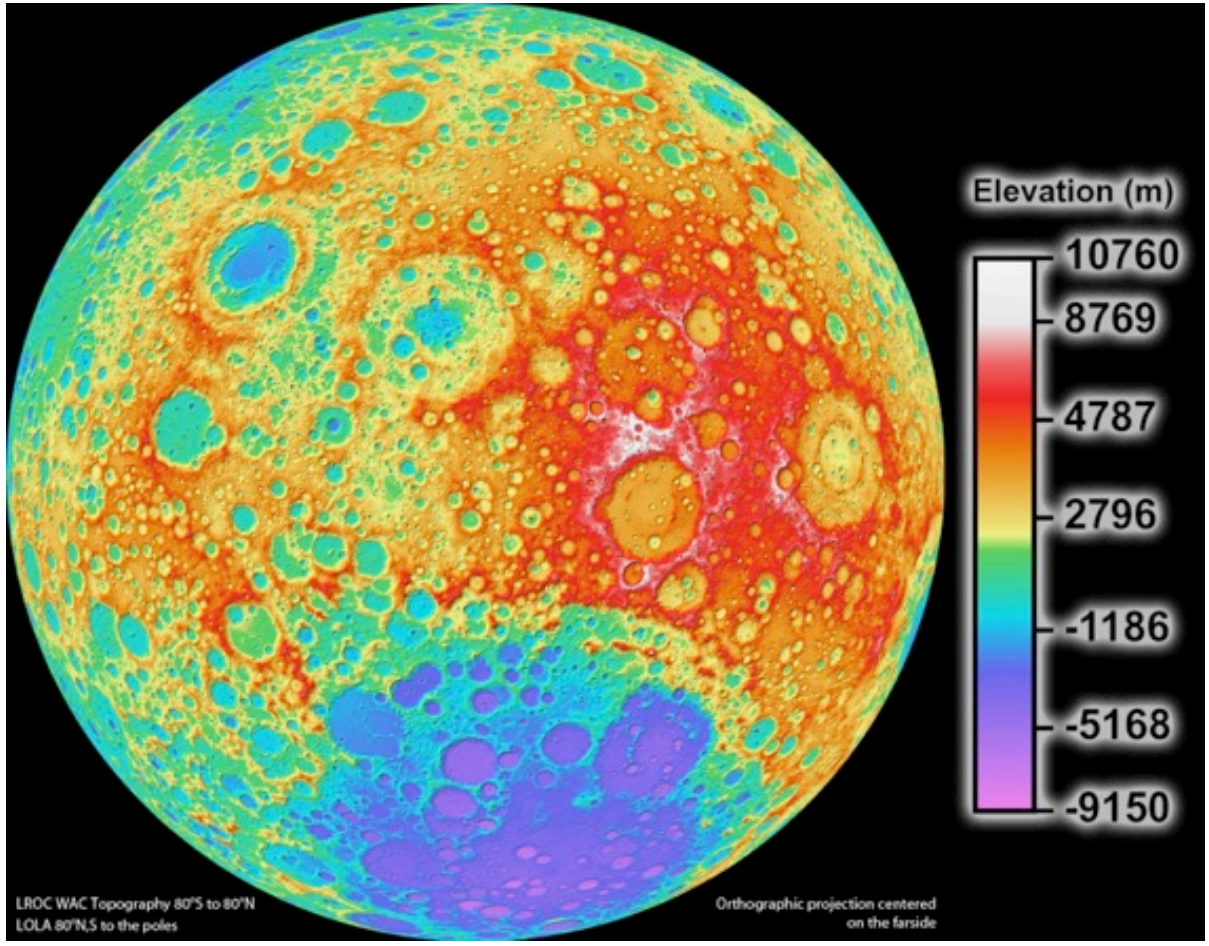
≠ order



✓ Length

Échelle de couleur

- Illustration



Elementary Graphical Perception tasks (Cleveland & McGill, 1980s)

- **Expériences contrôlées** sur des sujets humains pour mesurer à quel point il est possible de juger des changements dans les variables visuelles
- Focalisé sur des **informations quantitatives**
- **Variables** utilisées
  - Angle
  - Surface
  - couleur : hue, saturation, valeur
  - Longueur
  - Position
  - Pente
  - volume

Caractéristiques des indices pour la vision

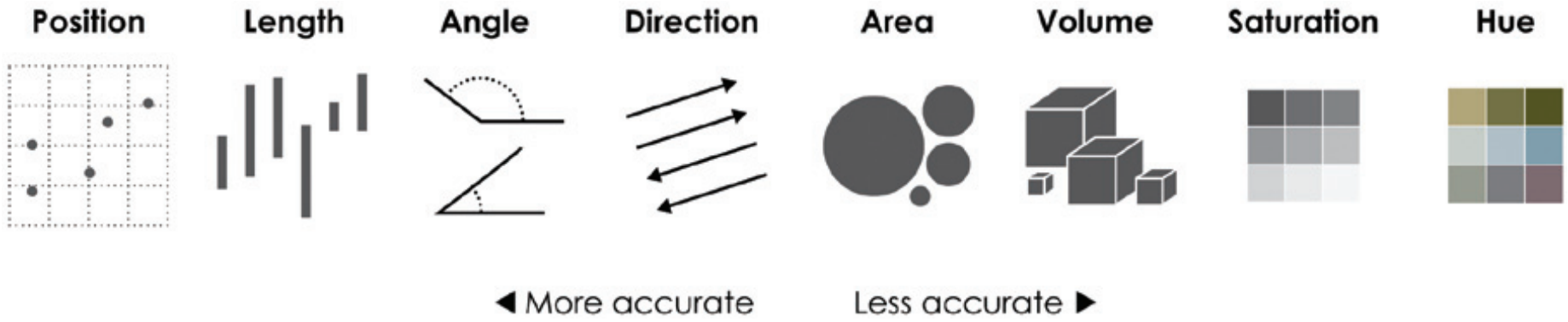
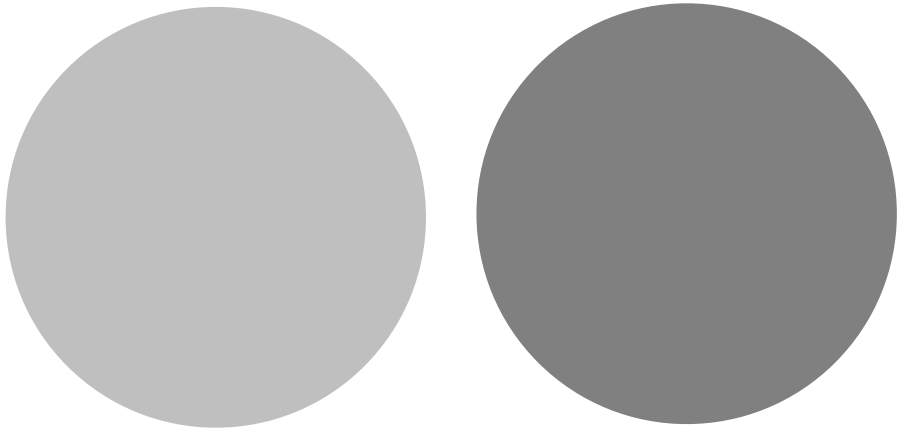


FIGURE 3-12 *Visual cues ranked by Cleveland and McGill*

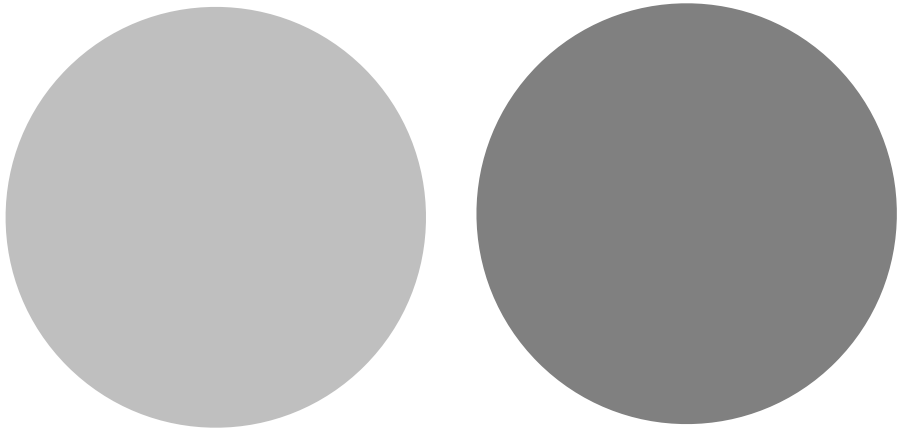
[Nathan YAU (2013)
« Data points. Visualization that means something », p. 104]

- What percentage in
  value is the right from the left
  (=100%) ?



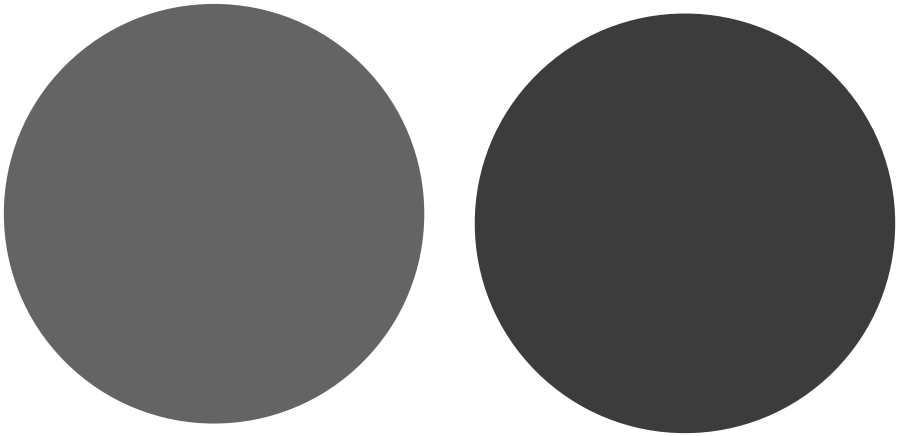
...

- What percentage in value is the right from the left (=100%) ?

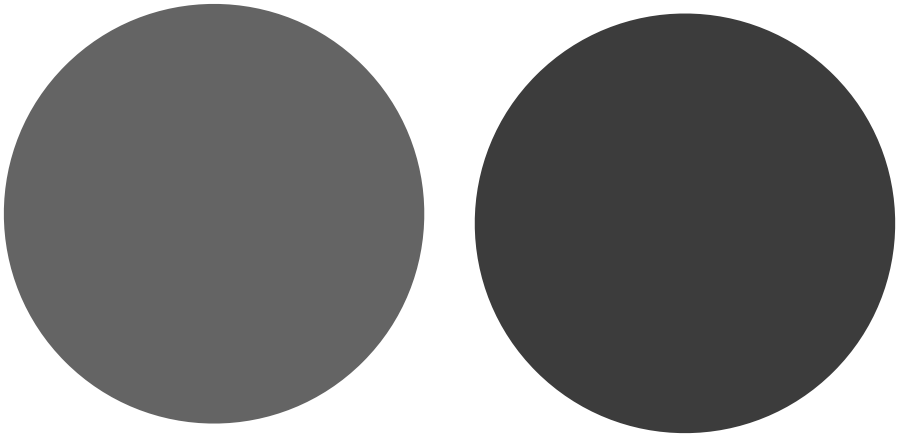


66 %

- What percentage in
  value is the right from the left
  (=100%) ?

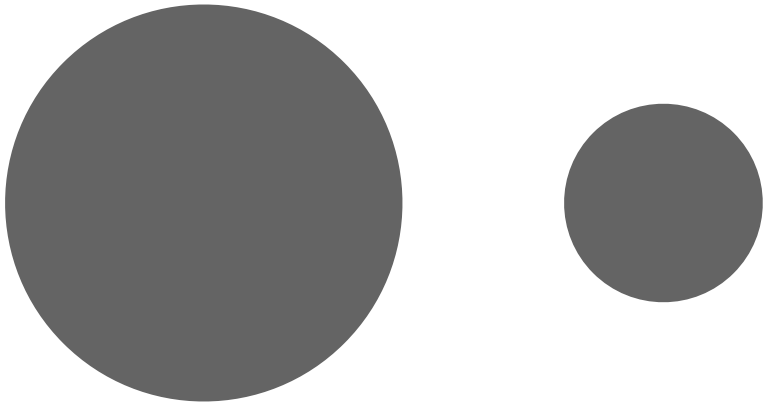


- What percentage in value is the right from the left (=100%) ?

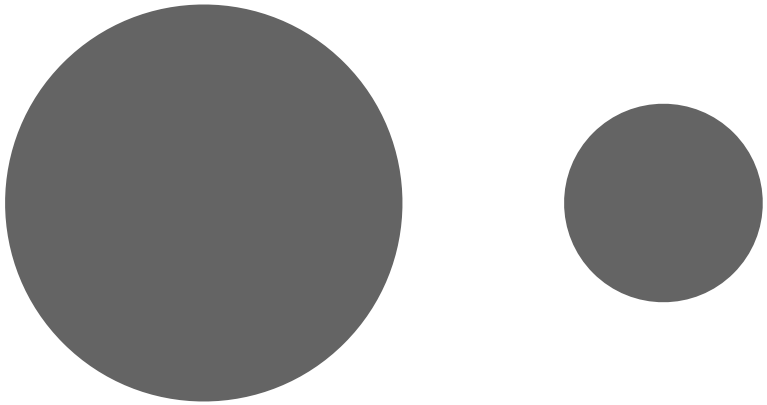


60 %

- What percentage in size is the right from the left (=100%)?

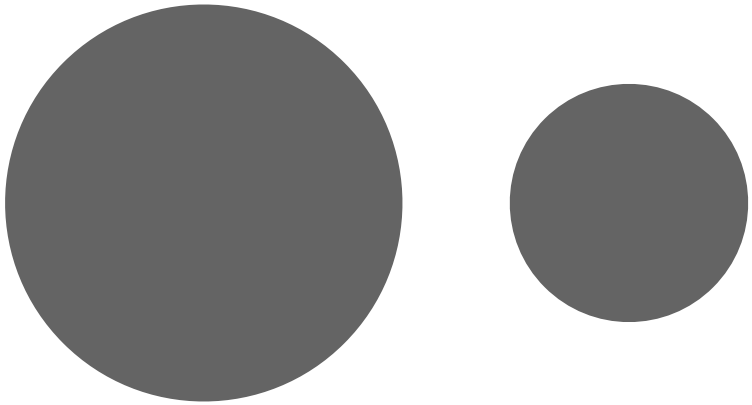


- What percentage in size is the right from the left (=100%)?

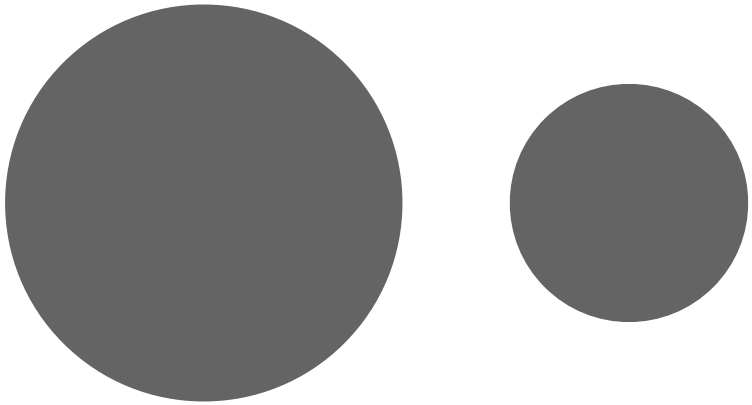


25 %

- What percentage in size is the right from the left (=100%)?



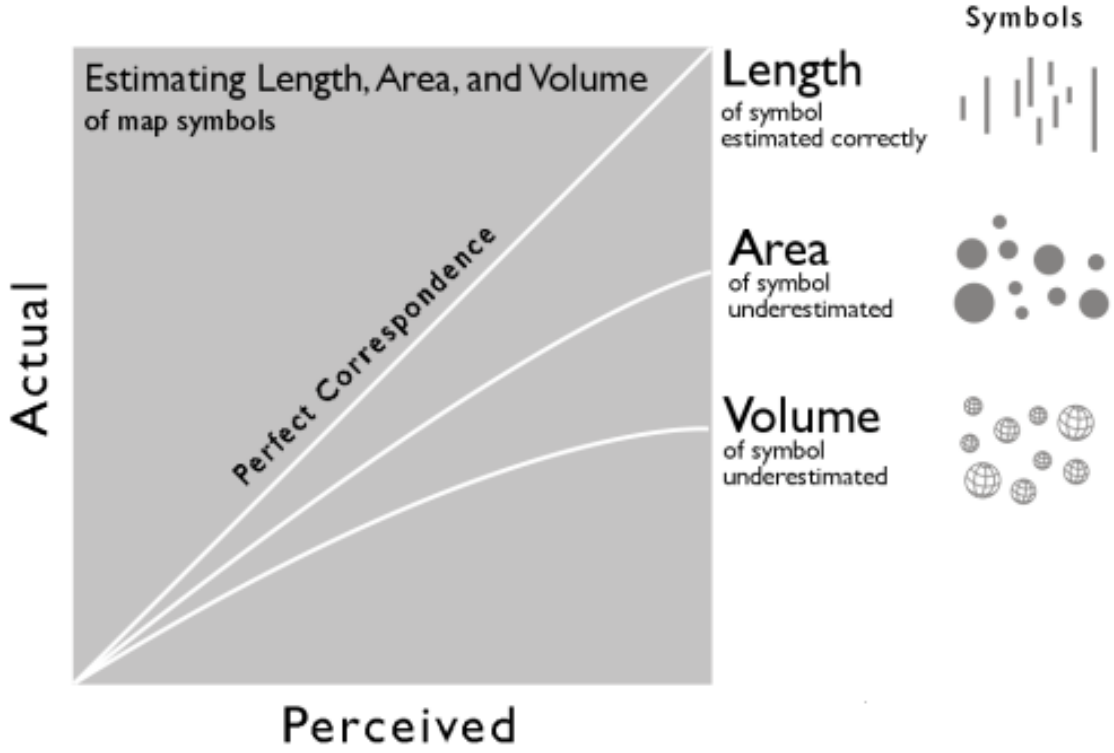
- What percentage in size is the right from the left (=100%)?



36 %

Visual Perception Accuracy

- People tend to **correctly estimate lengths**
- They tend to **underestimate areas and volumes.**
  - When asked to pick a circle that is two times the size of another most people would pick a circle ~1.8 times the size. This tendency gets worse with larger areas, and is worse in general for estimations of volumes.

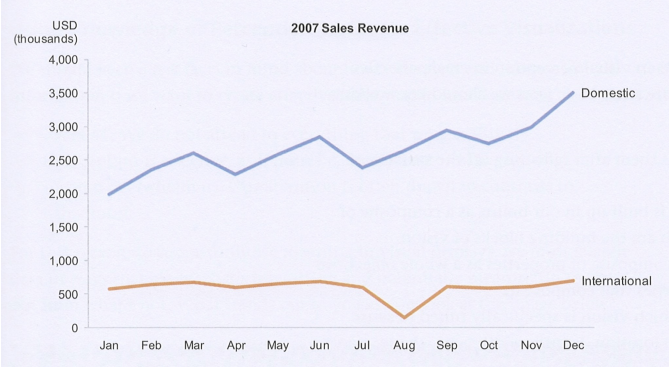


...

Comment représenter un processus temporel ?

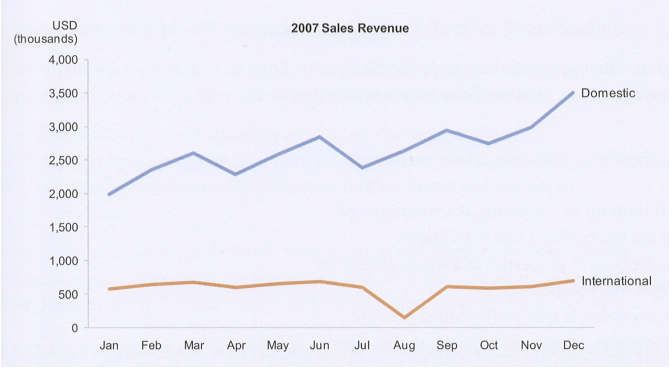
Comment représenter un processus temporel ?

