



# Data Visualization

**Vandy BERTEN**

Section Recherche

# Table des matières

---

Introduction

Les fondamentaux

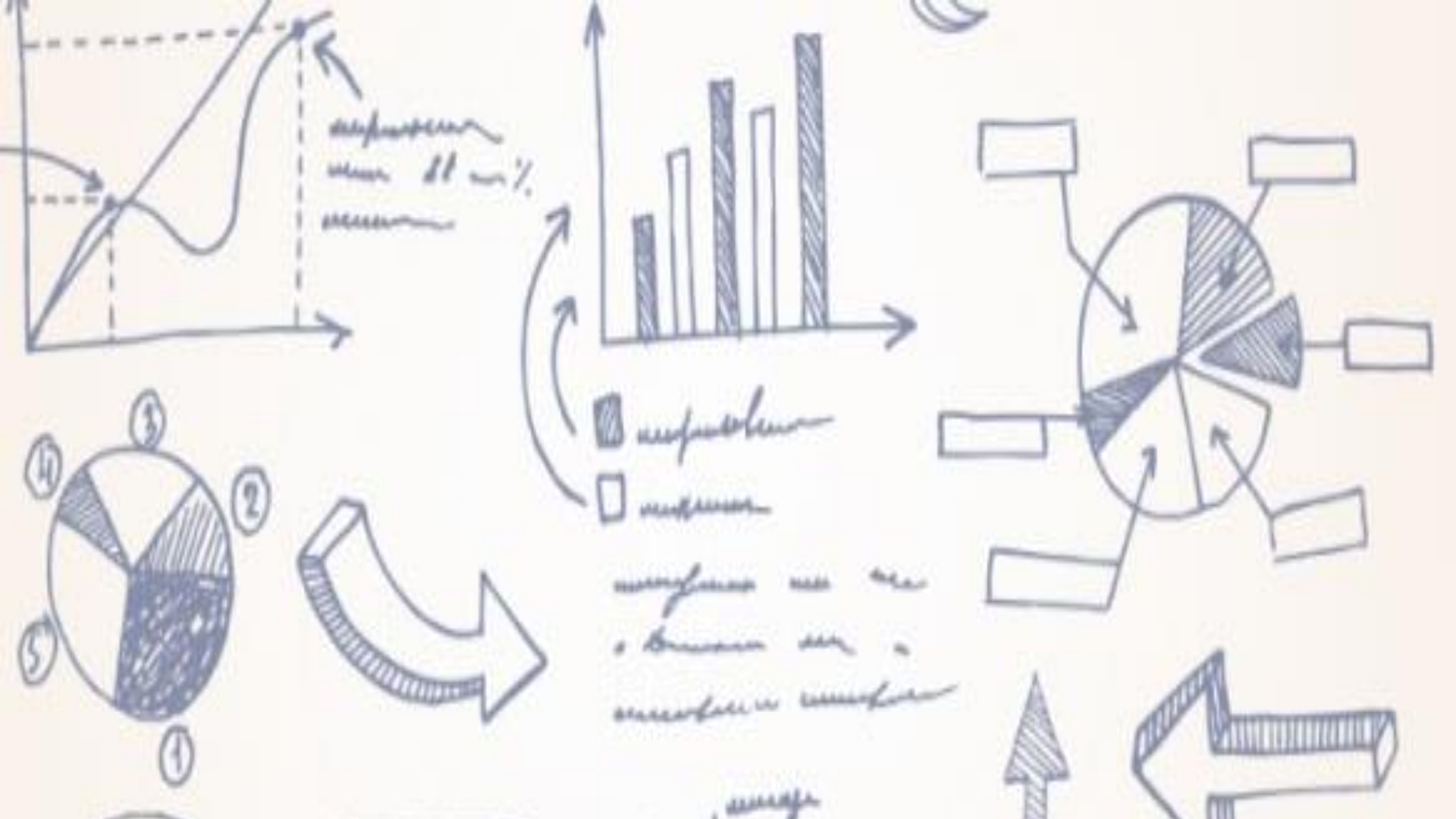
Choix de graphique

Visualisation géographique

Visual Analytics

Conclusions



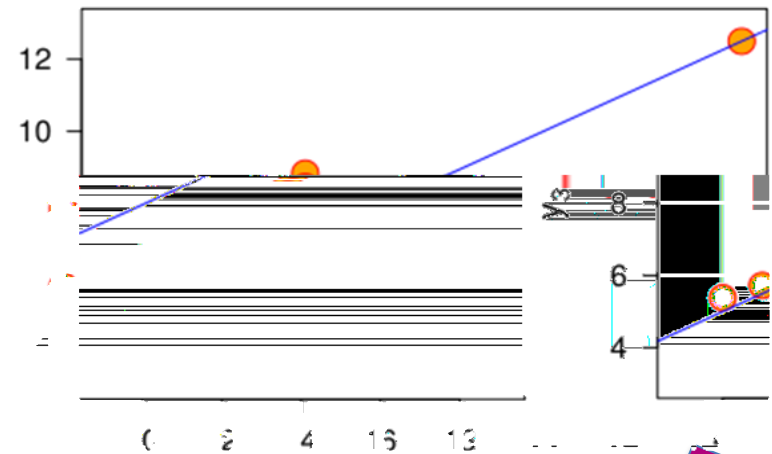
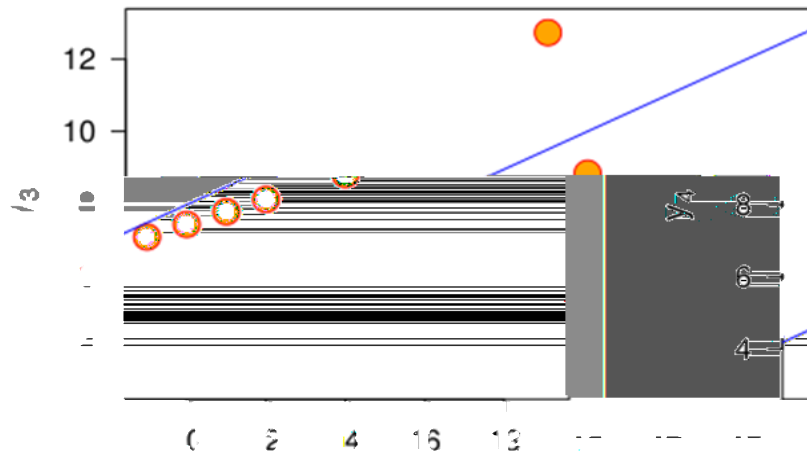
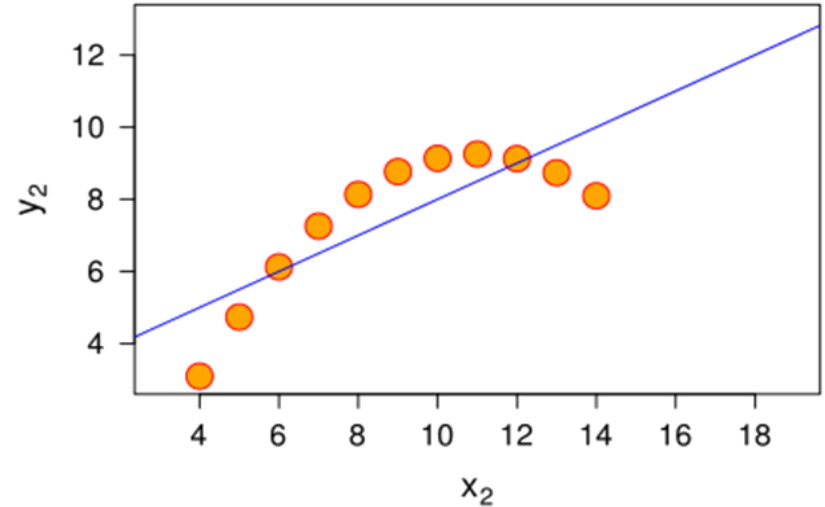
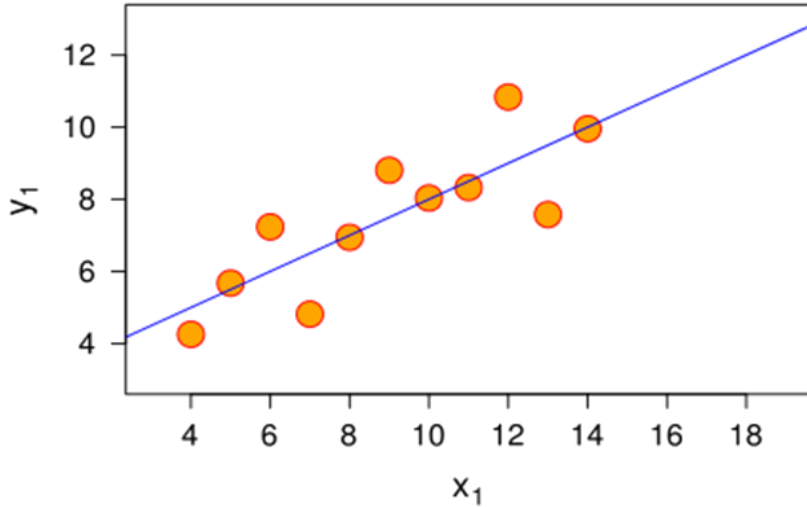


# Introduction

# Limites des statistiques (Anscombe)

$x_1$	$y_1$	$x_2$	$y_2$	$x_3$	$y_3$	$x_4$	$y_4$
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				8.0	7.71
9.0	8.81	9.0				8.0	8.84
11.0	8.33	11.0				8.0	8.47
14.0	9.96	14.0				8.0	7.04
6.0	7.24	6.0				8.0	5.25
4.0	4.26	4.0				19.0	12.50
12.0	10.84	12.0				8.0	5.56
7.0	4.82	7.0				8.0	7.91
5.0	5.68	5.0				8.0	6.89
			<b>Propriété</b>	<b>Valeur</b>			
			Moyenne des x	9.0			
			Variance des x	10.0			
			Moyenne des y	7.5			
			Variance des y	3.75			
			Corrélation entre les x et les y	0.816			
			Équation de la droite de régression linéaire	$y = 3 + 0.5x$			
			Somme des carrés des erreurs relativement à la moyenne	110.0			

# Limites des statistiques (Anscombe)

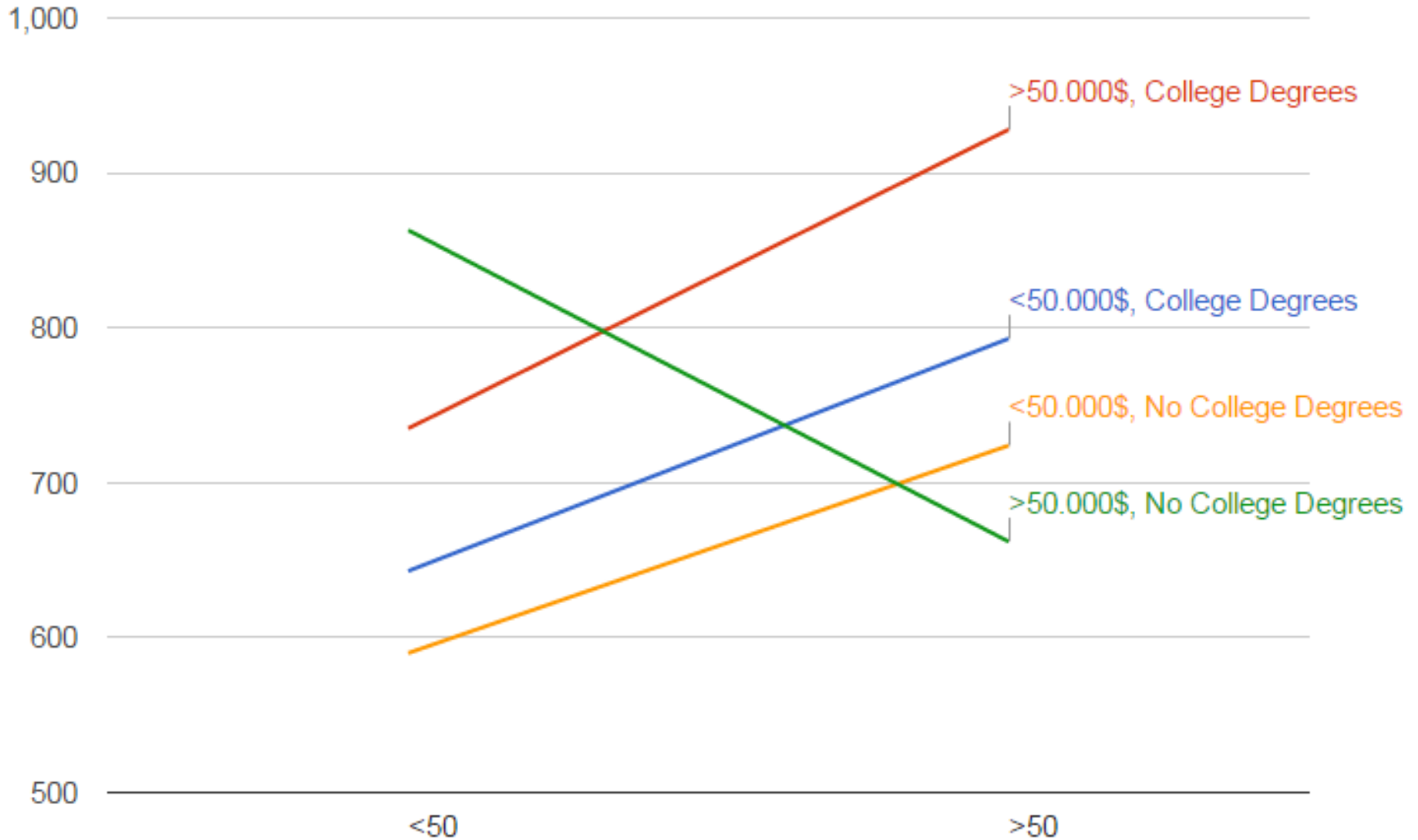


# Limites de la cognition

Satisfaction	College Degrees		No College degrees	
	≤50	>50	≤50	>50
Income/Age				
< 50,000\$	643	793	590	724
> 50,000\$	735	928	863	662

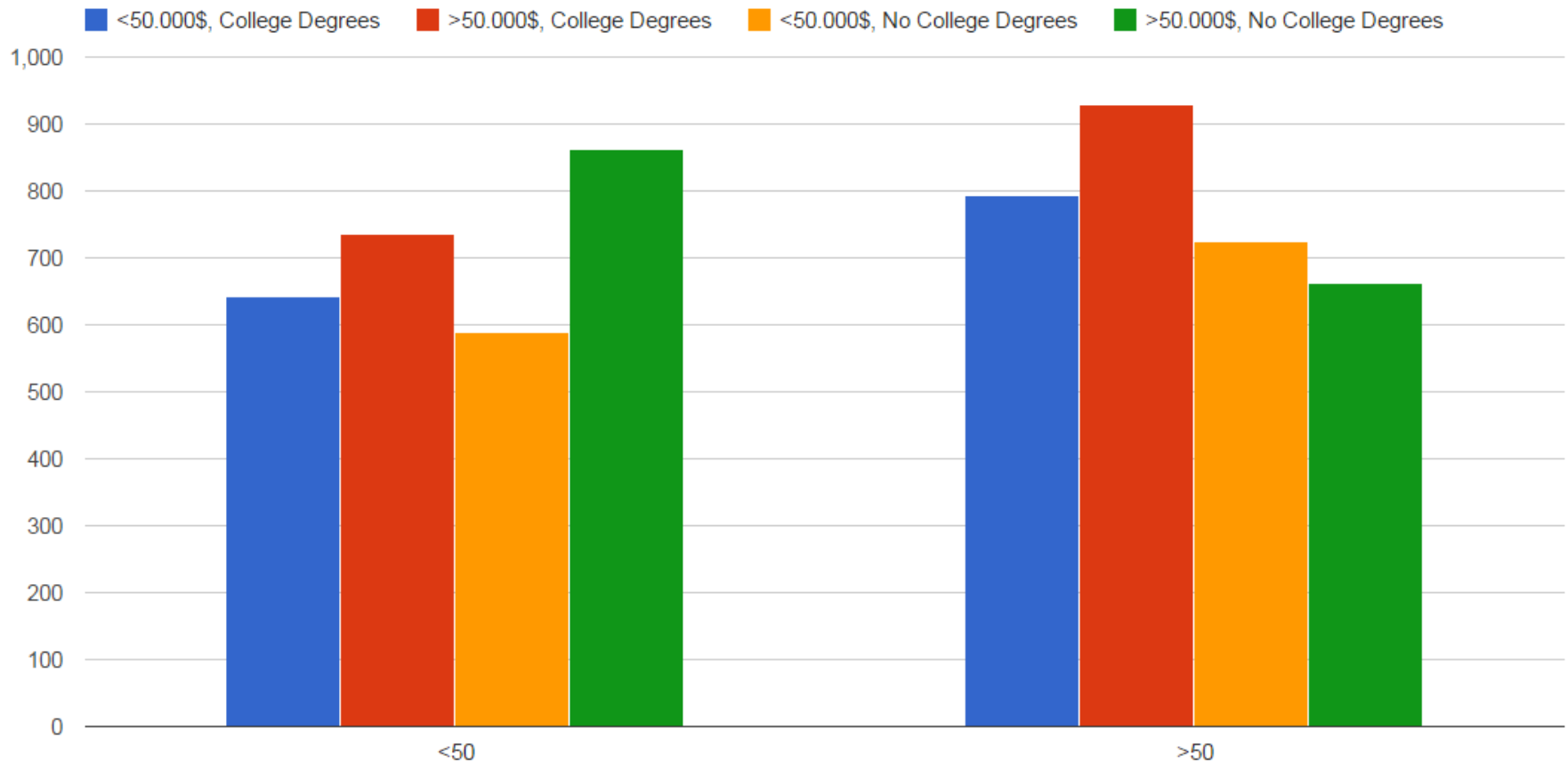
- Est-ce que la satisfaction de tous les employés augmente avec l'âge ?
- Un groupe a-t-il un *pattern* différent des autres ?

# Limites de la cognition



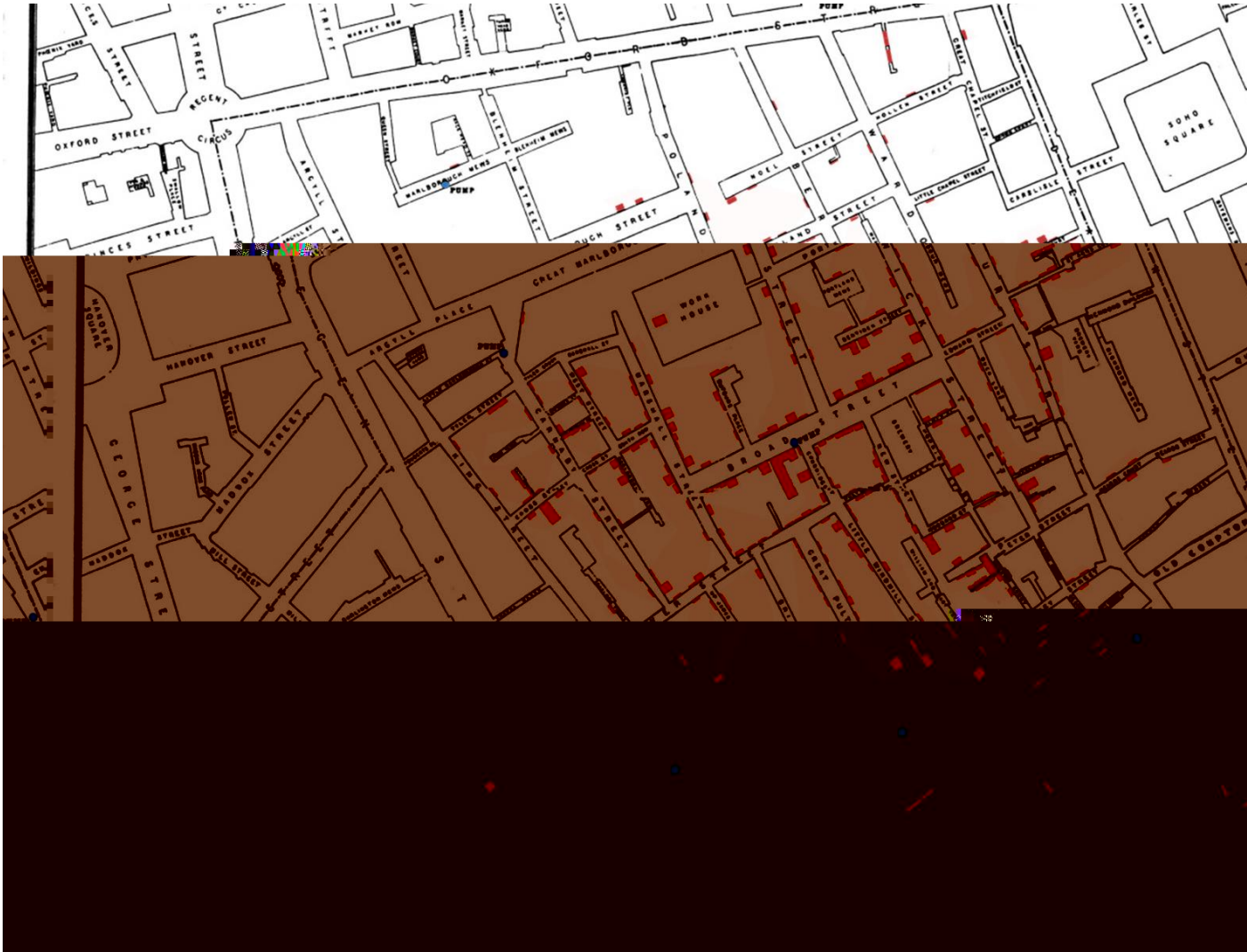
Outil : Google Chart

# Limites de la cognition

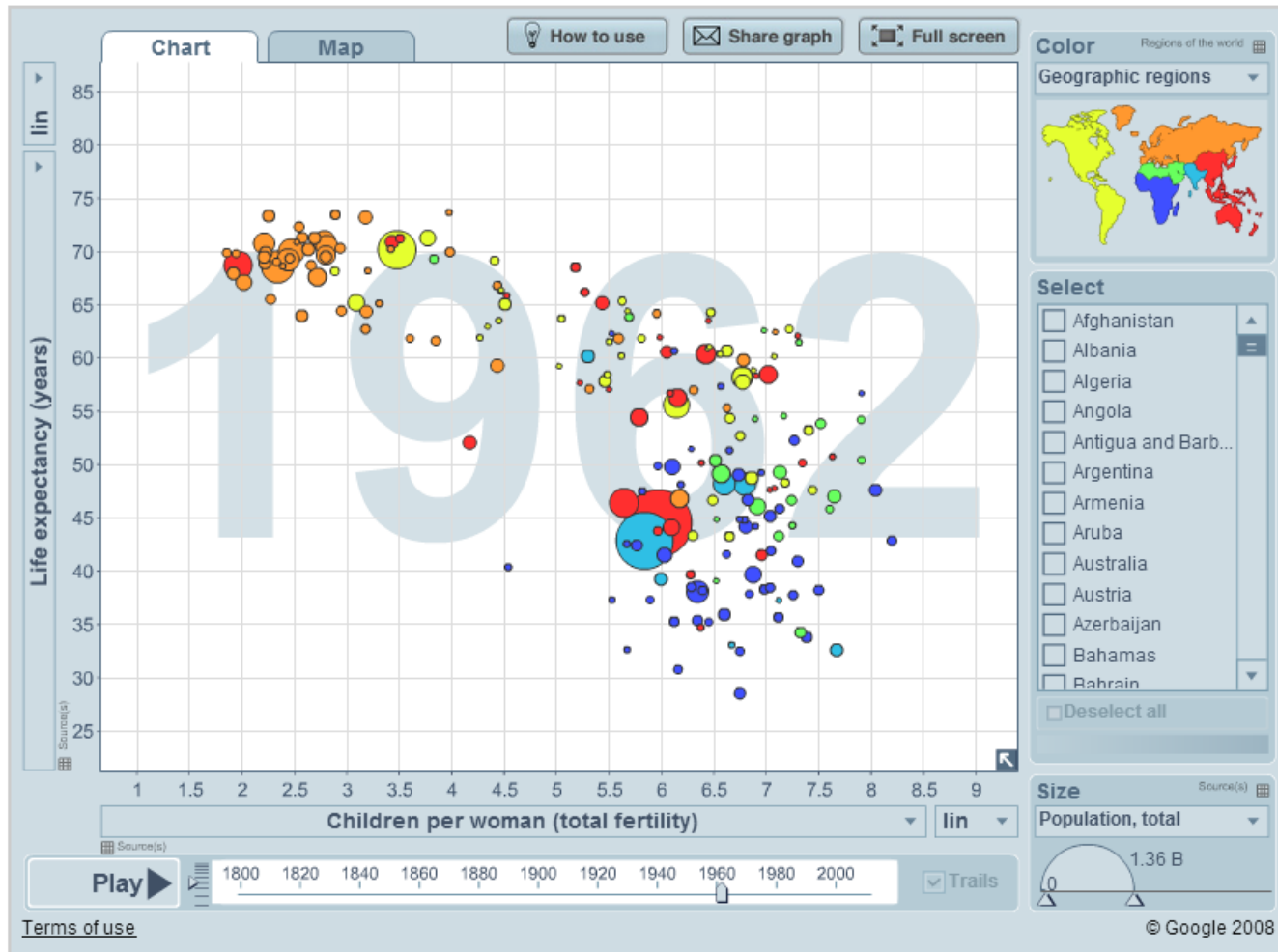


Outil : Google Chart

# Santé (John Snow, 1854)

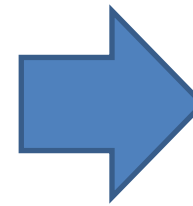
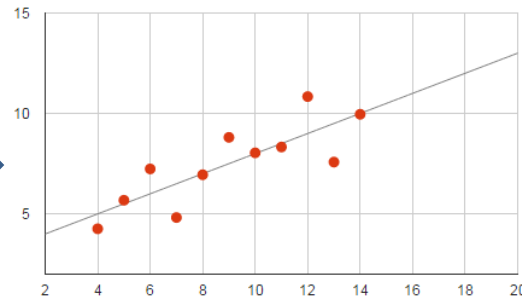
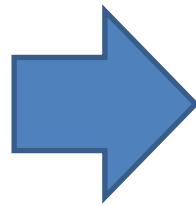


# Démographie (H. Rosling, Gapminder)



# Data visualization : objectifs

x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68



Données

Visualisation

Information

# Data visualization : objectifs

---

Enregistrer de  
l'information

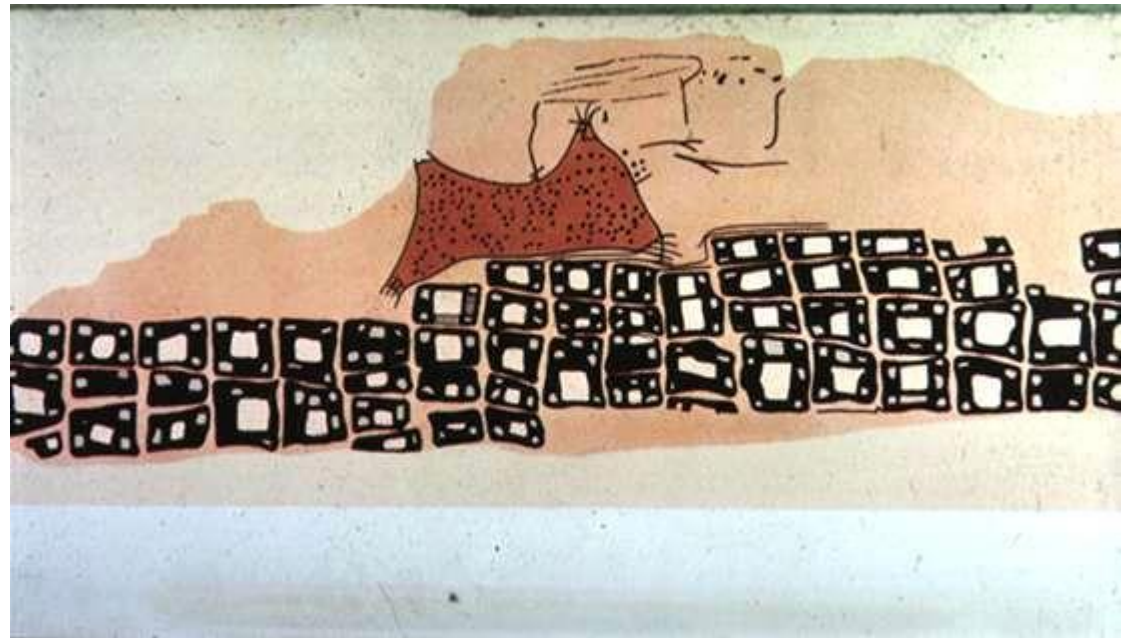
Communiquer  
des informations  
(visual explanation)

Analyser des  
données  
(visual analytics,  
visual exploration)

# Enregistrer de l'information



Anaximander's Map of the World



<http://www.datavis.ca/milestones/>







1	233
2	0120678888
3	014899
4	0000199
5	01222266779
6	000033456798
7	0123333455678999
8	112356778
9	0125

Occurrences

	ABC	1234	X45
Category		543,2109	7%
Group		45,67	45%
Unit		9876	98%
Class		123,78	12%

# Les fondamentaux

# Outline

---

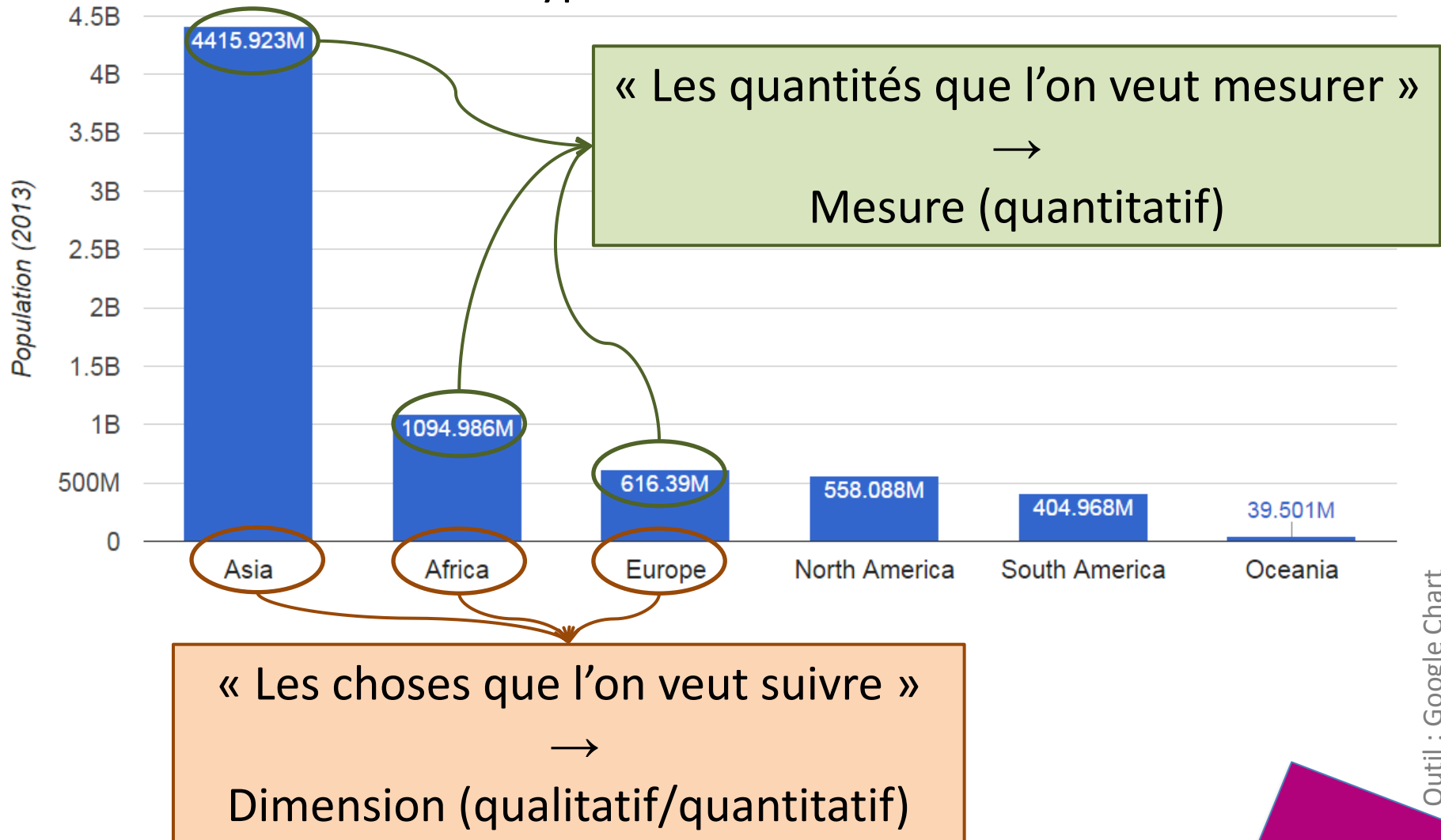
- Données :
  - Classification par fonction (Mesure & dimension)
  - Classification par nature (Quantitatif, qualitatif)
- Graphique :
  - Éléments graphiques (sémiologie)
  - Perception (Gestalt, processus pré-attentif)
- Qualité :
  - Mesures de qualité
  - Exemples

Les fondamentaux

# MESURES & DIMENSIONS

# Mesures & Dimensions

Deux types d'informations :



Outil : Google Chart

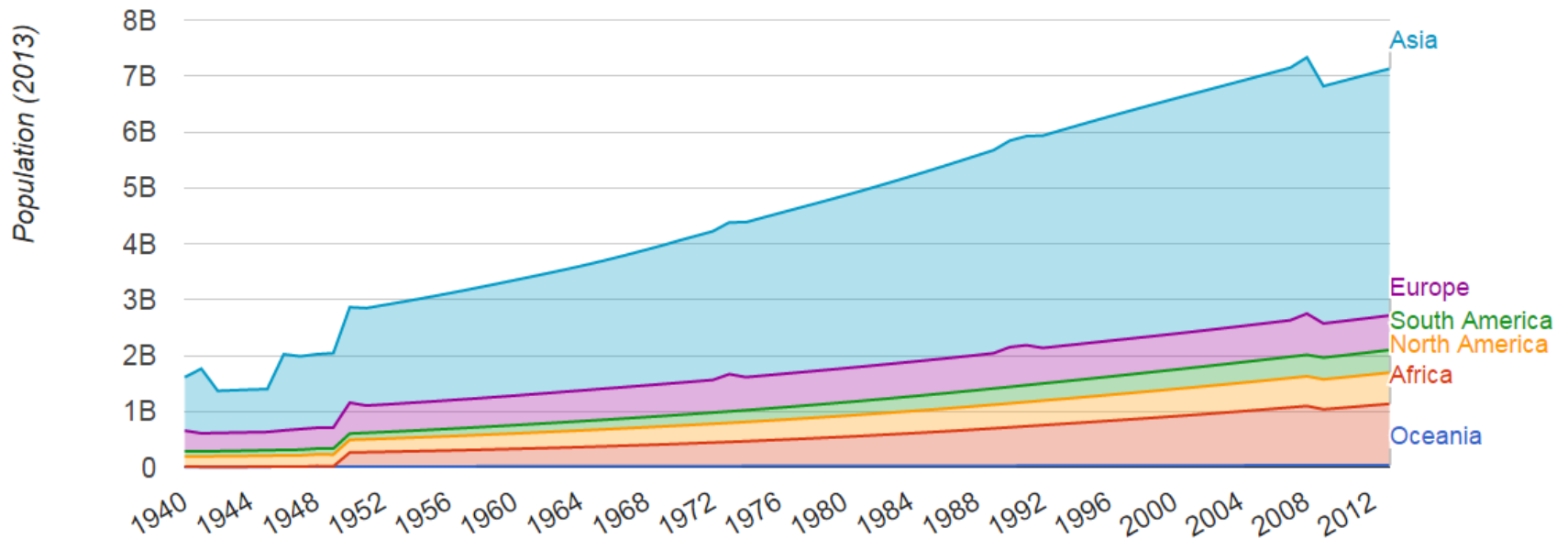
# Mesures & Dimensions

---

Mesures		Dimensions
La population	par	continent
Le PIB	par	pays
Le nombre de personnes	par	genre
Les bénéfices et les pertes	par	date
Les marges	par	trimestre et département