- Pourquoi orienter les graphes de la gauche vers la droite ?



Comment représenter un processus temporel ?

- Pourquoi orienter les graphes de la gauche vers la droite ?



- Pourquoi pour les *occidentaux*,

  le **passé** est **derrière** nous

  et le **futur** est **devant** nous ?

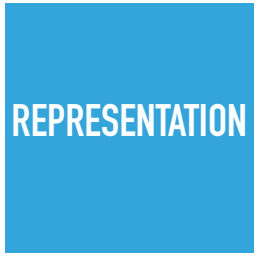
Plan

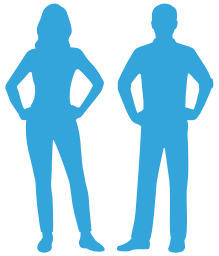
1. Puissance (et limites) de l'appareil visuel
2. Qu'est-ce qu'une représentation
3. Les primitives visuelles
4. Types de données et types de visualisation
   - Démarche
   - Données catégorielles
   - Séries temporelles
   - Données multi-variées
5. Outils et librairies

Démarche pour l'**exploration de données**

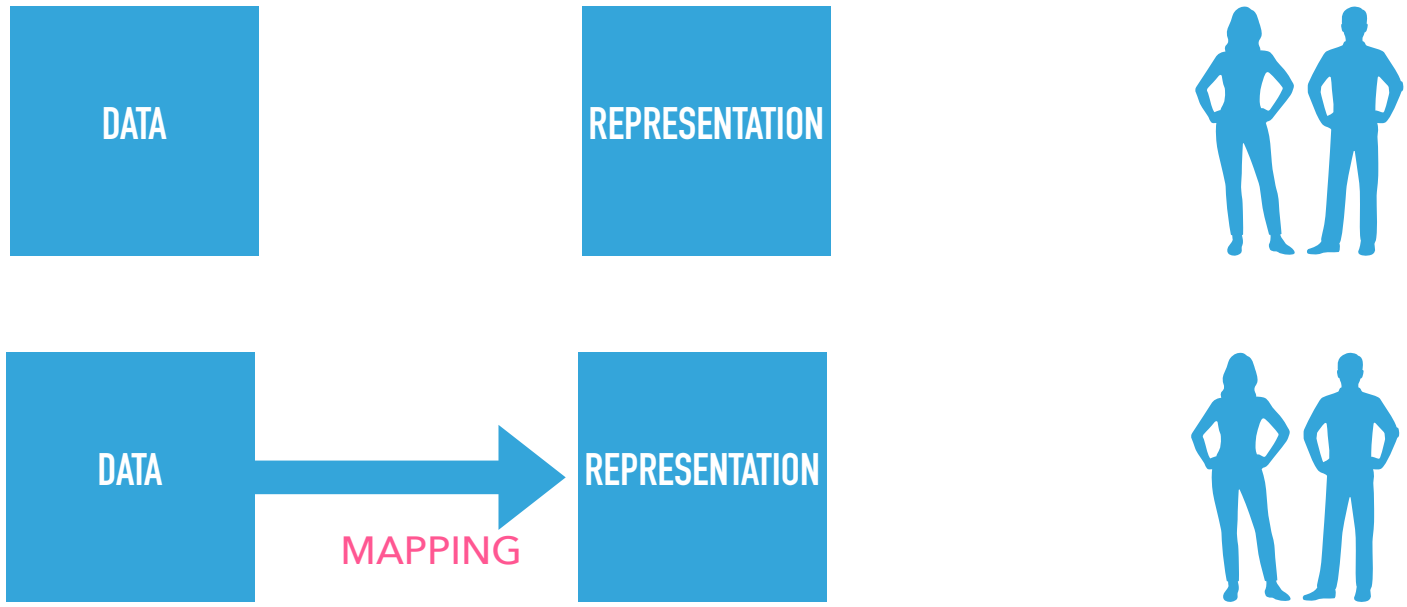
par visualisation


DATA

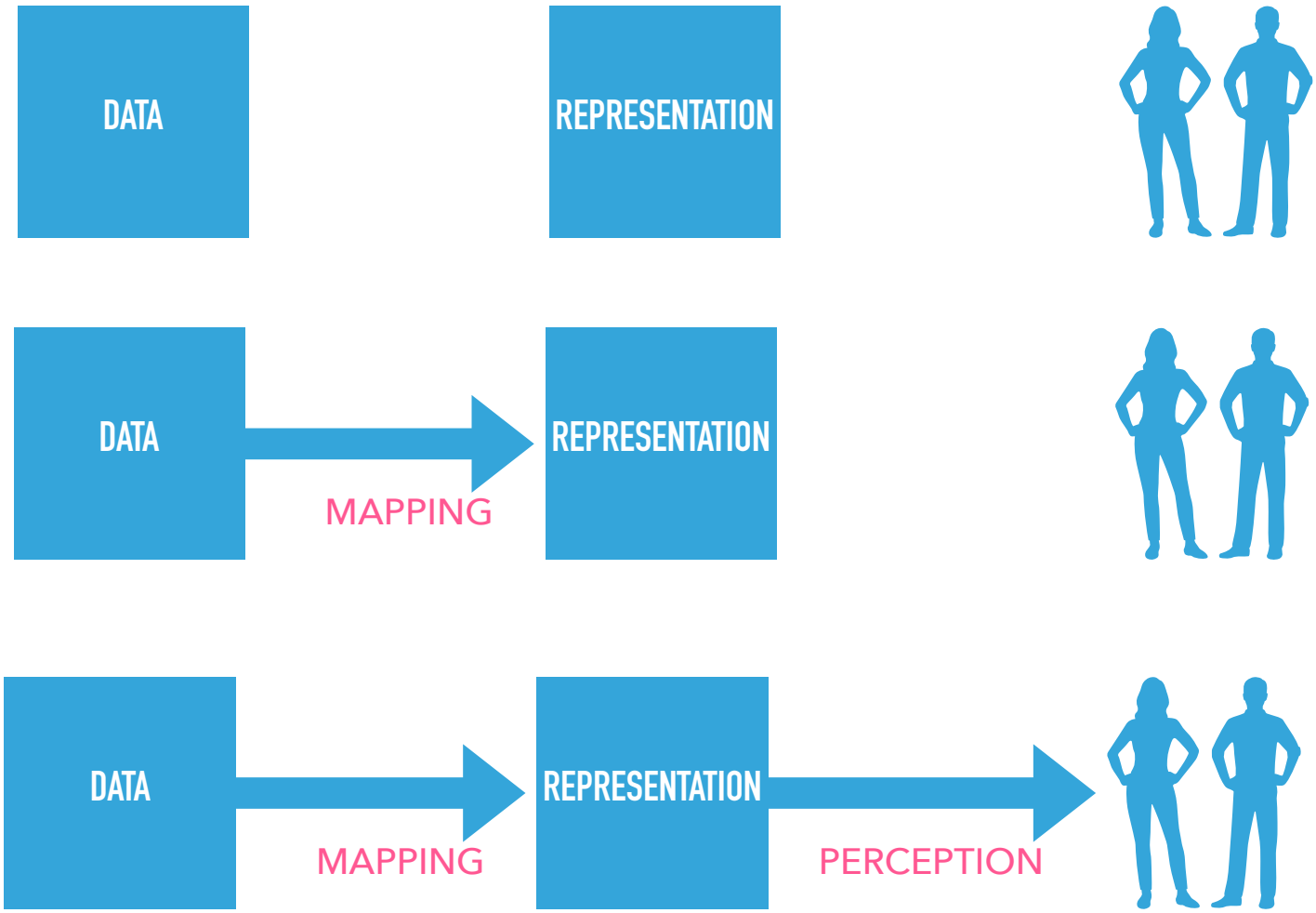

REPRESENTATION



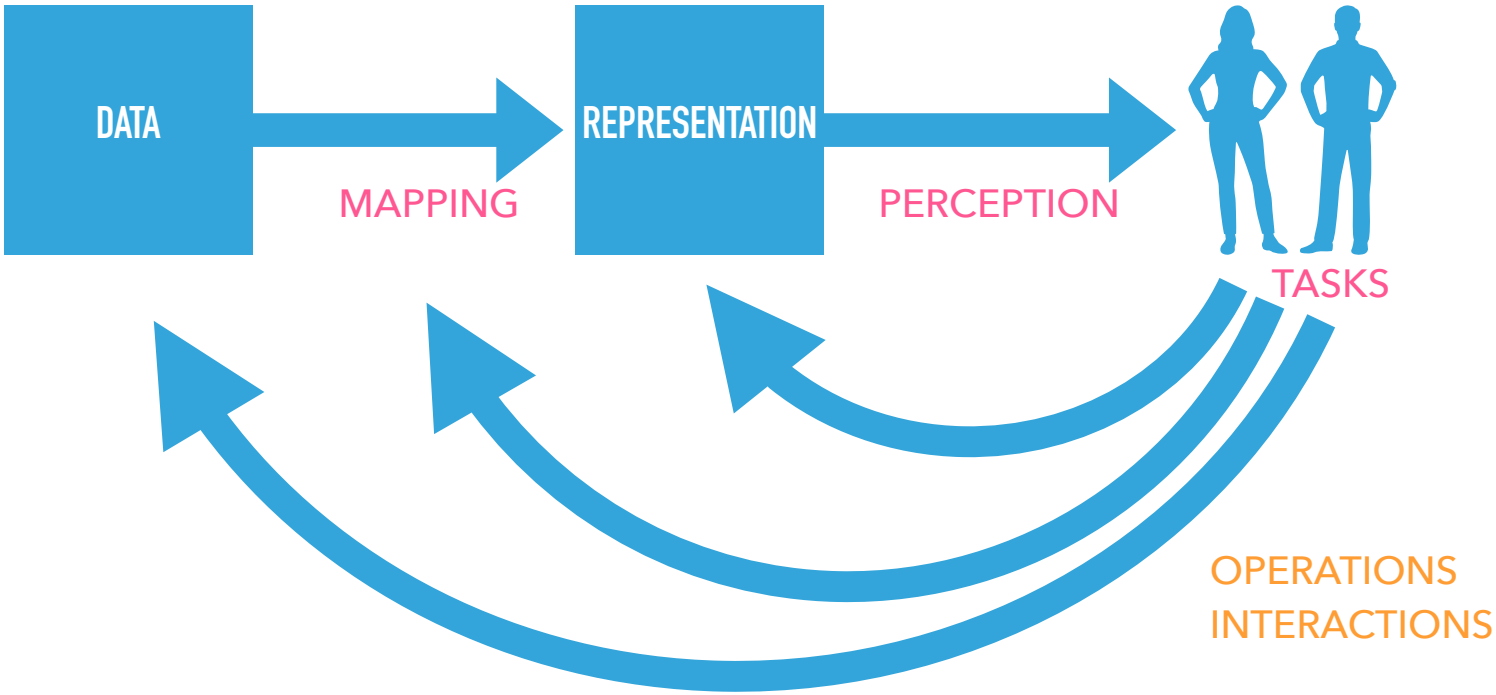
...


DATA
REPRESENTATION
DATA
MAPPING
REPRESENTATION

...


DATA
REPRESENTATION
DATA
MAPPING
REPRESENTATION
DATA
MAPPING
REPRESENTATION
PERCEPTION

...


DATA
MAPPING
REPRESENTATION
PERCEPTION
TASKS
OPERATIONS
INTERACTIONS

...

Démarche

1. Quelles **données** sont disponibles ?
2. Que **cherche**-t-on ?
3. Quelles **méthodes de visualisation** pourrait-on (devrait-on) utiliser ?
4. Que **voit-on** et cela a-t-il **un sens** ?

La collecte des données

Souvent long et long …

… et coûteux

- Chercher les données
  - À partir d'un site
  - À partir d'une API
  - Bases de données multiples
  - …

La préparation des données

Souvent long et long …

… et coûteux

1. « **Nettoyer** » les données
2. Remplir les **valeurs manquantes** (ou éliminer des données)
3. **Normaliser** les données (si utile)
4. Changer la **représentation**

Que voit-on ? Cela a-t-il un sens ?

- Parfois il faut **du temps pour voir**

- Se demander : ce que je vois a-t-il **un sens ?**
  - Est-ce **différent** de ce que produirait le **hasard** ?
  - **Quelle est l'incertitude** attachée aux valeurs mesurées ?
  - Quelle est la **fiabilité** des données ?

Quelles **méthodes**

pour quel type de **données**

- Voir par exemple

https://ft-interactive.github.io/visual-vocabulary/

Visual Vocabulary

Designing with data

There are so many ways to visualise data – how do we know which one to pick? Click on the coloured categories below to decide which data relationship is most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations

Inspired by the Graphic Continuum by Jon Schwabish and Severino Ribecca

Deviation | Correlation | Change v Time | Ranking | Distribution | Part to whole | Magnitude | Spatial | Flow

Change v Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries: Choosing the correct time period is important to provide suitable context for the reader

Examples of use

Share price movements, economic time series

Chart types

line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points

column-timeline

Columns work well for showing change over time - but usually best with only one series of data at a time

column-line-timeline

A good way of showing the relationship over time between an amount (columns) and a rate (line)

stock-price

Usually focused on day-to-day activity, these charts show opening/closing and hi/low points of each day

slope

Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story

area

Use with care. These are good at showing changes to total, but seeing change in components can be very difficult.

fan

Use to show the uncertainty in future projections - usually this grows the further forward to projection

scatterplot-line-timeline

A good way of showing changing data for two variables whenever there is a relatively clear pattern of progression. Connected scatterplot

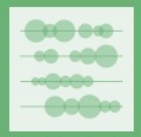
calendar-heatmap

A great way of showing temporal patterns (daily, weekly, monthly), at the expense of showing precision in quantity

priestley timeline

Great when date and duration are key elements of the story in the data

circles-timeline



Good for showing discrete values of varying size across multiple categories (eg earthquakes by contintent)

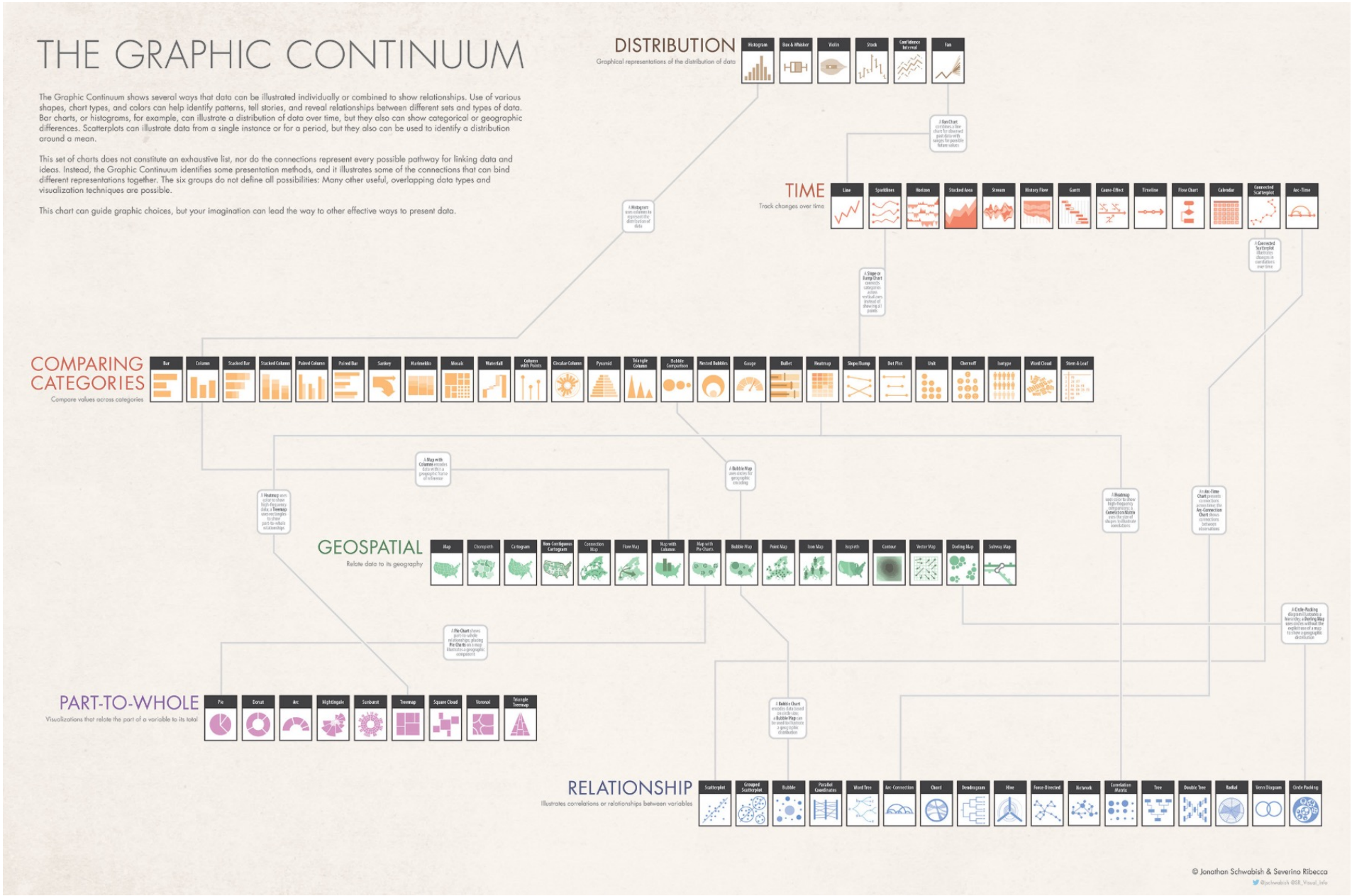
seismogram

Another alternative to the circle timeline for showing series where there are big variations in the data

...

- Voir par exemple (2)

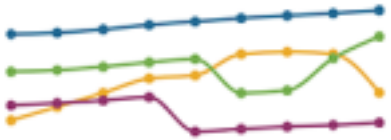
https://www.informationisbeautifulawards.com/showcase/611-the-graphic-continuum


THE GRAPHIC CONTINUUM
The Graphic Continuum shows several ways that data can be illustrated individually or combined to show relationships. Use of various shapes, chart types, and colors can help identify patterns, tell stories, and reveal relationships between different sets and types of data. Bar charts, or histograms, for example, can illustrate a distribution of data over time, but they also can show categorical or geographic differences. Scatterplots can illustrate data from a single instance or for a period, but they also can be used to identify a distribution around a mean.
This set of charts does not constitute an exhaustive list, nor do the connections represent every possible pathway for linking data and ideas. Instead, the Graphic Continuum identifies some presentation methods, and it illustrates some of the connections that can bind different representations together. The six groups do not define all possibilities: Many other useful, overlapping data types and visualization techniques are possible.
This chart can guide graphic choices, but your imagination can lead the way to other effective ways to present data.
DISTRIBUTION
Graphical representations of the distribution of data
TIME
Track changes over time
COMPARING CATEGORIES
Compare values across categories
GEOSPATIAL
Relate data to its geography
PART-TO-WHOLE
Visualizations that relate the part of a variable to its total
RELATIONSHIP
Illustrates correlations or relationships between variables
© Jonathan Schwabish & Severino Ribecca

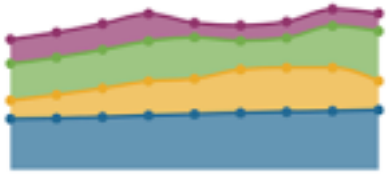
...

1. Change over time

If you mainly want to communicate change over time, we recommend these. Click on them to make one now!

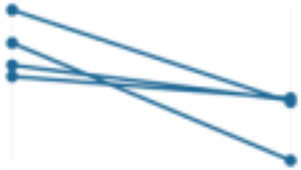


Line:
The standard way to show changing time series



Area:
Great at showing total change, though individual series are less clear

Column:
Another way to show change over time, especially for a single series



Slope:
Good for clearly conveying 'before and after' data

- From https://flourish.studio/2018/09/28/choosing-the-right-visualisation/

2. Size

Use these when you mainly want to communicate size comparisons, relative or absolute.