# Mesures & Dimensions

Une mesure  
(Population)



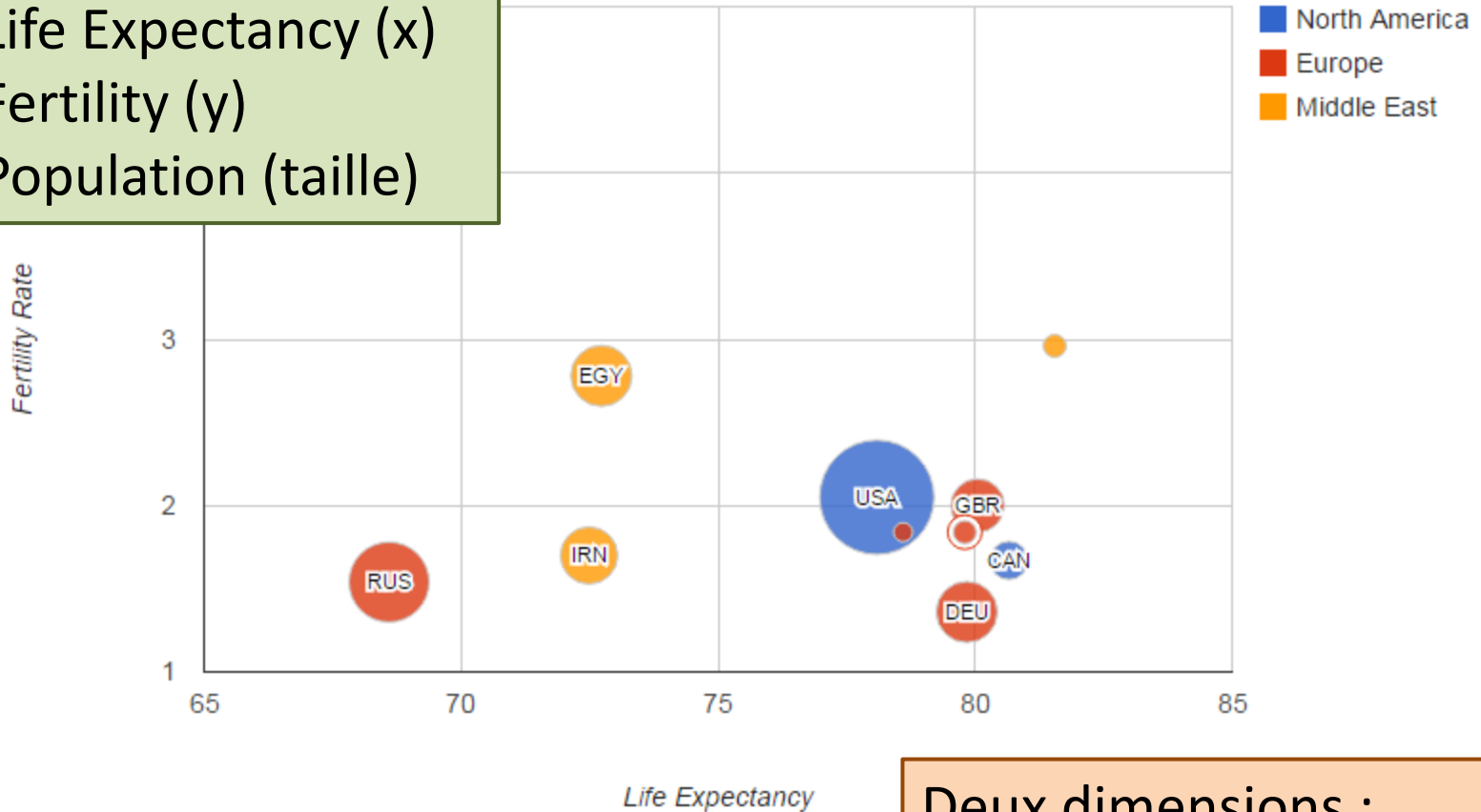
Deux dimensions (année, continent)

Outil : Google Chart

# Mesures & Dimensions

Trois mesures :

1. Life Expectancy (x)
2. Fertility (y)
3. Population (taille)



Deux dimensions :

1. Pays (point)
2. Continent (couleur)



Les fondamentaux

# ***SCALE OF MEASUREMENT***



# Scale of measurement

## Qualitative

### Nominal

No order

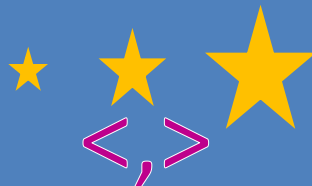


=, ≠

### Ordinal

Intrinsic order

1r, 2d, 3e



## Quantitative

### Intervalle

Gap comparison  
Conventional zero

22°C

1/1/2011, 3:50

50°55'21,2''

-

=, ≠, <, >

### Ratio

Size comparison  
Intrinsic zero

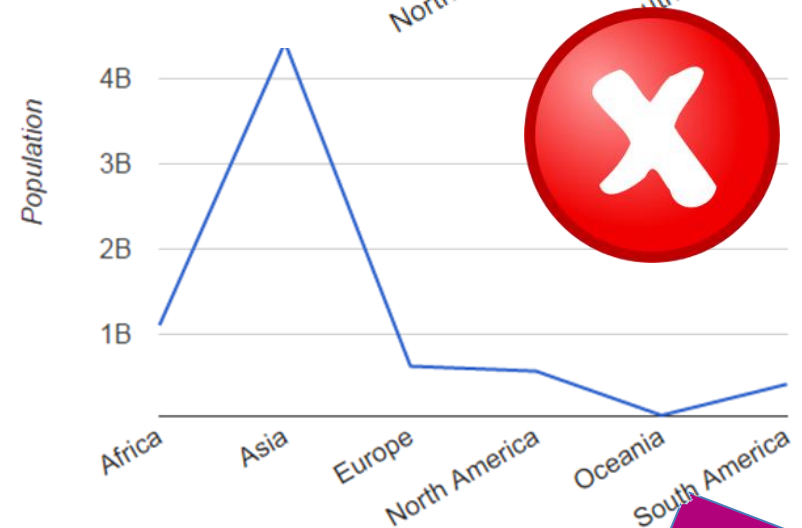
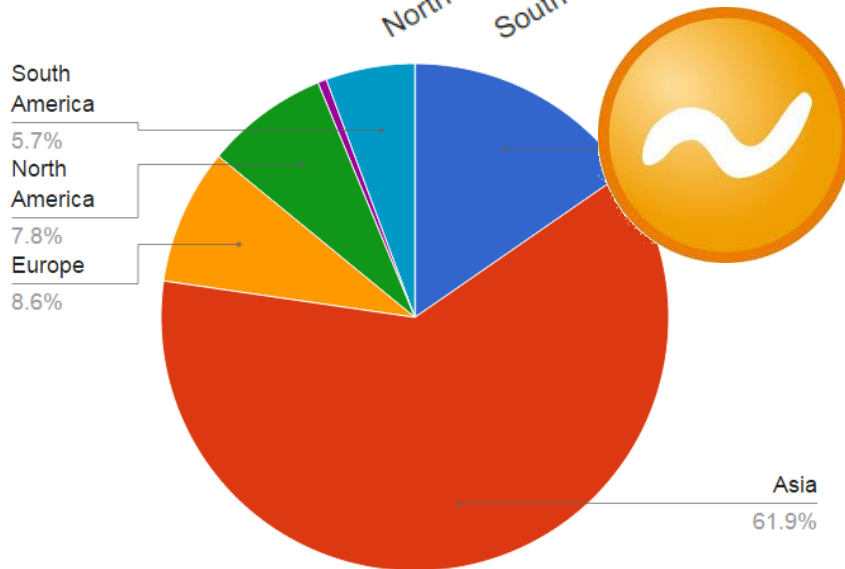
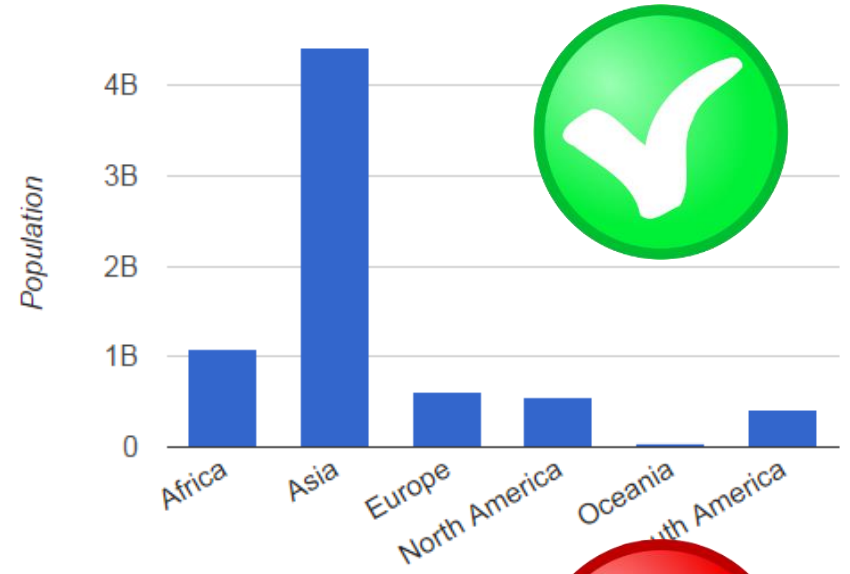
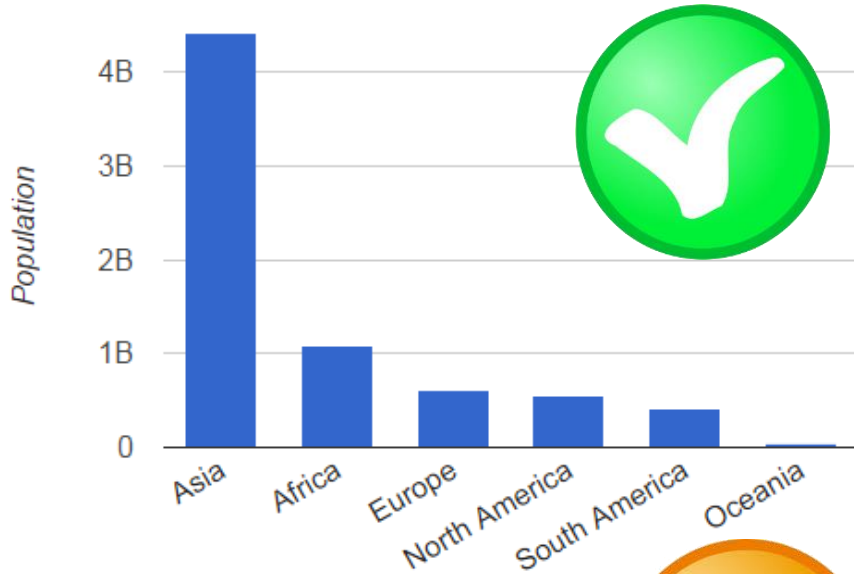
22Kg

25 secondes

105€

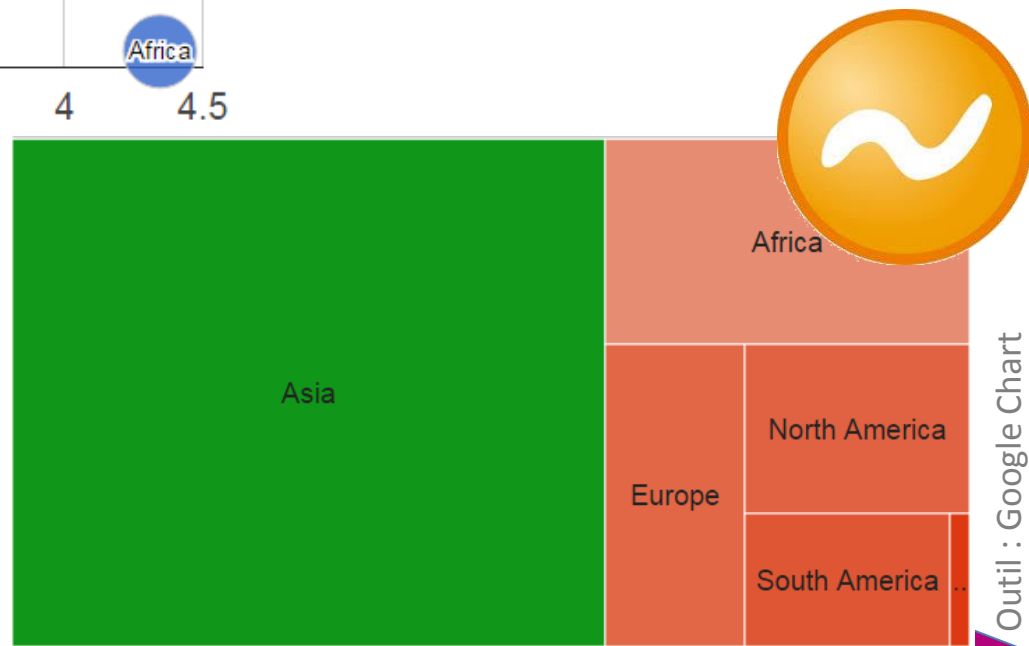
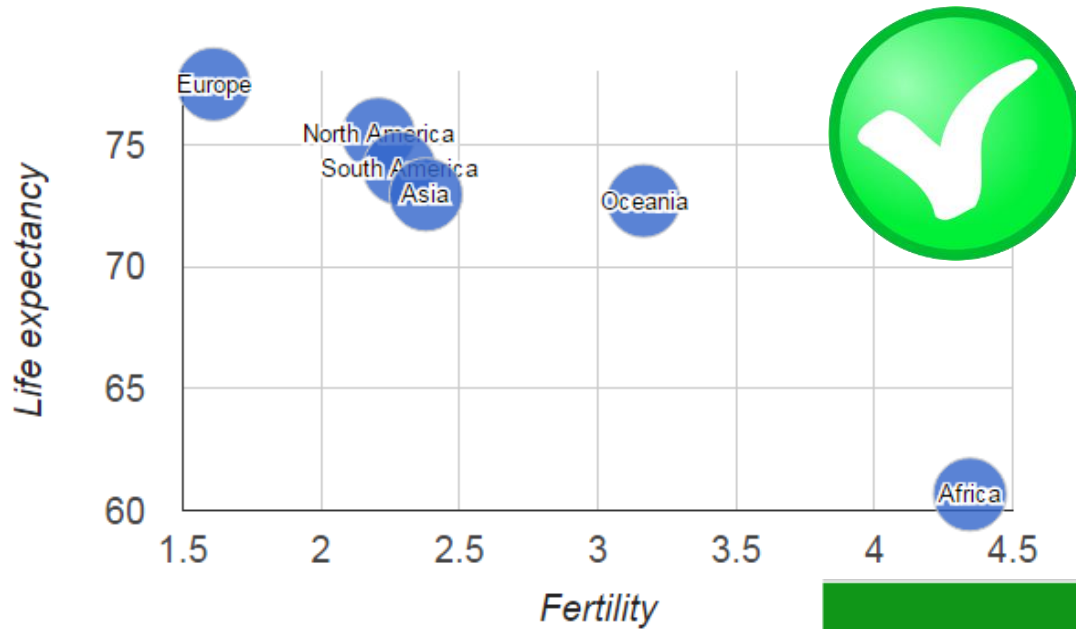
-, +, /

# Nominal : Dimension



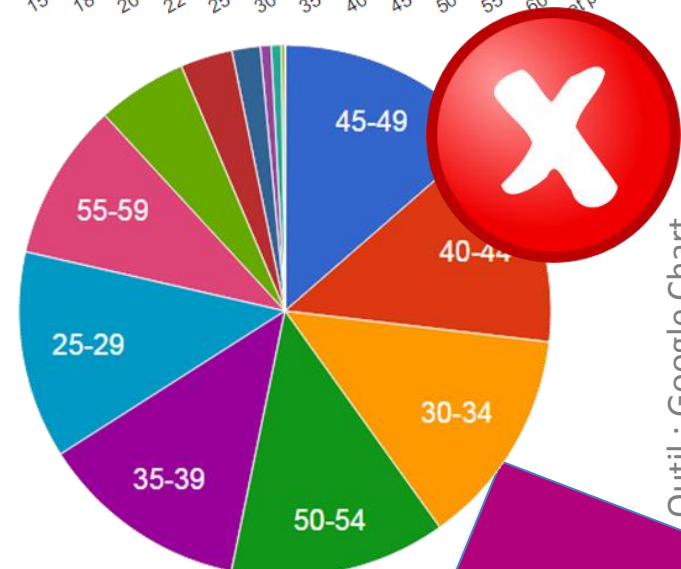
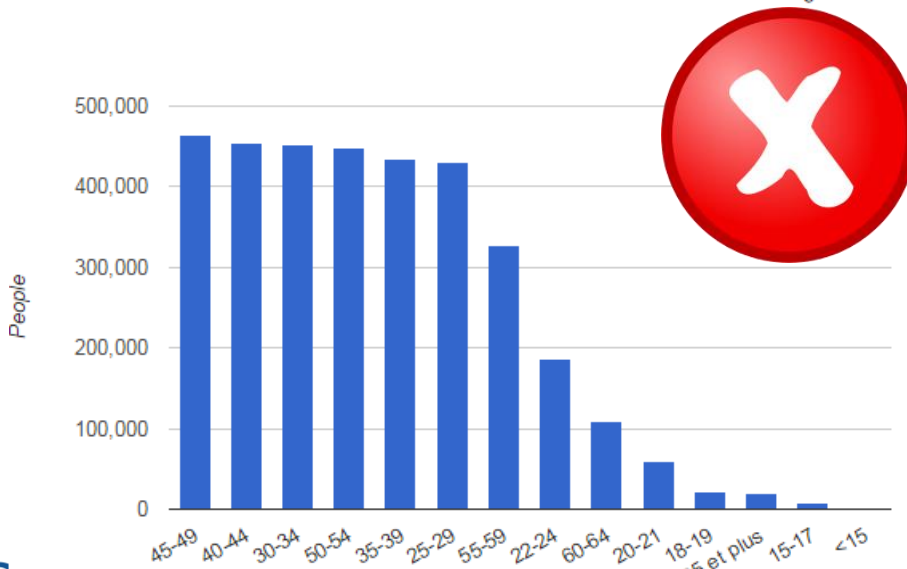
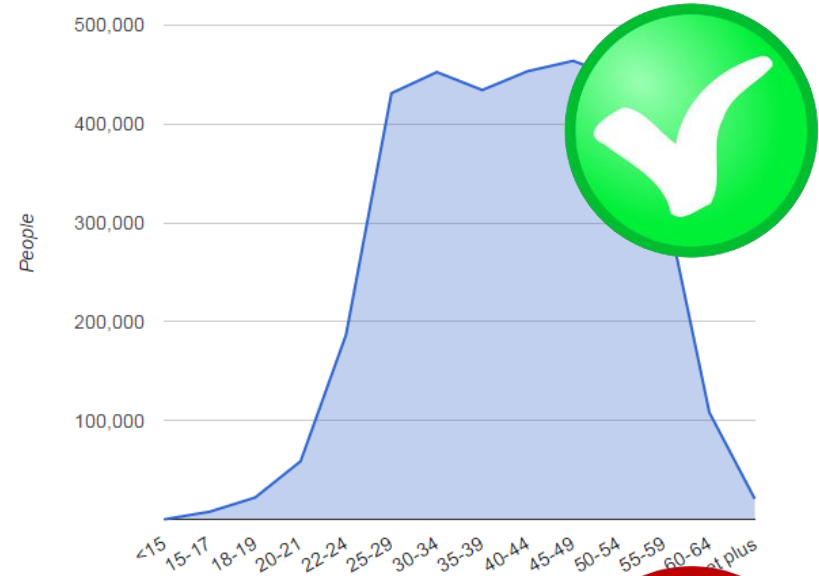
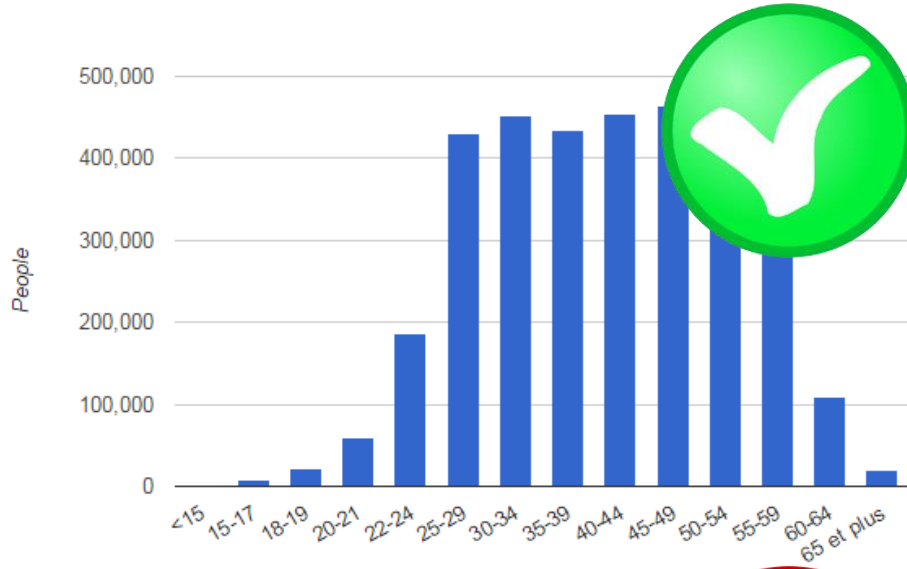
Outil : Google Chart

# Nominal : Dimension



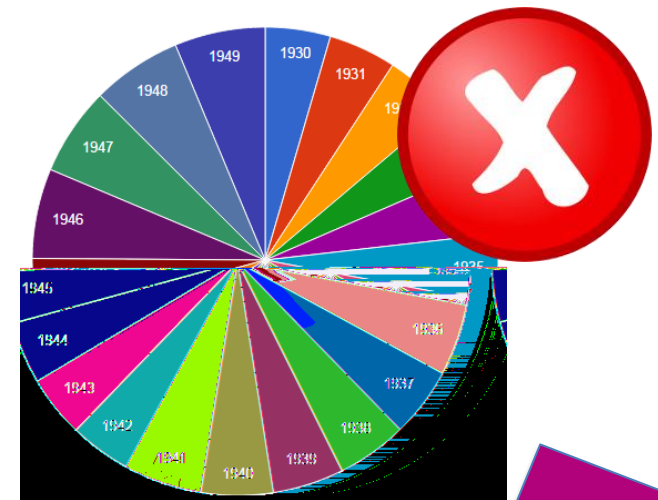
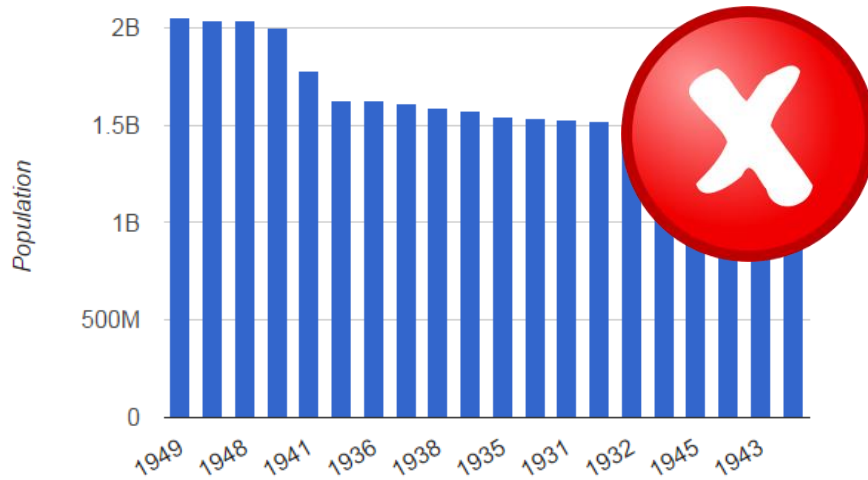
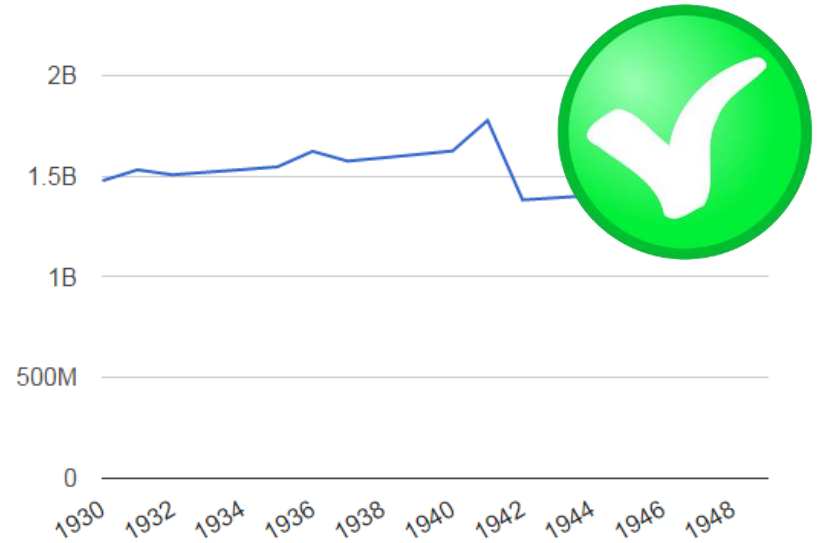
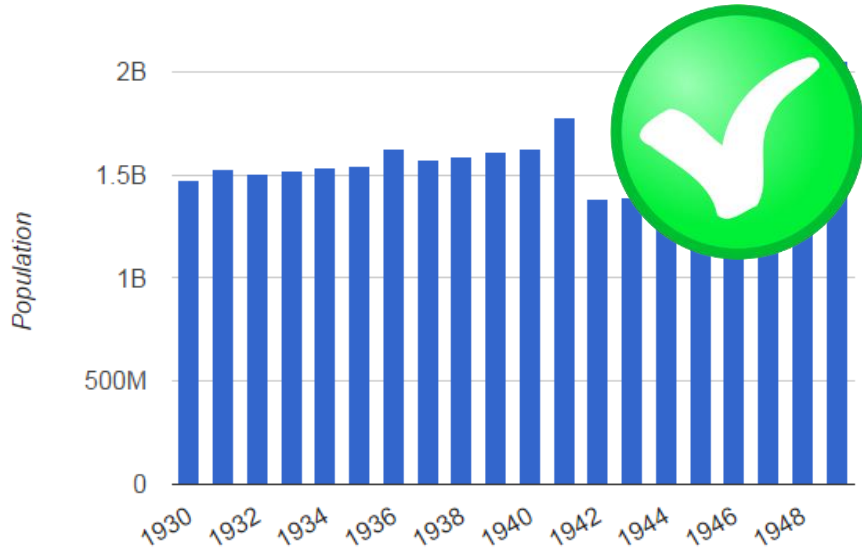
Outill : Google Chart

# Ordinal : Dimension



Outil : Google Chart

# Intervalle : dimension



# Scale of Measurement

		Dimension			Measure	
		Sort by	Use	Don't use	Use	Don't use
Qualitative	Nominal	Alpha, Measure	Bar, Pie, Scatter, Tree map	Line	Count	Average Median Max Sum
	Ordinal	Intrinsic order	Bar, Line, Scatter	Pie	Idem + Median Max	Average Sum
Quantitative	Interval	Value	Bar, Line	Pie	Idem + Average	Sum
	Ratio	Value	Bar, Line, Histogram	Pie	Idem + Sum	

Inspiration : H. Cronström, Qlik

Les fondamentaux

# SÉMIOLOGIE GRAPHIQUE

# Sémiologie graphique

---

- Nombreuses façons de visualiser des valeurs
- Quantitatives :
  - Position
  - Longueur
  - Couleur
  - Angle
  - Orientation
  - Surface
  - Vitesse
- Qualitatives :
  - Forme
  - Couleur
  - Texture
- Toutes n'ont pas la même efficacité
- Sémiologie : Science des signes

# Sémiologie graphique : Position

---

- A, B et C sont alignés, distinguables
- A-B est deux fois plus long que B-C

+ A

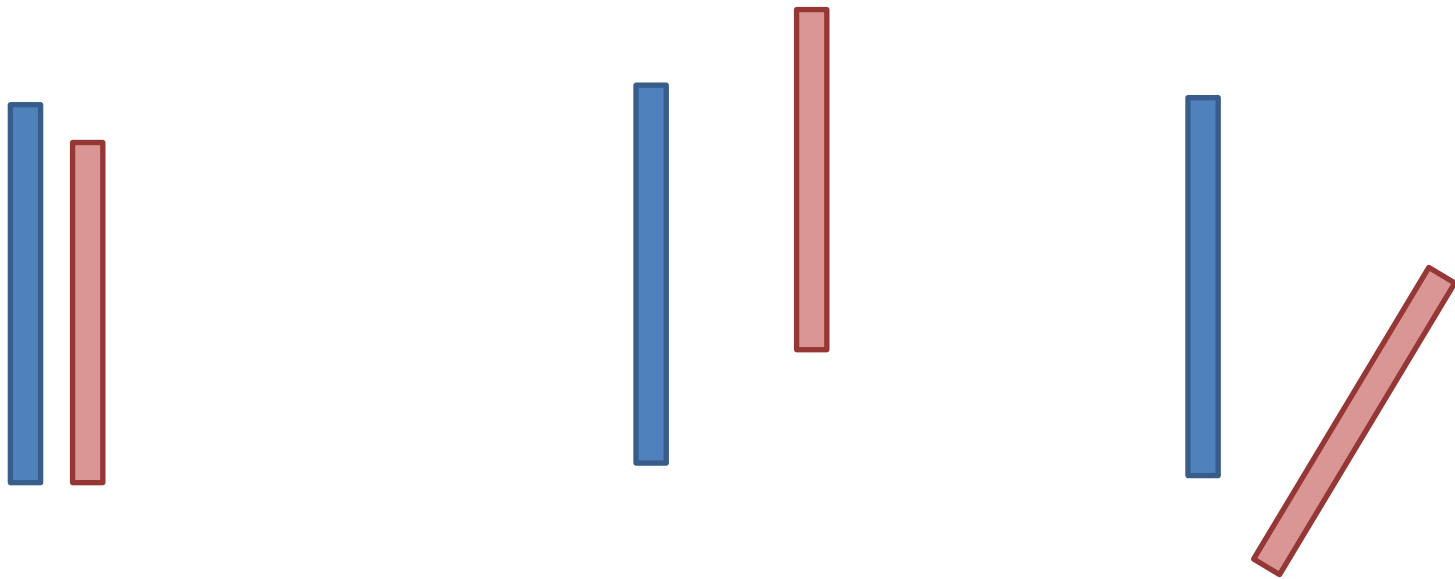
+ B

+ C

# Sémiologie graphique : Longueur

---

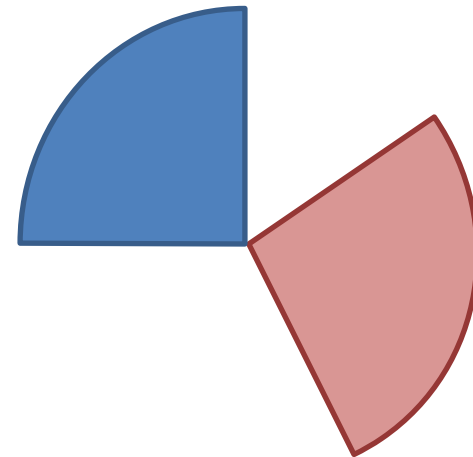
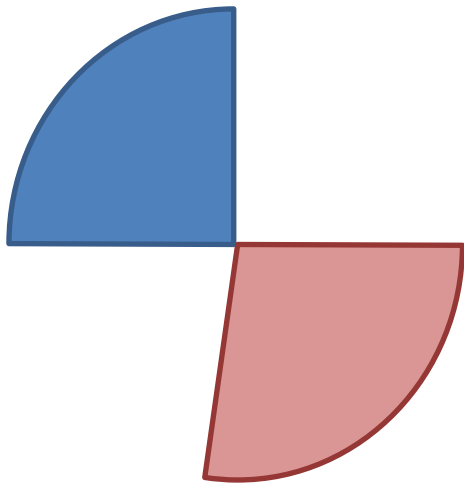
- Quelle est la plus longue barre (ratio : 9/10)?



# Sémiologie graphique : Angle

---

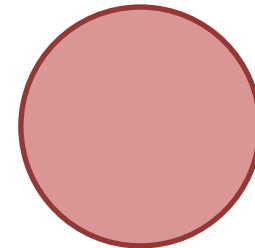
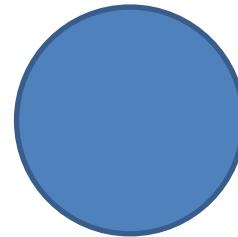
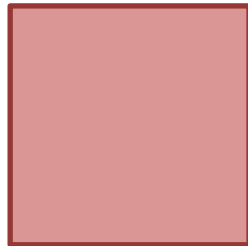
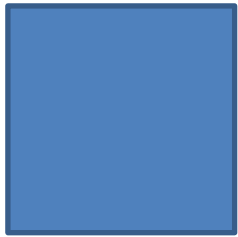
- Quelle tranche est la plus grande ?