Column:
The standard way to compare the size of things; y-axis should always start at 0

Bar:
Like columns, especially when the data are not time series, or axis labels are long

Grouped column:
As per column, but for multiple series

Grouped bar:
As per bar, but for multiple series

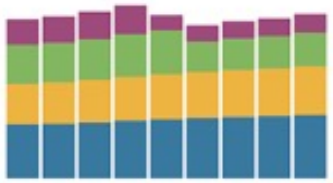
- From https://flourish.studio/2018/09/28/choosing-the-right-visualisation/

3. Parts of a whole

Use these to show how an entity breaks down into its components. If you're mainly interested in the absolute size of the components, consider one of the size charts above instead.

Pie:
Common, though hard to accurately compare the segments

Donut:
Similar to a pie, though the center can be used to convey additional information



Stacked column:
Good for combining with change over time, though can be hard to read

Treemap:
Use for hierarchical part-to-whole relationships

Sunburst:
Attractive alternative to treemap, though may be harder to read

Packed circles:
Attractive alternative to treemap, though may be harder to read

- From https://flourish.studio/2018/09/28/choosing-the-right-visualisation/

4. Correlation

Use correlation visualizations when you want to show the relationship between two or more variables.

Scatterplot:
The standard way to show the relationship between two continuous variables

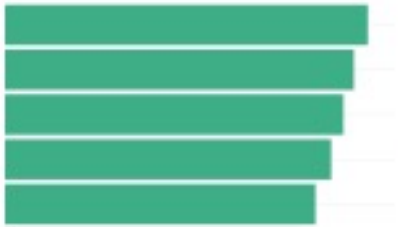
Bubble chart:
Like a scatterplot, but sizes the circles by a third variable

Connected scatterplot:
Aka Rosling chart: show how relationship has changed over time

- From https://flourish.studio/2018/09/28/choosing-the-right-visualisation/

5. Ranking

Use where an item's position in an ordered list is the most important thing you want to show.



Ordered bar:
A simple bar chart. Just order your data as you want it to display

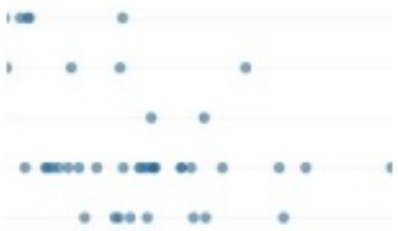
Horserace:
Also known as a bumps chart. Show more detailed comparison over time

Slope chart:
Perfect for showing effectively how ranks have changed over time

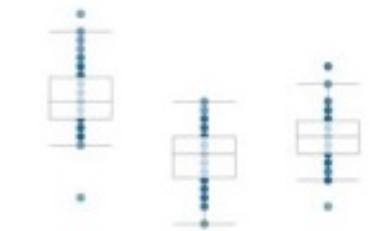
- From https://flourish.studio/2018/09/28/choosing-the-right-visualisation/

6. Distribution

Show values in a dataset and how often they occur. The shape of a distribution is often useful to see. Coming soon: histograms!



Dot plot:
A simple way to show raw values in the data, across categories



Box plot:
Summarize multiple distributions by showing the median & range of the data

- From https://flourish.studio/2018/09/28/choosing-the-right-visualisation/

7. Flows and relationships

Use these to show volume, movement or connections between two or more states.

Sankey:
Show changes in volume between different conditions

Chord:
A complex but powerful way to illustate two-way flows

Network:
Use to show the strength and connectness of relationships

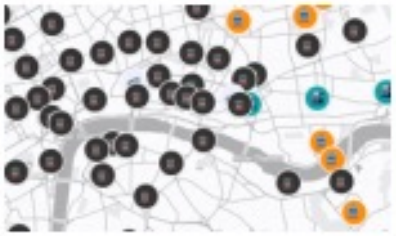
Directed network:
With arrows to show directed relationships

- From https://flourish.studio/2018/09/28/choosing-the-right-visualisation/

8. Geospatial

Use when geospatial patterns in your data are more important than anything else.

Choropleth:
Use with your own GeoJSON data



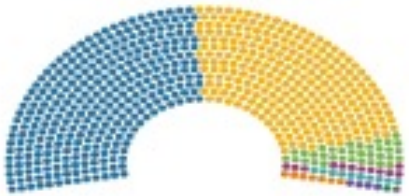
Icon map:
Simple point map, using icons or emojis

Connections globe:
Unique way to show flow between countries

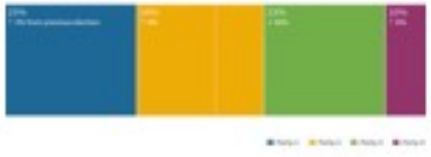
- From https://flourish.studio/2018/09/28/choosing-the-right-visualisation/

9. Flourish special: Election visualizations

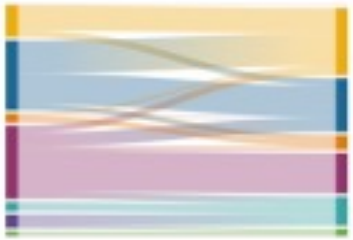
This category is specifically for people wanting to visualize the outcome of elections.



Parliament chart: Aka arcmap, often used for election results



Election chart: Searchable bar chart ideal for election results, includes a coalition builder feature

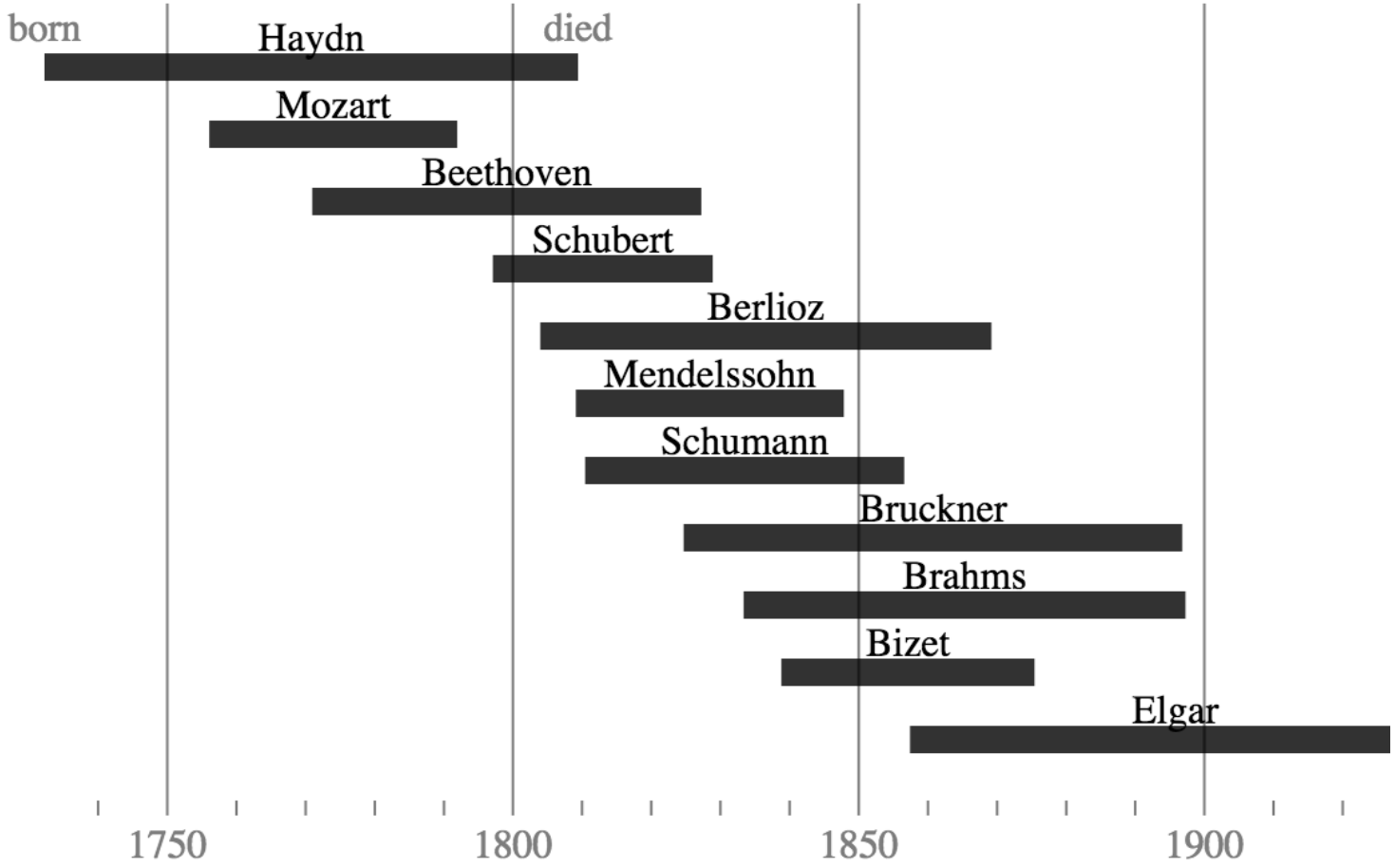


Sankey: Also good for showing voting changes between elections

- From https://flourish.studio/2018/09/28/choosing-the-right-visualisation/

Exemple

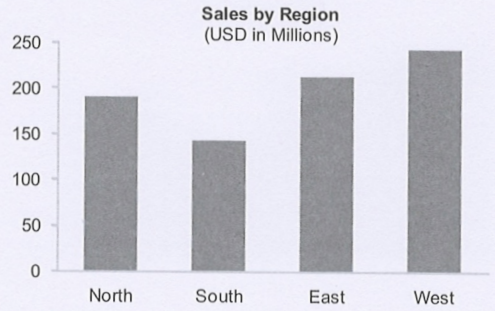
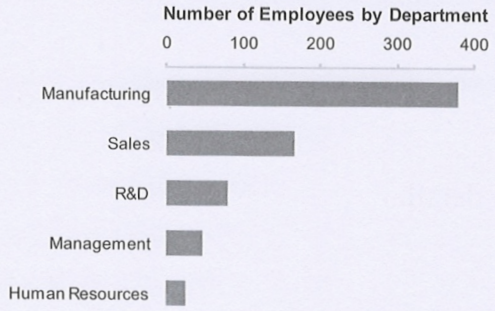
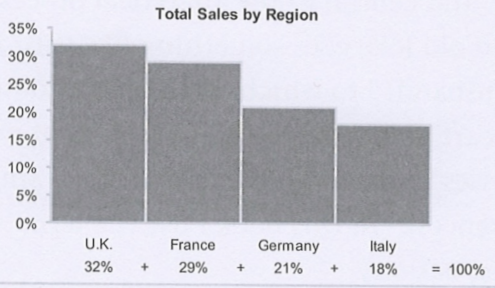
Comparaison de dates



Interactions avec les données

- Comparer
- Trier
- Ajouter des variables
- Filtrer
- Mettre en relief
- Agréger
- Ré-exprimer
- Changer de visualisation
- Zoomer et orienter son attention sur certaines zones
- Changer d'échelle
- Accéder aux détails à la demande
- Annoter
- Garder un historique des manipulations effectuées

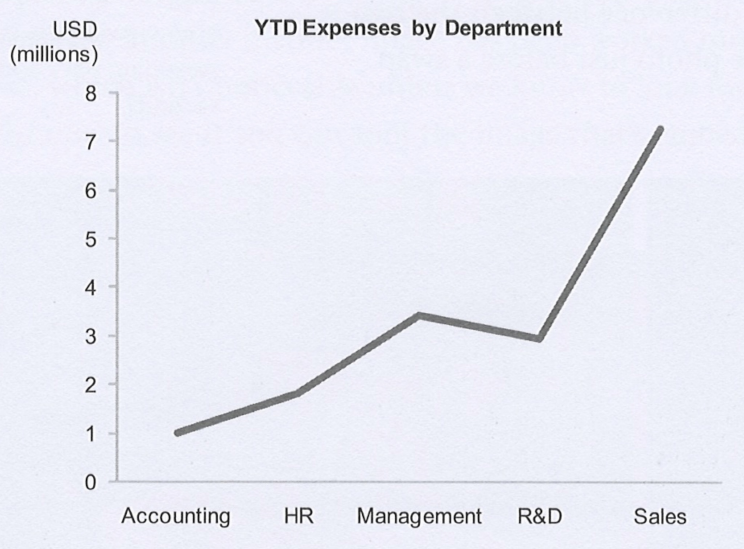
Comparer

Type	Description	
Nominal	Comparing values that have no particular order	
Ranking	Comparing values that are arranged by magnitude, from low to high or high to low	
Part-to-Whole	Comparing values that when combined make up parts of a whole	

[Stephen FEW (2009)
« Now you see it », p. 56]

Exemples

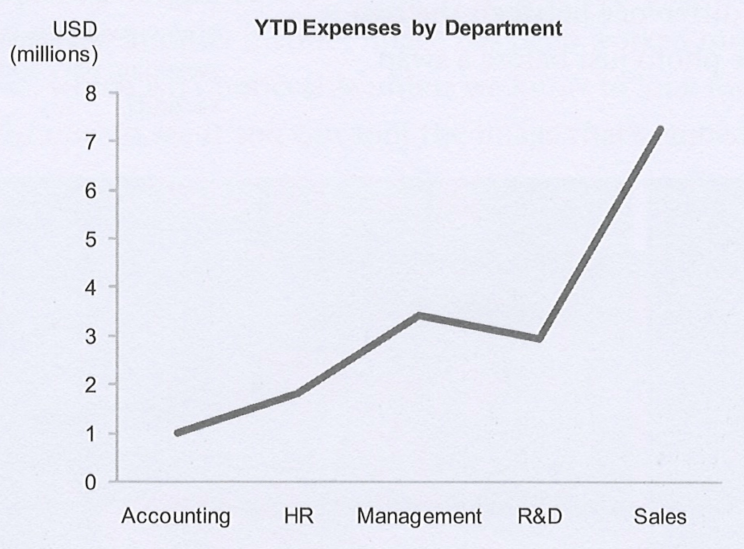
Ne pas faire n'importe quoi

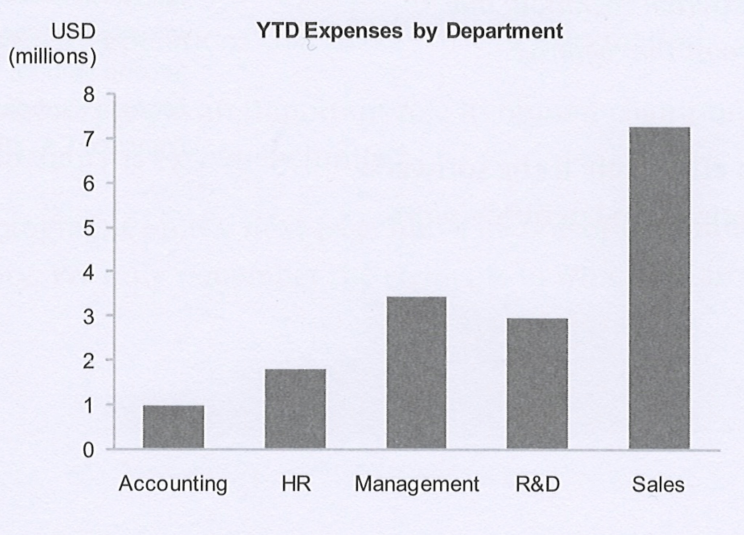


[Stephen FEW (2009)
« Now you see it », p. 36]

Dépenses par secteurs

Ne pas faire n'importe quoi





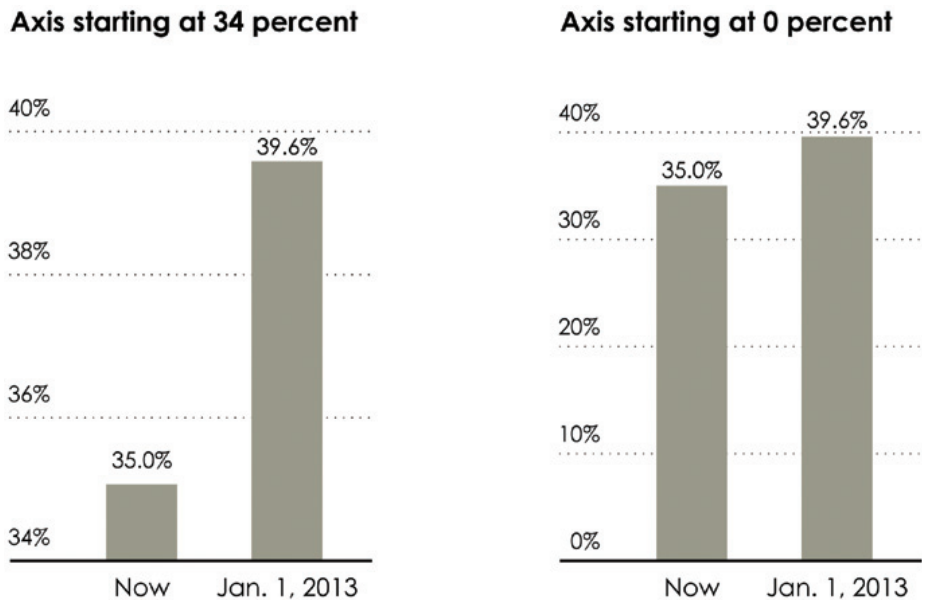
[Stephen FEW (2009)
« Now you see it », p. 36]

Dépenses par secteurs

Attention ! Axe des ordonnées

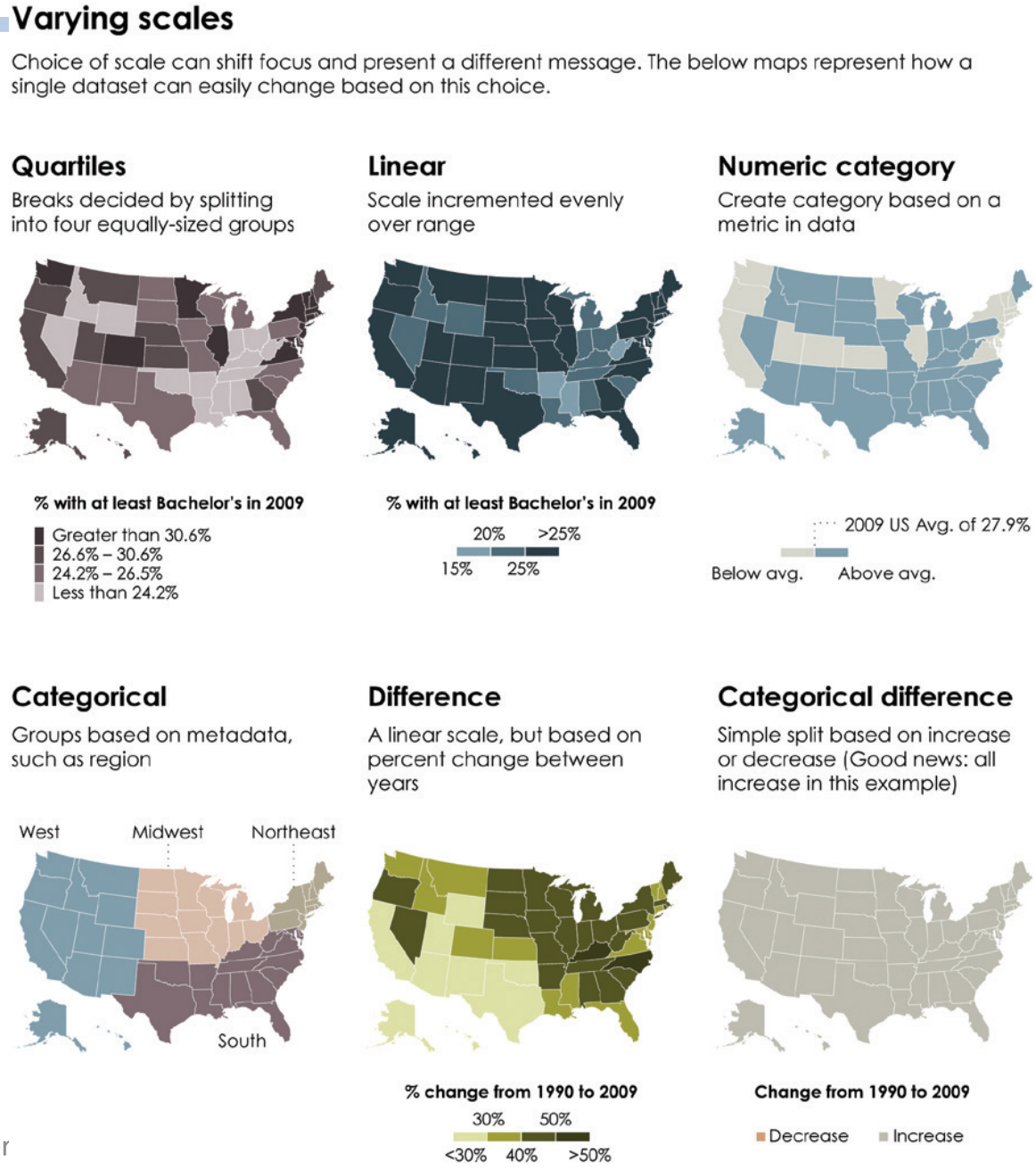
Augmentation d'une taxe

[Nathan YAU (2013)
« Data points. Visualization that means something », p. 97]



The difference between the two values looks like a huge increase—the length of the right bar is about five times the length as the other—because the value axis starts at 34 percent. The chart on the right shows the change when the axis starts at zero, which looks less dramatic. Of course, you can always look at the axis to verify what you see (and you always should), but that defeats the purpose of showing the values with length, and if the chart is shown quickly on television, most people won't notice the misstep.

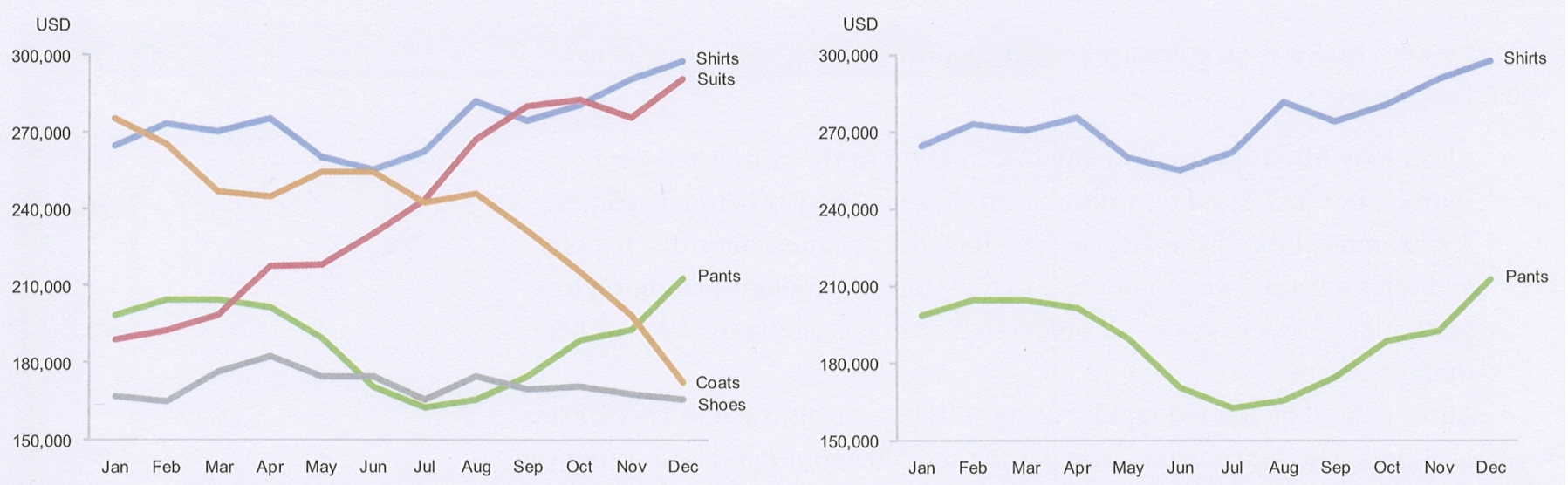
Effet de la discrétisation des valeurs



[Nathan YAU (2013)
« Data points. Visualization that means something », p. 131]

Filtrer

[Stephen FEW (2009)
« Now you see it », p. 65]

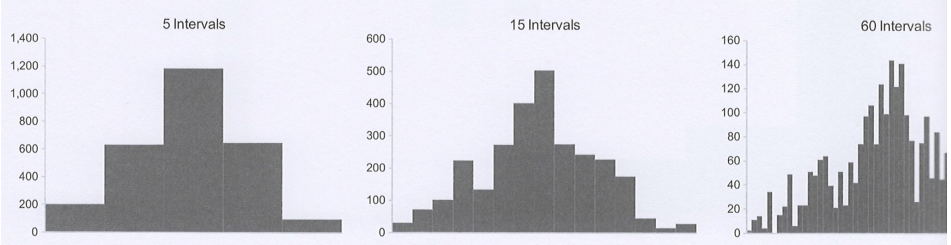


Retirer les informations inutiles à un moment donné de l'analyse

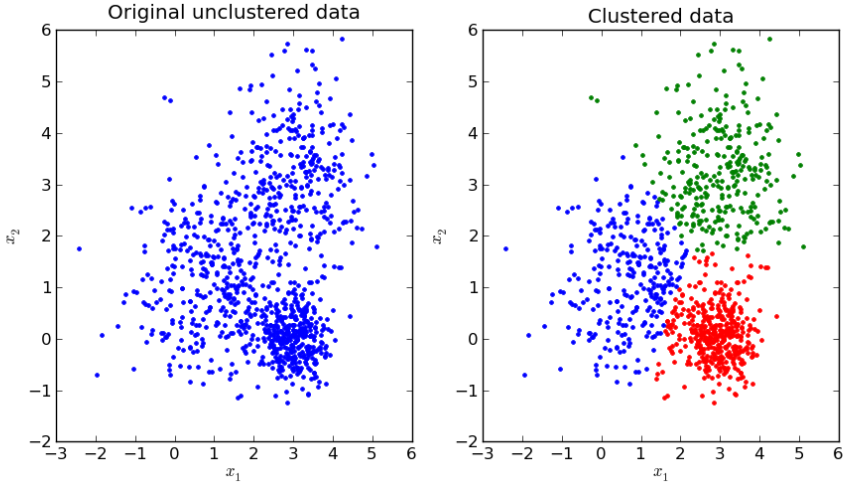
Mettre en relief

1. Changement du niveau de détail

[Stephen FEW (2009)
« Now you see it », p. 69]

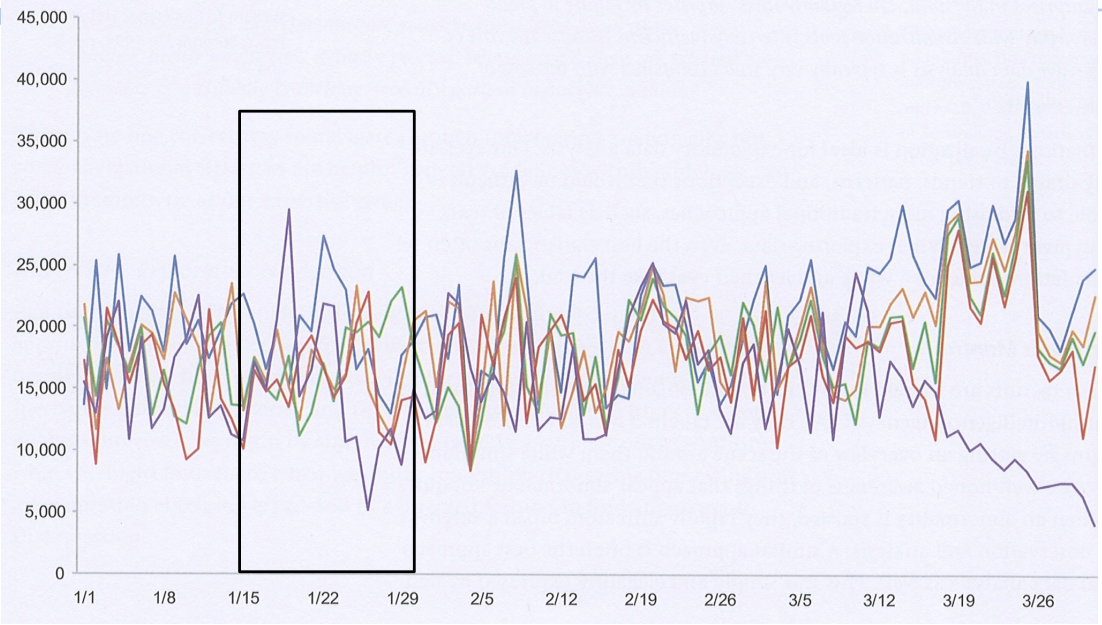


2. Clustering



Etc. …

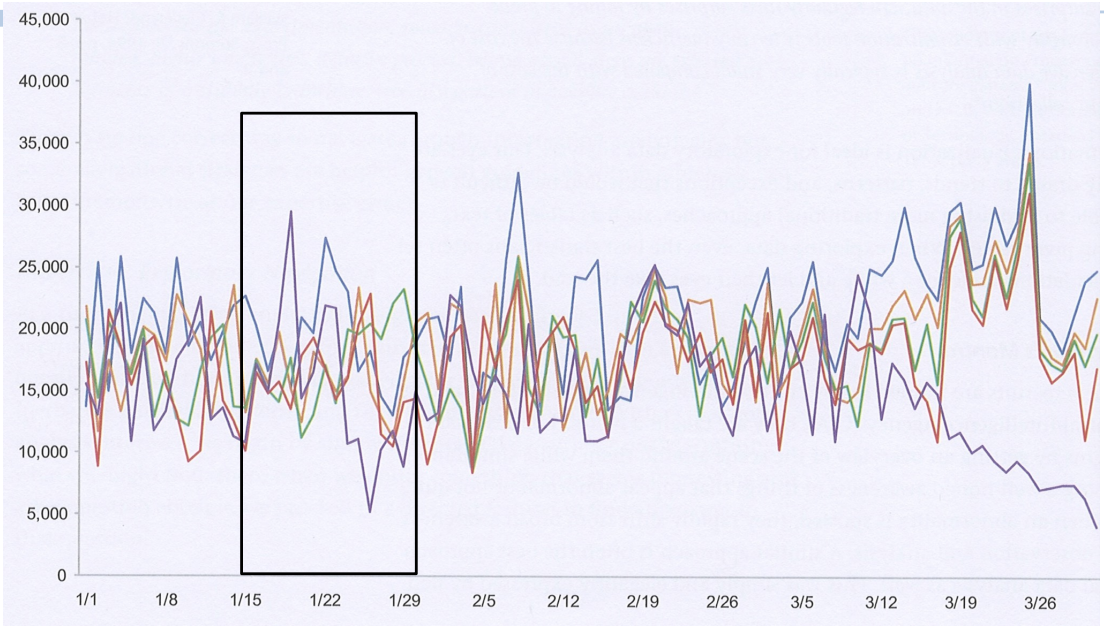
Zoomer et orienter son attention sur certaines zones



[Stephen FEW (2009)
« Now you see it », p. 75-76]

~

Zoomer et orienter son attention sur certaines zones



[Stephen FEW (2009)
« Now you see it », p. 75-76]

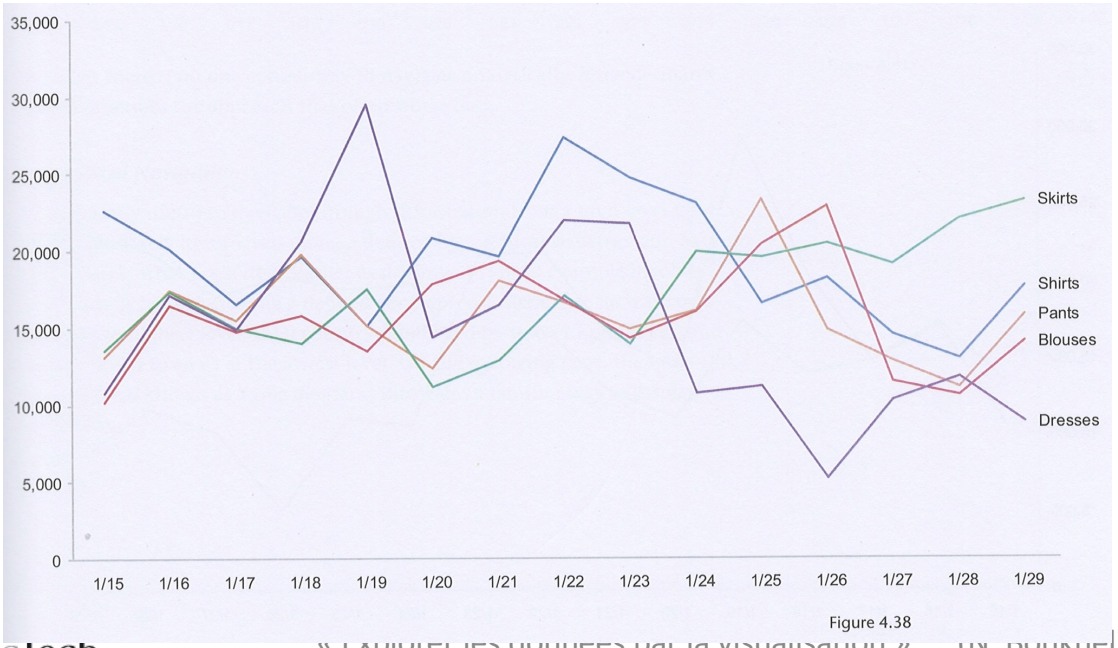
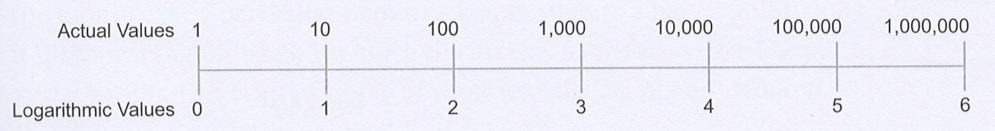


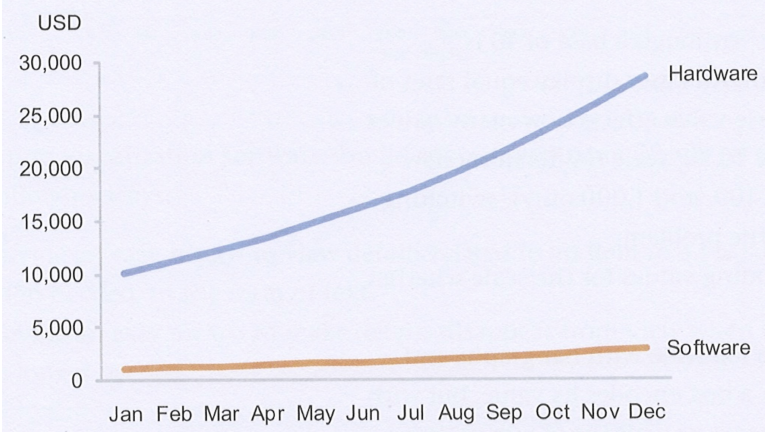
Figure 4.38

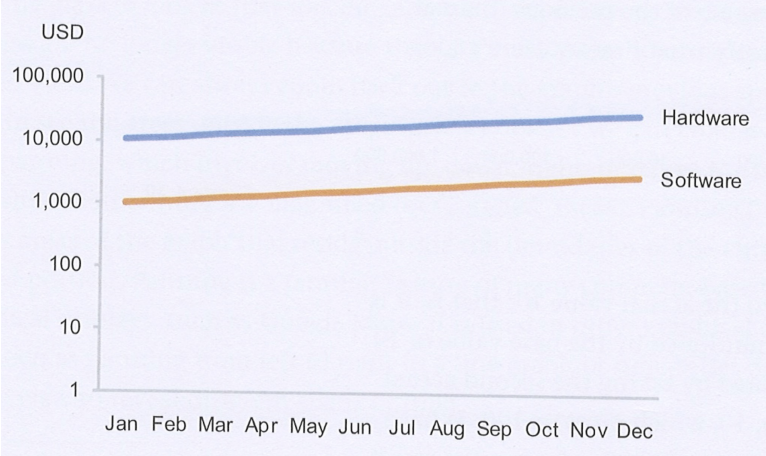
~

Changer d'échelle



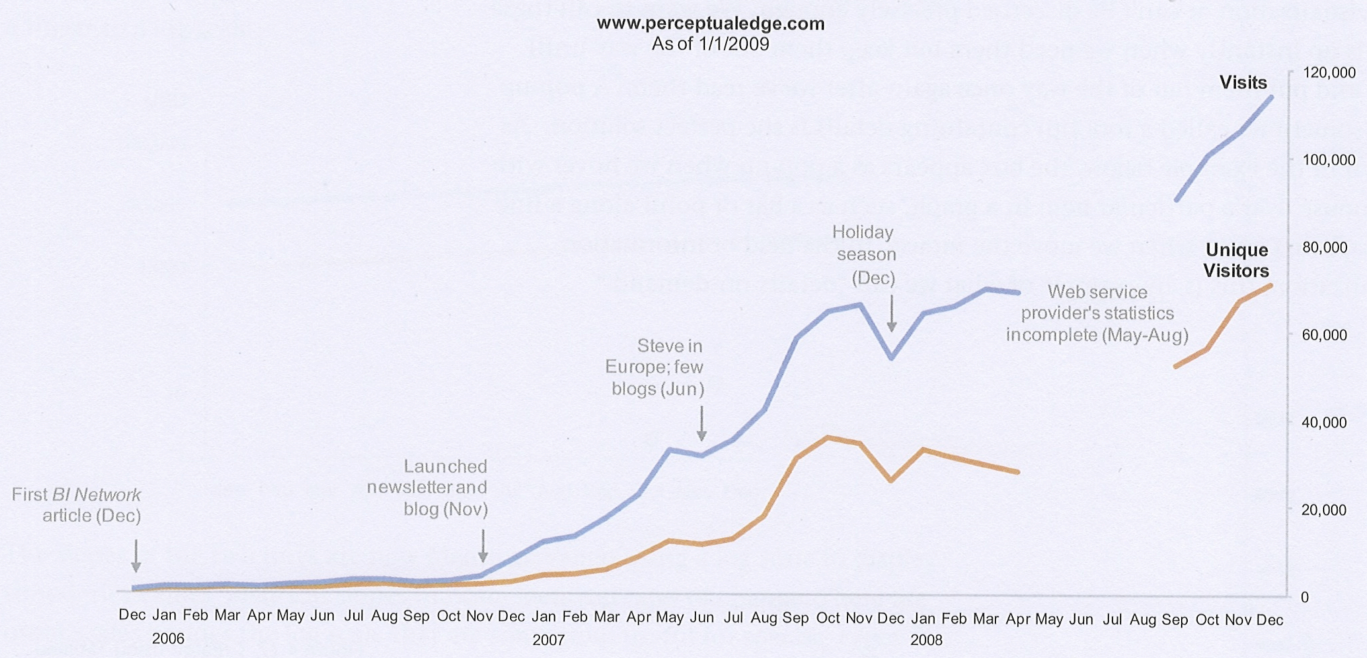
[Stephen FEW (2009)
« Now you see it », p. 77-78]