# Sémiologie graphique : Surface

---

- Quelle est la plus grande surface (ratio : 9/10)

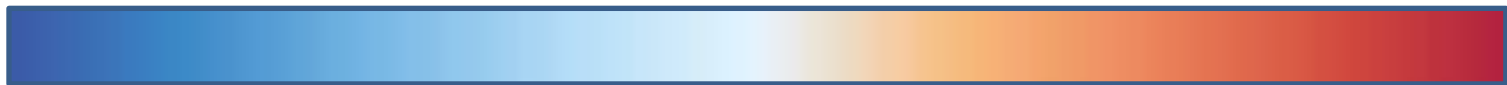
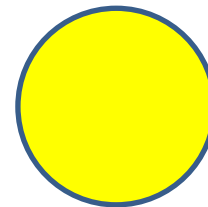
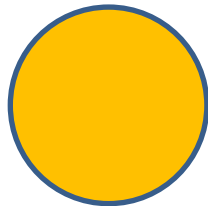


# Sémiologie graphique : Couleur

---

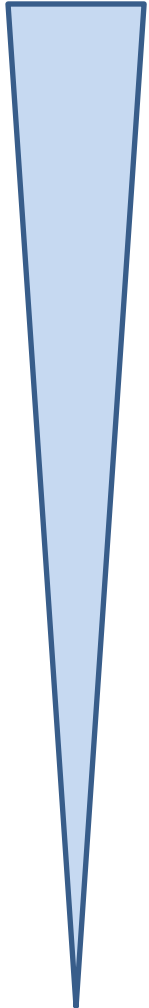


Combien représentent ces boules :



# Sémiologie graphique

Efficacité



Position



Longueur



Pente



Angle



Surface



Intensité



Couleur



Forme

Plus adapté pour :

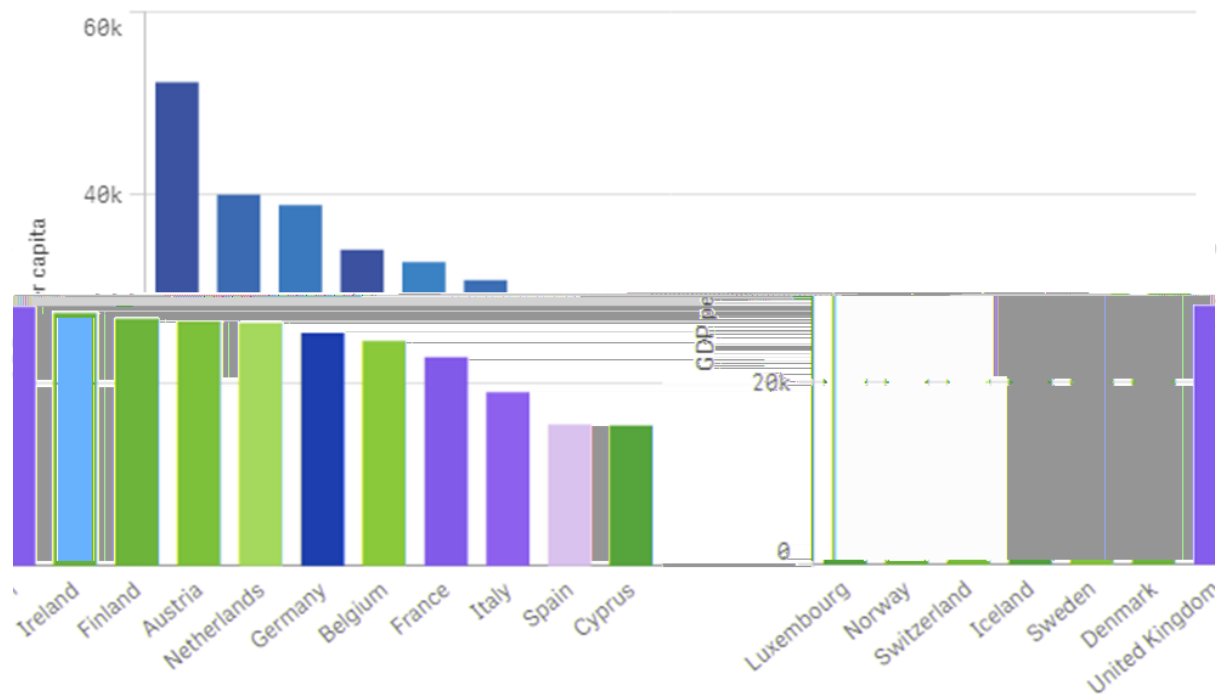
Quantitative

Ordinal

Nominal

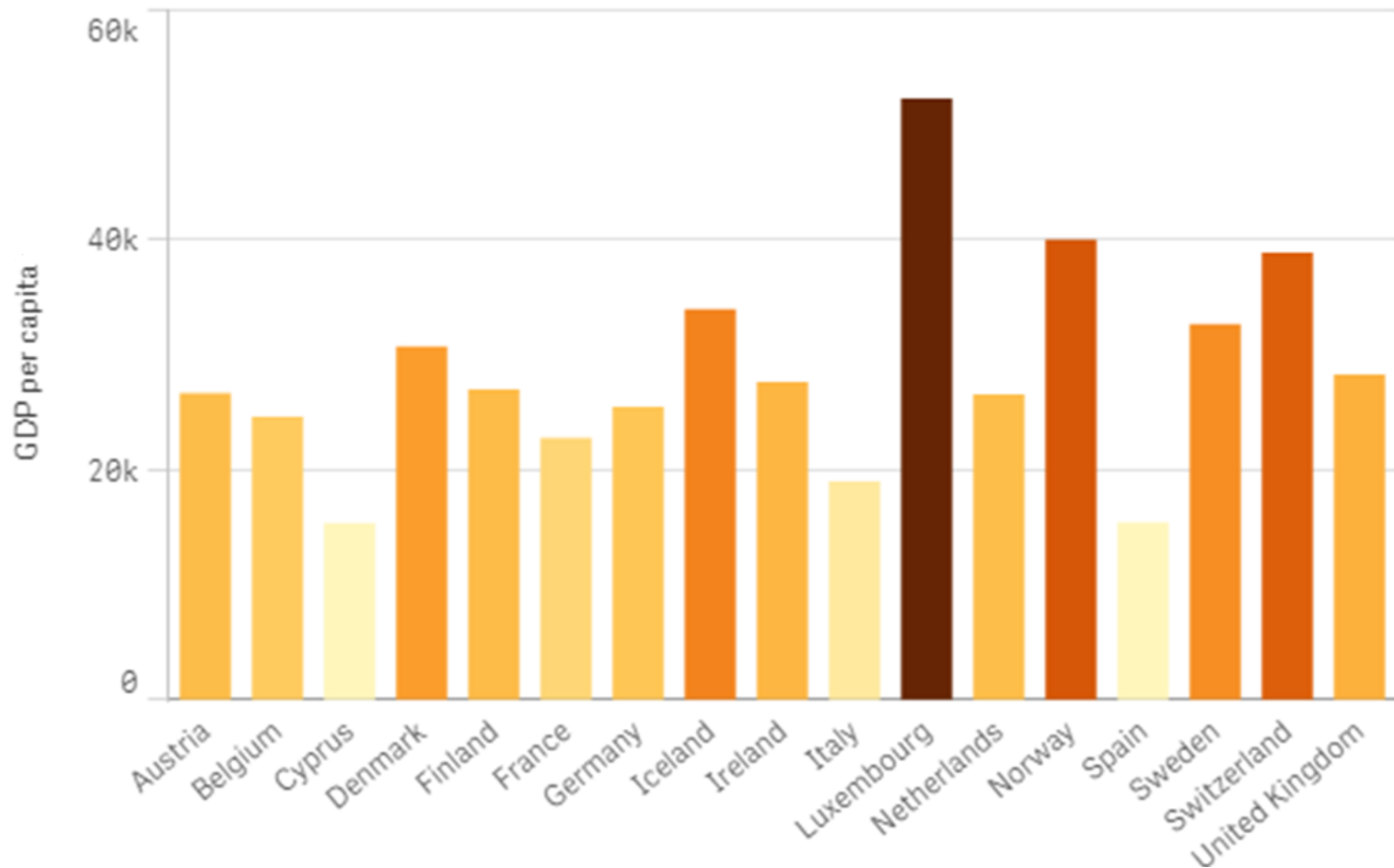
# Sémiologie graphique

- En règle générale, on utilise plusieurs « canaux » dans un graphique
- On veillera à utiliser le canal le plus efficace pour l'information la plus importante !



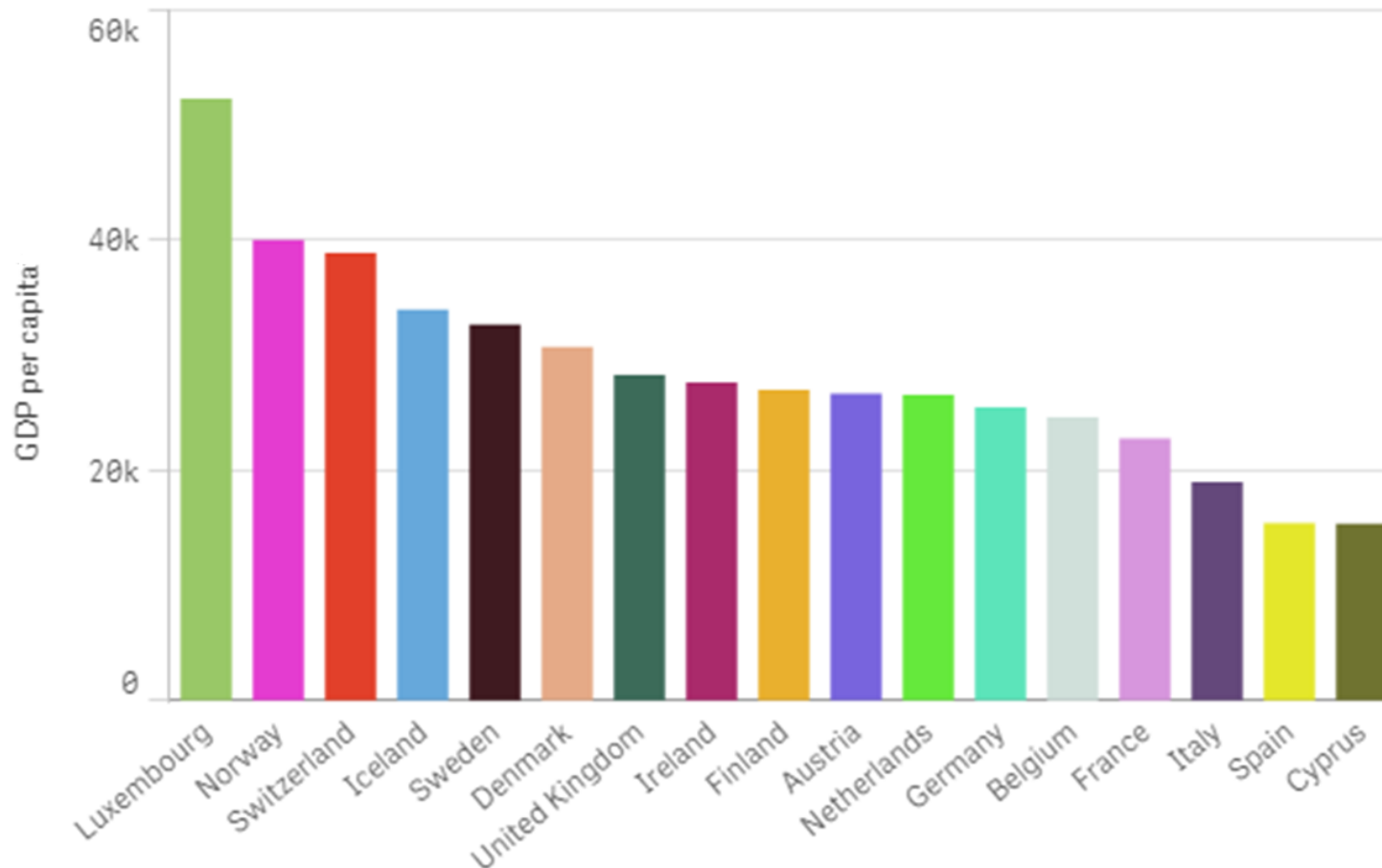
# Sémiologie graphique

- Éventuellement, duplication des canaux :



# Sémiologie graphique

- À éviter : canal « inutile »



Les fondamentaux

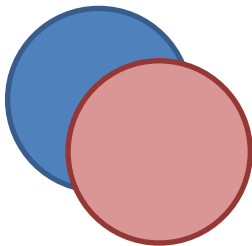
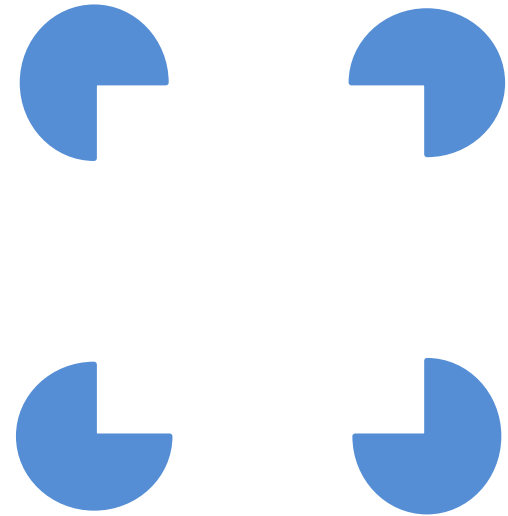
# LOIS DE GESTALT

# Lois de Gestalt

---

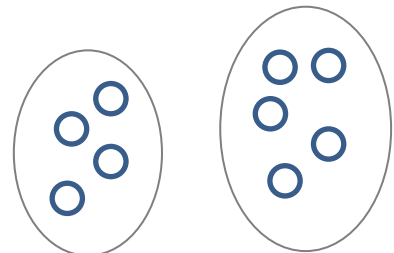
- Gestalt : « **Pattern** » en allemand
- Émises par un groupe de psychologues allemands
- Elles décrivent comment le **cerveau** « **complète** » automatiquement tout ce qu'il perçoit
- Ces mécanismes permettent de **comprendre un graphique**

# Gestalt Laws : Closure



Fermeture : Terminaison  
des contours manquants

# Gestalt Laws : Proximity



Proximité :  
regroupement des  
entités proches

# Gestalt Laws : Proximity

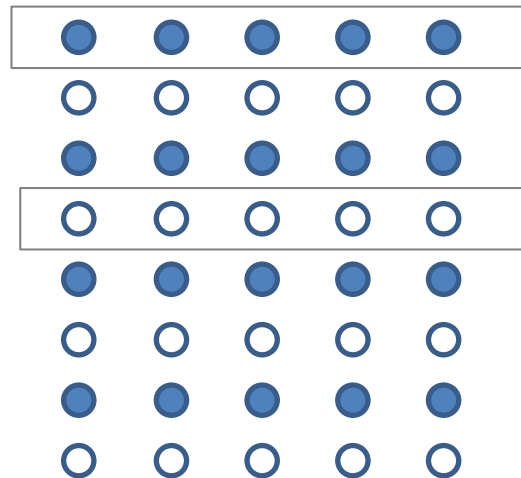
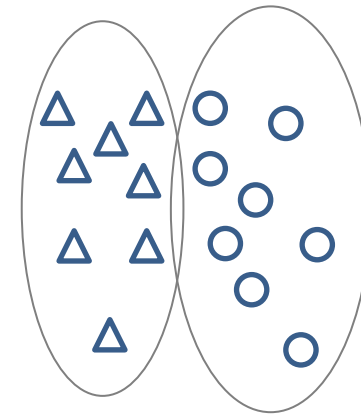
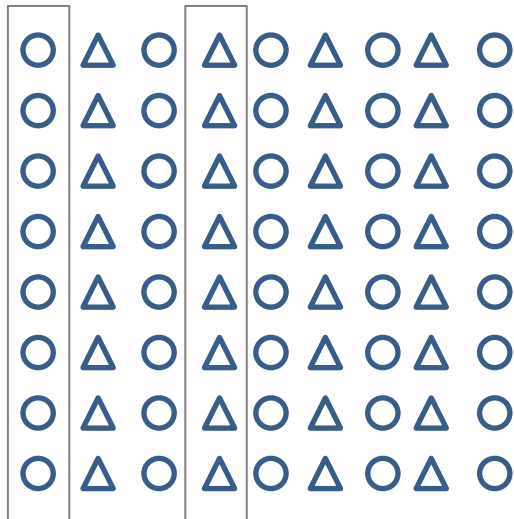
Mise en évidence  
des colonnes :

1	5	3	1	7
2	5	4	7	4
5	7	3	5	8
6	8	4	3	2
7	7	9	7	1

Mise en évidence  
des lignes :

1	5	3	1	7
2	5	4	7	4
5	7	3	5	8
6	8	4	3	2
7	7	9	7	1

# Gestalt Laws: Similarity



Similarité :  
regroupement des  
entités similaires

# Gestalt Laws : Proximity

Mise en évidence  
des colonnes :

1	5	3	1	7
2	5	4	7	4
5	7	3	5	8
6	8	4	3	2
7	7	9	7	1

1	5	3	1	7
2	5	4	7	4
5	7	3	5	8
6	8	4	3	2
7	7	9	7	1

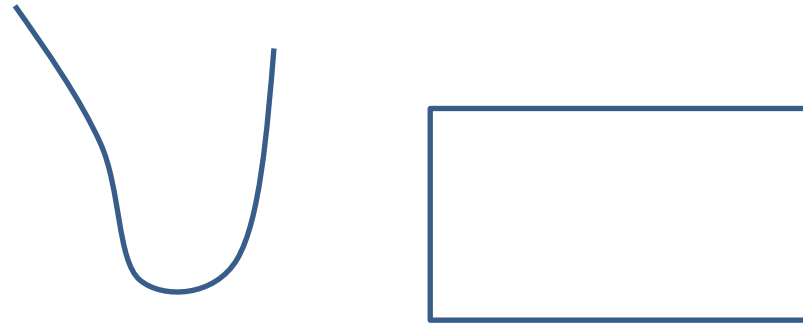
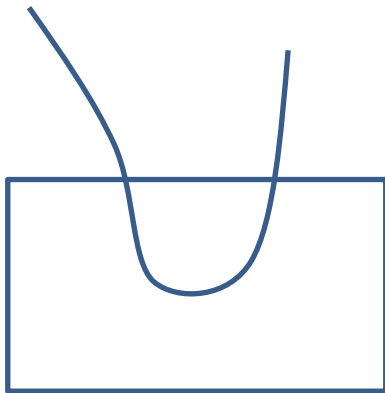
Mise en évidence  
des lignes :

1	5	3	1	7
2	5	4	7	4
5	7	3	5	8
6	8	4	3	2
7	7	9	7	1

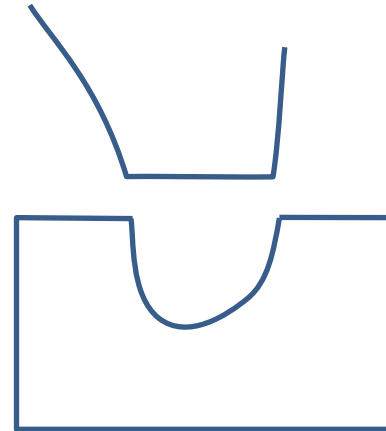
1	5	3	1	7
2	5	4	7	4
5	7	3	5	8
6	8	4	3	2
7	7	9	7	1

# Gestalt Laws : Continuity

---

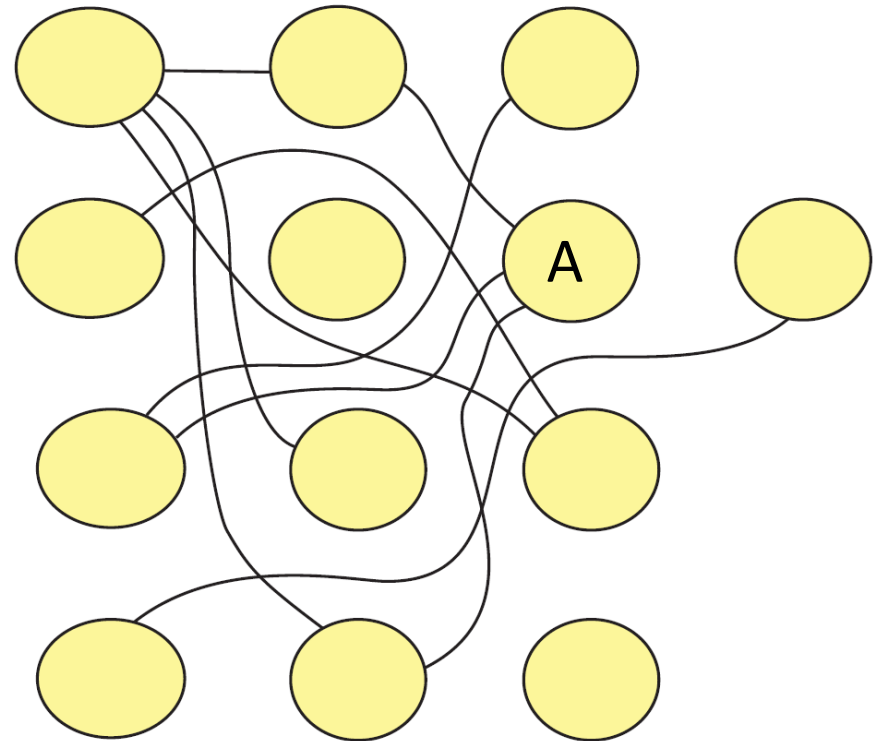
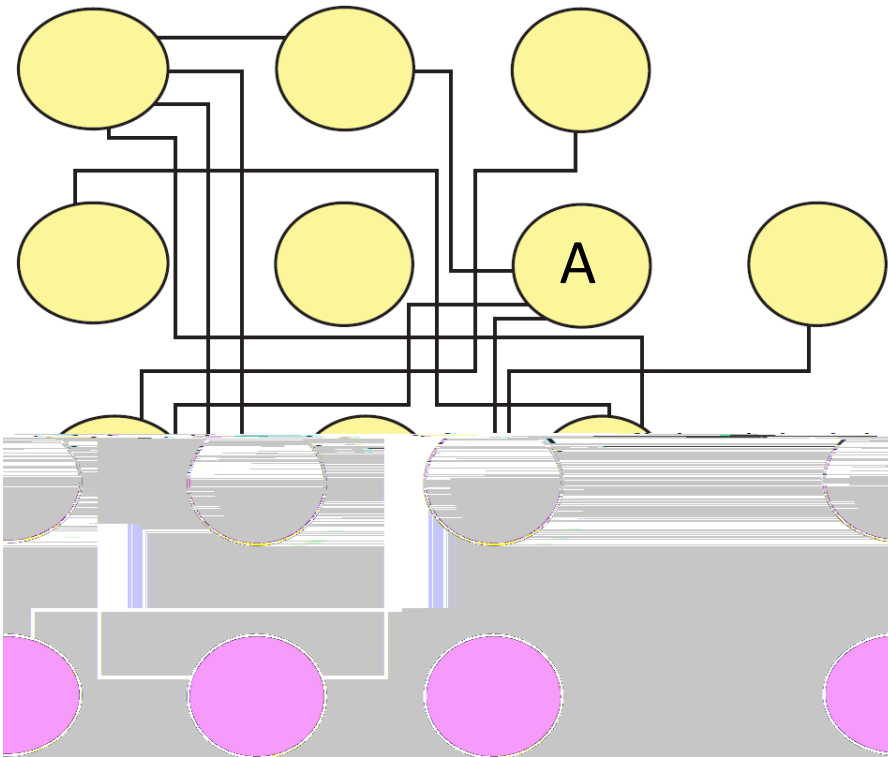


Continuité :  
Construction sur  
base d'éléments  
lisses et continus



# Gestalt Laws : Continuity

- À quoi est connecté A ?



Les fondamentaux

# PERCEPTION PRÉ-ATTENTIVE

# Perception pré-attentive

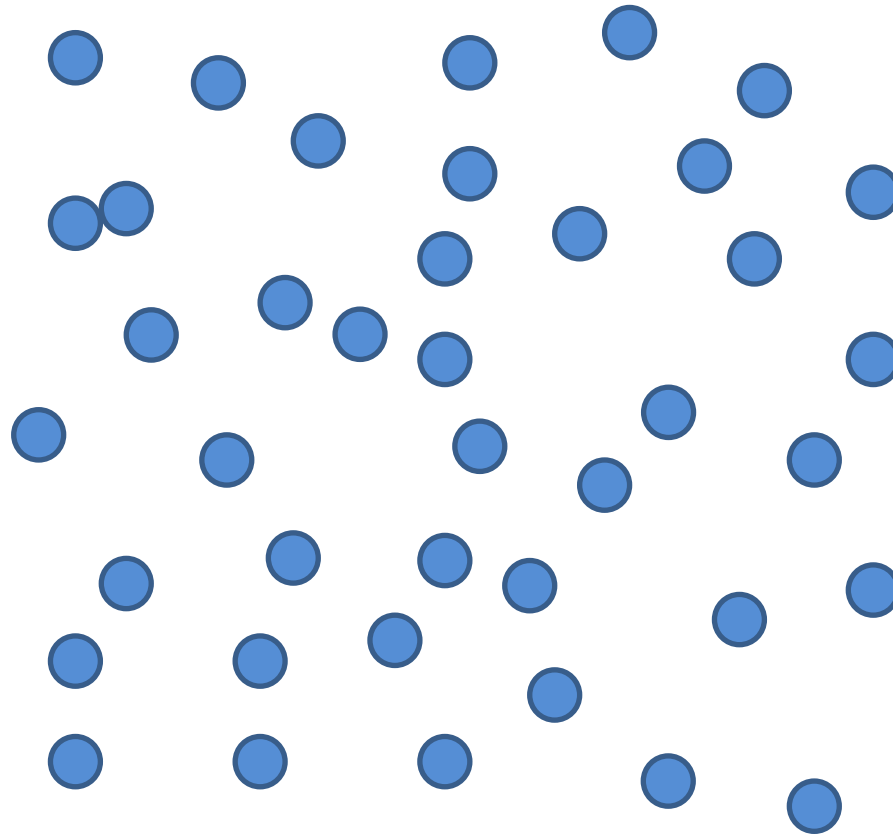
---

- Certains éléments sont identifiés  
« *instantanément* »
- Trouvez le 3 dans la série :

6546496874540668654687845487684846864  
85257687684689897897554684568621**3**5768  
7861687196876868178676786876876714546

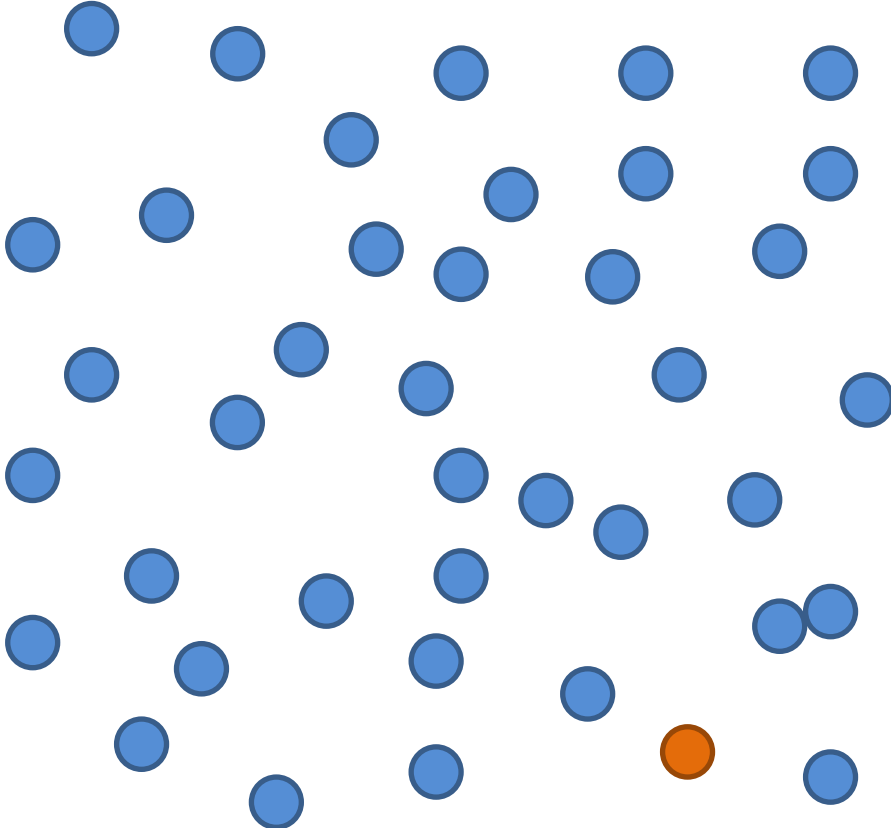
# Perception pré-attentive : Couleur

---



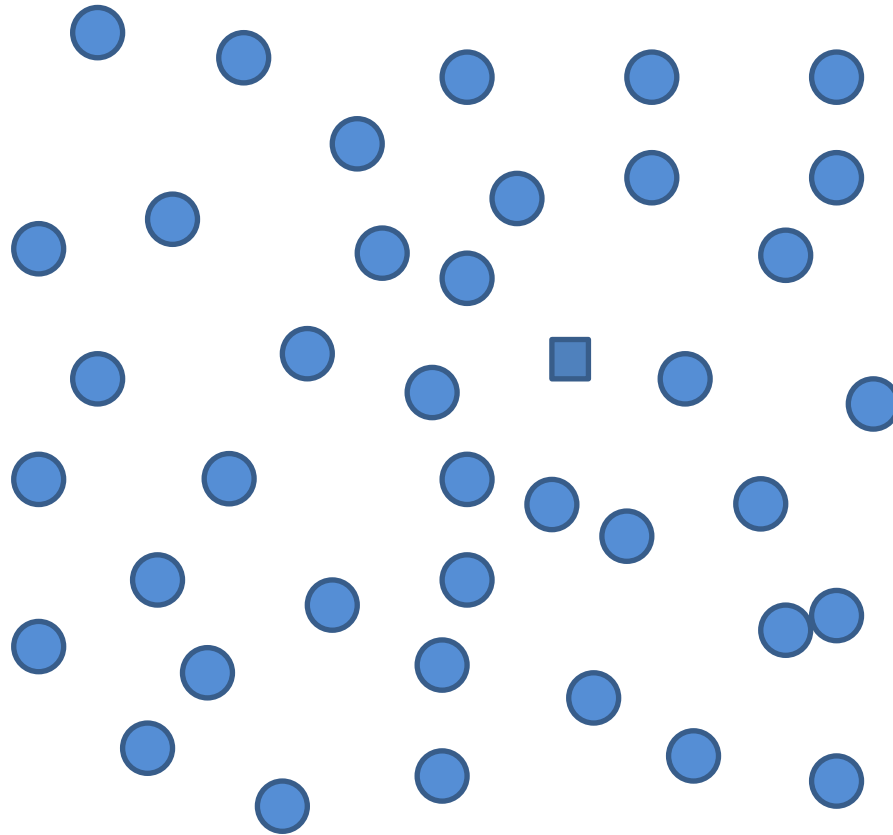
# Perception pré-attentive : Couleur

---



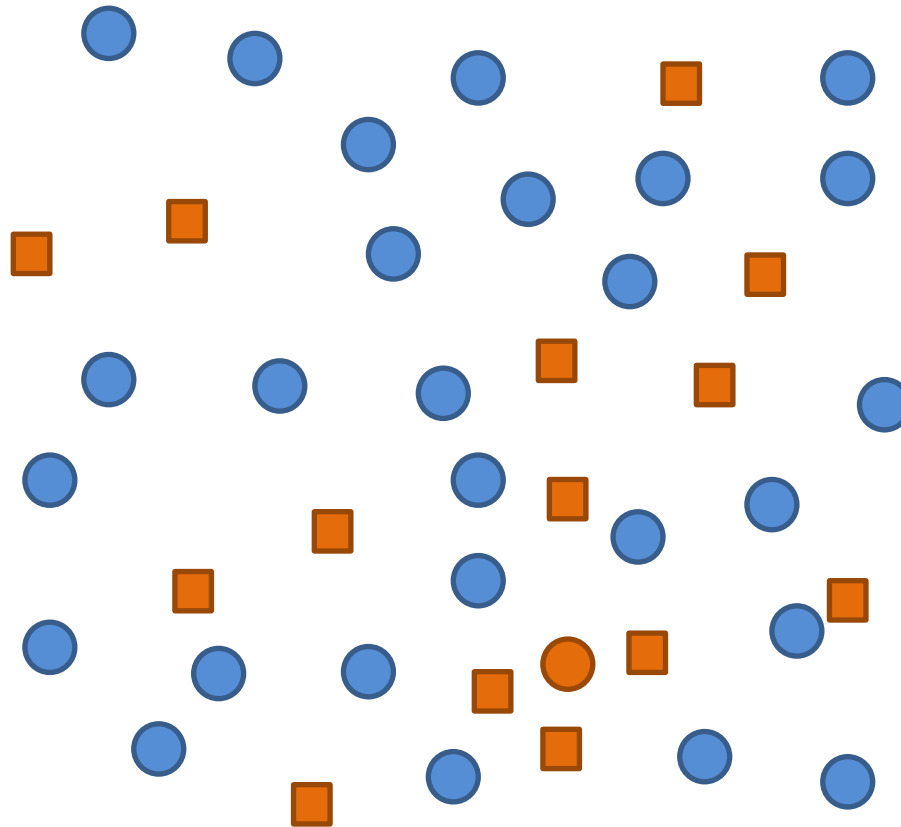
# Perception pré-attentive : Forme

---



# Perception pré-attentive : Combinaison

---

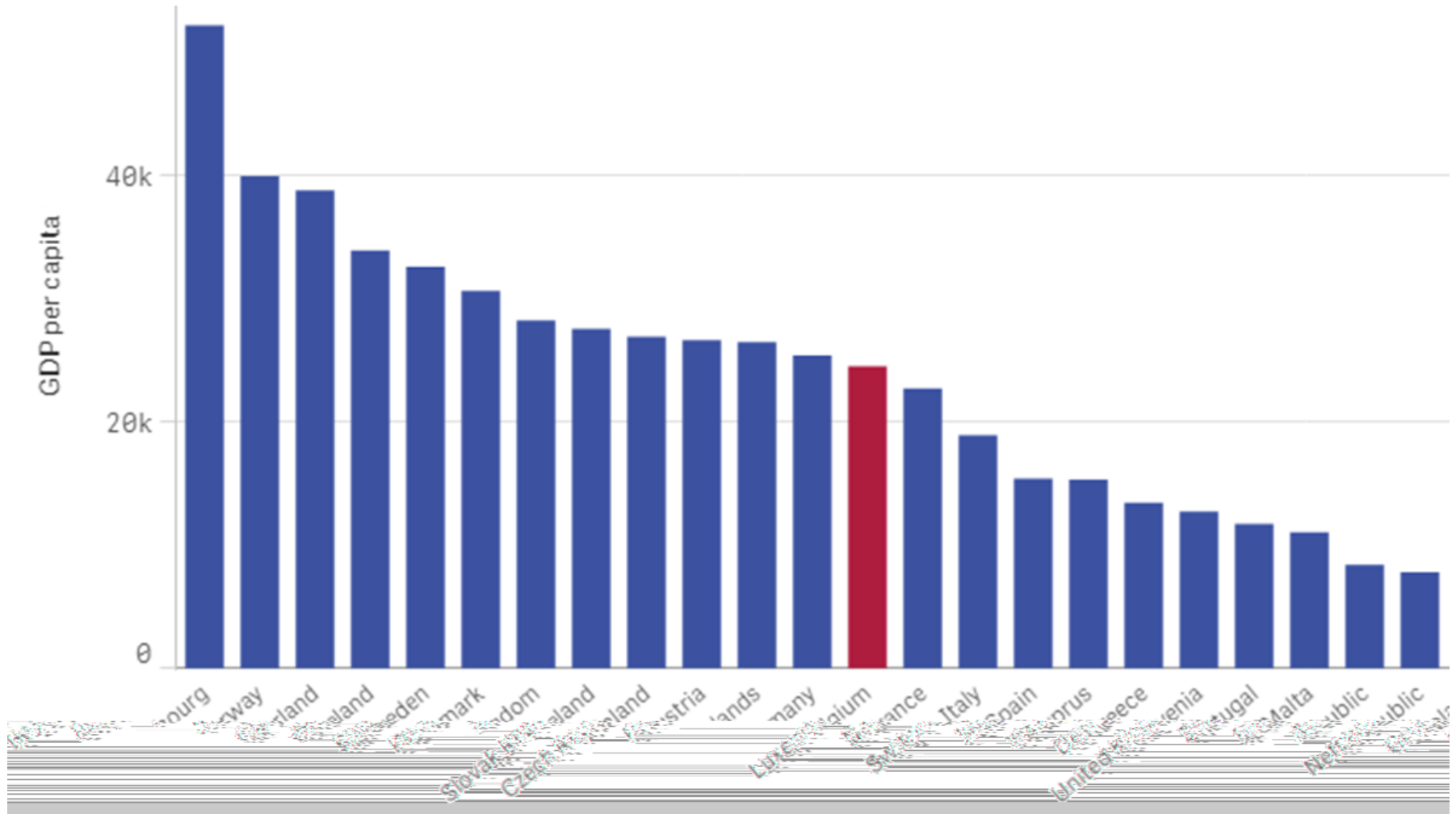


# Perception pré-attentive

---

- On peut utiliser la perception pré-attentive pour mettre fortement un élément en évidence :
  - Couleur (le plus efficace)
  - Forme
  - Orientation
  - Taille
  - ...
- Une seule méthode à la fois !

# Perception pré-attentive



Outil : Qlik Sense

Les fondamentaux

# MESURE DE QUALITÉ

# Mesure de qualité

---

Tufte a proposé 3 métriques pour mesurer la qualité d'un graphique :

- **Lie Factor** :

À quel point le graphique est **fidèle aux données**

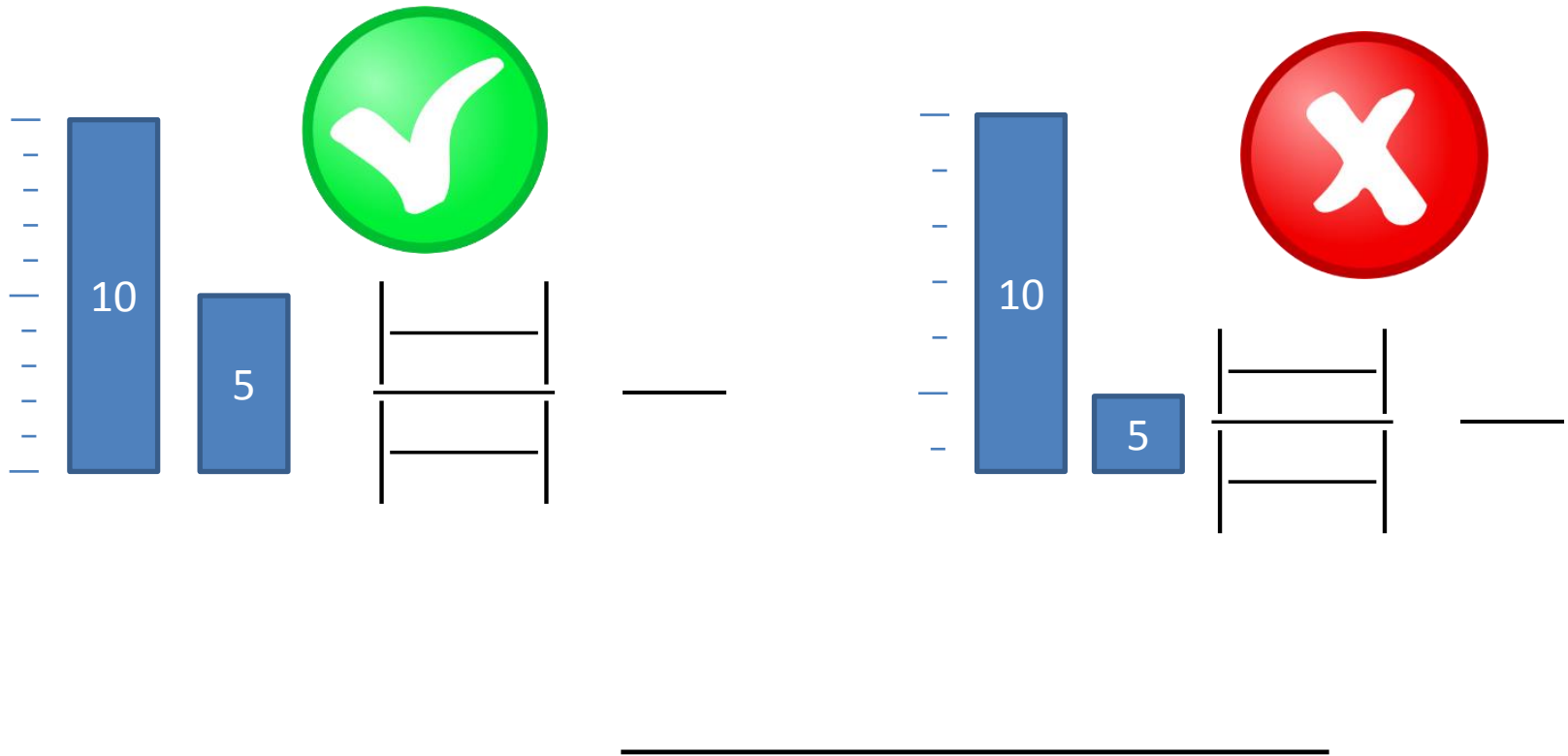
- **Data-ink factor** :

À quel point **l'encre** est utilisée efficacement (et la distraction diminuée)

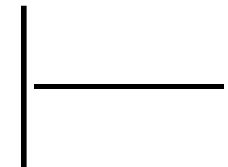
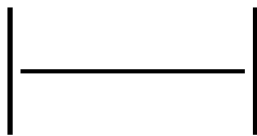
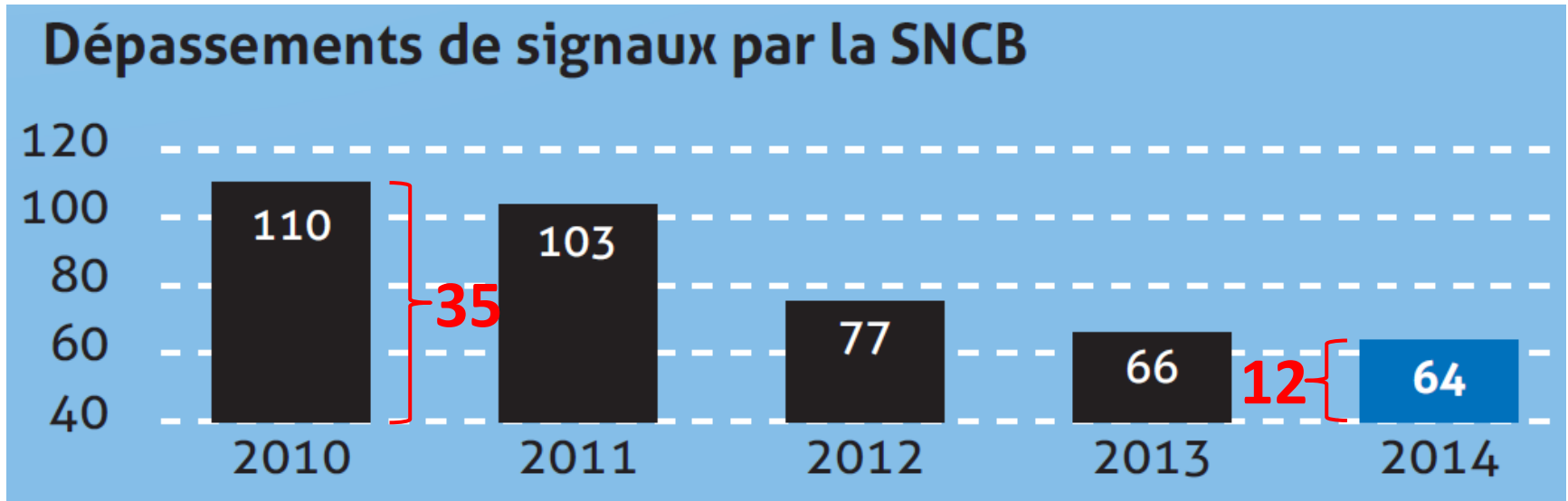
- **Data density** :

À quel point **l'espace** est utilisé efficacement

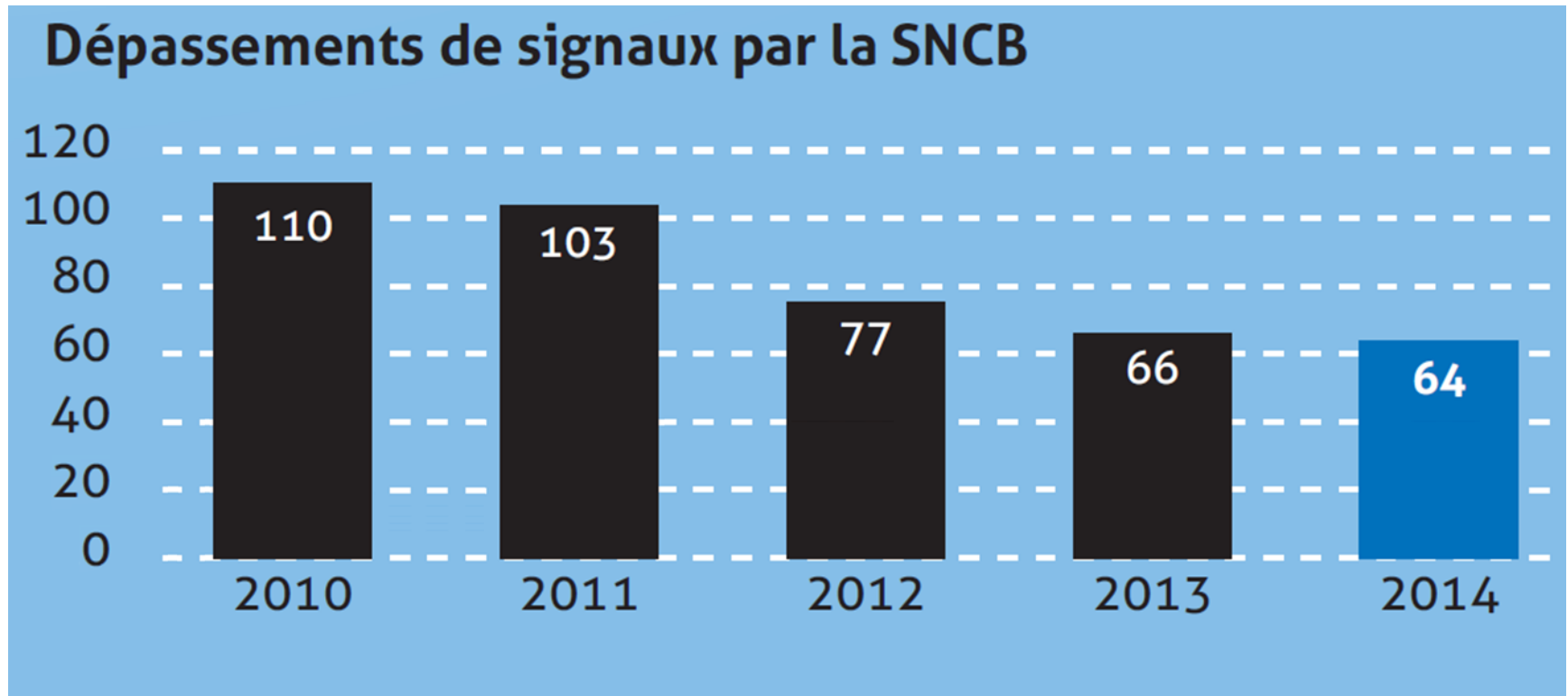
# Lie Factor



# Lie Factor : Exemple



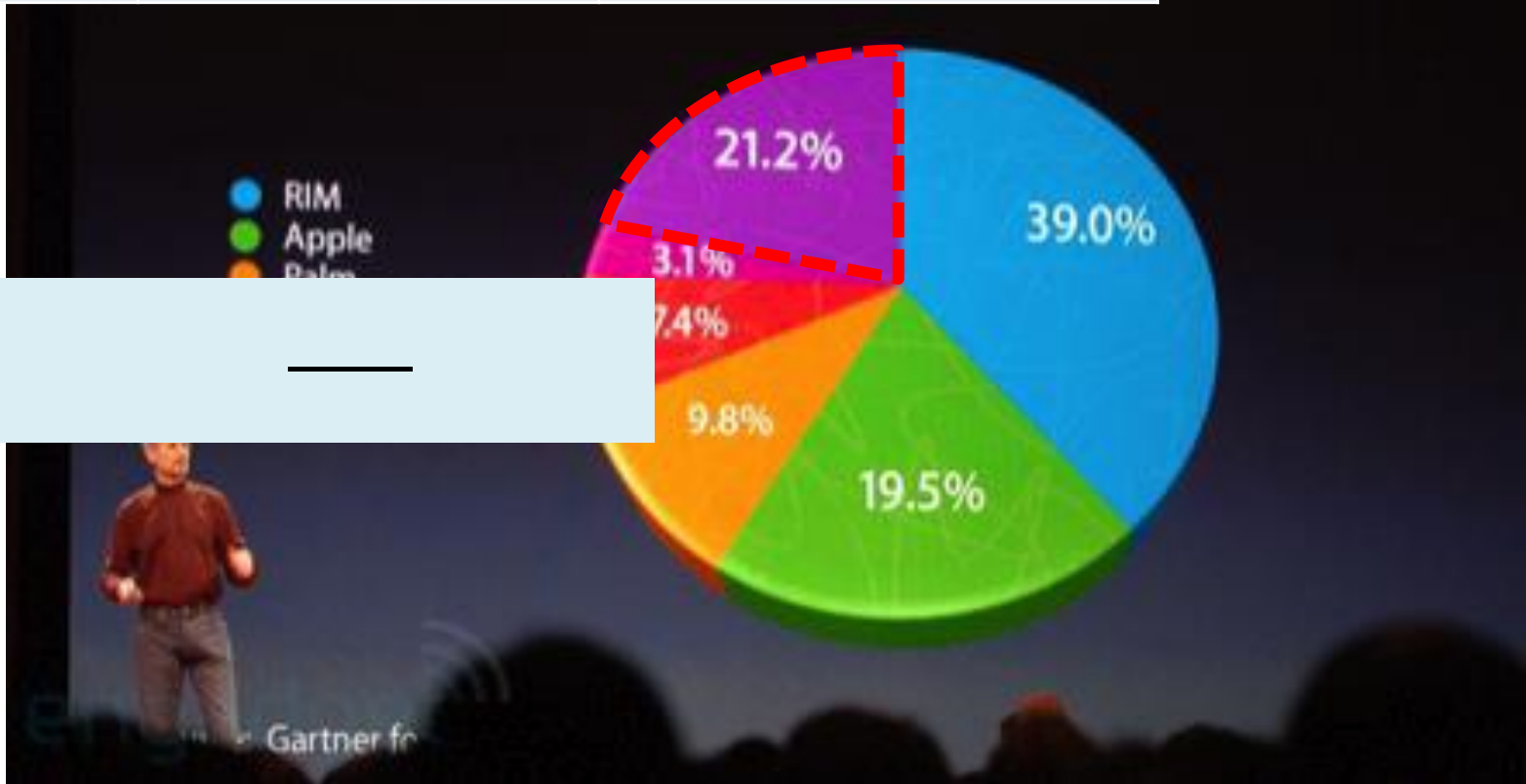
# Lie Factor : Exemple



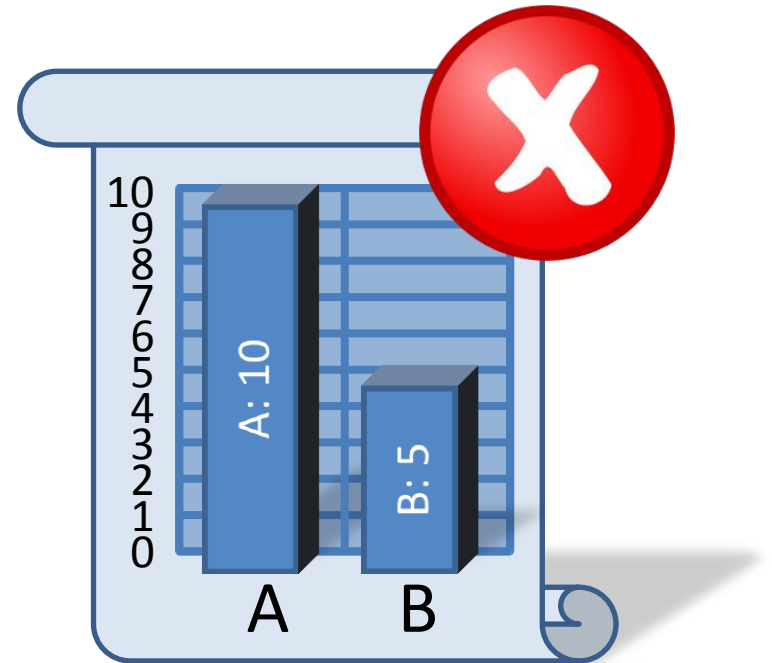
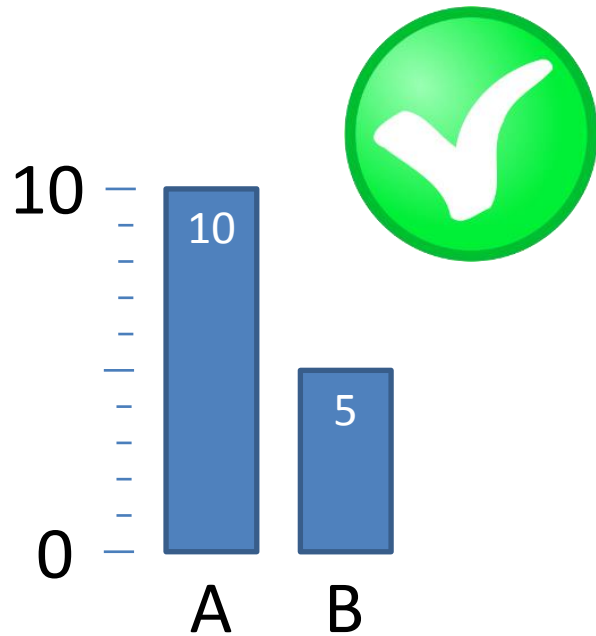
# Lie factor : Exemple

	Données	Graphique
1: Other	21,5%	3838 px
2: Apple	19,5%	6914 px

-8%      +80 % !

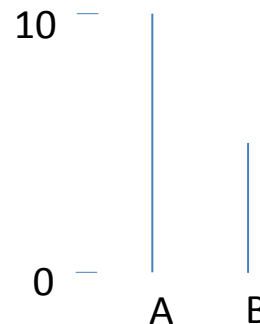
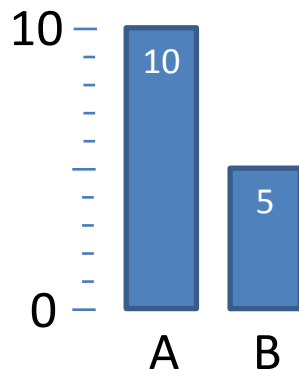


# Data-ink ratio



# Data-ink ratio

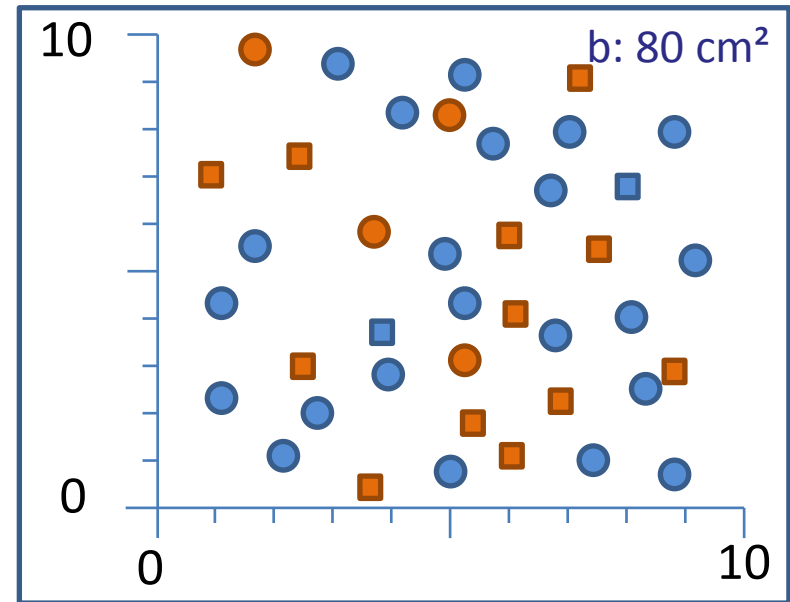
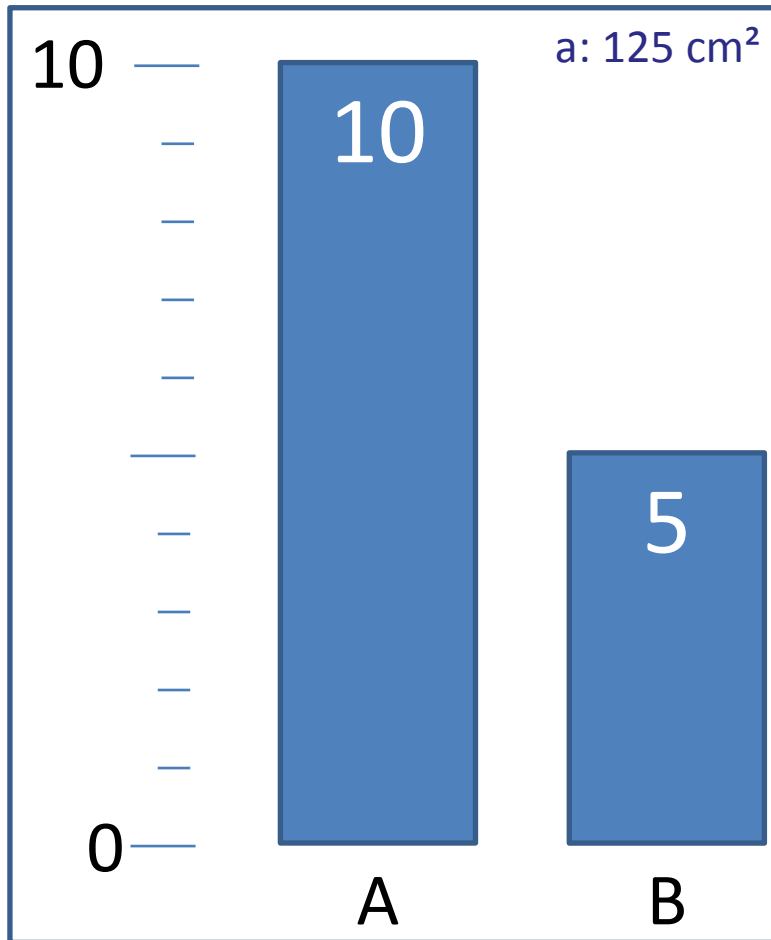
- Principe : enlever ce qui ne contient pas d'info
- Éliminer tout ce qui pourrait distraire le lecteur (*chart-junk, non-data ink, redundant data-ink*) :
  - Fond de couleur
  - Grille trop dense, visible, épaisse
  - Axe avec trop de « tics »...
  - Effet d'ombre, 3D...
- Notion subjective ! Attention aux excès !



« Above else,  
show the data »

E. Tufte

# Data density



# Data density

a: 125 cm<sup>2</sup>

Pt	val
A	10
B	5

b: 80 cm<sup>2</sup>

Pt	x	y	coul	form
1	10	6	B	R
2	5	2	O	R
3	8	5	O	C
...	...	...	...	...
40	1	1	B	R

—

—

# Data density

---

- L'objectif n'est **pas** d'augmenter la densité à tout **prix** !



- Le graphique n'utilise-t-il pas une **place inutile** ?
- Un autre graphique peut-il représenter la même chose **plus densément** ?
- Un autre graphique peut-il représenter **plus de données** ?
- Qu'apporte le graphique par rapport à un **simple tableau** ?

# Mesures de qualité : en pratique

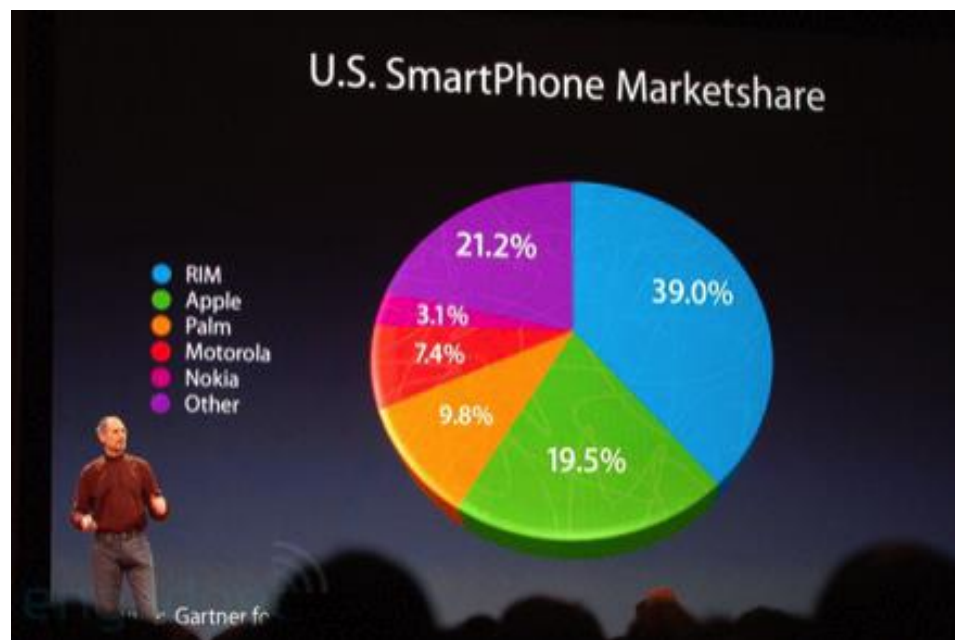
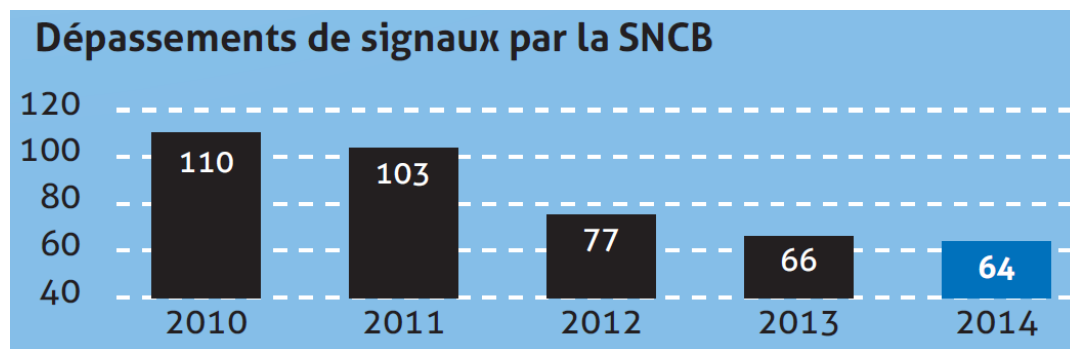
---

- Viser un « lie factor » de 1
- Minimiser le « data-ink » ratio avec raison :
  - Supprimer la « décoration » (3D, ombre, fond...)
  - Limiter la redondance
- Maximiser la « data-density » sans perdre en lisibilité

Les fondamentaux

# ERREURS & MANIPULATIONS

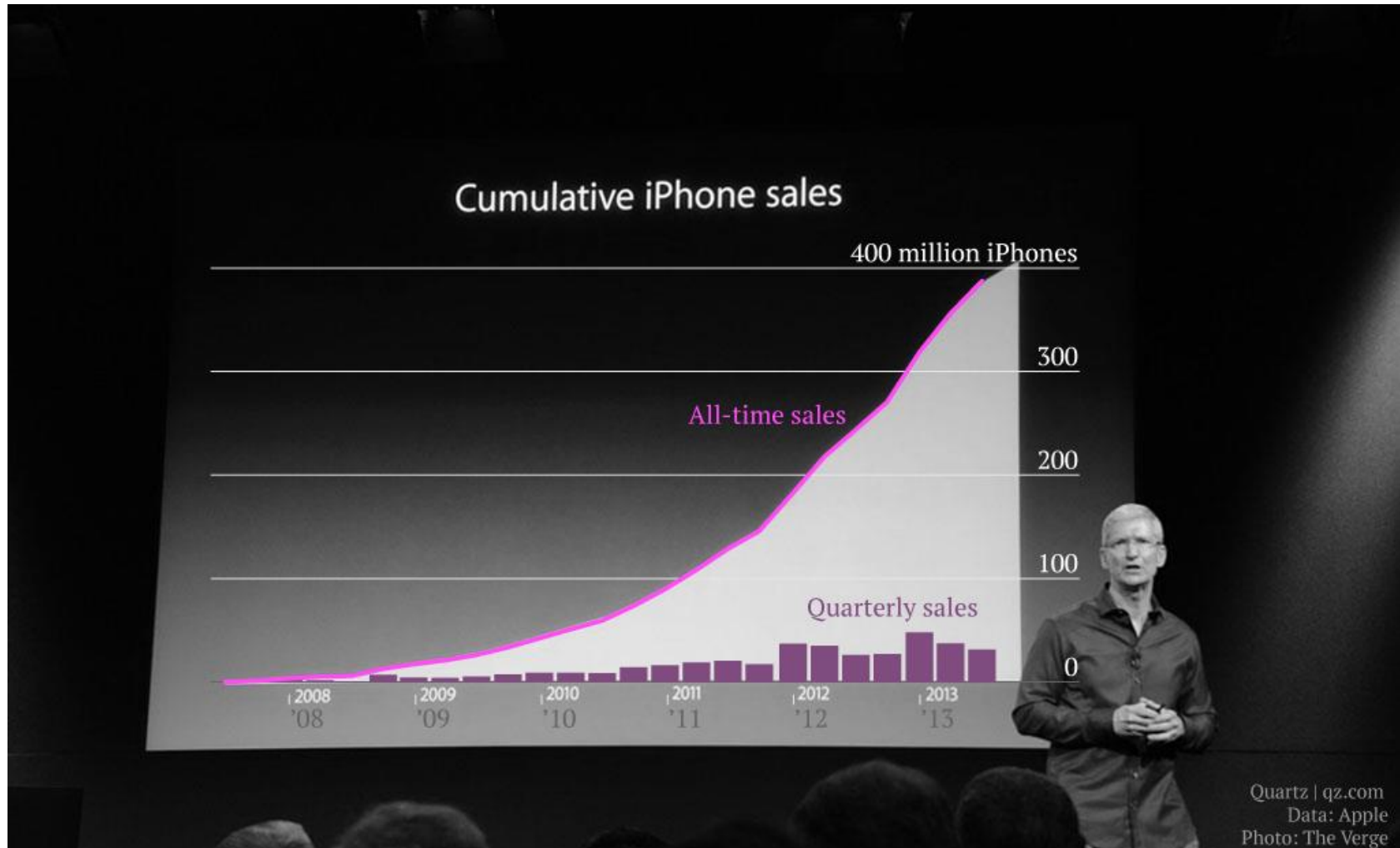
# Tronquer un axe, utiliser la perspective