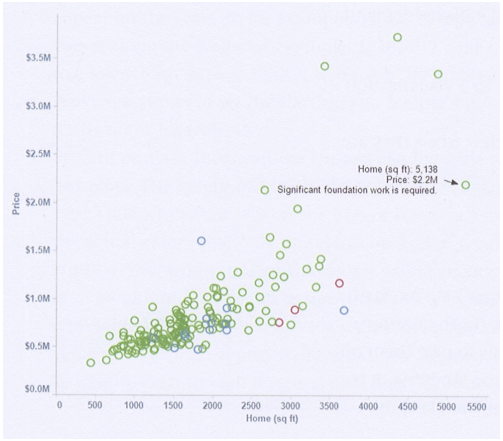
Changement du référentiel

Annoter



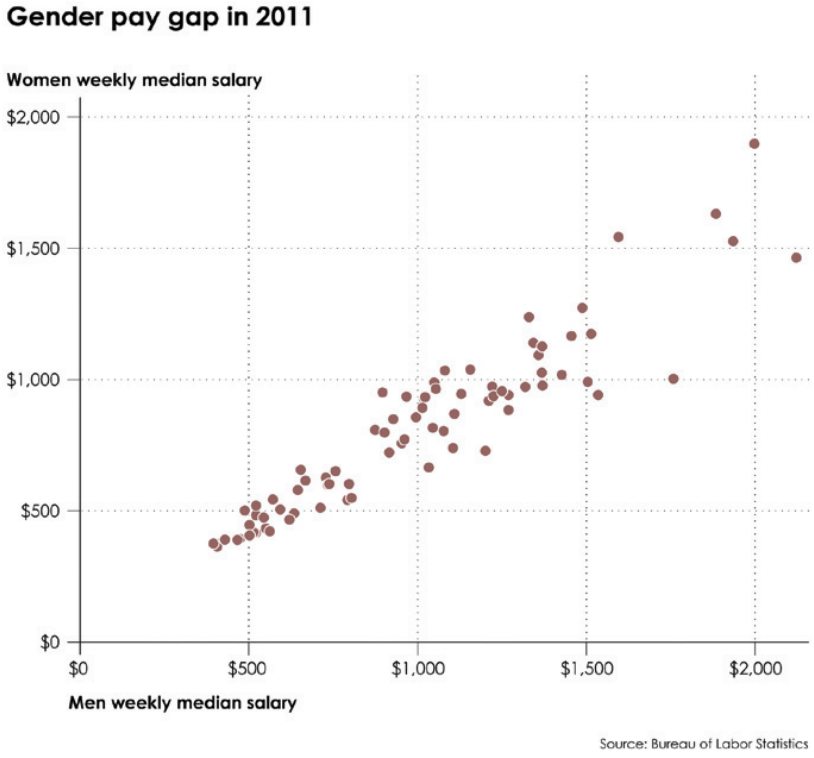
[Stephen FEW (2009)
« Now you see it », p. 80]

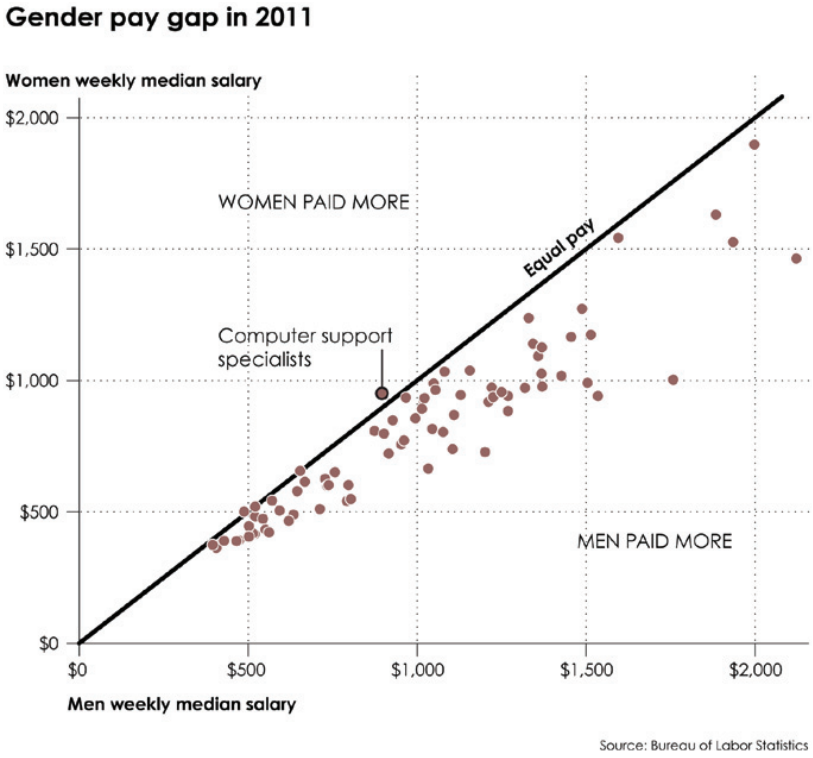
Commenter sur la figure



L'intérêt d'annoter

Exemple





[Nathan YAU (2013)
« Data points. Visualization that means something », p. 231-232]

Données **catégorielles**

Catégories

Sous-parties et tris

Données catégorielles

Personnes, lieux, objets, …

- Histogrammes

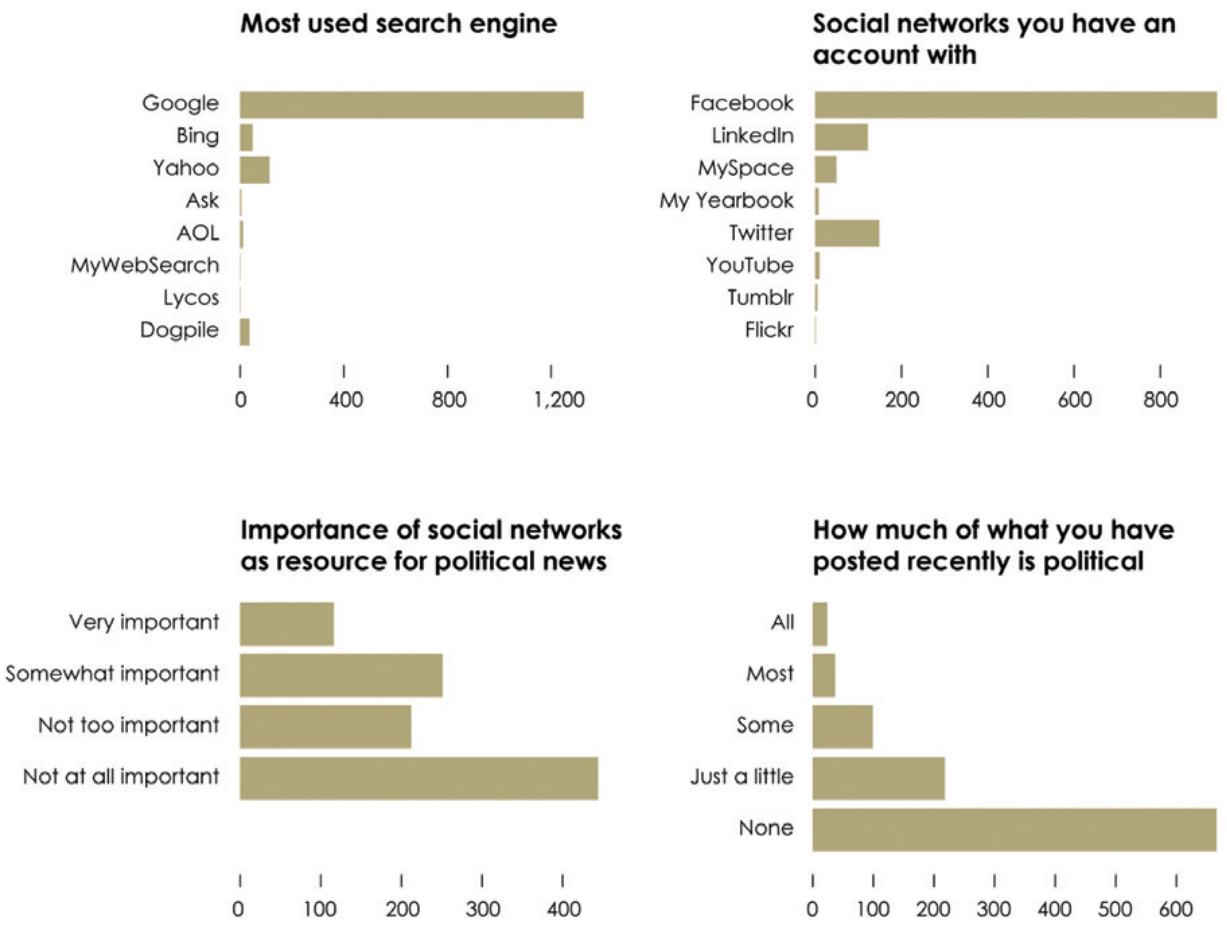
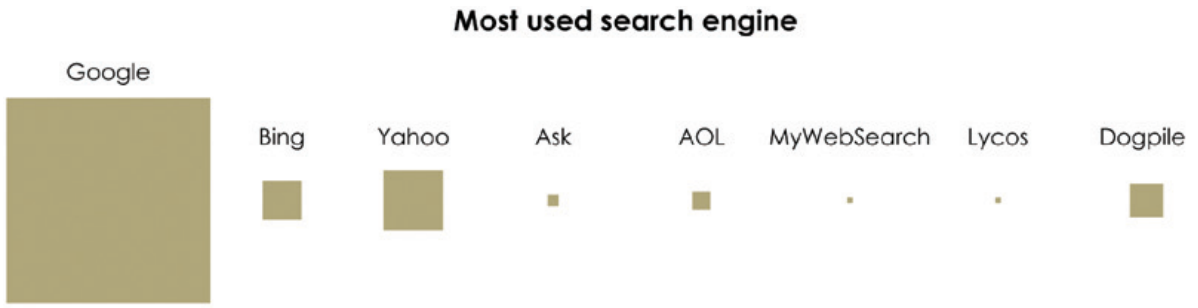


FIGURE 4-6 *Bar graphs for survey results*

[Nathan YAU (2013)
« Data points. Visualization that means something », p. 147]

Données catégorielles

- Histogrammes
- Surfaces



[Nathan YAU (2013)
« Data points. Visualization that means something », p. 148]

Données catégorielles

- Histogrammes
- Cercles
- Pie charts
- Treemaps
- Mosaic plots

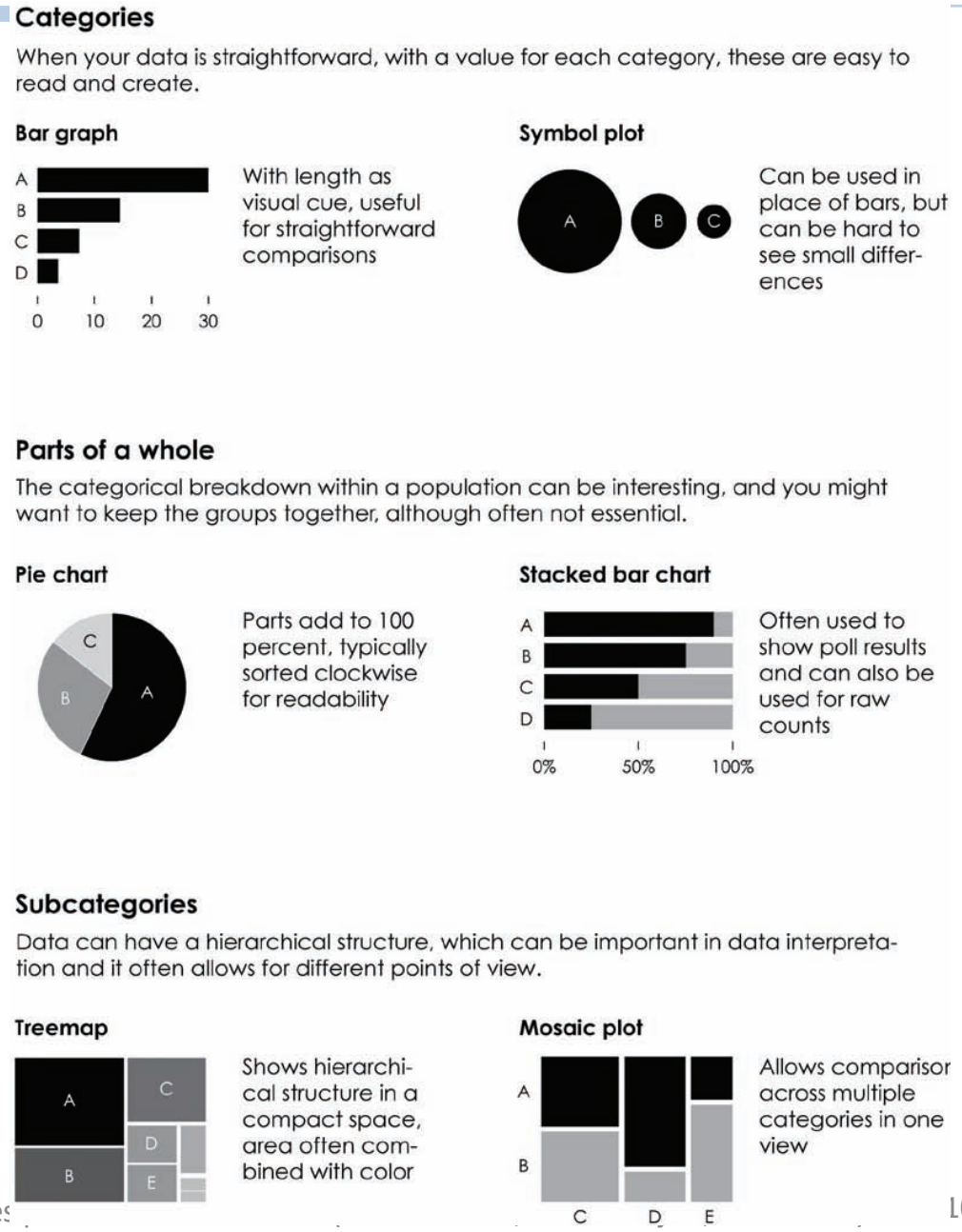
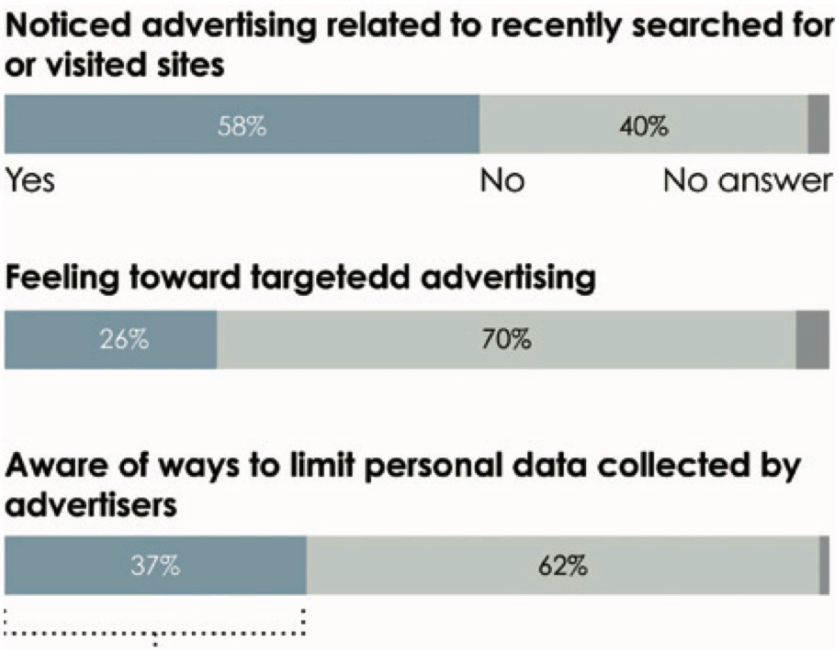


FIGURE 4-5 *Visualizing categorical data*

[Nathan YAU (2013)
« Data points. Visualization that means something », p. 146]

Données **sous**-catégorielles

- Histogrammes
- Surfaces
- Pie charts
- Subcategories



[Nathan YAU (2013)
« Data points. Visualization that means something », p. 150]

Analyse de corrélations

entre catégories

Bar plots

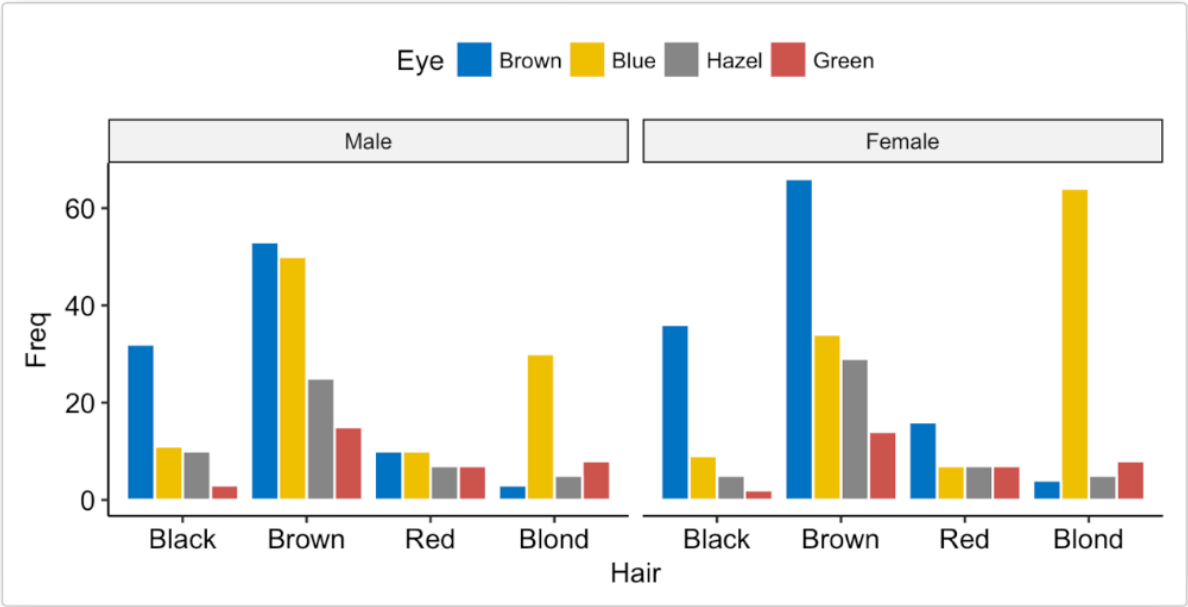
- Illustration en R

```
library(ggplot2)
library(ggpubr)
theme_set(theme_pubr())
data("HairEyeColor")
df <- as.data.frame(HairEyeColor)
head(df)
```

```
##     Hair   Eye  Sex Freq
## 1  Black Brown Male   32
## 2  Brown Brown Male   53
## 3    Red Brown Male   10
## 4  Blond Brown Male    3
## 5  Black  Blue Male   11
## 6  Brown  Blue Male   50
```

- Create the bar graph:
  - Hair color on x-axis
  - Change bar fill by Eye color
  - Split the graph into multiple panel by Sex

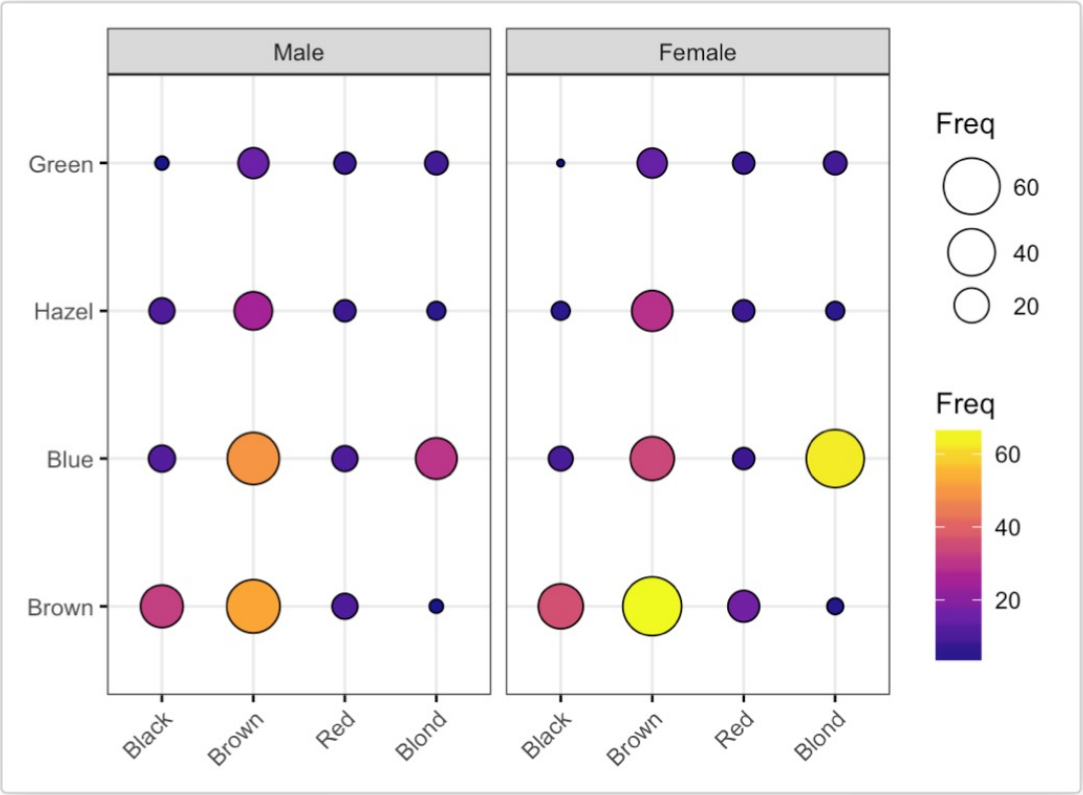
```
ggplot(df, aes(x = Hair, y = Freq))+
  geom_bar(
    aes(fill = Eye), stat = "identity", color = "white",
    position = position_dodge(0.9)
    )+
  facet_wrap(~Sex) +
  fill_palette("jco")
```