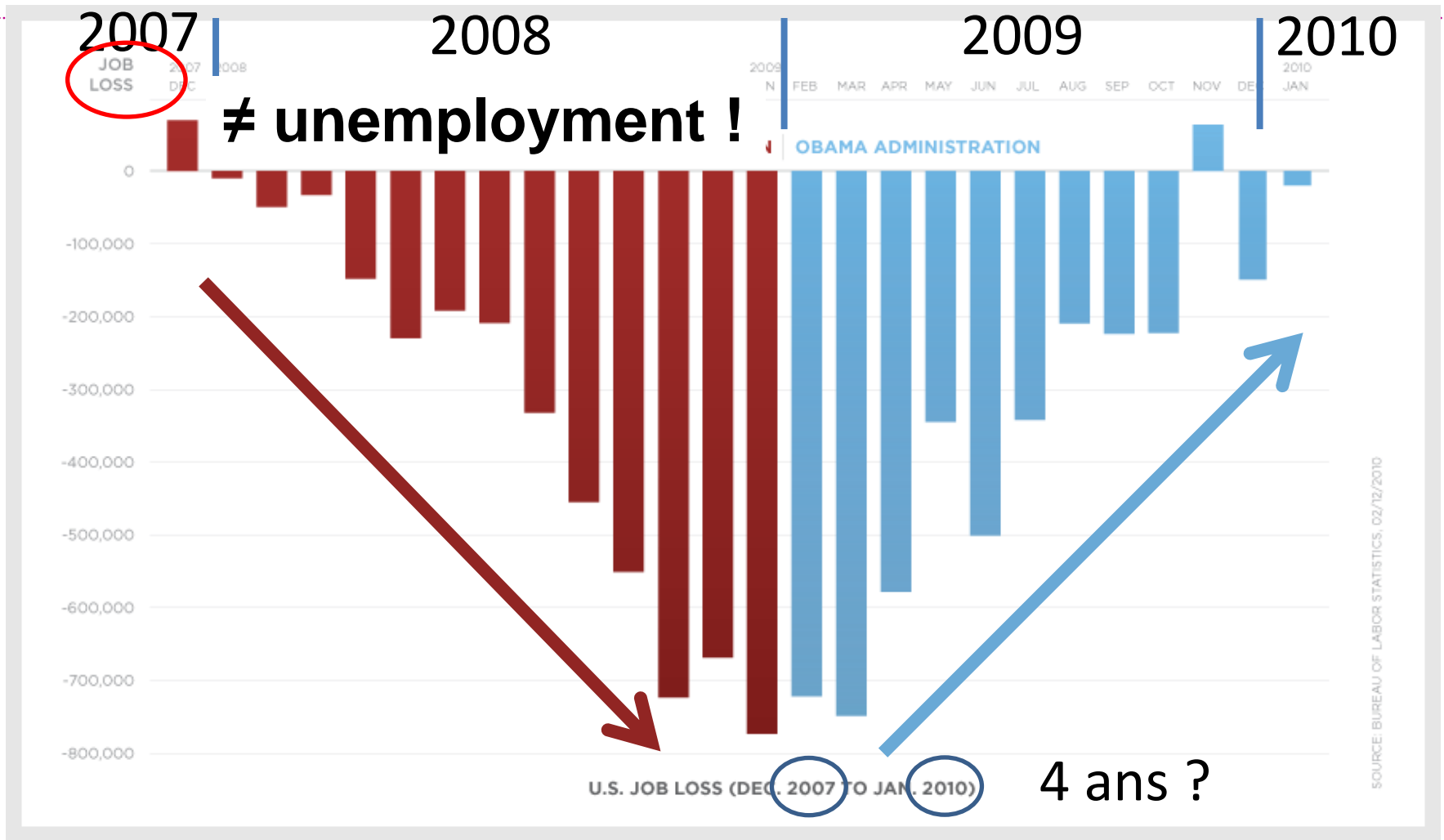
# Valeurs cumulatives



# Valeurs cumulatives



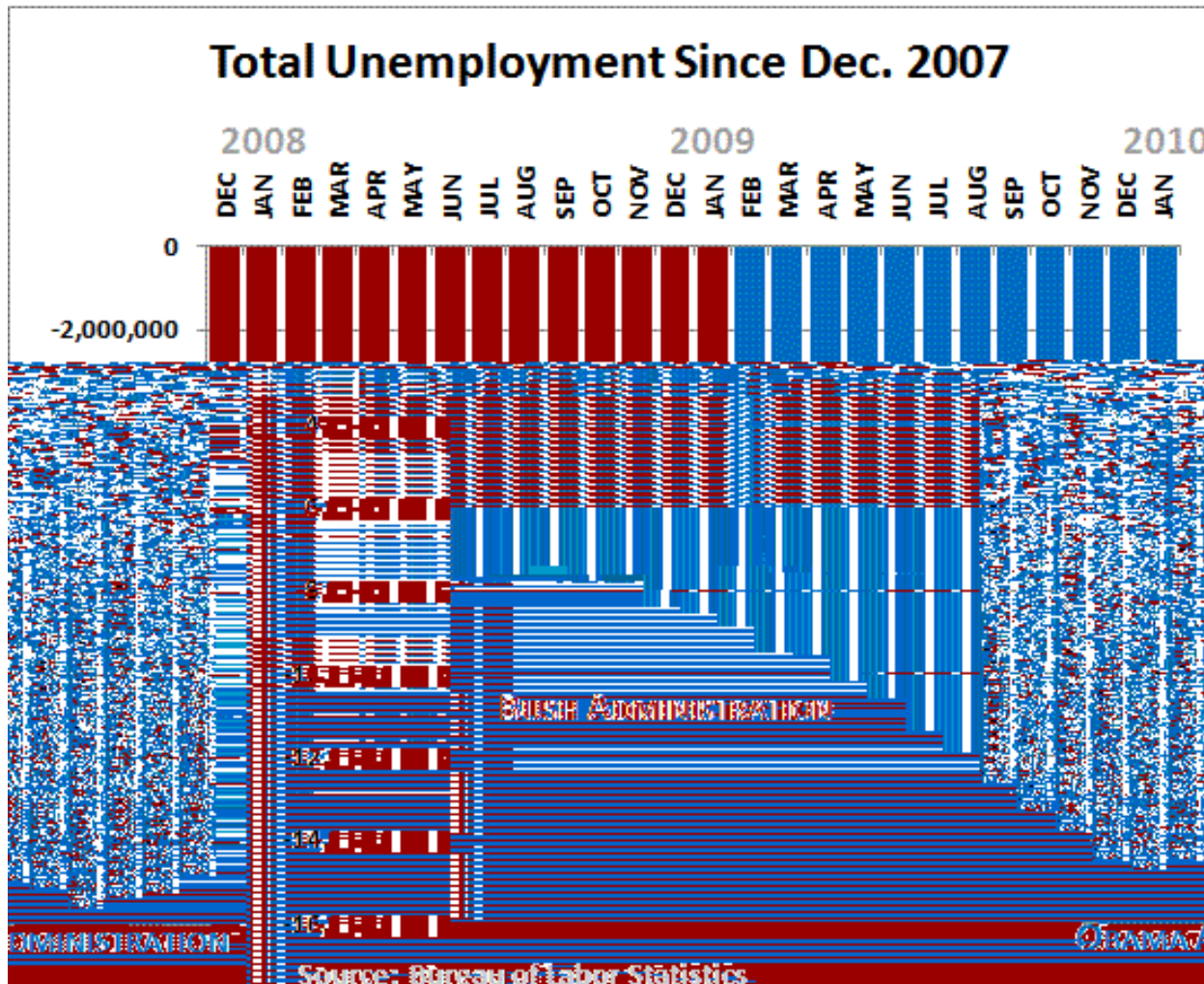
# Le contraire



Luminance : **Bush : 57%**

**Obama: 84%**

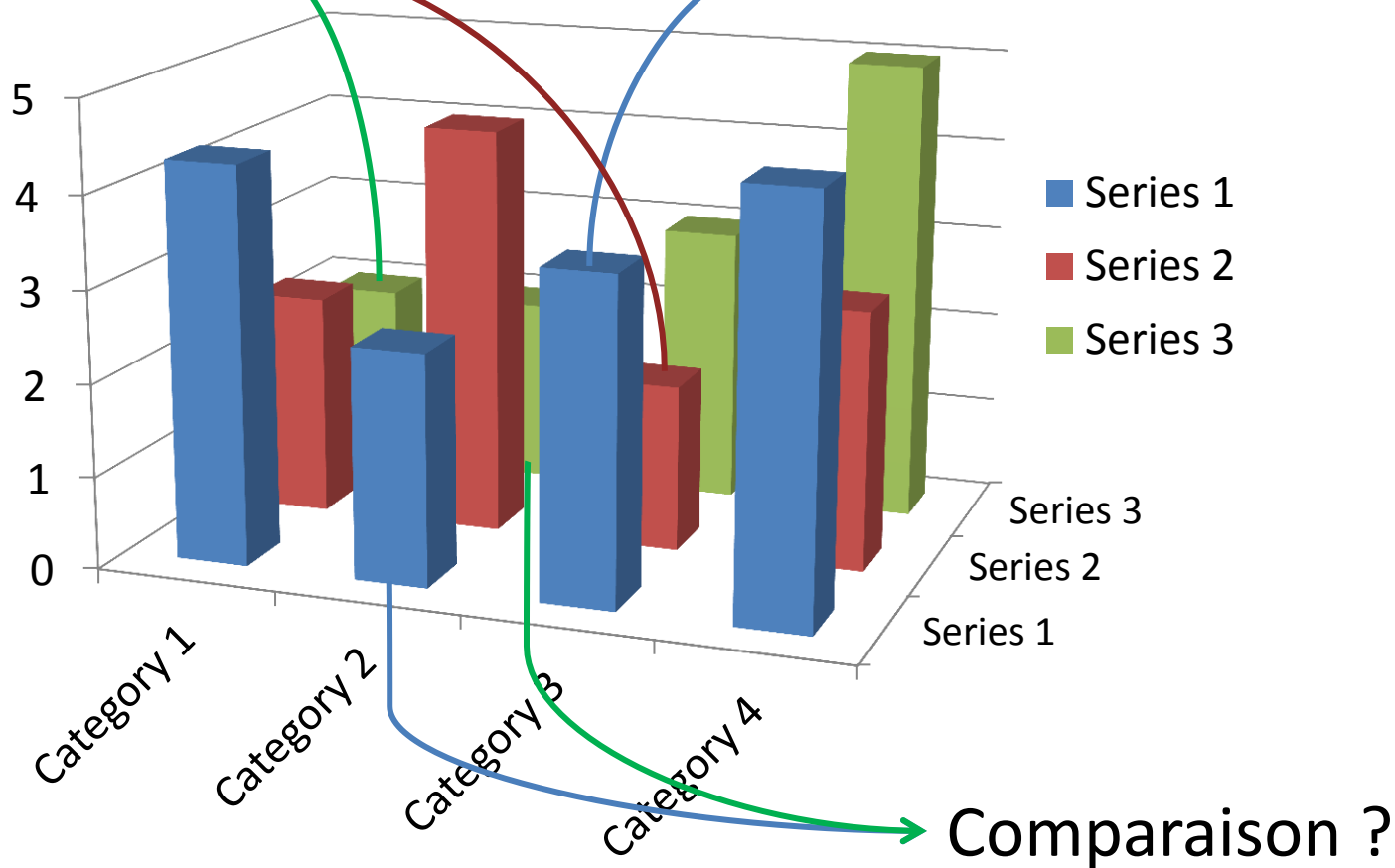
# Le contraire



# Perspectives & 3D

Comparaison ?

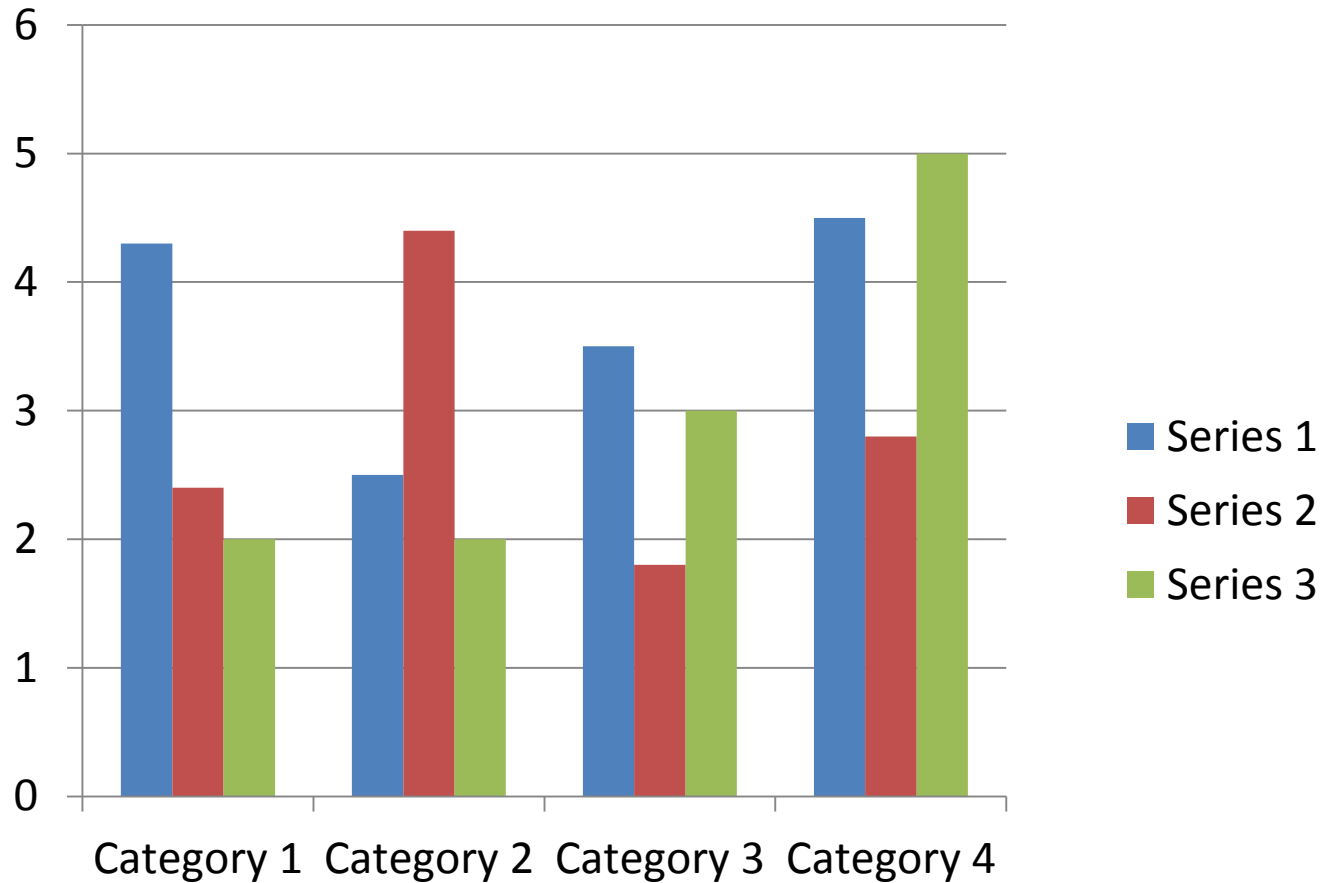
Valeur ?



Outil : Excel

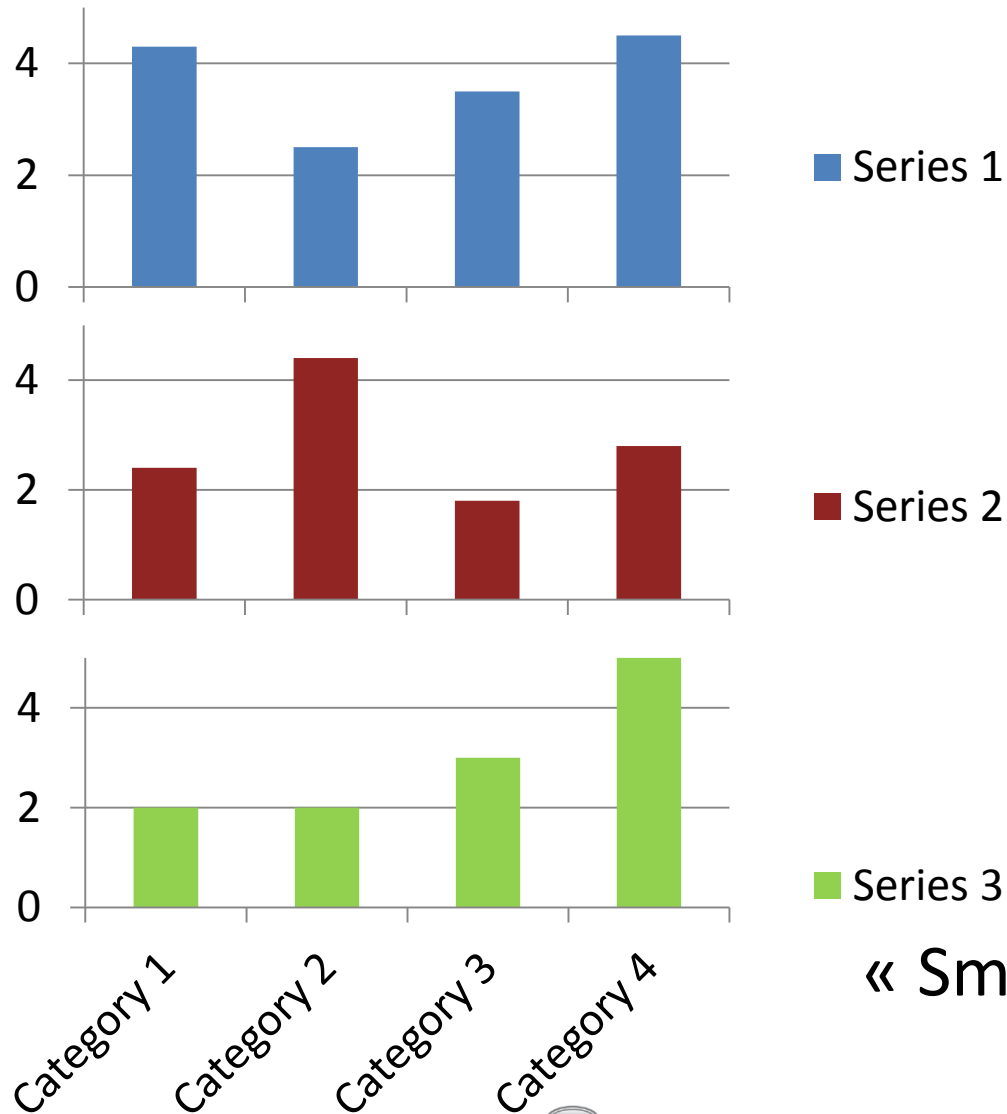
## Excel, « Histogramme 3D » par défaut

# Perspectives & 3D



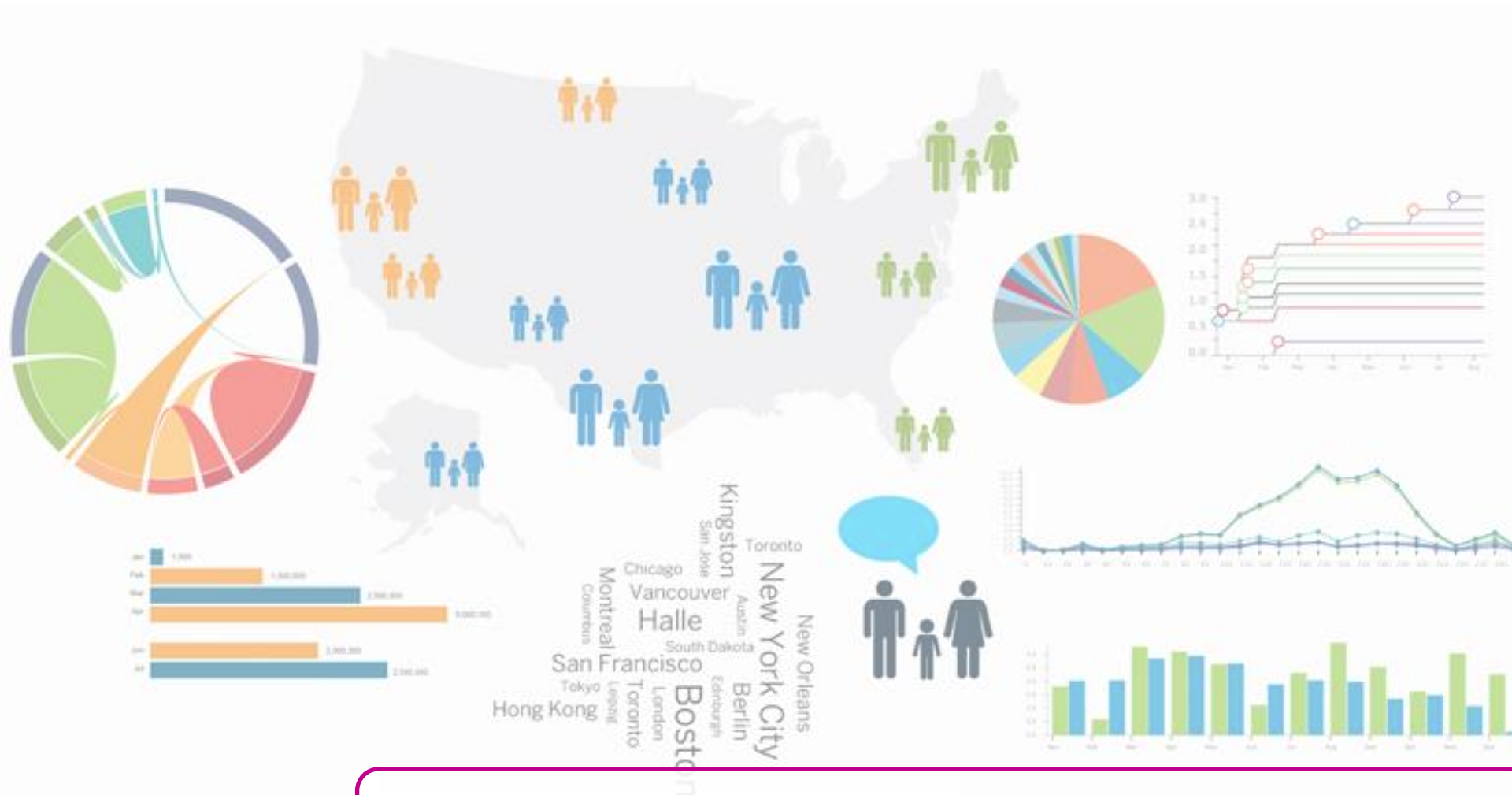
Outil : Excel

# Perspectives & 3D



« Small multiples »

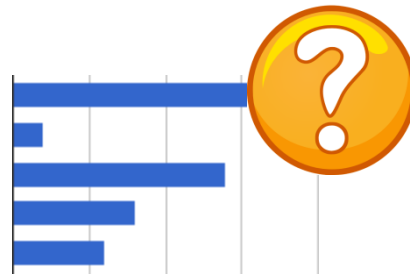
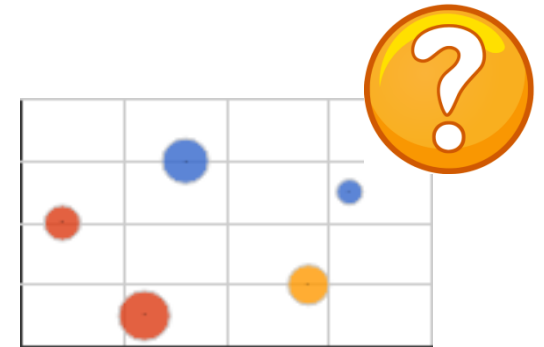
Outil : Excel



# Choix des graphiques

# Choix des graphiques

Key	Value
A	8
B	1
C	7
D	4
E	3



- Pas une science exacte !
- Dépend de beaucoup de paramètres
- Quelques *guidelines* peuvent aider

# Choix des graphiques

---

## Comparaison

Quelle province est la plus active ? Comment évolue le taux de chômage de 2014 à 2015 ?

## Composition

Comment les régions participent au résultat ? De quoi composé le budget de l'État ?

## Corrélation

Quel est le lien entre l'âge et répartition des genres ? La nationalité et le secteur ?

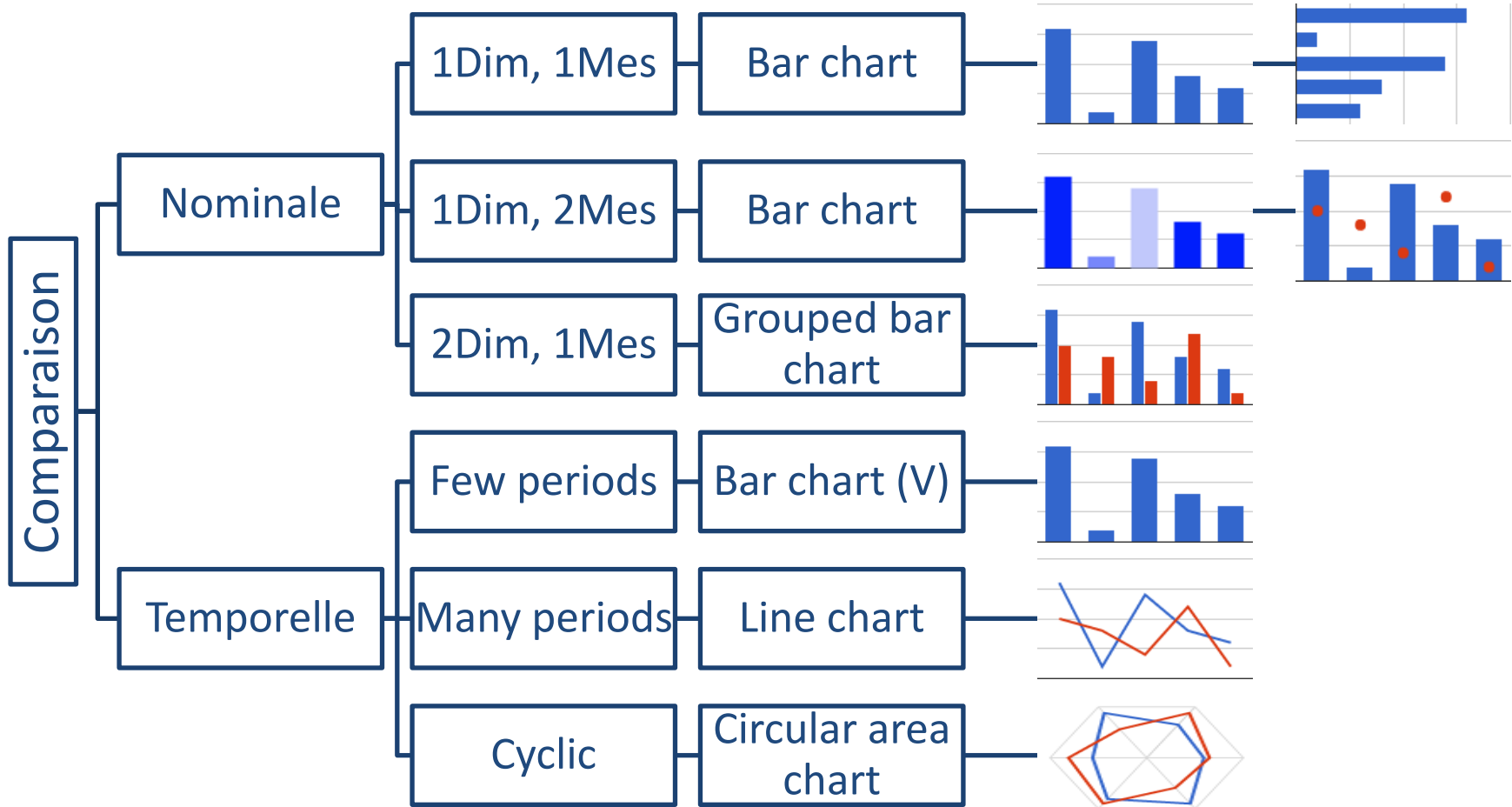
## Distribution

Quelle est la forme de la durée de chômage ? Combien de travailleurs par groupe d'âge ?

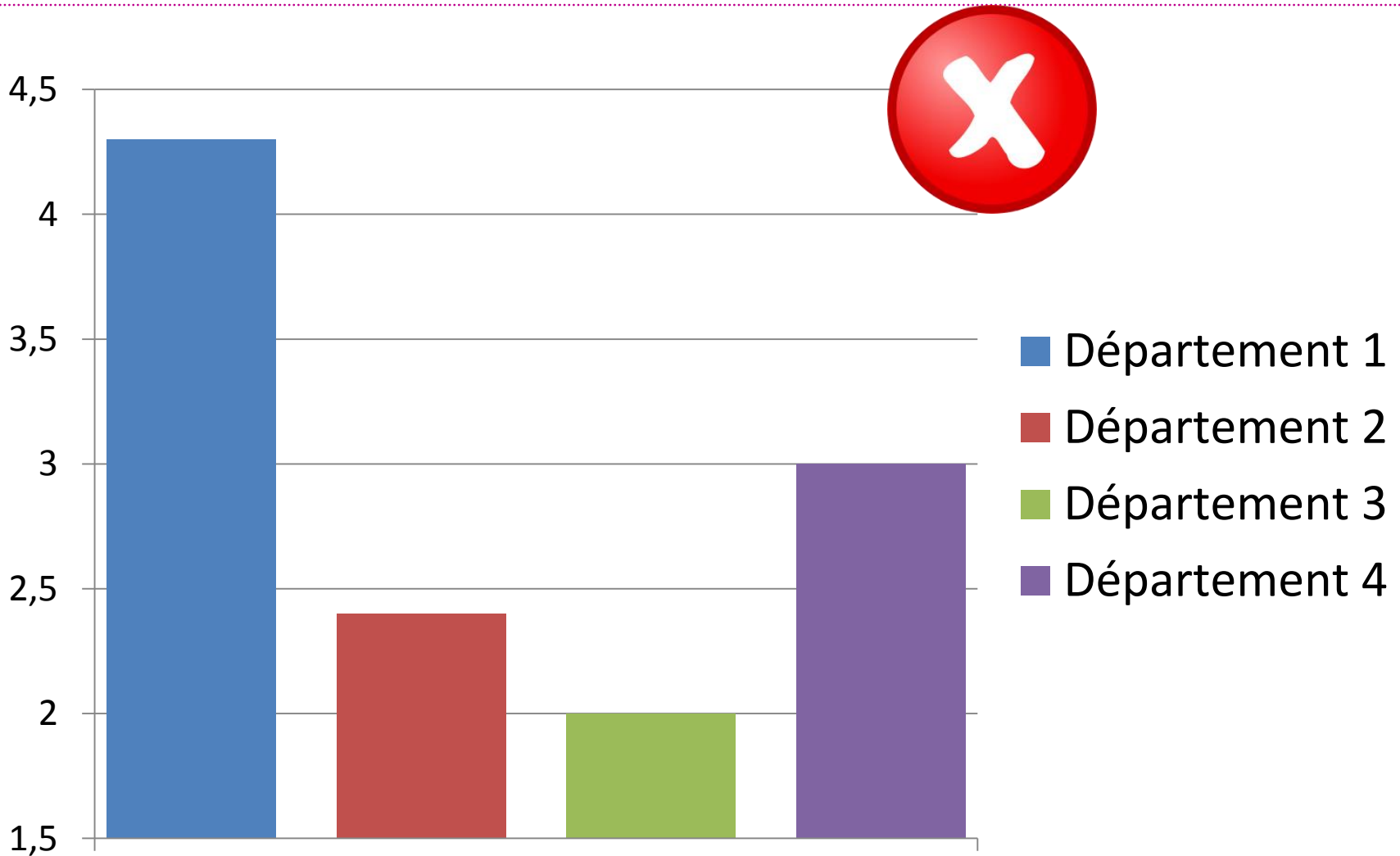
Choix des graphiques

# COMPARAISON

# Comparison



# Bar chart

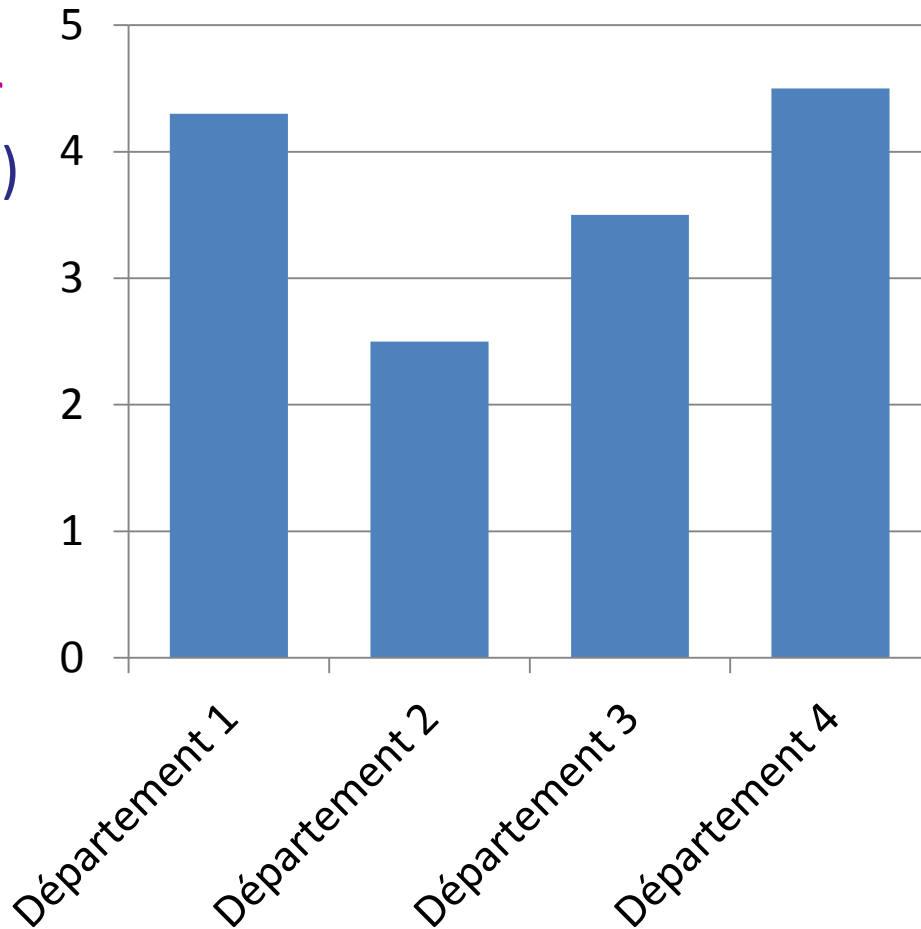


# Bar chart

Perception : Longueur  
(2x plus long = 2x plus)

Var. quantitative ratio

Tronquer = tromper !