Balloon plots

- Illustration en R

```
df <- as.data.frame(HairEyeColor)
ggballoonplot(df, x = "Hair", y = "Eye", size = "Freq",
              fill = "Freq", facet.by = "Sex",
              ggtheme = theme_bw()) +
  scale_fill_viridis_c(option = "C")
```



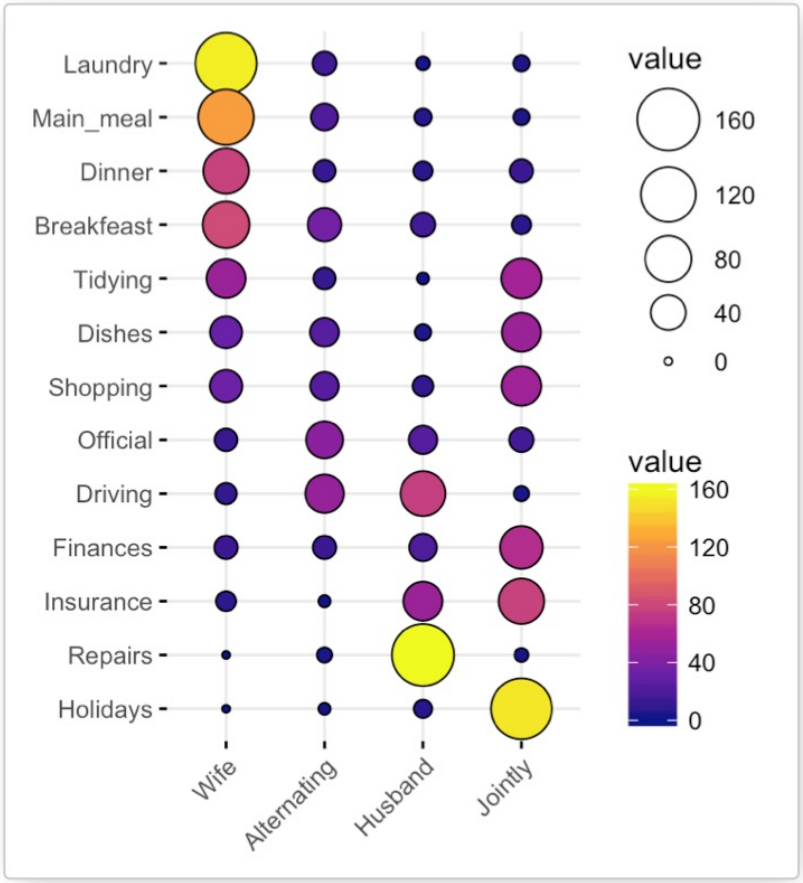
Balloon plots

- Illustration en R

```
housetasks <- read.delim(
  system.file("demo-data/housetasks.txt", package = "ggpubr"),
  row.names = 1
  )
head(housetasks, 4)
```

```
##                Wife Alternating Husband Jointly
## Laundry         156           14       2       4
## Main_meal       124           20       5       4
## Dinner           77           11       7      13
## Breakfeast       82           36      15       7
```

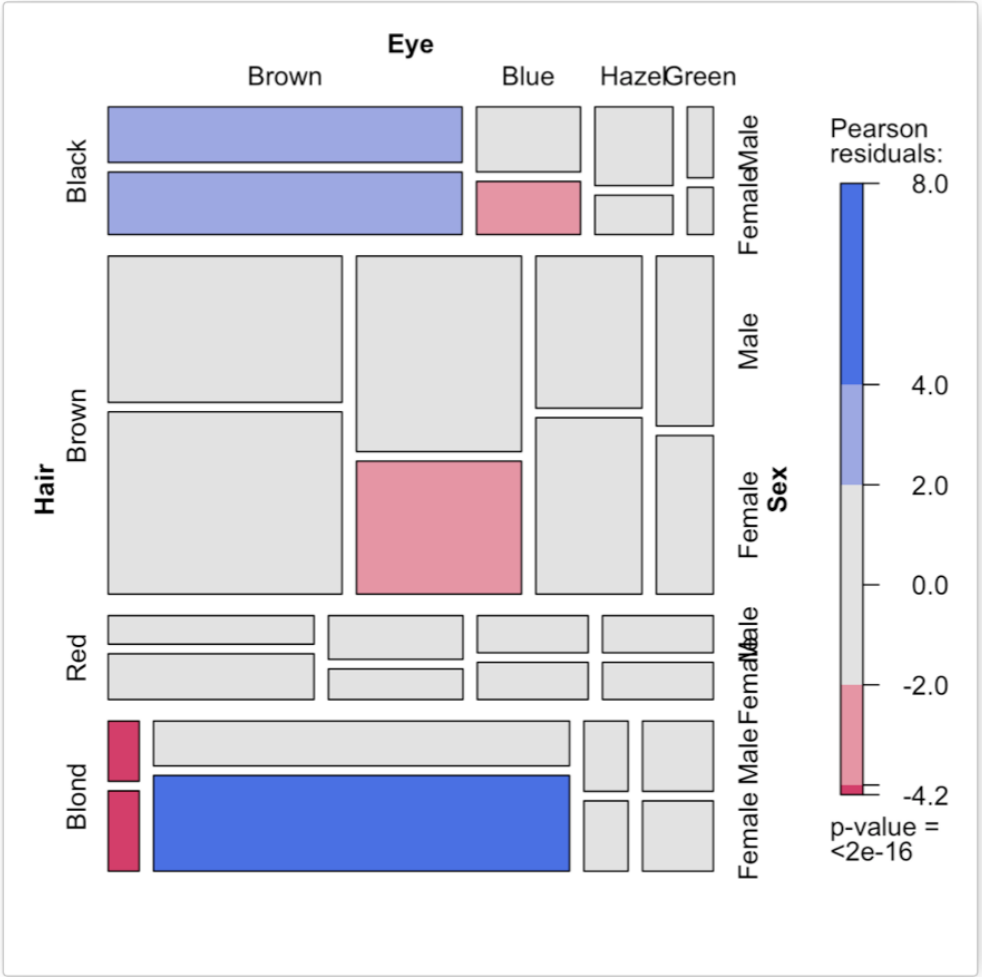
```
ggballoonplot(housetasks, fill = "value")+
  scale_fill_viridis_c(option = "C")
```



Mosaic plots

- Illustration en R

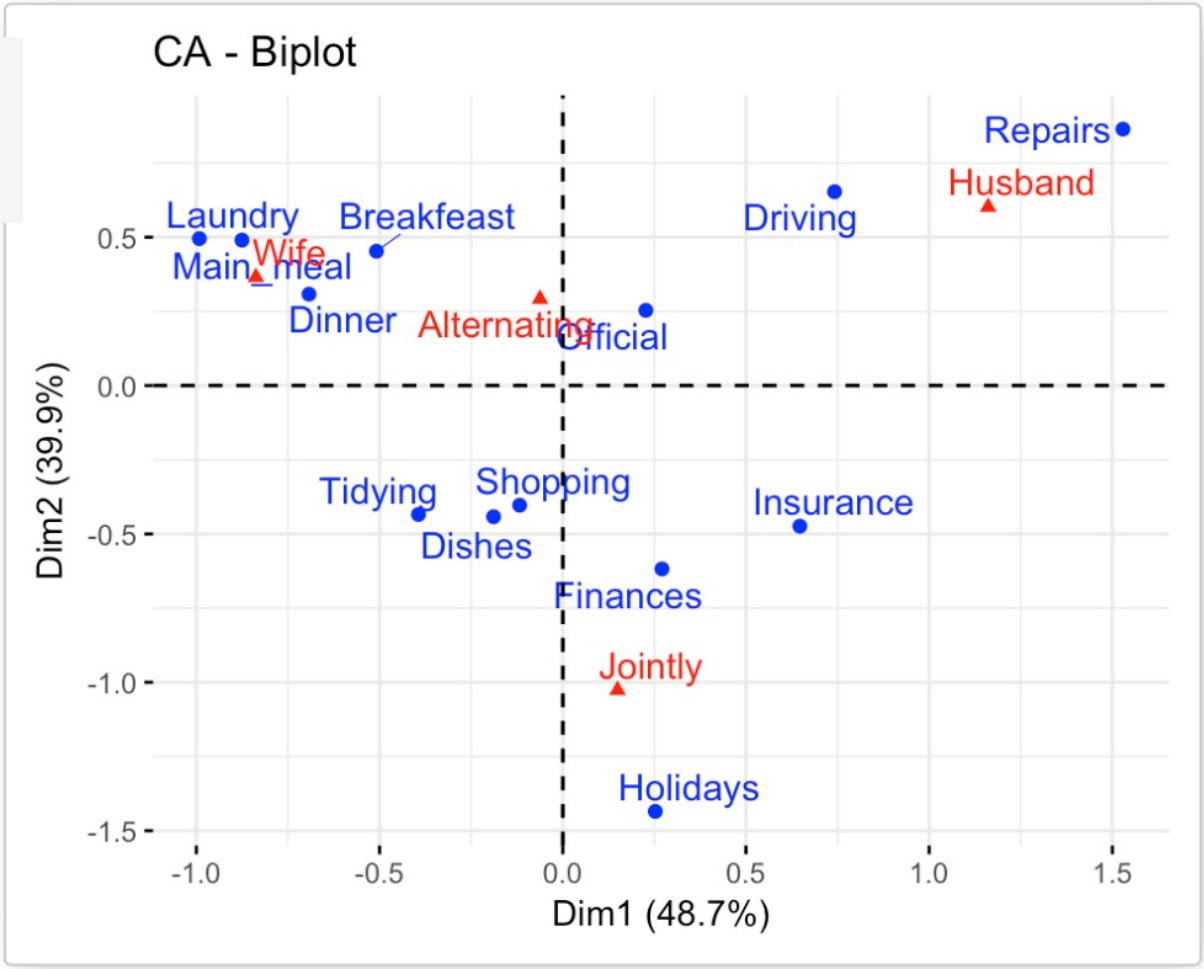
```
library(vcd)
mosaic(HairEyeColor, shade = TRUE, legend = TRUE)
```



Correspondence analysis

- Illustration en R

```
library(FactoMineR)
library(factoextra)
res.ca <- CA(housetasks, graph = FALSE)
fviz_ca_biplot(res.ca, repel = TRUE)
```



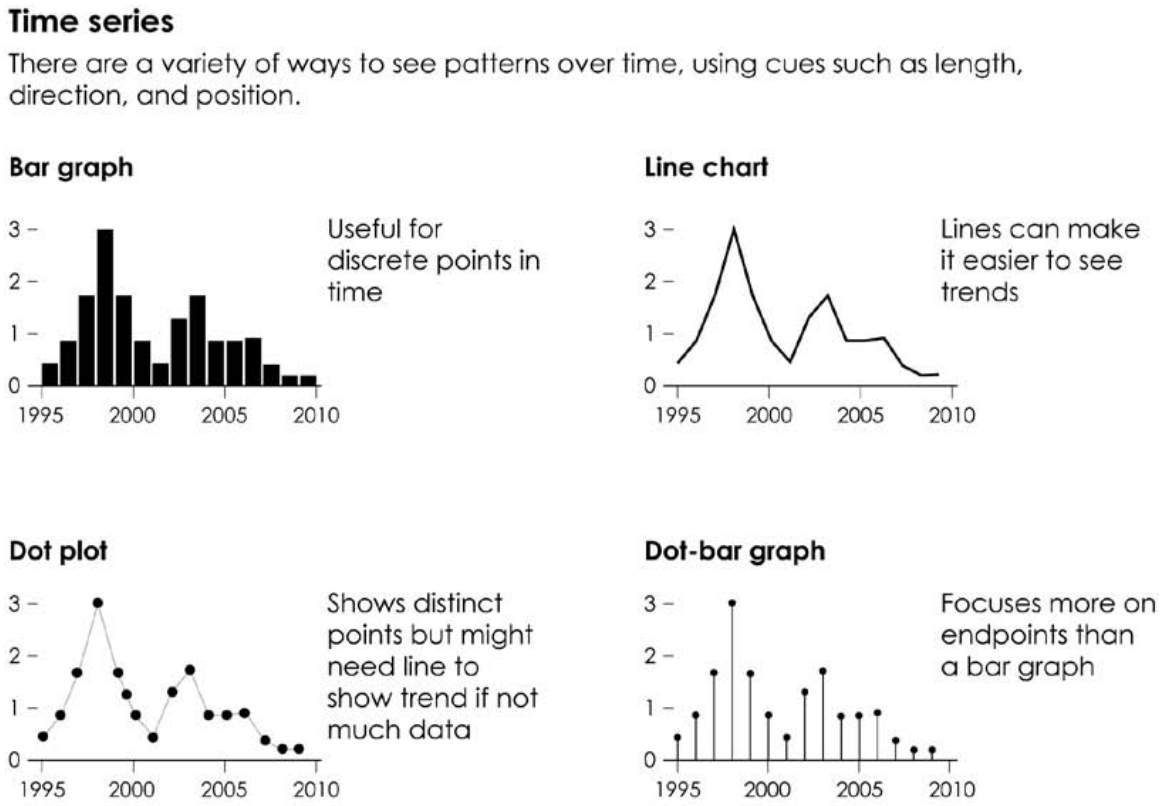
From the graphic above, it's clear that:

- Housetasks such as dinner, breakfeast, laundry are done more often by the wife
- Driving and repairs are done more frequently by the husband

Analyse de séries temporelles

Données temporelles

- Histogrammes
- Graphe
- Dot plot
- Dot-bar graphs



[Nathan YAU (2013)
« Data points. Visualization that means something », p. 155]

Données temporelles

Peu de points

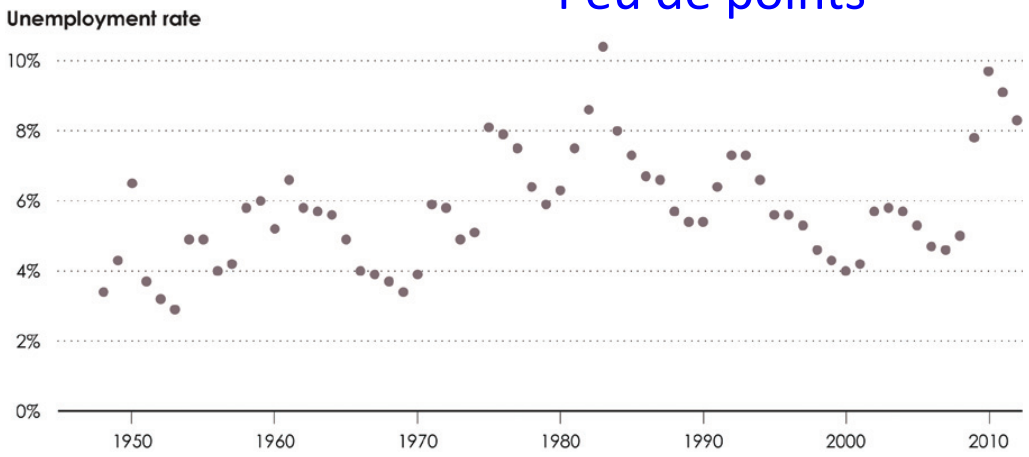


FIGURE 4-19 *Sparse dot plot*

Les points connectés

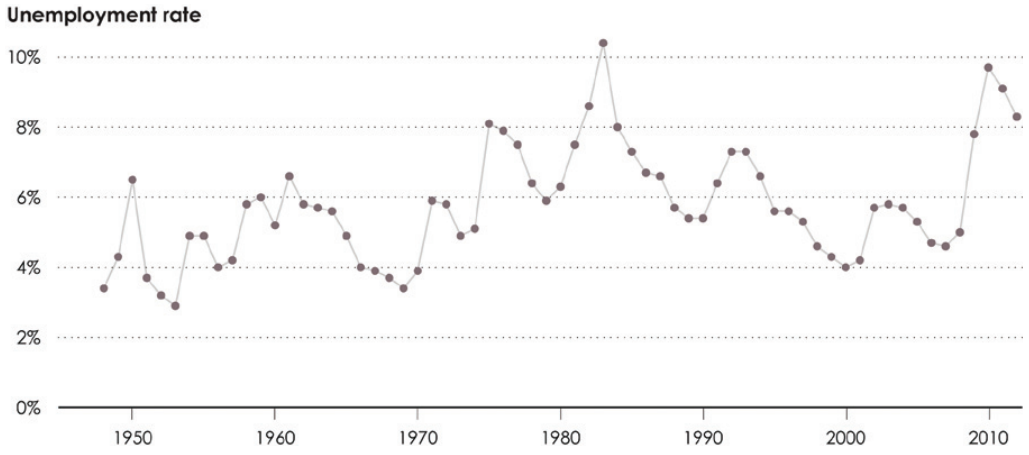
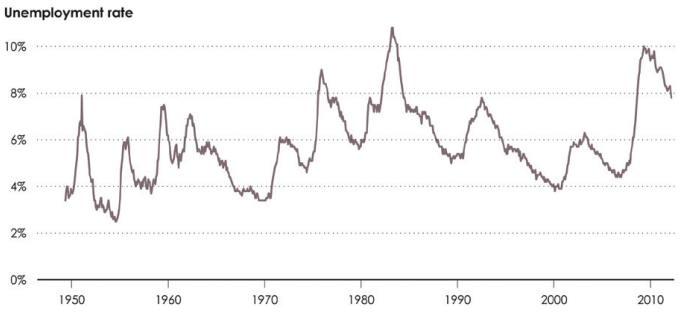


FIGURE 4-20 *Sparse dot plot with connecting line*

[Nathan YAU (2013)
« Data points. Visualization that means something », p. 160]

Données temporelles



Données lissées par LOESS

LOESS ou **LOWESS** (*Locally Weighted Scatterplot smoothing*) : interpolation par une fonction polynomiale