Variable  
qualitative  
(nominale ou  
ordinaire)

Outil : Excel

# Line chart

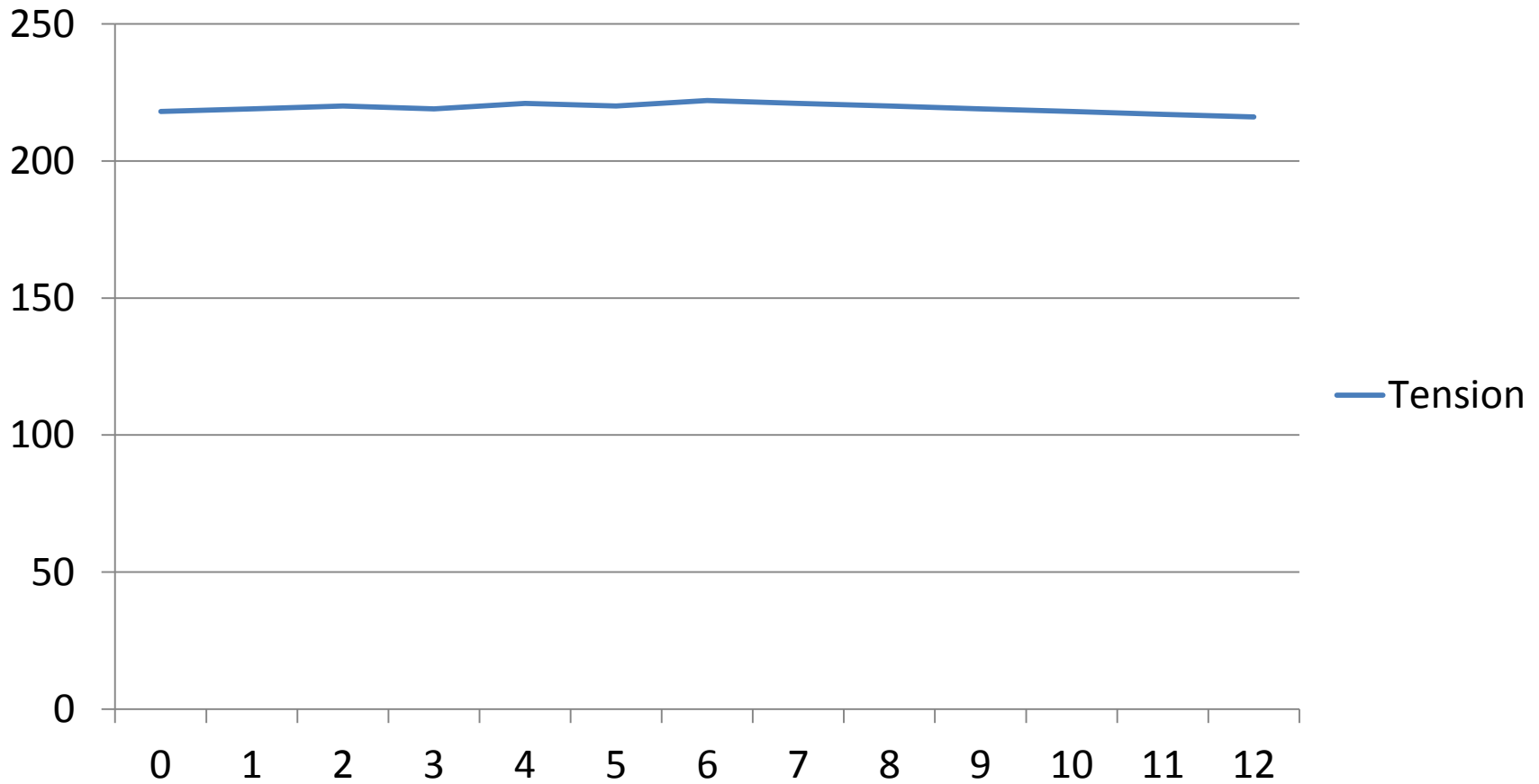
---

- Axe **vertical** :
  - Ce que l'œil perçoit, c'est la **position** des points
  - L'origine de l'axe n'a pas (peu) d'importance, peut être arbitraire... en étant « honnête »
- Axe **horizontal** : notion de **continuité**
  - Les valeurs **intermédiaires** ont un sens
  - Pas de variable **nominale** (rouge, vert, bleu...)
  - Variable **ordinaire** (petit, moyen, grand...) : à éviter

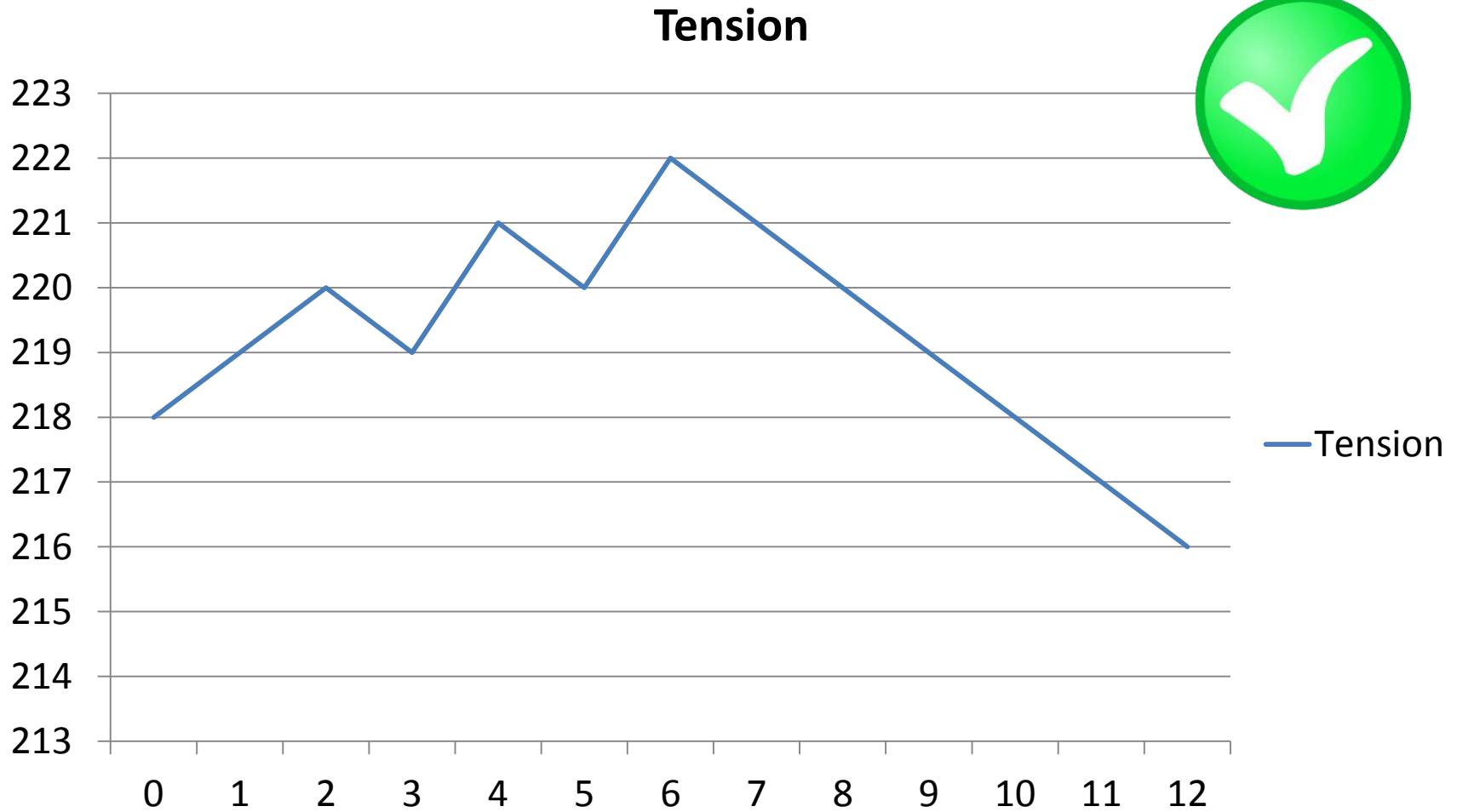


# Line chart

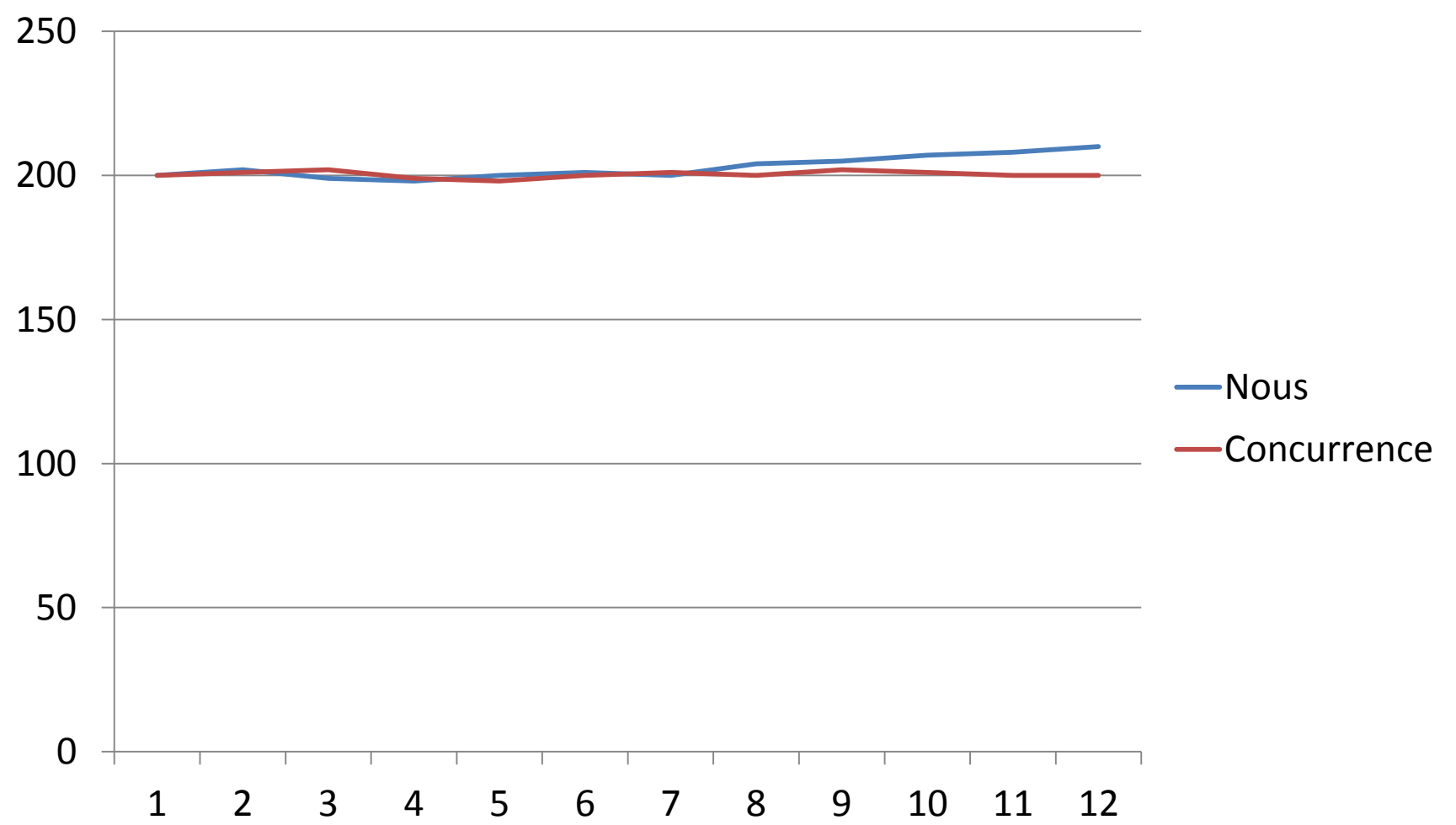
## Tension



# Line chart

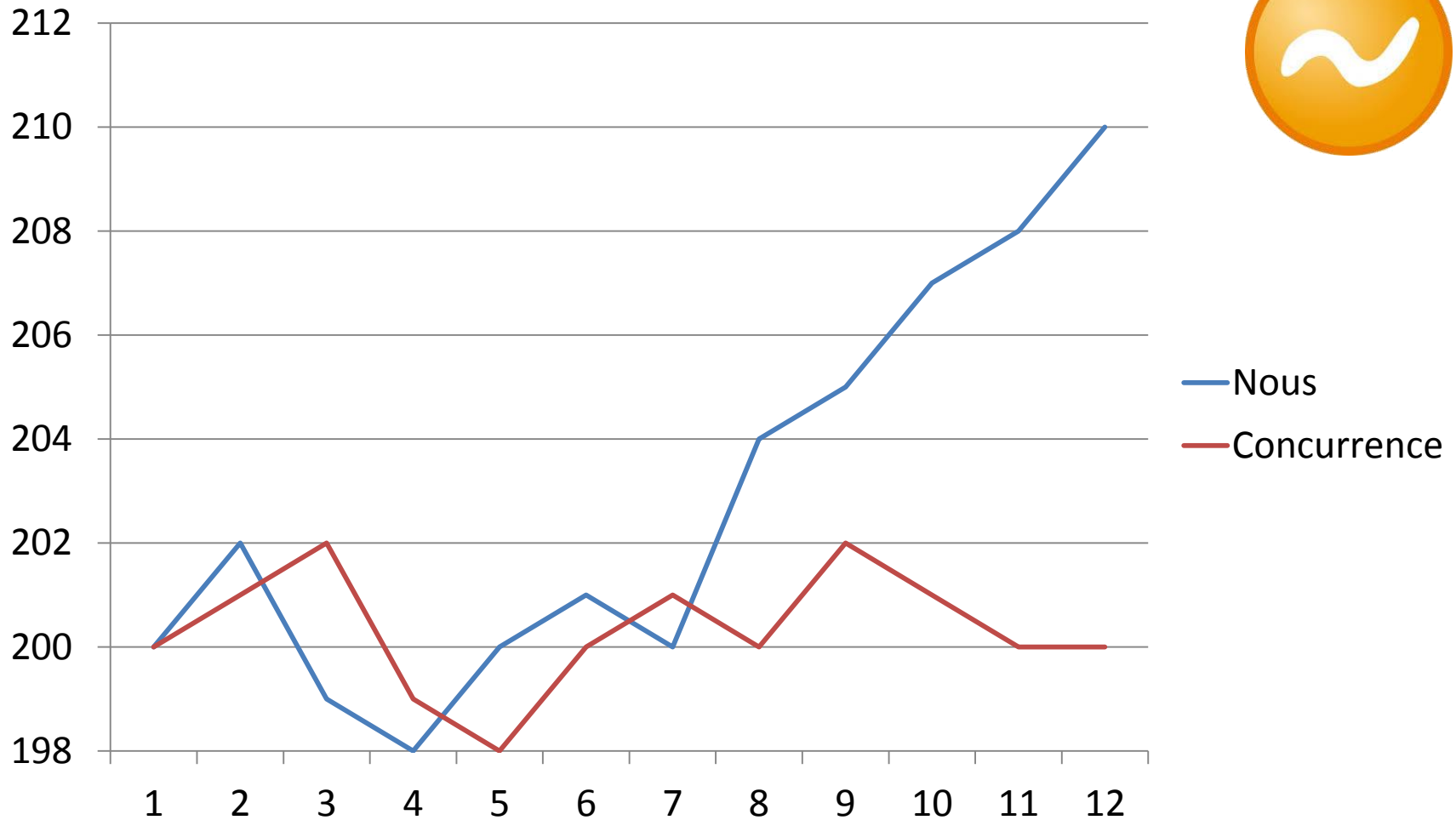


# Line Chart



Outil : Excel

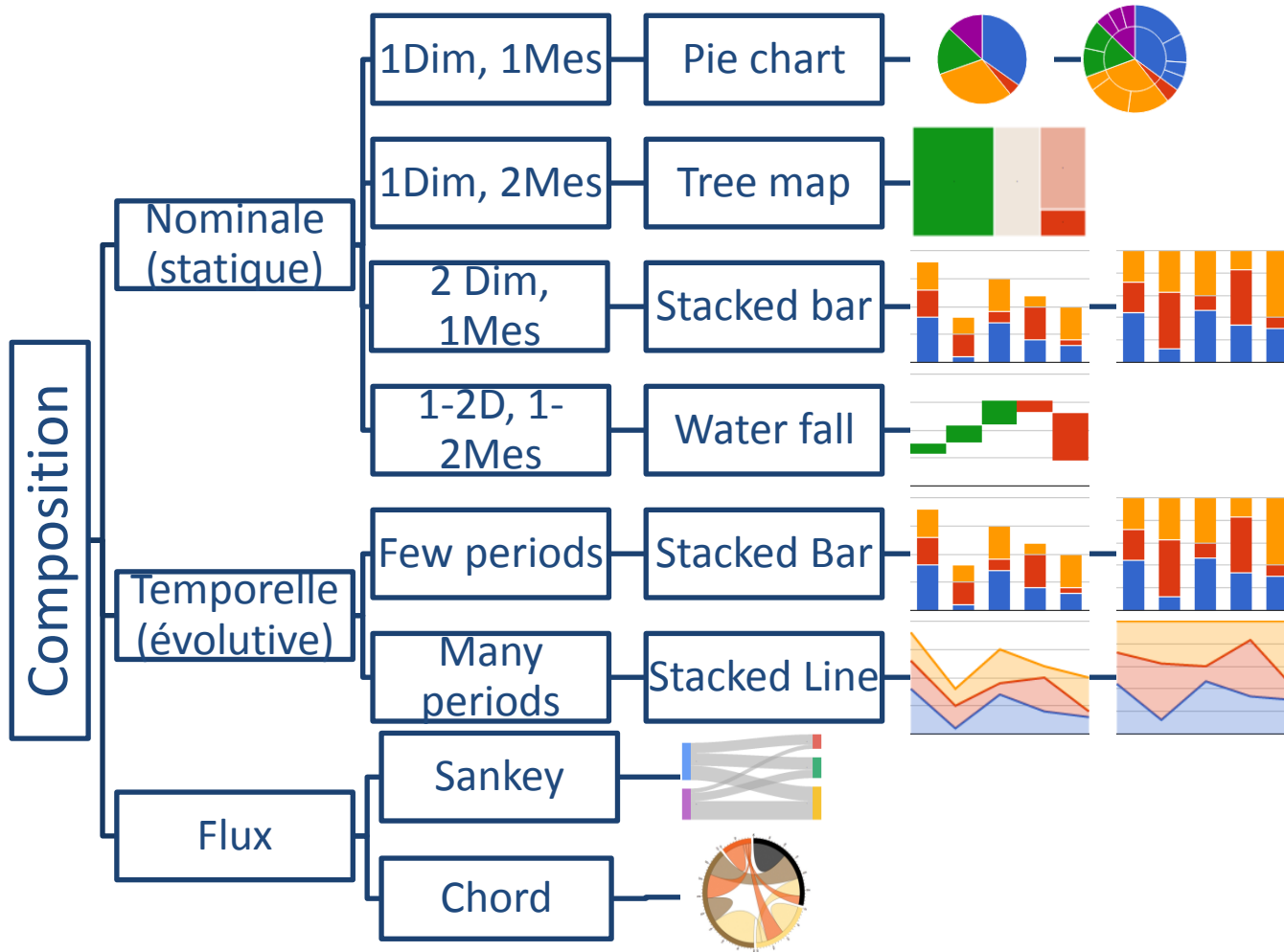
# Line Chart



Choix des graphiques

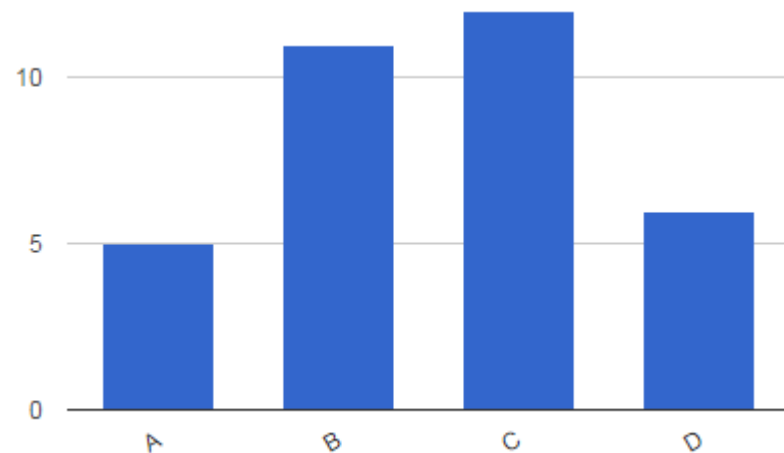
# COMPOSITION

# Composition



# Pie Chart

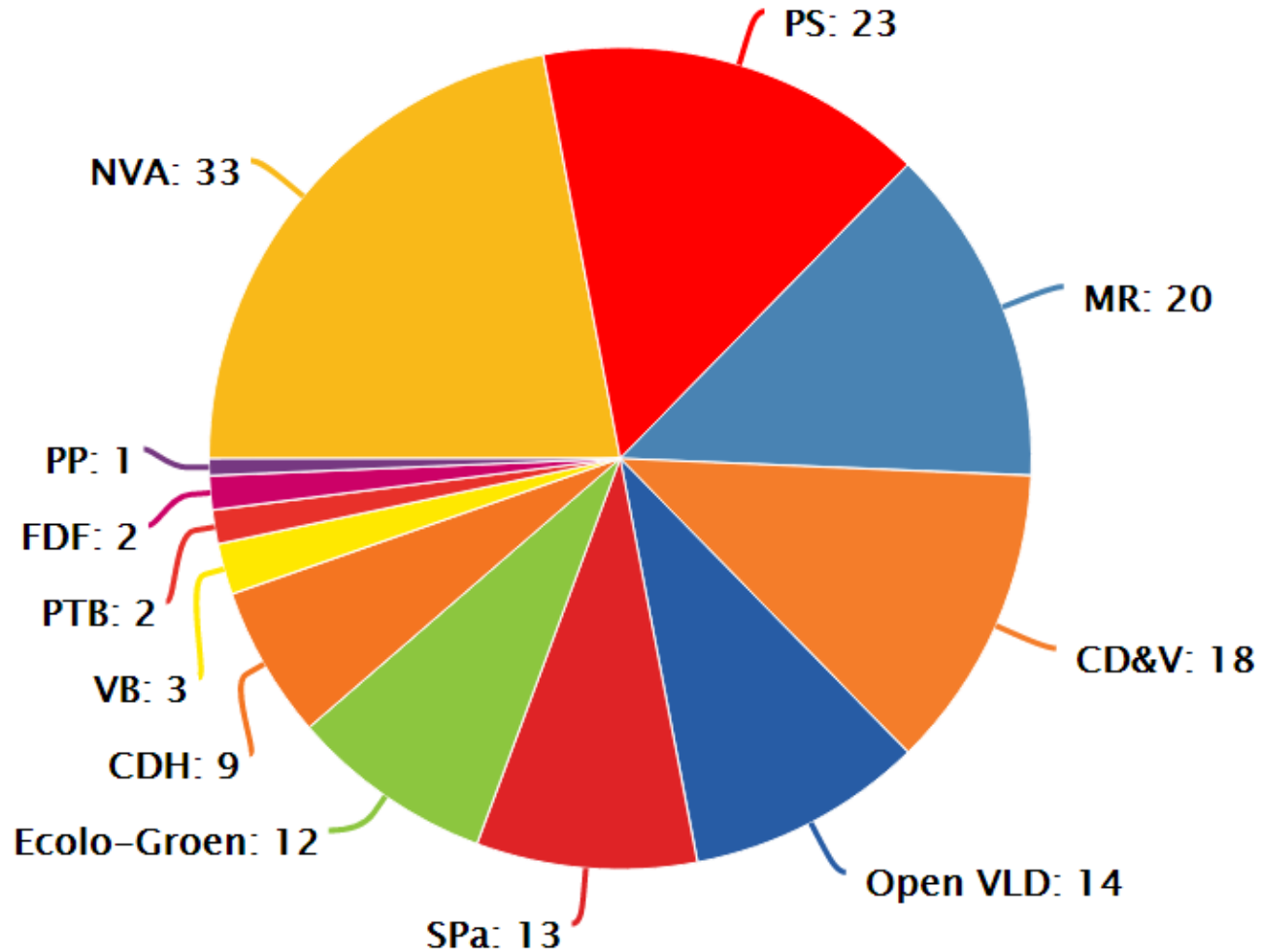
- Avantages :
  - Intuitif, facile à comprendre
  - Donne une bonne vue d'ensemble de la composition :  $D+A \approx B \approx C$
- Inconvénients :
  - Peu précis !  $A > D ? B > C ? A$  vs  $B$
  - Peu adapté au-delà de 3-4 tranches