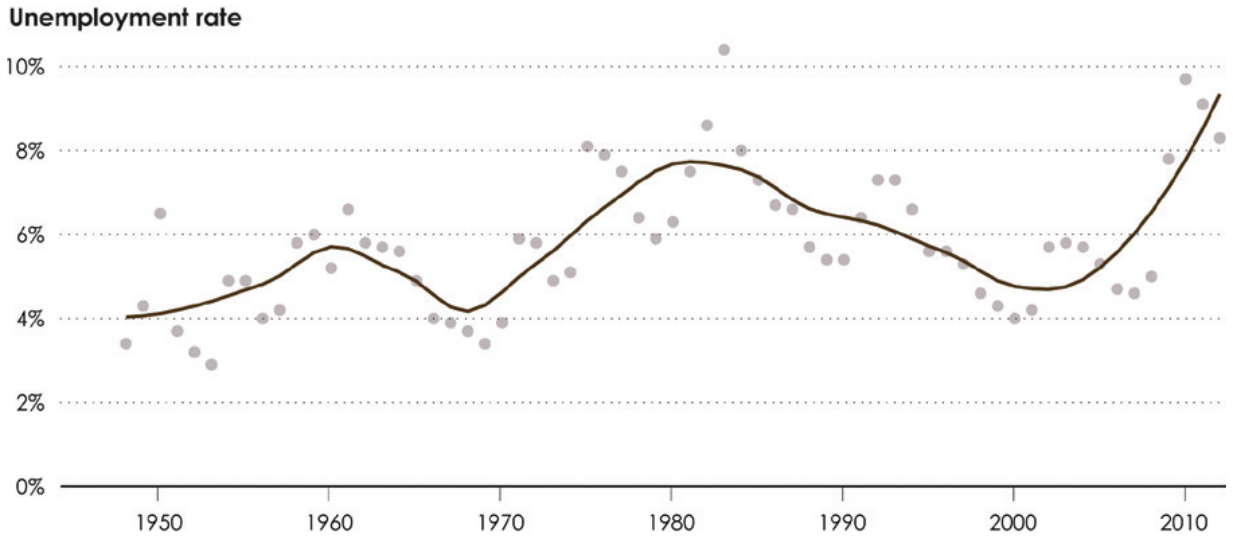


FIGURE 4-21 *Fitted LOESS curve*

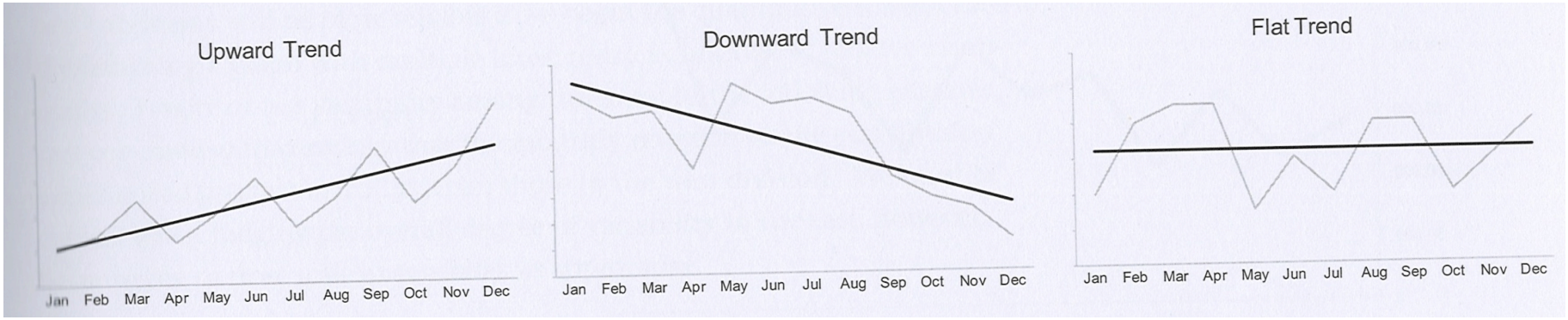
[Nathan YAU (2013)
« Data points. Visualization that means something », p. 160]

Six types de patterns pour les séries temporelles

1. Tendances
2. Variabilité
3. Taux de changement
4. Co-variation
5. Cycles
6. Exceptions

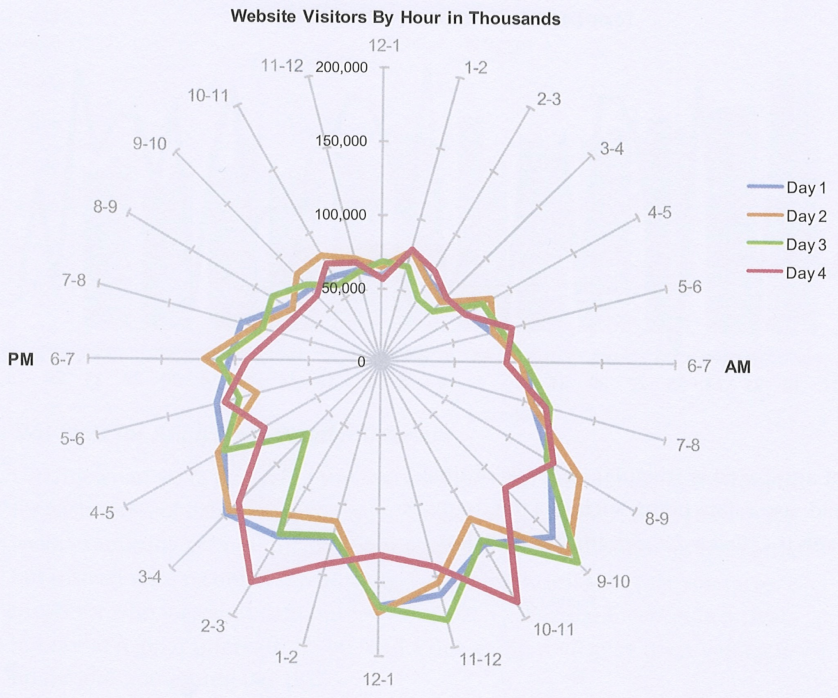
Tendances

[Stephen FEW (2009)
« Now you see it », p. 143-144]

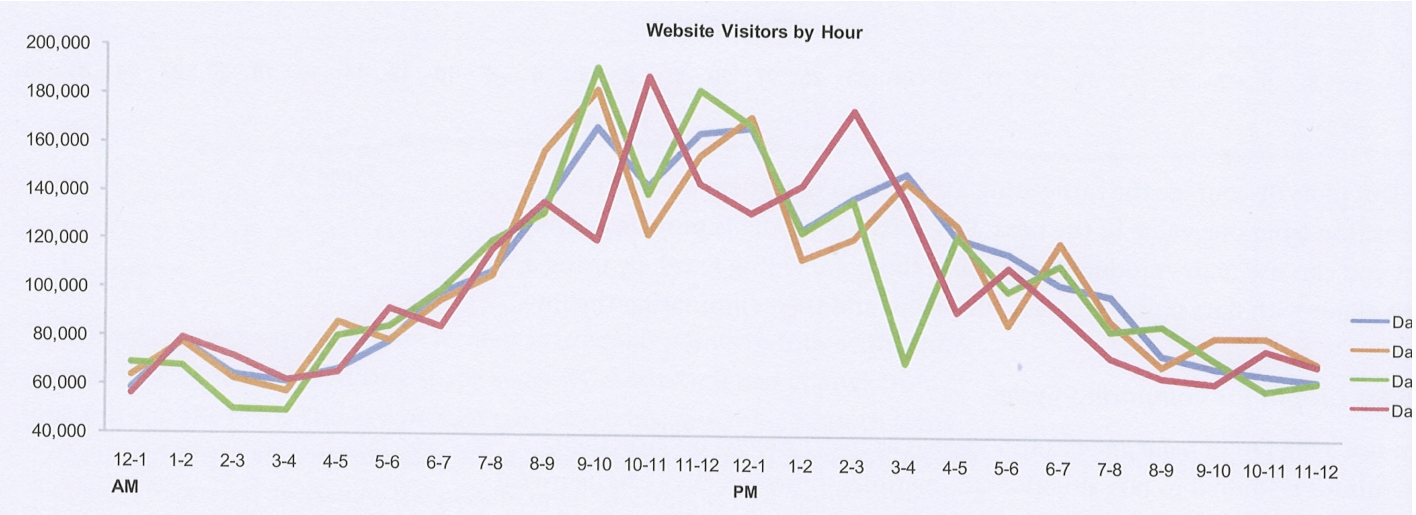


Exemple

Cycles



[Stephen FEW (2009)
« Now you see it », p. 153-155]



Exemple

Analyser des distributions

Formes de distribution

[Stephen FEW (2009)
« Now you see it », p. 219-224]

1. Courbées ou plates ?
2. Si courbée, convexe ou concave ?
3. Si courbée, avec un pic simple ou pics multiples ?
4. Si un seul pic, symétrique ou biaisé
5. Des concentrations ?
6. Exceptions ?

Exemple

Distributions

In 2010, Gawker Media, which runs large blogs like Lifehacker and Gizmodo, was hacked, and 1.3 million usernames and passwords were leaked. They were downloadable via BitTorrent. The passwords were encrypted, but the hackers cracked about 188,000 of them, which exposed more than 91,000 unique passwords.

What would you do with that kind of data?

[Nathan YAU (2013)
« Data points. Visualization that means something », p. 40]

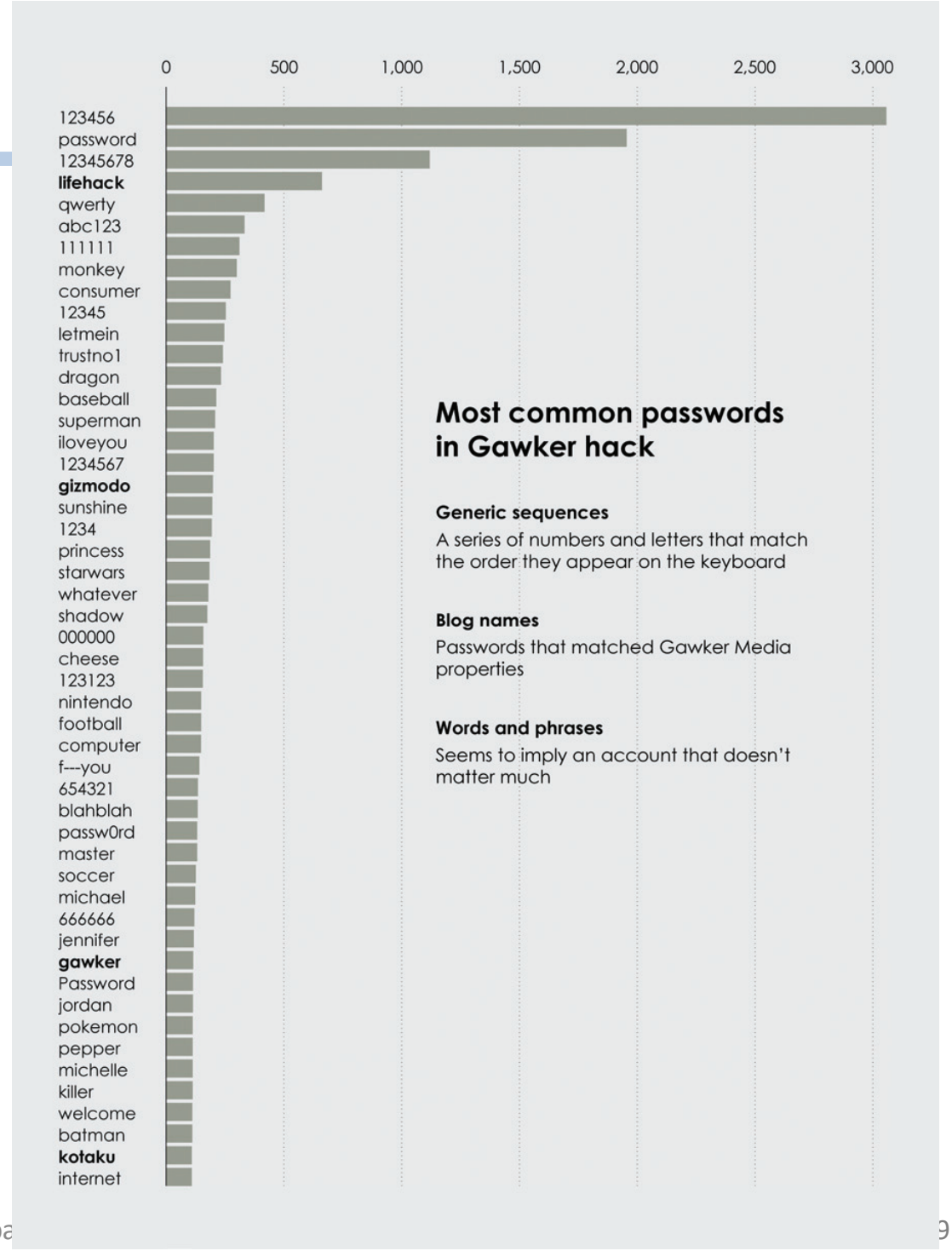
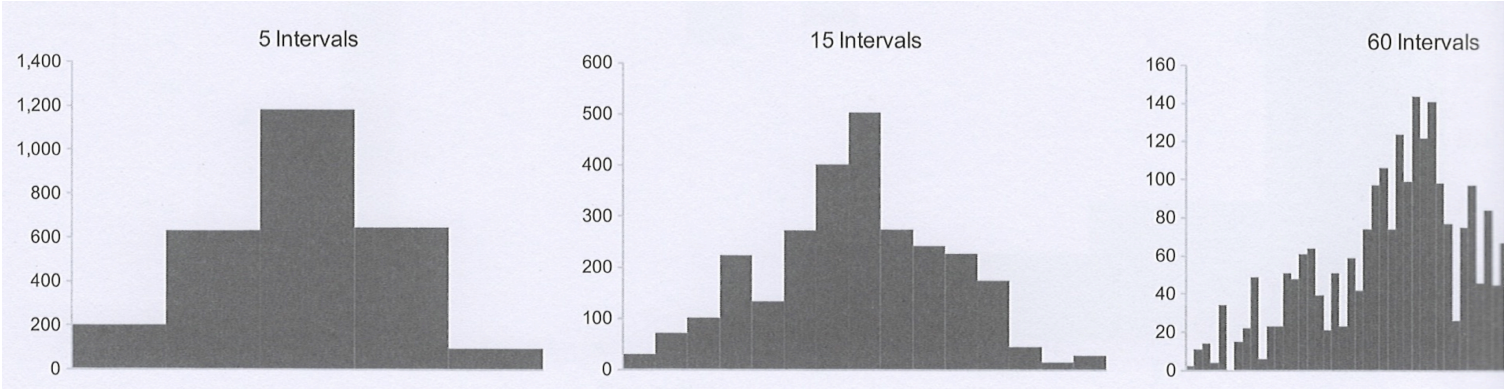


FIGURE 1-27 *Commonly used passwords in Gawker hack*

Choix du meilleur intervalle

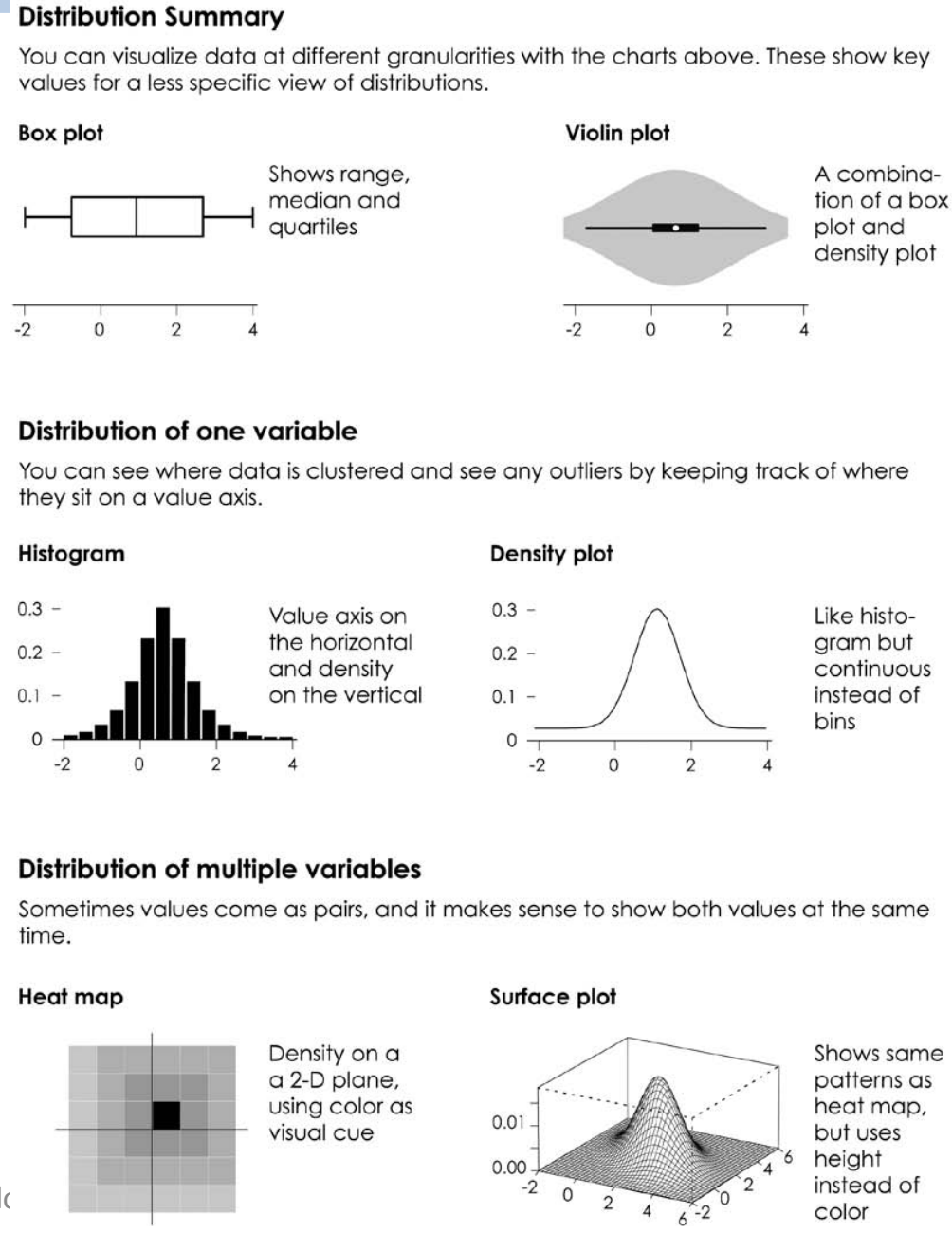
[Stephen FEW (2009)
« Now you see it », p. 242]



Exemple

Analyse de distributions

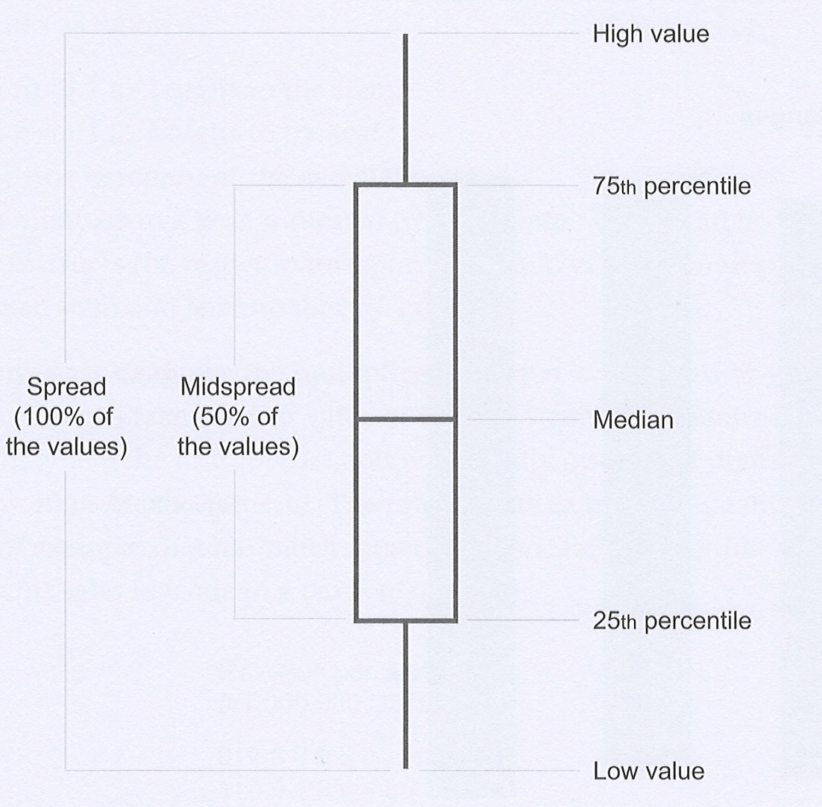
- Visualisations



[Nathan YAU (2013)
« Data points. Visualization that means something », p. 195]

Boîtes à moustaches (*box-and-whistler plots*)

[Stephen FEW (2009)
« Now you see it », p. 232-234]



Exemple

Analyse Multivariée

Analyser plusieurs **données** selon plusieurs **attributs** à la fois

Analyse multidimensionnelle

Nombreuses

variables

- Heatmaps

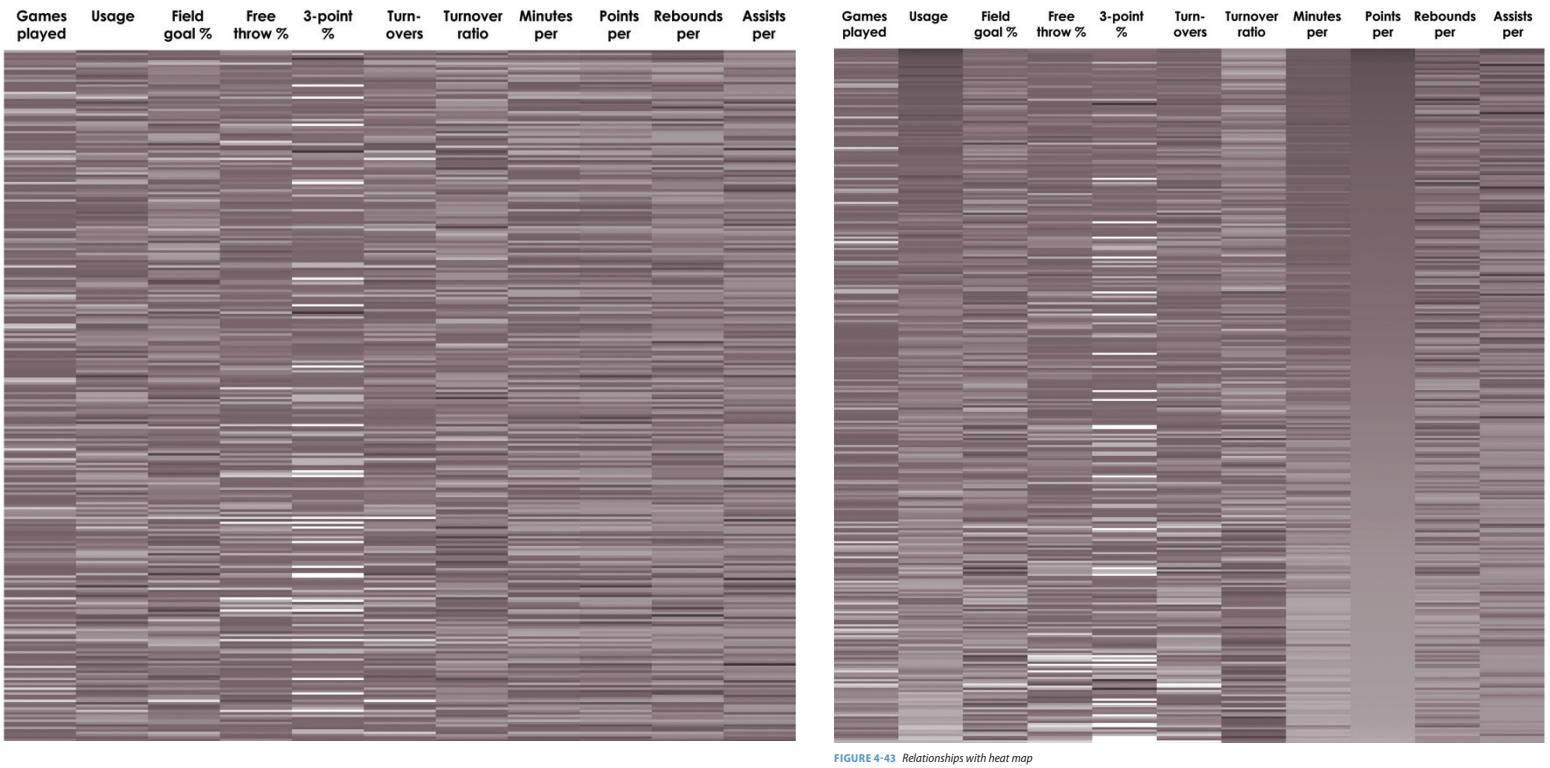


FIGURE 4-43 *Relationships with heat map*

Recherche de corrélations par rapport aux points gagnés

[Nathan YAU (2013)
« Data points. Visualization that means something », p. 183-184]

Analyse multidimensionnelle

Nombreuses

variables

- Parallel coordinates

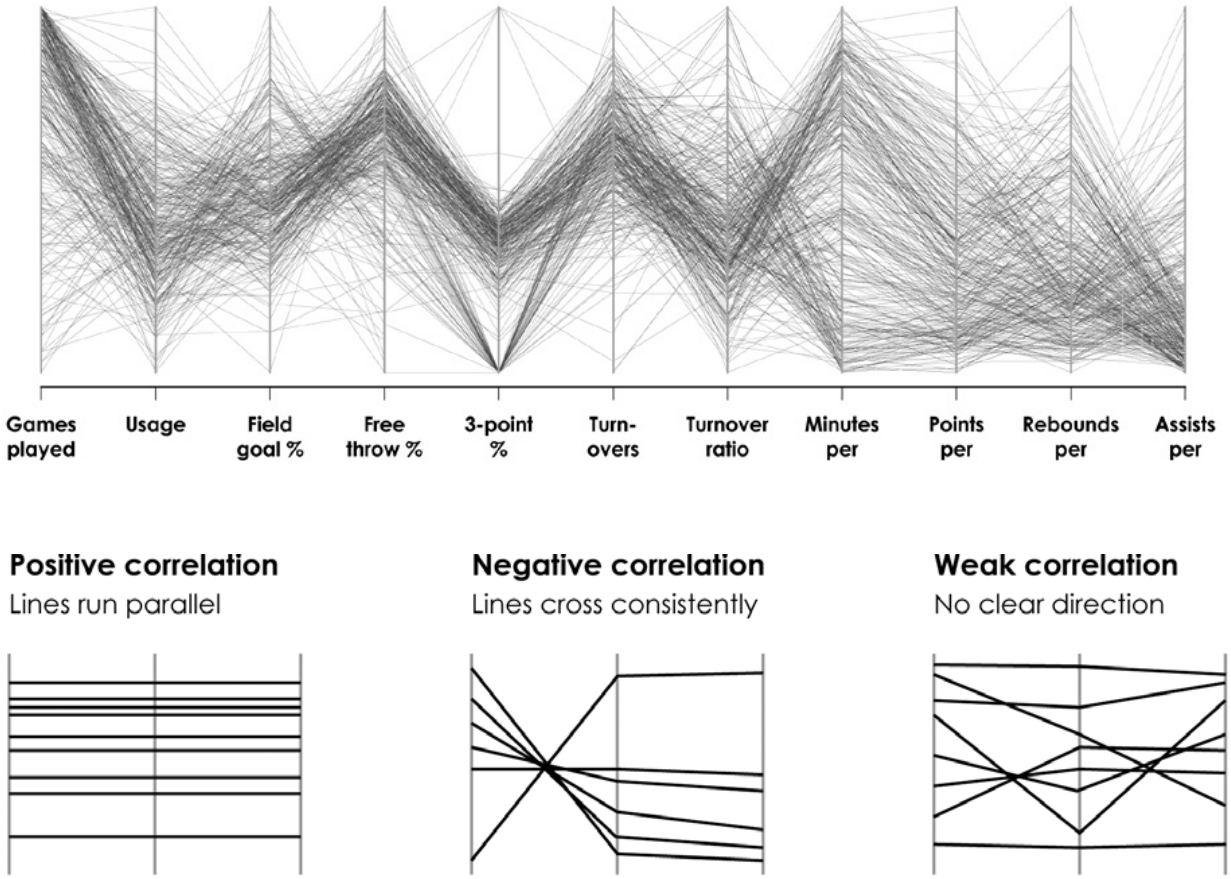


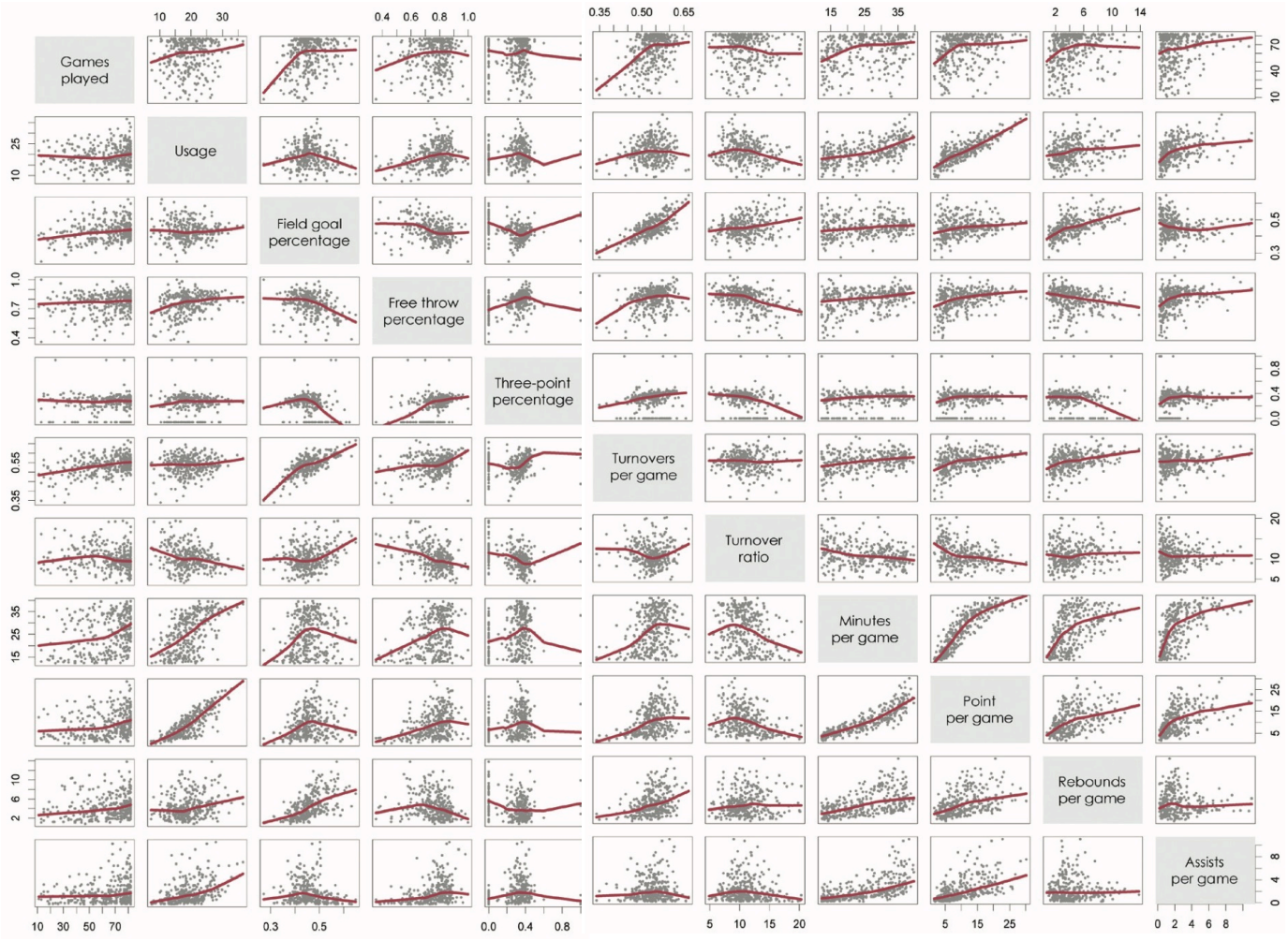
FIGURE 4-45 *Relationships with parallel coordinates plot*

[Nathan YAU (2013)
« Data points. Visualization that means something », p. 185]

Analyse multidimensionnelle

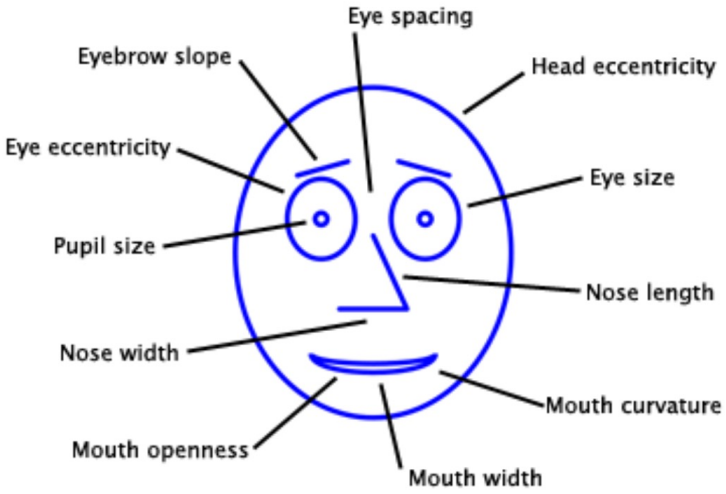
Analyse de **corrélations**

- Matrice de corrélation

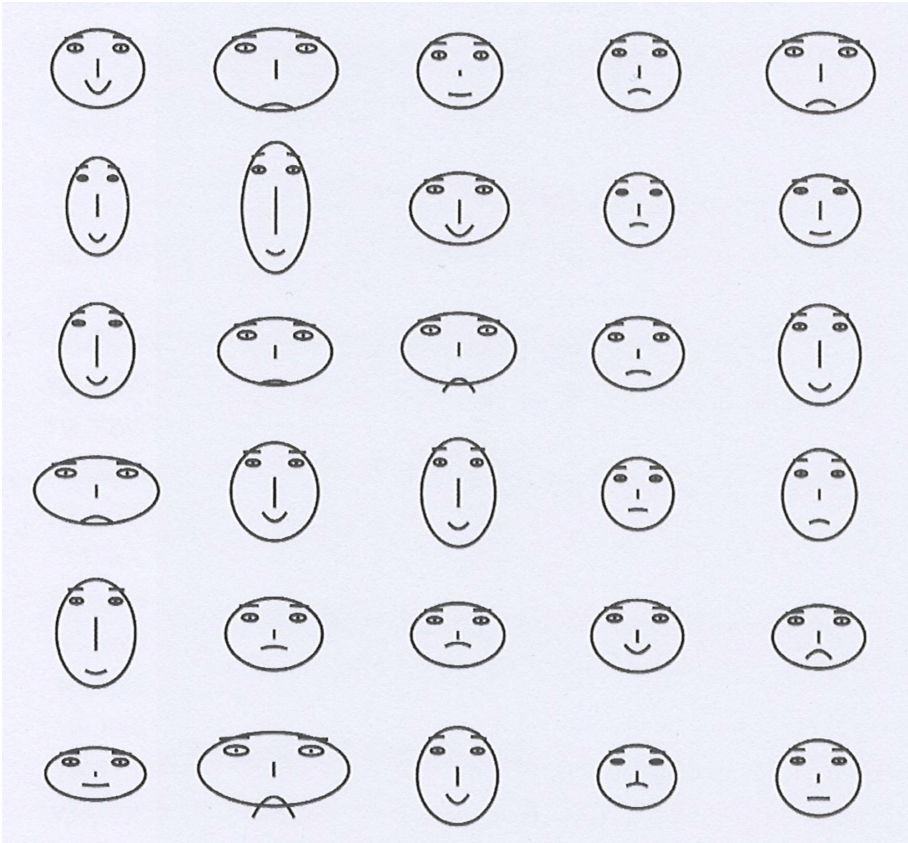


[Nathan YAU (2013)
« Data points. Visualization that means something », p. 191-192]

Les visages de Chernoff (1973)



[Stephen FEW (2009)
« Now you see it », p. 283]

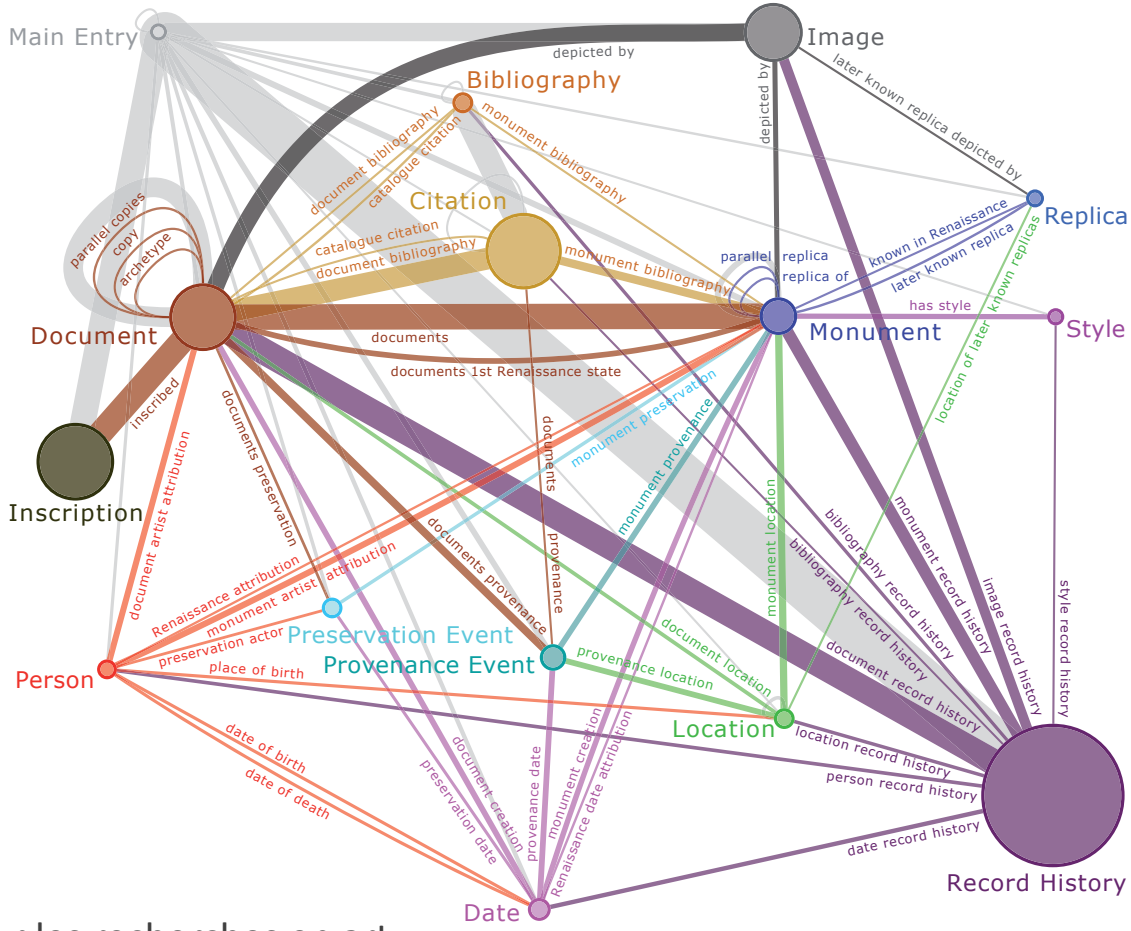


Pas certain que ça marche bien

Visualisations

- Tag explorer

Visualisations



- Base de données sur les recherches en art
- Les **nœuds** (types d'objets étudiés)
- Les **liens** correspondent à des relations entre nœuds
- Outil de dessin de graphe : Cytoscape

Plan

1. Puissance (et limites) de l'appareil visuel
2. Qu'est-ce qu'une représentation
3. Les primitives visuelles
4. Types de données et types de visualisation
   - Démarche
   - Données catégorielles
   - Séries temporelles
   - Données multi-variées
5. Outils et librairies

outils et librairies

Logiciels et langages

- Microsoft Excel
- Google Spreadsheets
- Tableau Software
- IBM Many Eyes
- ImagePlot
- Treemap

- **R**
- **Python**. Librairies pandas, seaborne
- …

Références

- Stephen FEW (2009) : Now you see it. Simple visualization techniques for quantitative analysis. Analytic Press, 2009.
- Alexandru TELEA (2007) : Data visualization. Principles and practice. CRC Press, 2007.
- Nathan YAU(2011) : Visualize this. The FlowingData guide to design, visualization, and statistics. Wiley, 2011.
- Nathan YAU(2013) : Data points. Data visualization that means something. Wiley, 2013.