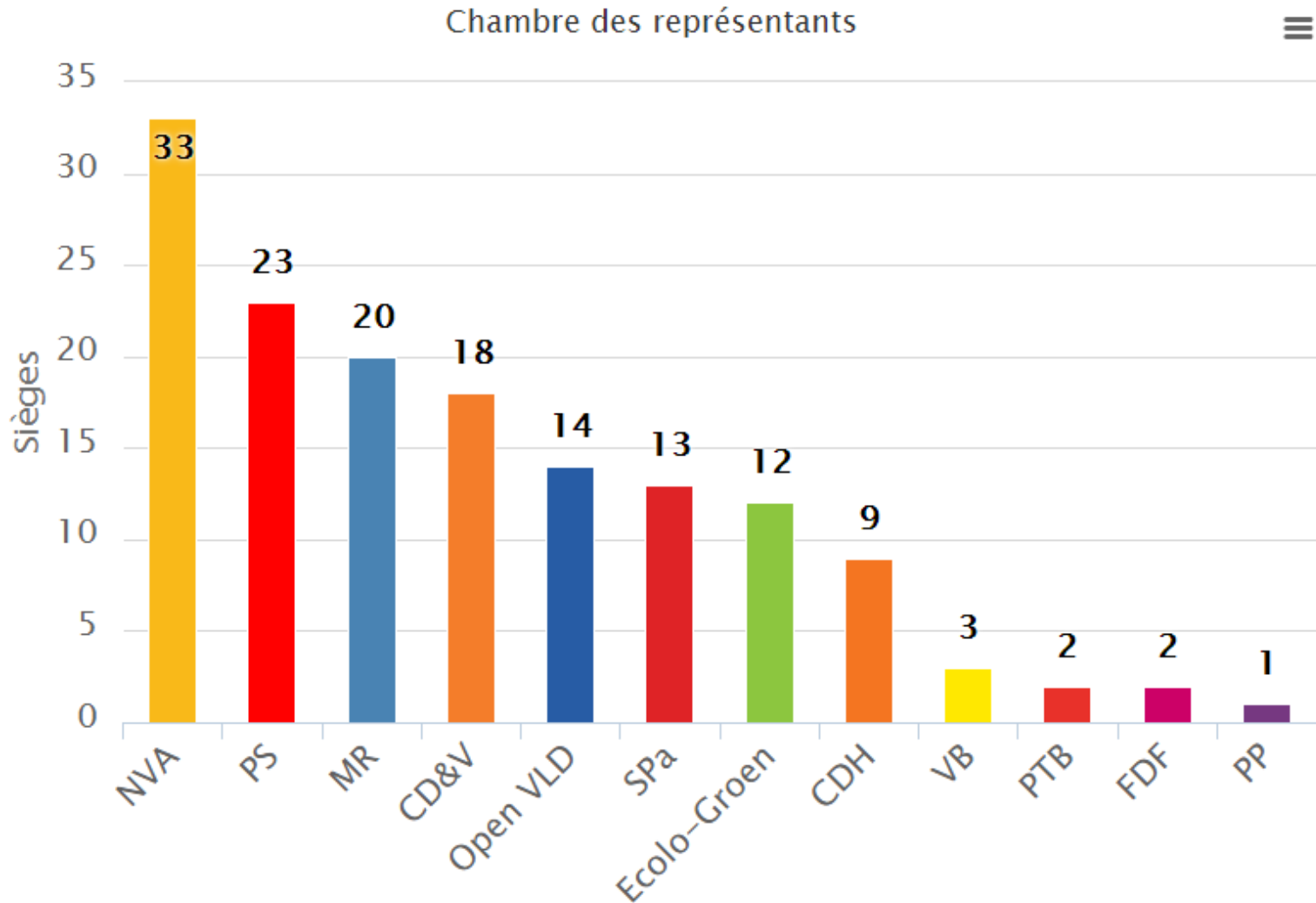
Outil : Google Charts

# Pie Chart

Chambre des représentants

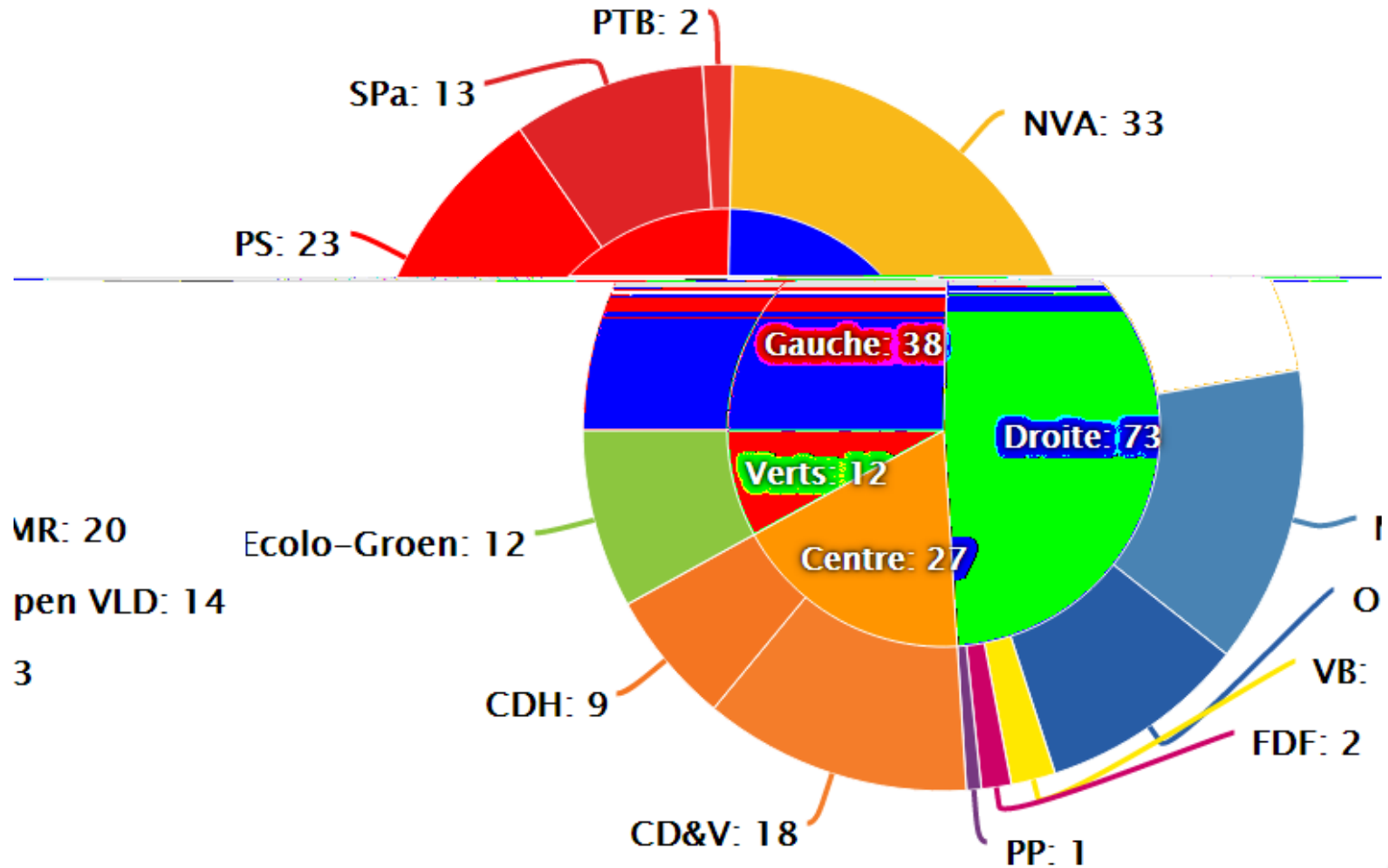


# Pie Chart vs Bar Chart



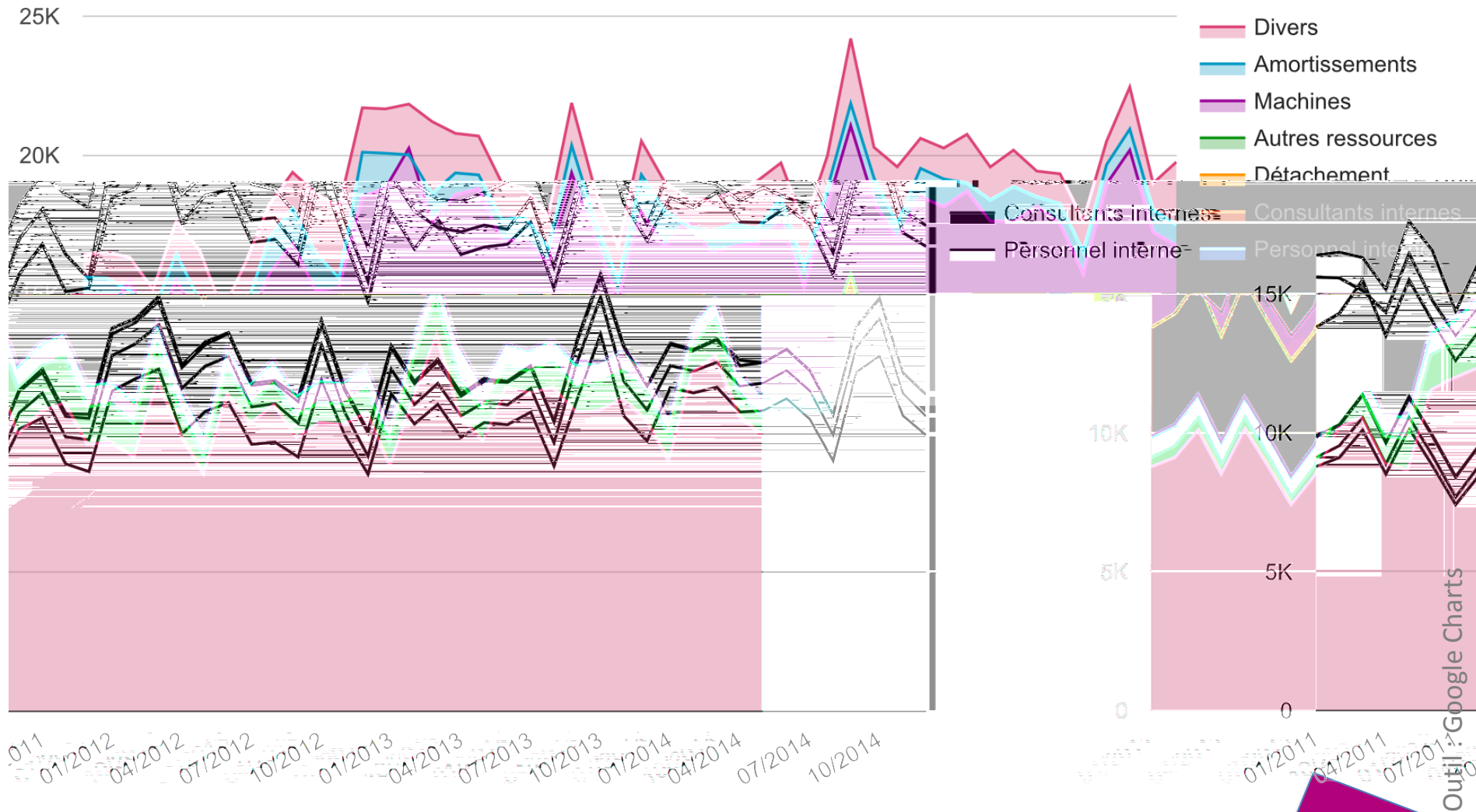
# Sun Burst

Chambre des représentants, Belgique

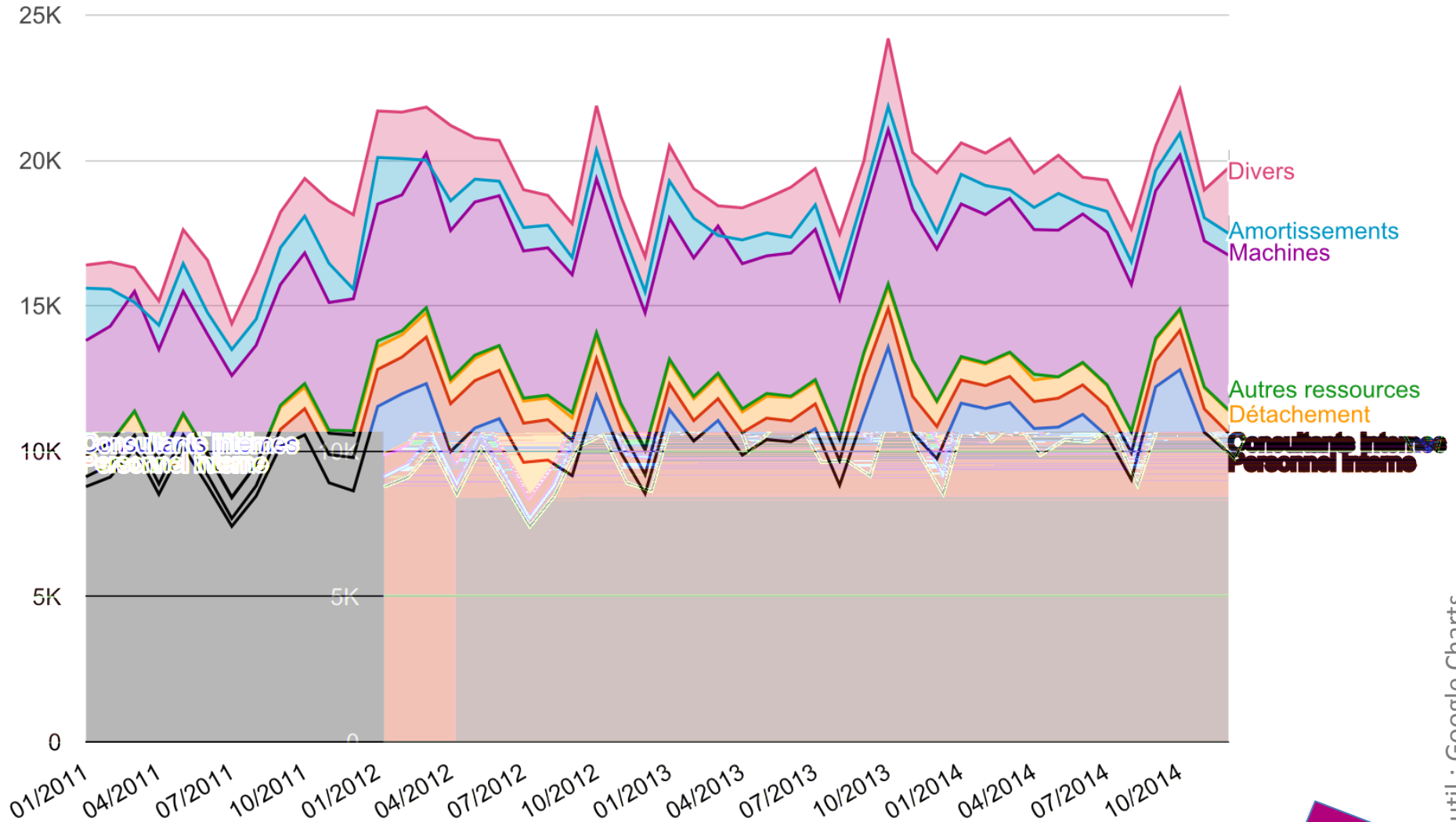


Outil : HighCharts

# Stacked line chart

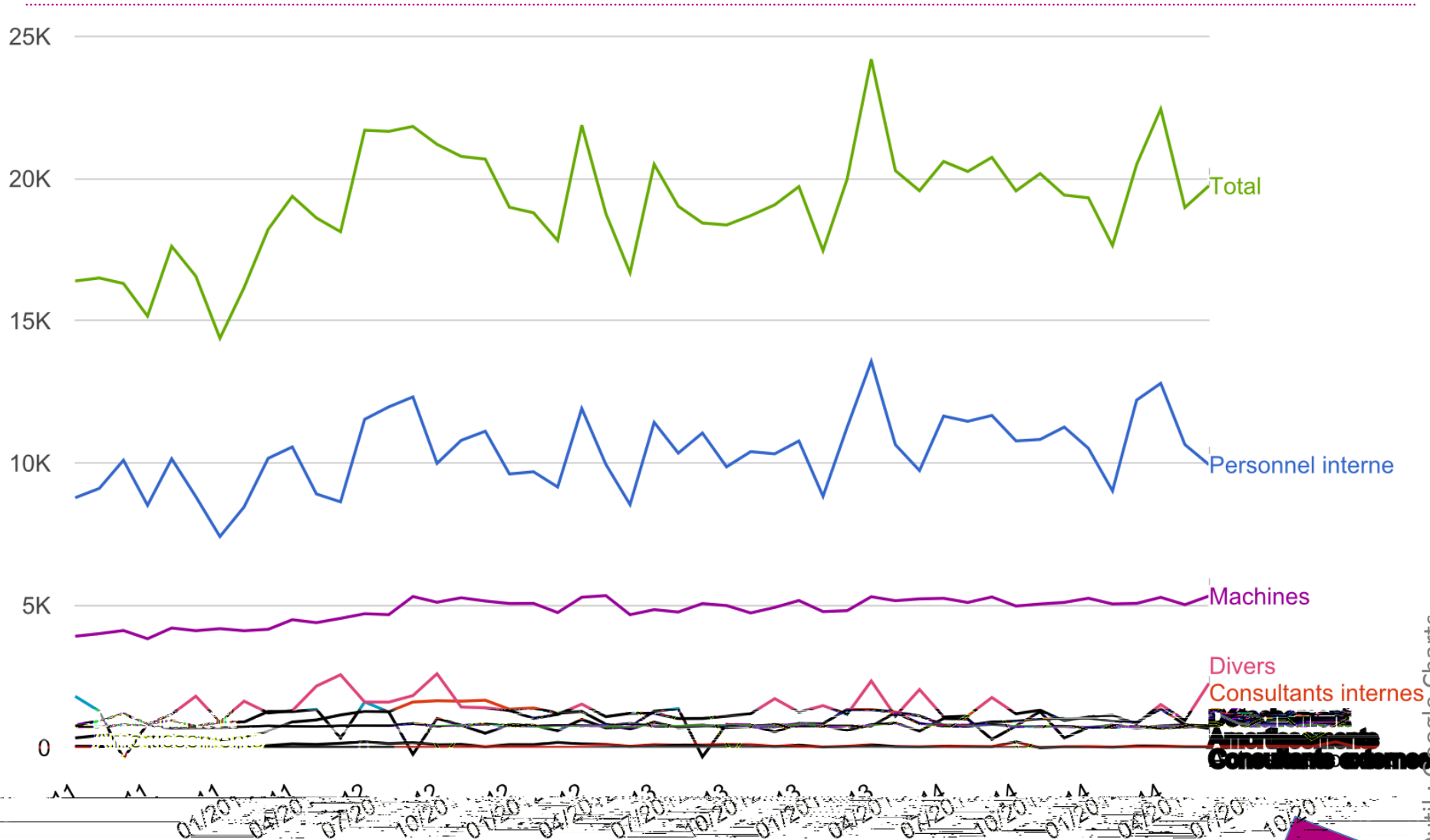


# Stacked line chart



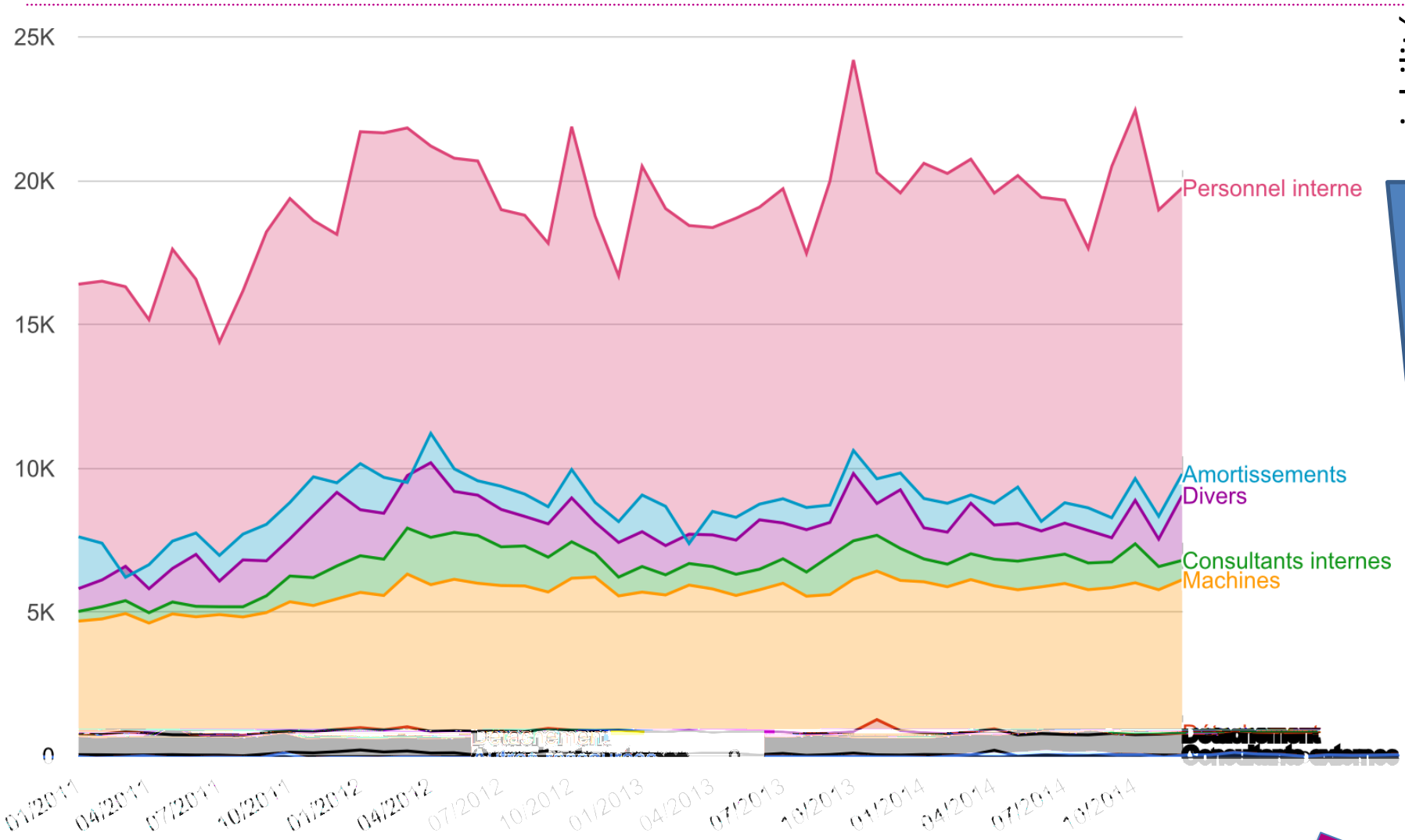
Outil : Google Charts

# Stacked line chart



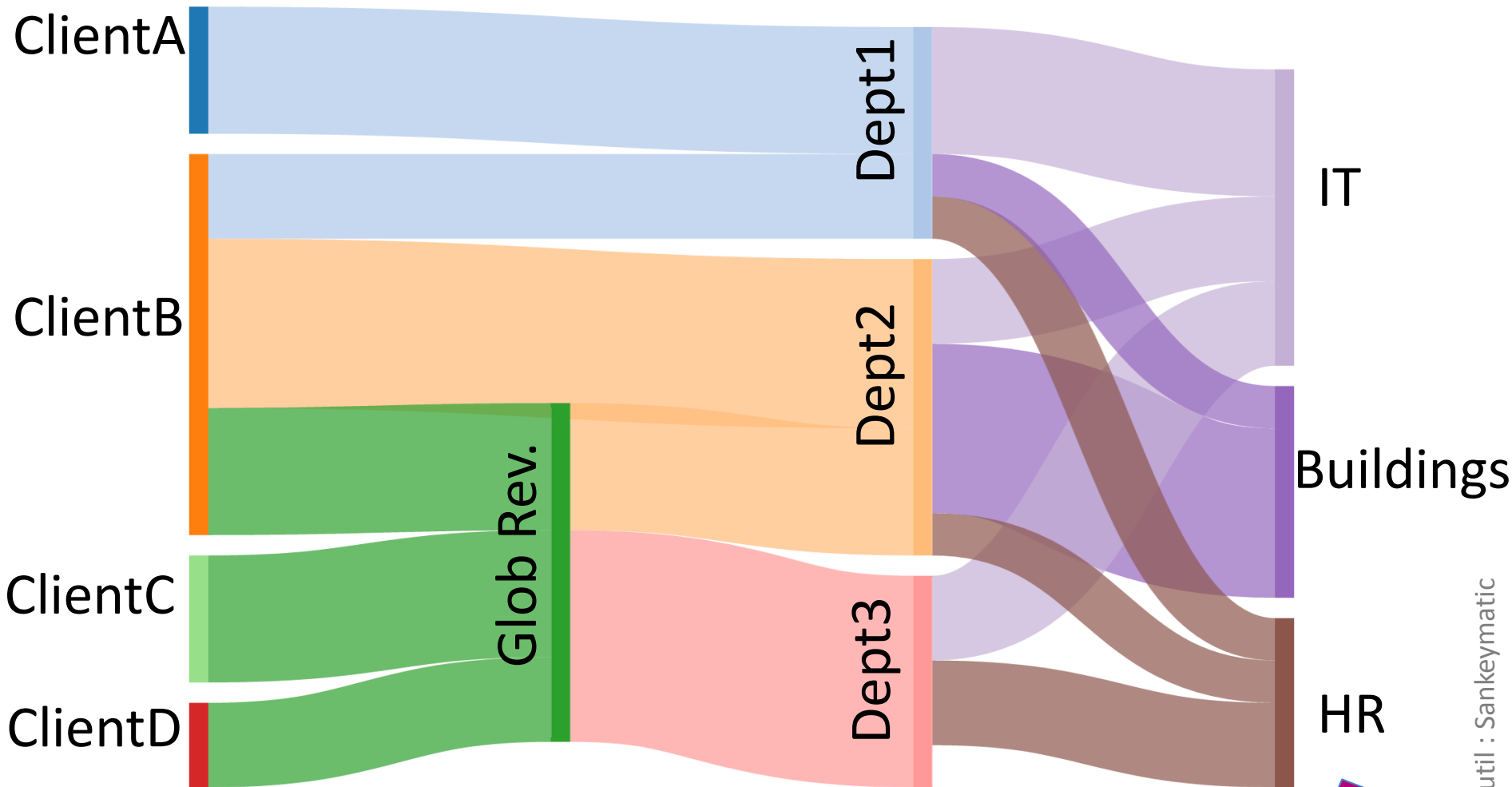
Outil : Google Charts

# Stacked line chart



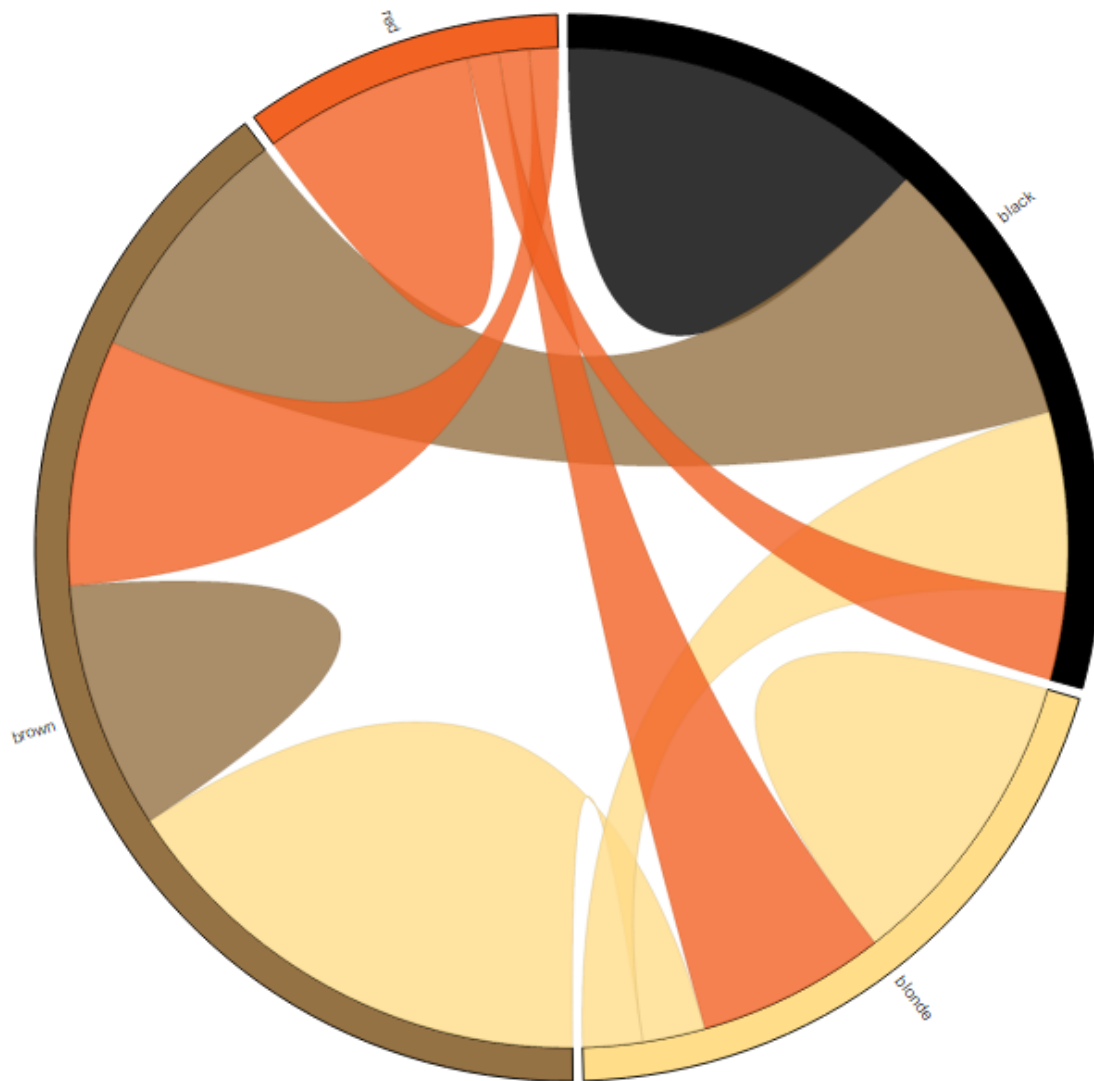
Outil : Google Charts

# Diagramme de Sankey



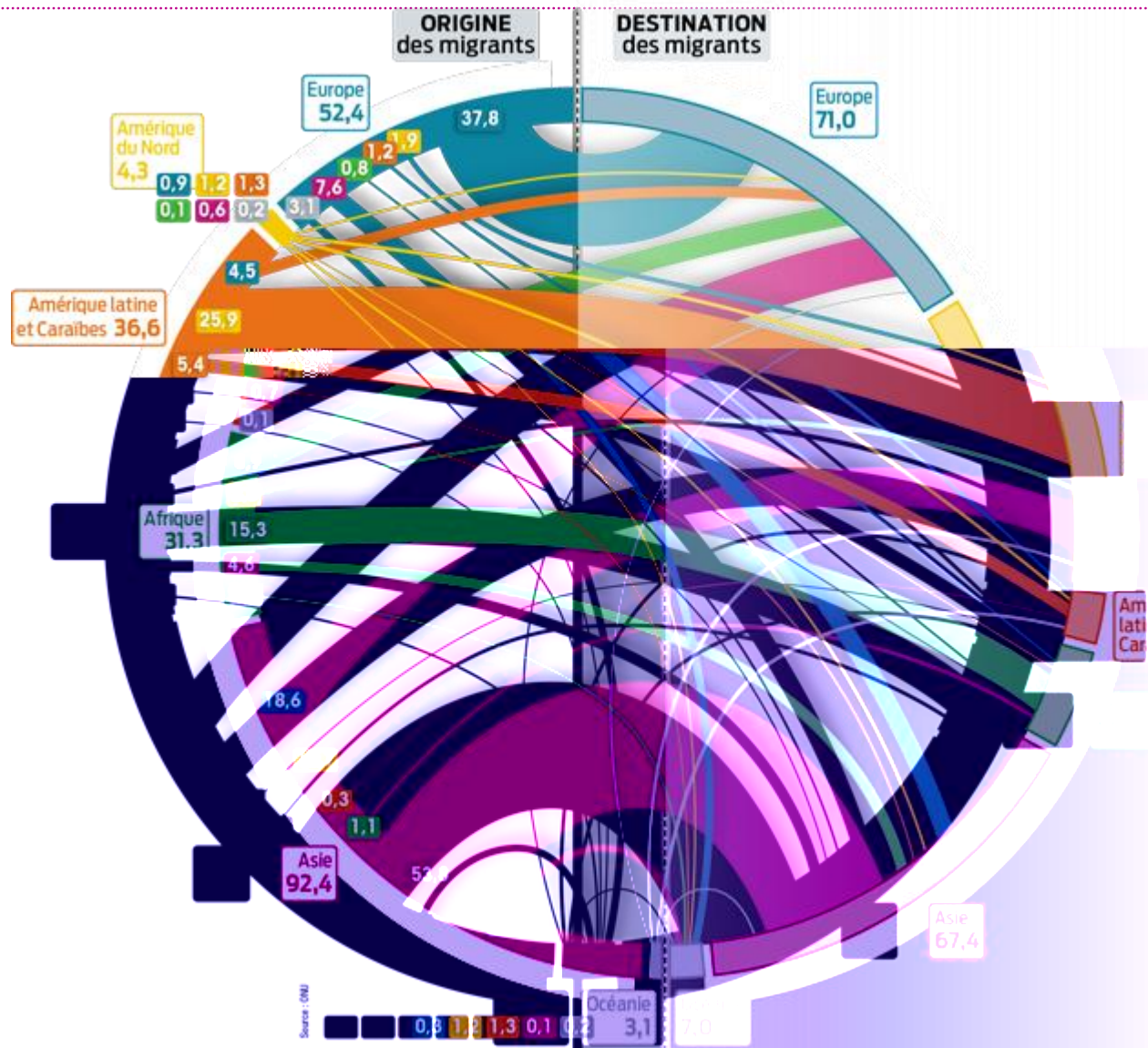
Outil : Sankeymatic

# Diagramme de Chord



<http://www.delimited.io>

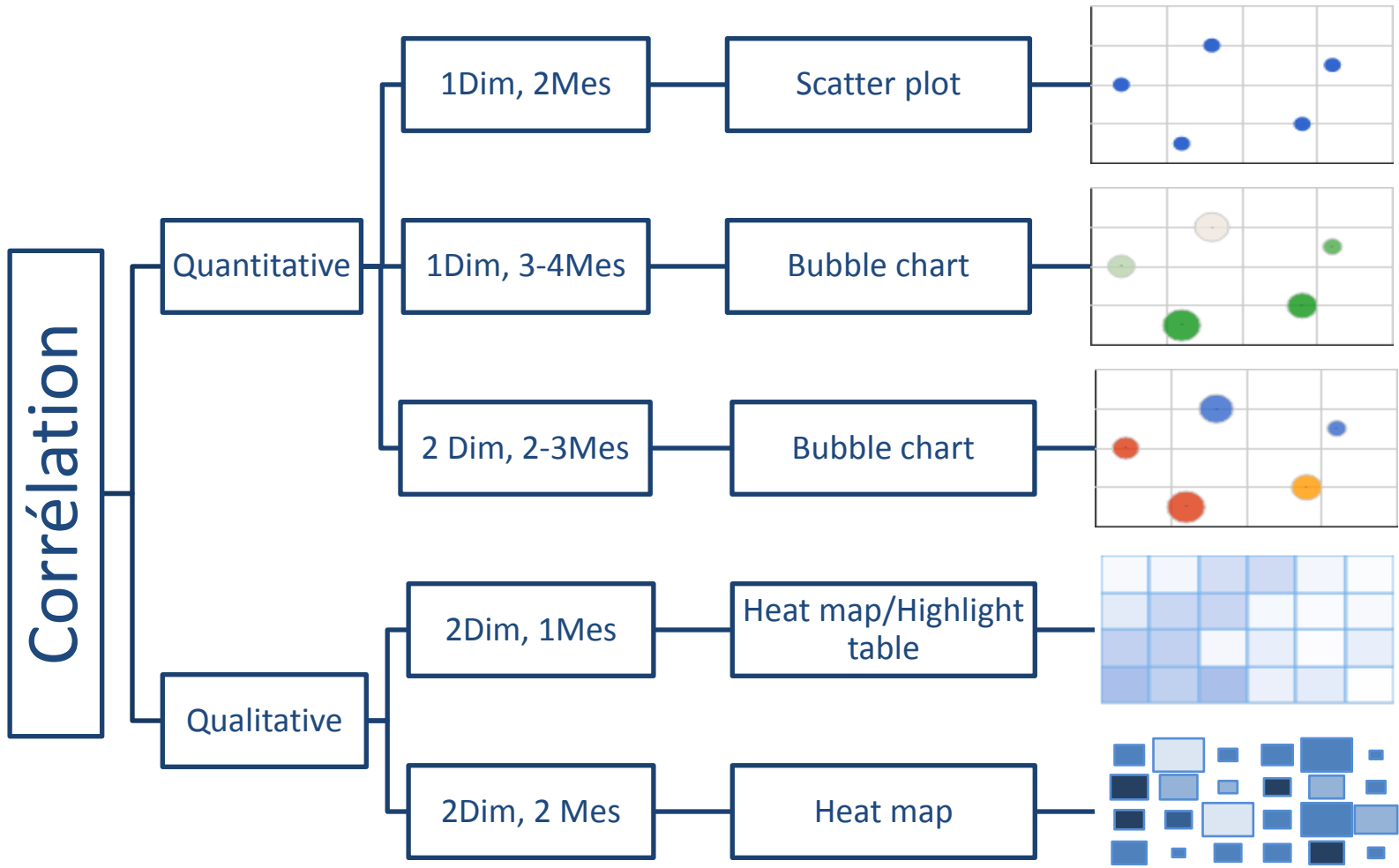
# Diagramme de Chord : migrations 2013



Choix des graphiques

# **CORRÉLATION**

# Corrélation



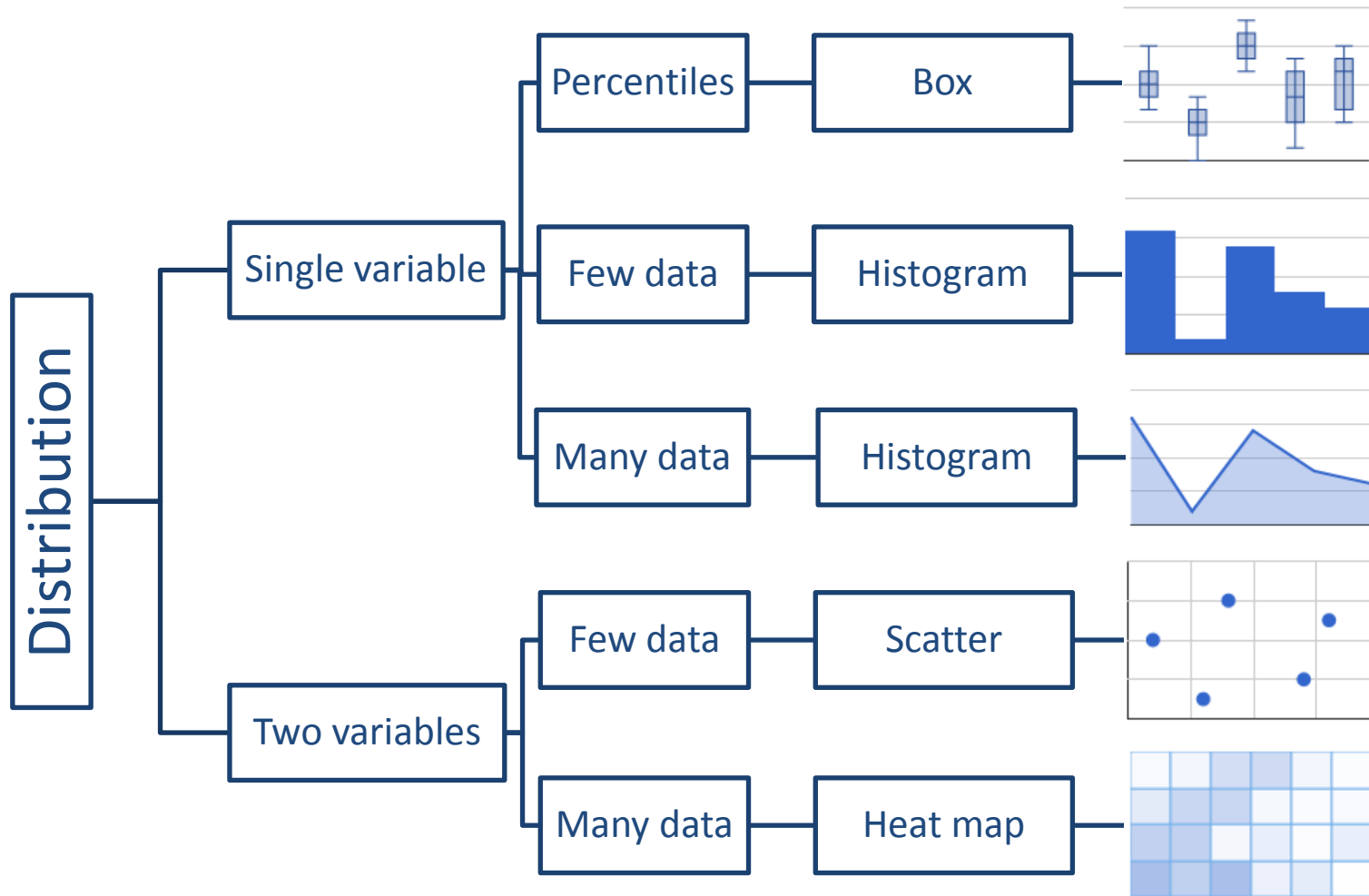
# Bubble Chart



Choix des graphiques

**DISTRIBUTION**

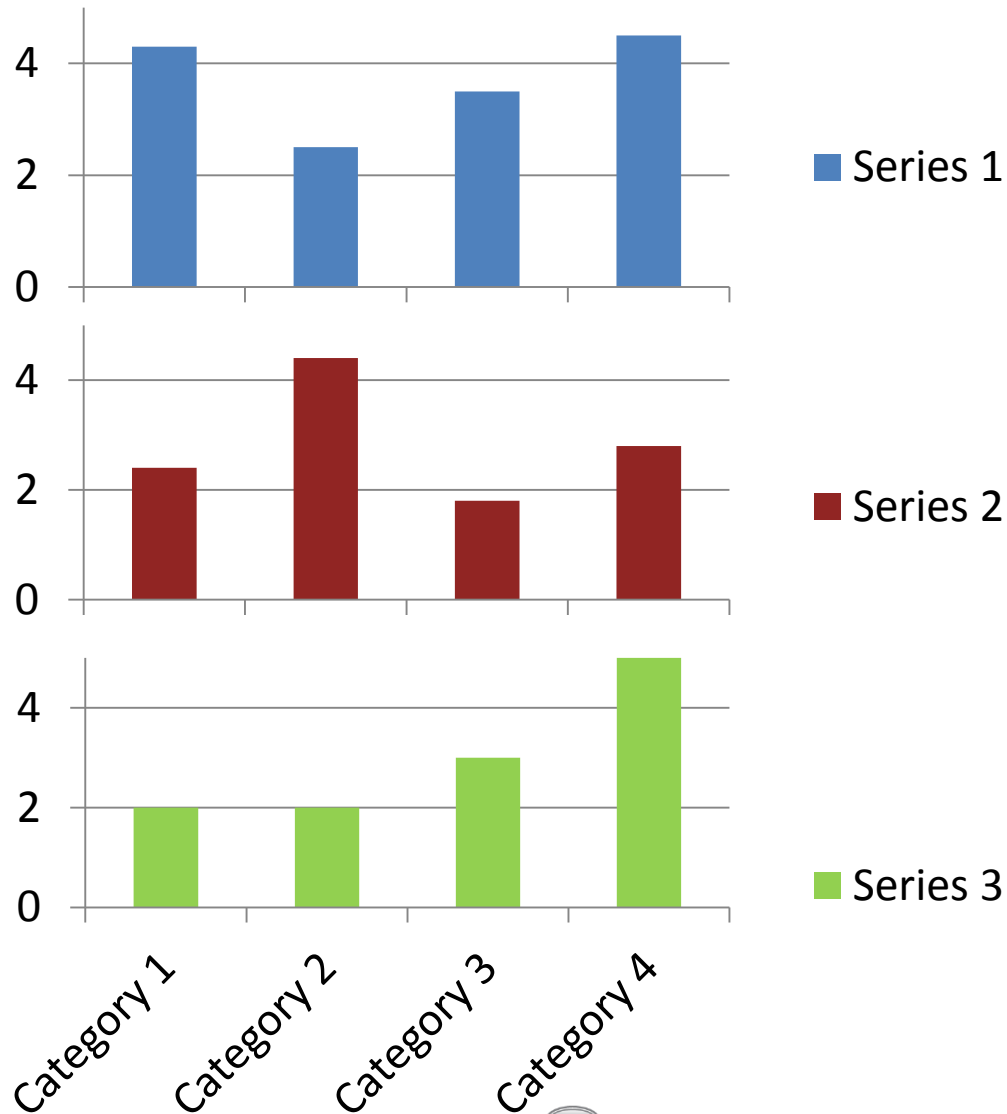
# Distribution



Choix des graphiques

**NOMBREUSES VARIABLES**

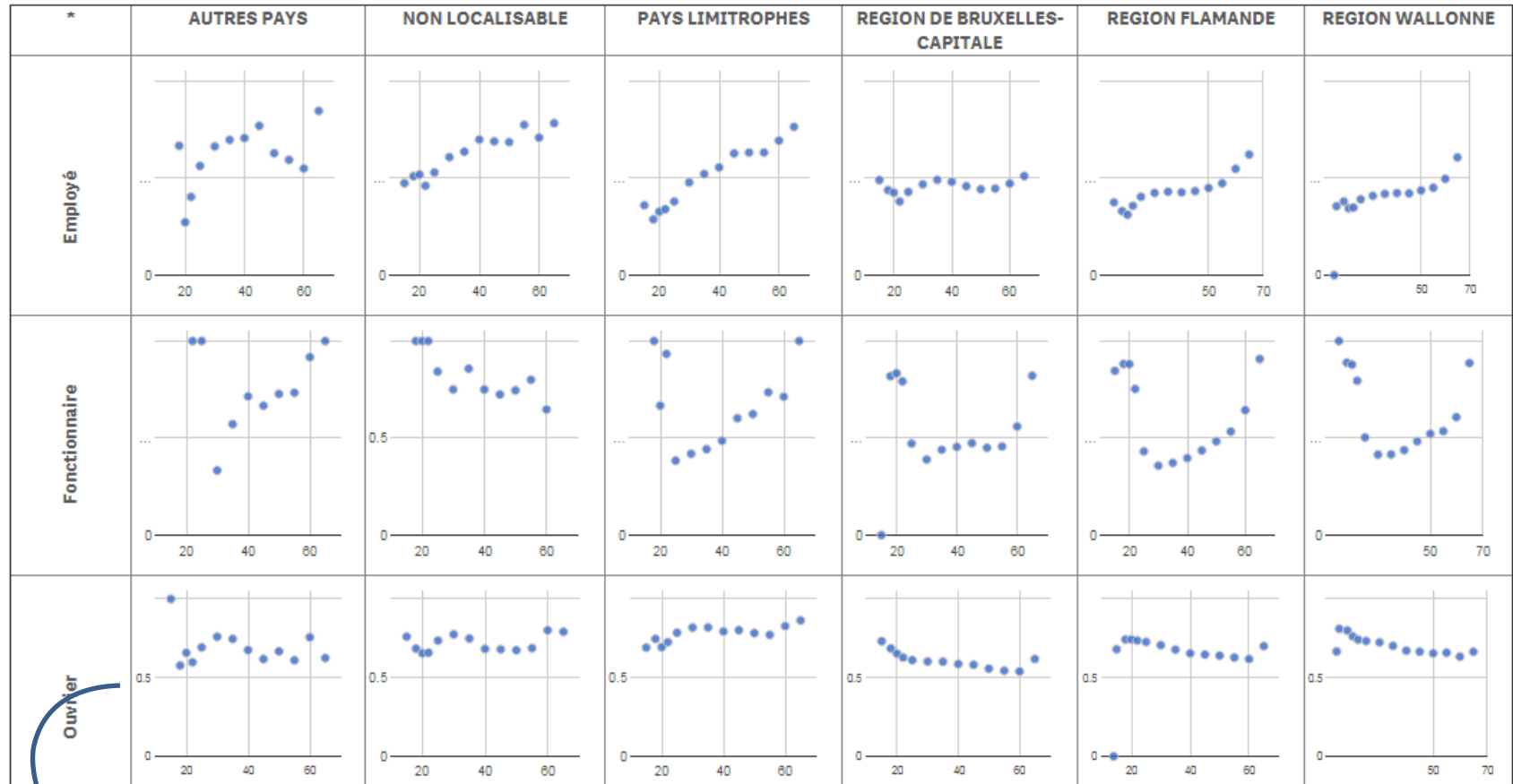
# Small multiple



# Matrix Plot

## Valeurs d'une dimension B

Valeurs d'une dimension A



Ratio ♂/♀

Age

Choix de graphiques

# RÉSEAUX

# Réseaux

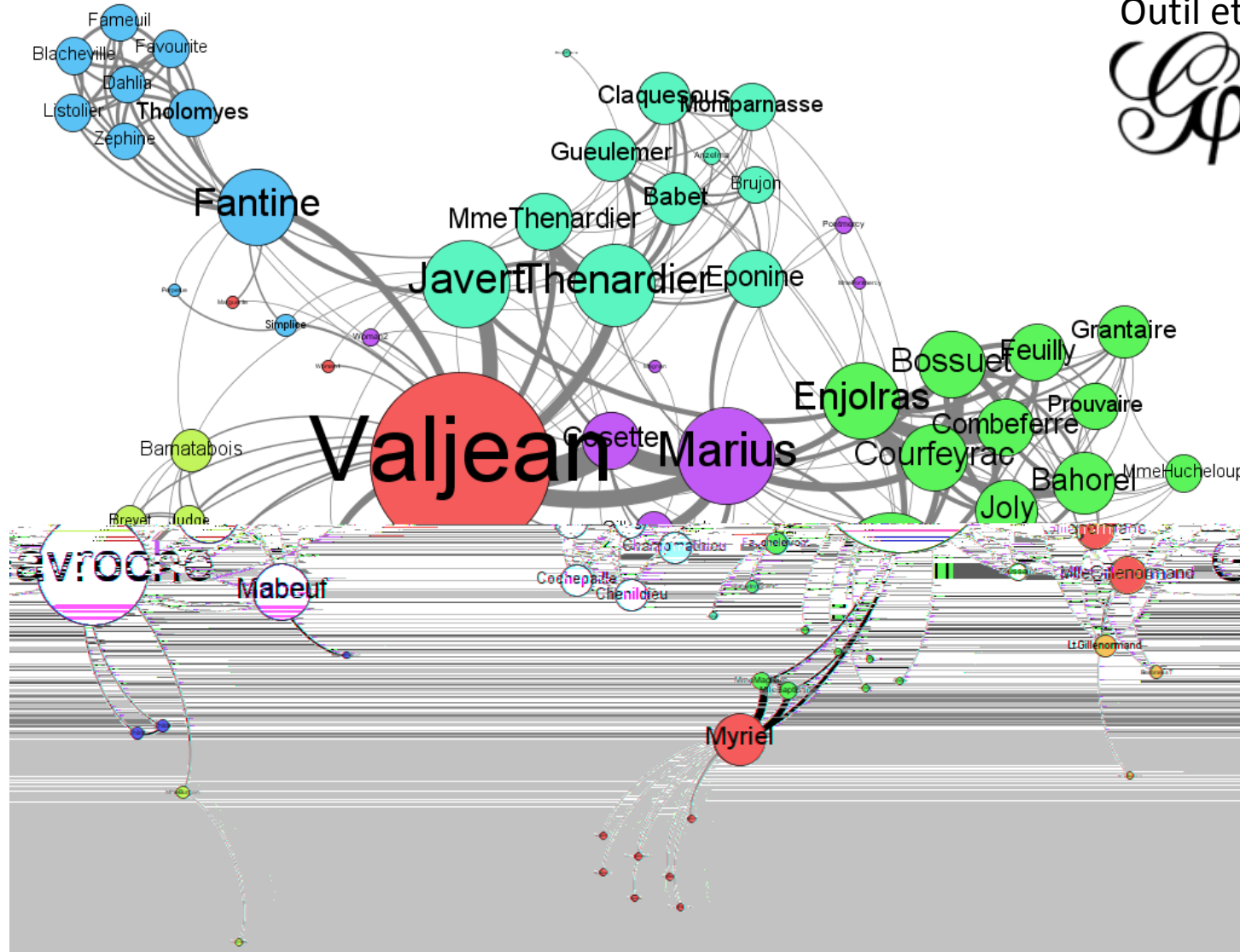
---

- Représentation d'entités (nœuds) + relations entre elles (arcs)
- Réseau social : personnes + liens d'amitié/intérêt
- Pages web + hyperliens
- Entreprises/Fournisseurs + liens contractuels
- Arbre généalogique
- Organigramme
- ...



# Réseaux : les misérables (V. Hugo)

Outil et données :

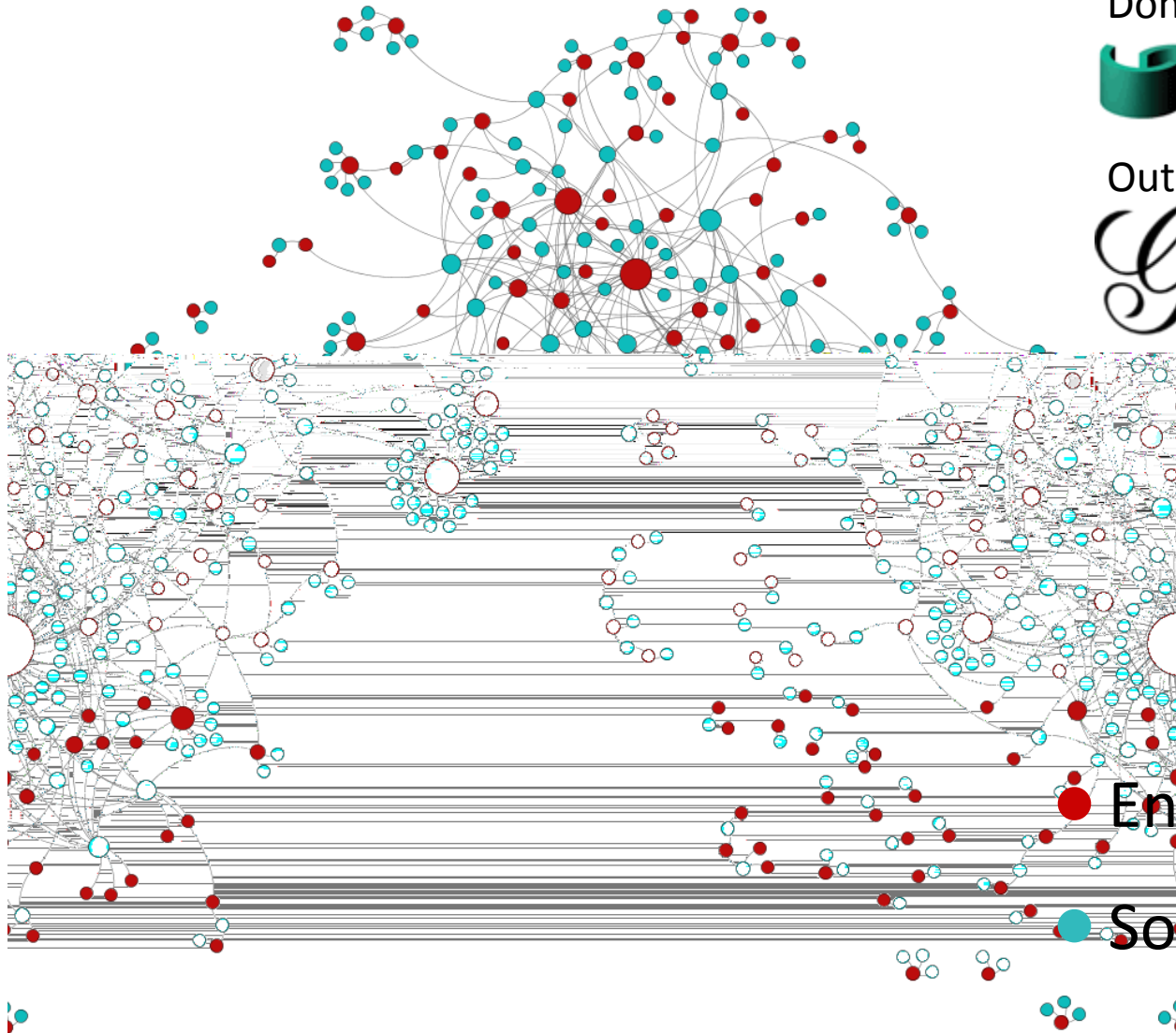


# Réseaux : Entrepreneur – sous-traitant

Données :



Outil :



Entrepreneur

Sous-traitant



# Questions ?

---



# Pause !

---



Choix des graphiques

**OUTILS**

# Outils

---

- Outils avec interface « user-friendly », données déjà agrégées
  - Composant d'une suite (Excel et autres suites bureautiques), souvent statique
  - Outils spécialisés, en ligne, interactif
- Outils avec interface « user-friendly », données brutes
  - Outils de type « visual analytics »
- Bibliothèques :
  - Javascript
  - Python
  - R (Shiny), gnuplot



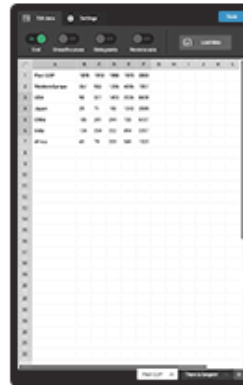
# Outils spécialisés : infogr.am



1

## Choose a Template

We have multiple templates you can choose from



2

## Visualize your data

Add charts, maps, videos, images, icons and more

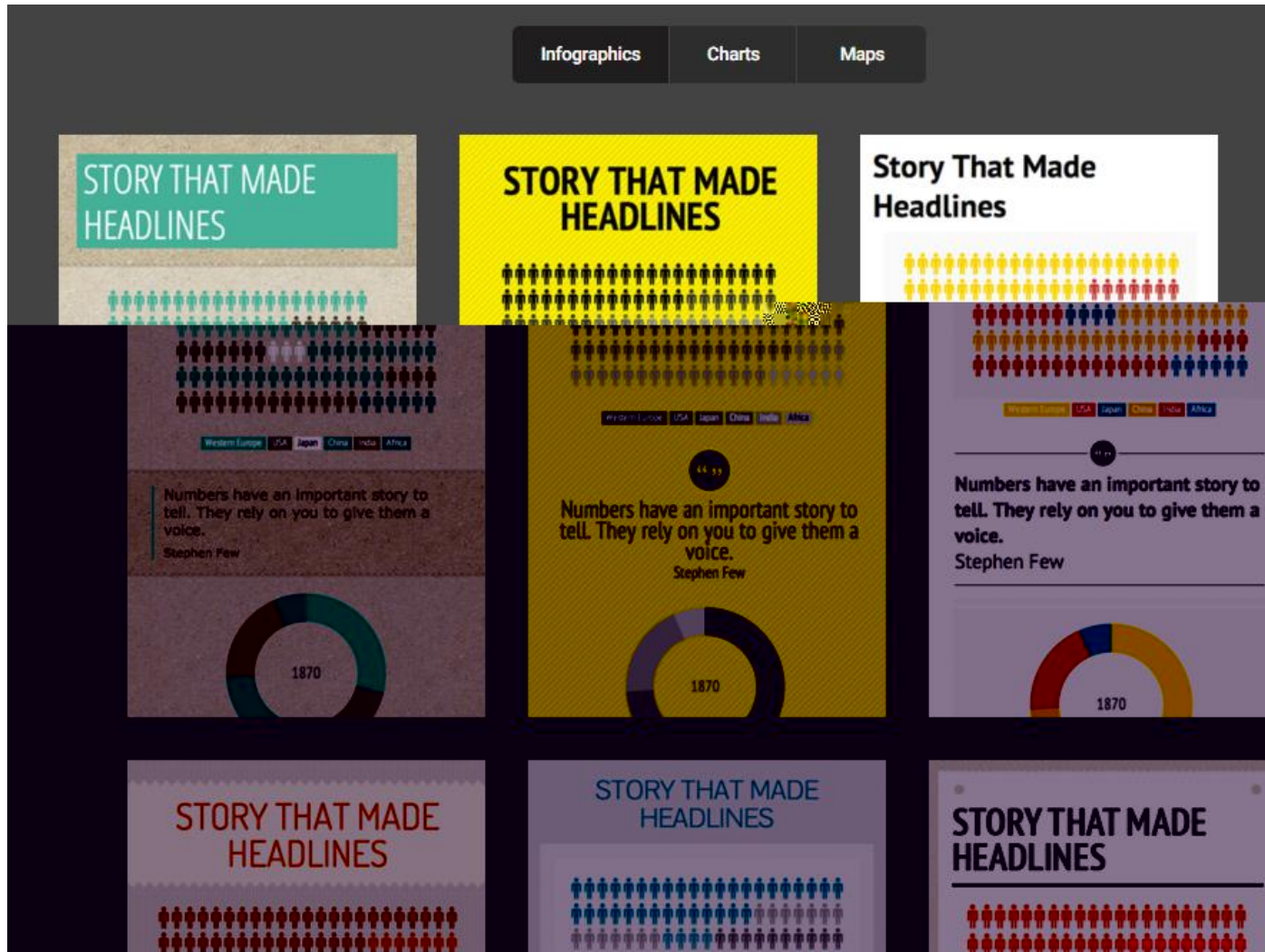


3

## Publish & Share

Use our one-click share buttons or easily generate an embed code to post on your blog or website

# Outils spécialisés : infogr.am



# Outils spécialisés : infogr.am



# Outils spécialisés : infogr.am

The screenshot displays the infogr.am interface. On the left, a data table is visible with columns A through J and rows 1 through 19. The data is as follows:

	A	B	C	D	E	F	G	H	I	J
1		1870	1913	1950	1973	2003				
2	Europe	367	902	1396	4096	7857				
3	USA	98	517	1455	3536	8430				
4	Japan	25	71	160	1242	2699				
5	China	189	241	244	739	6187				
6	India	134	204	222	494	2267				
7	Africa	45	79	203	549	1322				
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										

On the right, a stacked bar chart is displayed with the title "Double-click to add title". The chart shows GDP values for six regions (Europe, USA, Japan, China, India, Africa) across five years (1870, 1913, 1950, 1973, 2003). The x-axis represents GDP values from 0 to 25,000. The legend indicates the following colors for the regions: Europe (yellow), USA (red), Japan (teal), China (orange), India (dark red), and Africa (blue). Below the chart, there is a text prompt: "What's the story you're telling with your data? Write more here, or delete this object." At the bottom right, there are buttons for "Share", "Copyright notice", "Create infographics", and a help icon.


# Outils spécialisés : infogr.am

## Publishing

Test


test

 Publish for everyone 

 Share privately 

Cancel

## Share

 This infographic is published on the web!  
[Change it to only accessible with a private link.](#)

Change type: **Interactive** ▾

<https://infogr.am/test-095164169455>

[View on web](#)



Embed

Email

Wordpress

Responsive 

Fixed 

```
<script id="infogram_0_test-095164169455"
src="//e.infogr.am/js/embed.js?ZVa"
type="text/javascript"></script><div
style="width:100%;border-top:1px solid
```

Close

# Outils spécialisés : infogr.am

---

- Version gratuite (10 infographies, tout public, cartes limitées, pas de *live connections*)
- Version « pro » à partir de 15\$/mois
  
- Alternatives : [plot.ly](http://plot.ly), [chartblocks.com](http://chartblocks.com)

# Librairies JavaScript

---

- De nombreuses librairies, pas toutes de la même qualité !
- Panoplie de graphiques ?
- Interactivité ? Tooltip ? Action sur clics ?
- Possibilité de zoom, de « save as » ?
- Facilité de paramétrisation ?
- Compatible AngularJS ou autre ?
- Open Source ? Extensible ?
- Taille de la librairie ?
- Prix ?



# Librairies javascript : D3

---

- « Data-Driven Documents »
- Librairie (gratuite) pour produire en « SVG »
- Permet de tracer des formes (carrés, cercles, lignes...), et de gérer des événements
- Beaucoup d'exemples pour tout type de graphique
- Très puissant, mais très bas niveau
- À la base de beaucoup de librairies





# Librairies javascript : D3

```

<!DOCTYPE html> <meta charset="utf-8">
<style>
.bar { fill: steelblue; }
.bar:hover { fill: brown; }
.axis { font: 10px sans-serif; }
.axis path,
.axis line { fill: none; stroke: #000; shape-rendering:
crispEdges; }
.x.axis path { display: none; }
</style>
<body>
<script
src="https://cdnjs.cloudflare.com/ajax/libs/d3/3.5.5/d3.min.js"></script>
<script>
var margin = {top: 20, right: 20, bottom: 30, left: 40},
    width = 960 - margin.left - margin.right,
    height = 500 - margin.top - margin.bottom;
var x = d3.scale.ordinal() .rangeRoundBands([0, width], .1);
var y = d3.scale.linear() .range([height, 0]);
var xAxis = d3.svg.axis() .scale(x) .orient("bottom");
var yAxis = d3.svg.axis() .scale(y) .orient("left") .ticks(10,
"%");
var svg = d3.select("body").append("svg")
    .attr("width", width + margin.left + margin.right)
    .attr("height", height + margin.top + margin.bottom)
    .append("g")
    .attr("transform", "translate(" + margin.left + "," +
margin.top + ")");

```

```

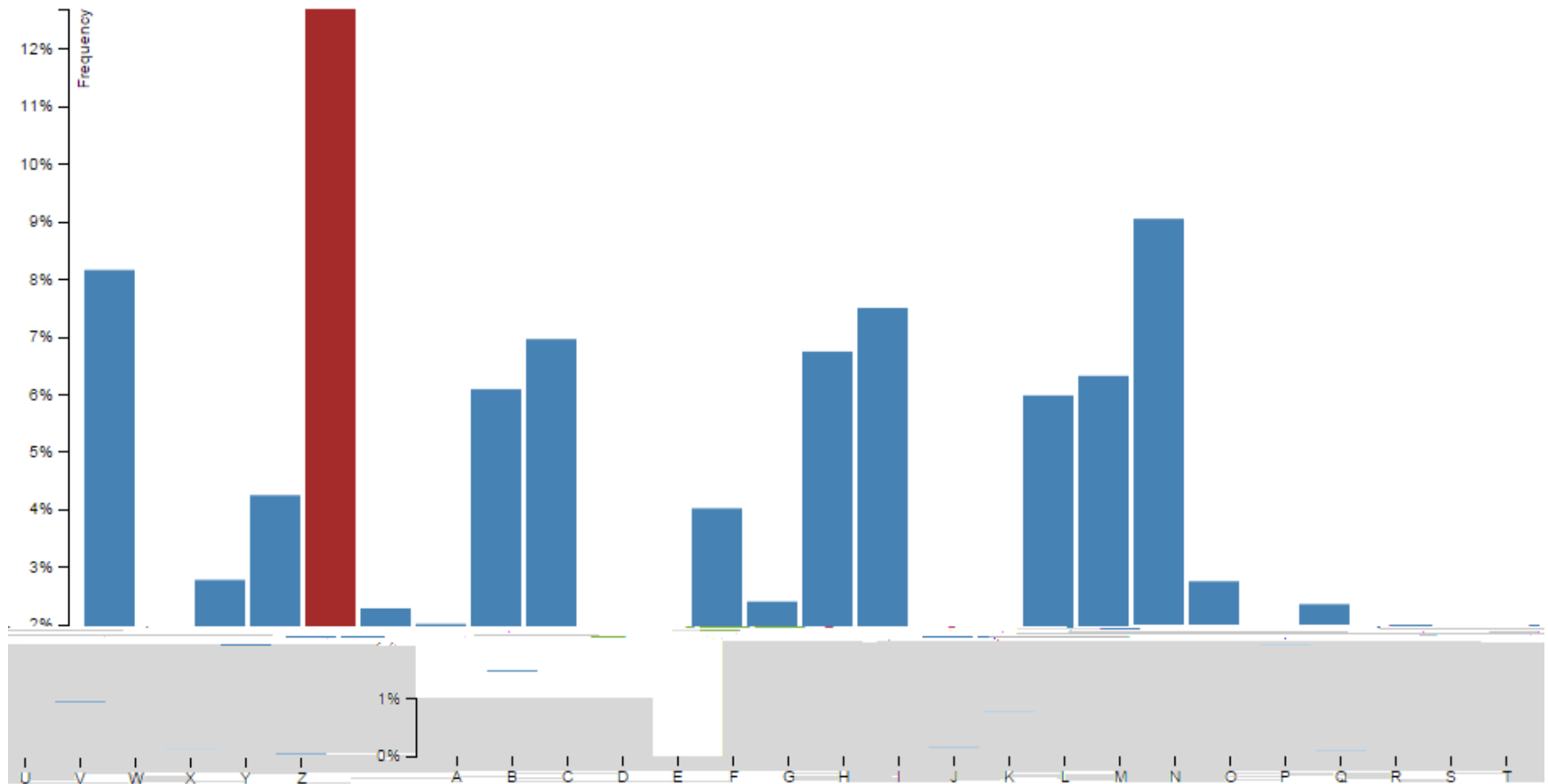
d3.tsv("data.tsv", type, function(error, data) {
    if (error) throw error;
    x.domain(data.map(function(d) { return d.letter;
    }));
    y.domain([0, d3.max(data, function(d) { return
d.frequency; })]);
    svg.append("g")
        .attr("class", "x axis")
        .attr("transform", "translate(0," + height + ")")
        .call(xAxis);
    svg.append("g") .attr("class", "y axis")
        .call(yAxis)
        .append("text")
        .attr("transform", "rotate(-90)")
        .attr("y", 6)
        .attr("dy", ".71em")
        .style("text-anchor", "end")
        .text("Frequency");
    svg.selectAll(".bar")
        .data(data)
        .enter()
        .append("rect")
        .attr("class", "bar")
        .attr("x", function(d) { return x(d.letter); })
        .attr("width", x.rangeBand())
        .attr("y", function(d) { return y(d.frequency);
    })
        .attr("height", function(d) { return height -
y(d.frequency); }); });
function type(d) { d.frequency = +d.frequency; return
d; } </script>

```

letter	frequency
A	.08167
B	.01492
C	.02782
D	.04253
E	.12702
F	.02288
G	.02015
...	

http://bl.ocks.org/mbostock/3885304

# Librairies javascript : D3



# NVD3

```
<div id="chart"> ·  
  <script>  
    data = [{ value  
      .2782 }, {x: "D", y:  
      "G", y: .2015 }]]
```

```
nv.addGraph(function() {  
  var chart = nv.models.discreteBarChart();
```

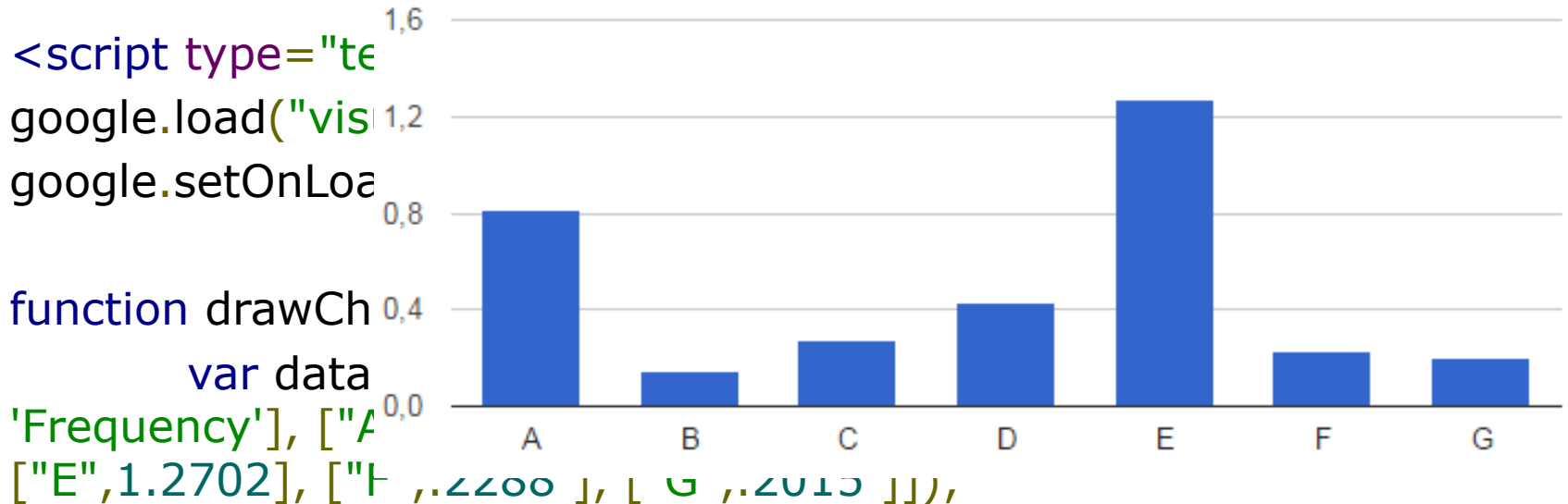
```
    d3.select('#chart svg')  
      .datum(data)  
      .call(chart);  
    nv.utils.windowResize(chart.update);  
    return chart;
```

```
  });  
</script>
```



Autre lib. basée sur D3: C3.js

# Google Charts



```
var div = document.getElementById('chart_div');
var chart = new google.visualization.ColumnChart(div);
chart.draw(data);
```

```
}
```

```
</script>
```

```
<div id="chart" style="width: 100%; height: 100%;"></div>
```

# Librairies javascript : HighCharts

---

- Librairie payante, gratuite pour usage « non commercial »
- Très grand nombre de visualisations (y compris cartes)
- Cartes : pays du monde, provinces belges (+ GeoJSON)
- Mobile ready (pinch-to-zoom...)
- OpenSource

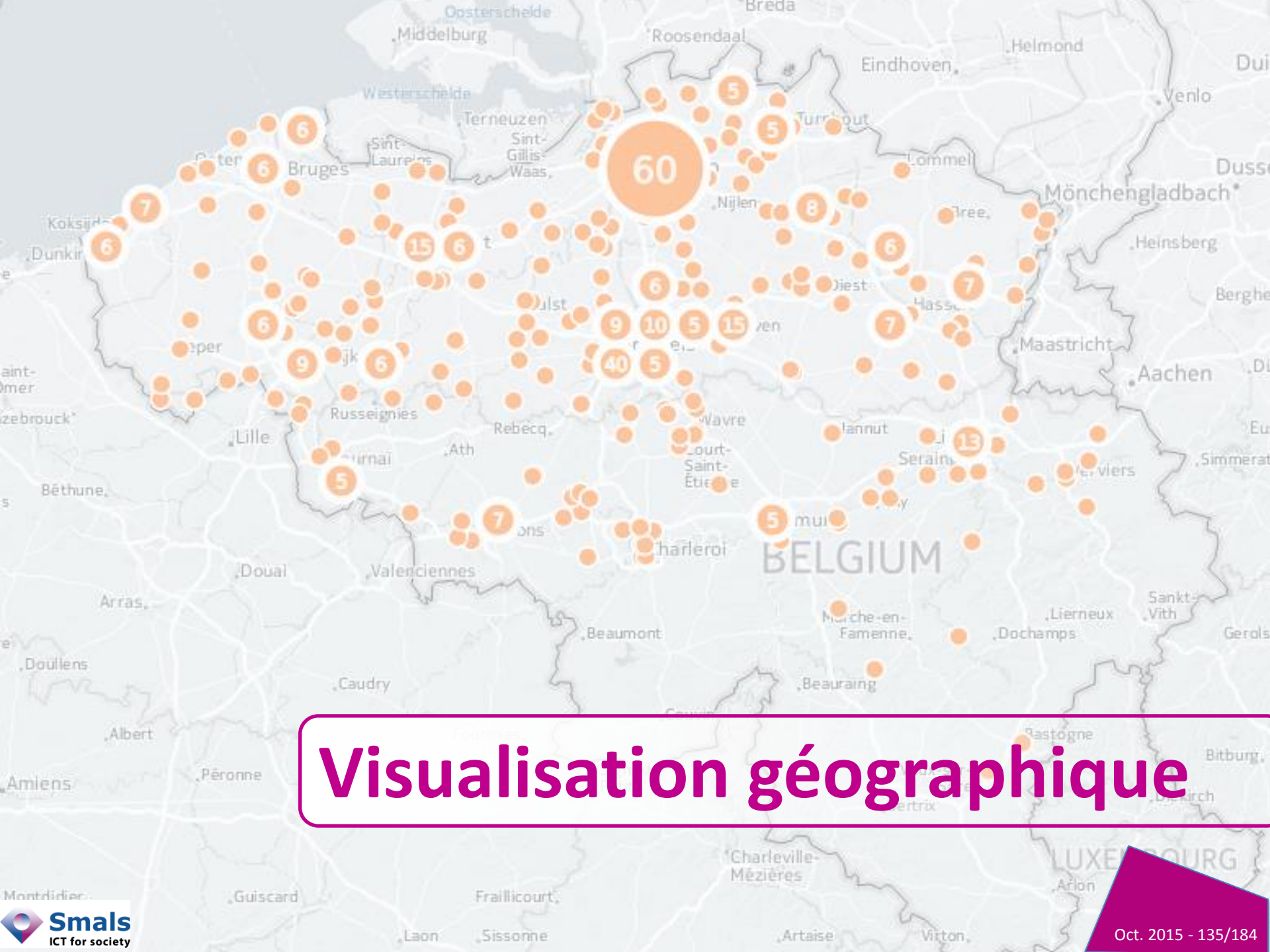


# Librairies javascript : AmCharts

---

- Librairie payante, gratuit avec « watermark »
- Niveau similaire à HighCharts
- Source « minifié »



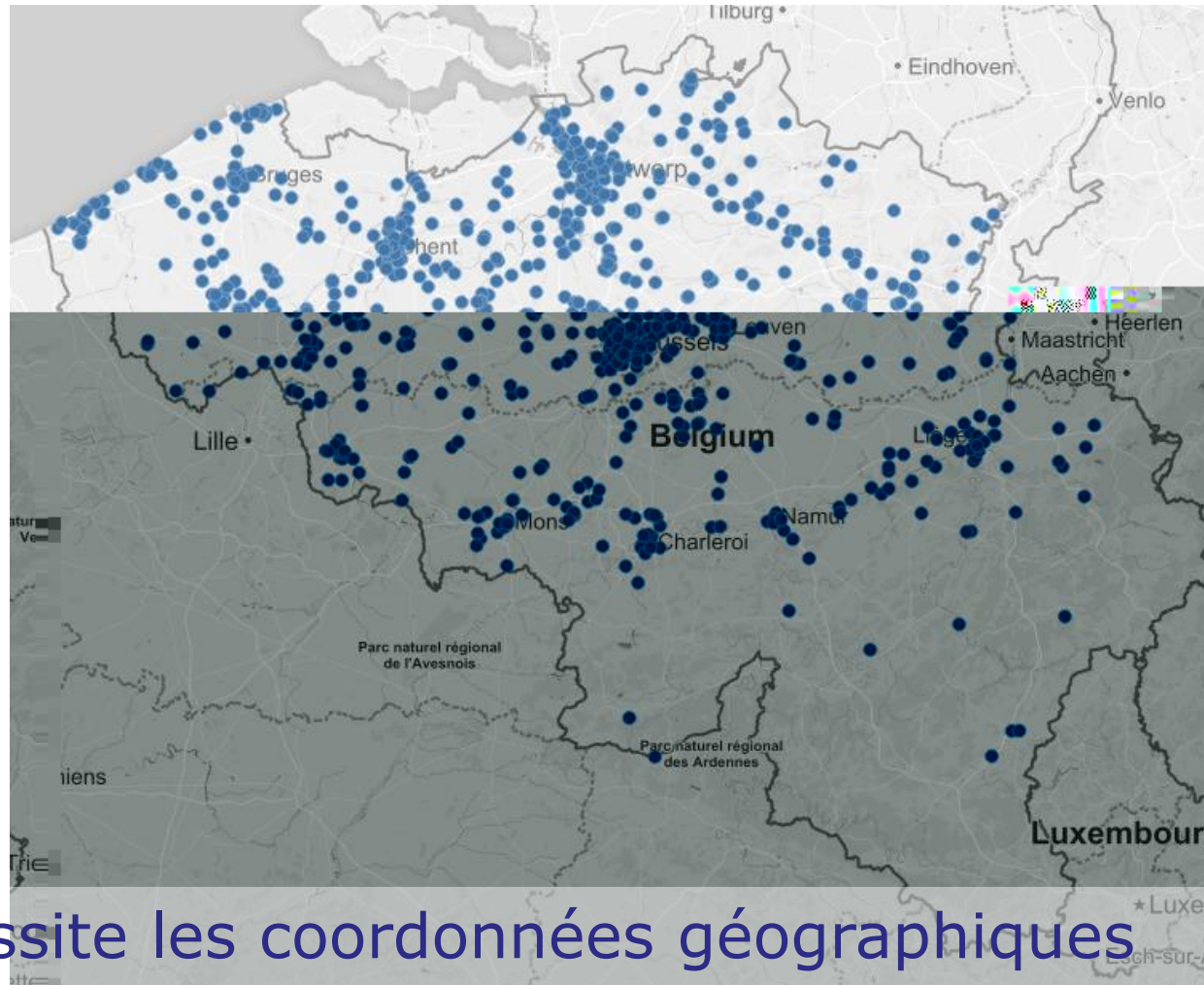


# Visualisation géographique

Visualisation géographique

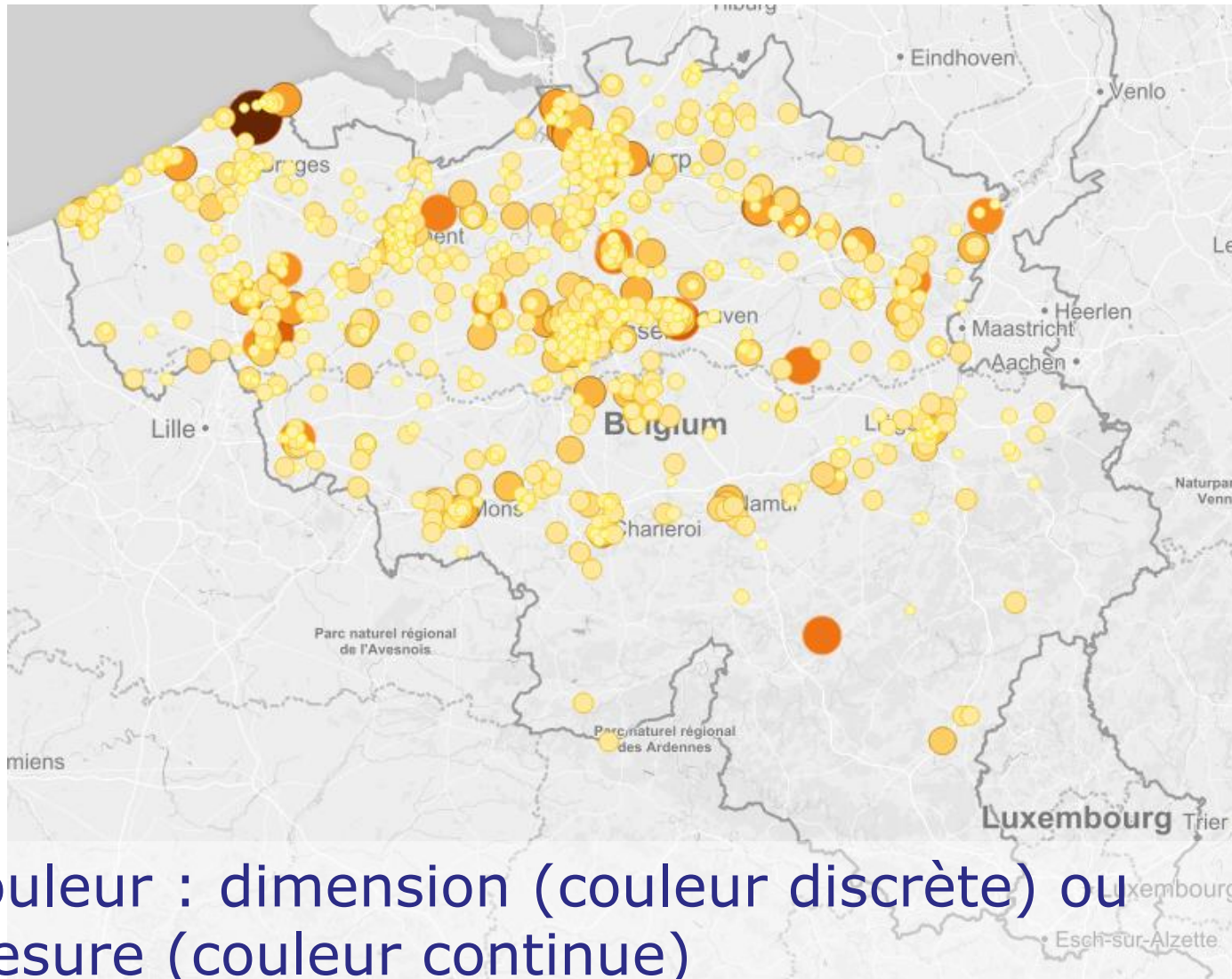
# TYPES DE VISUALISATION

# Point & bubble



- Nécessite les coordonnées géographiques

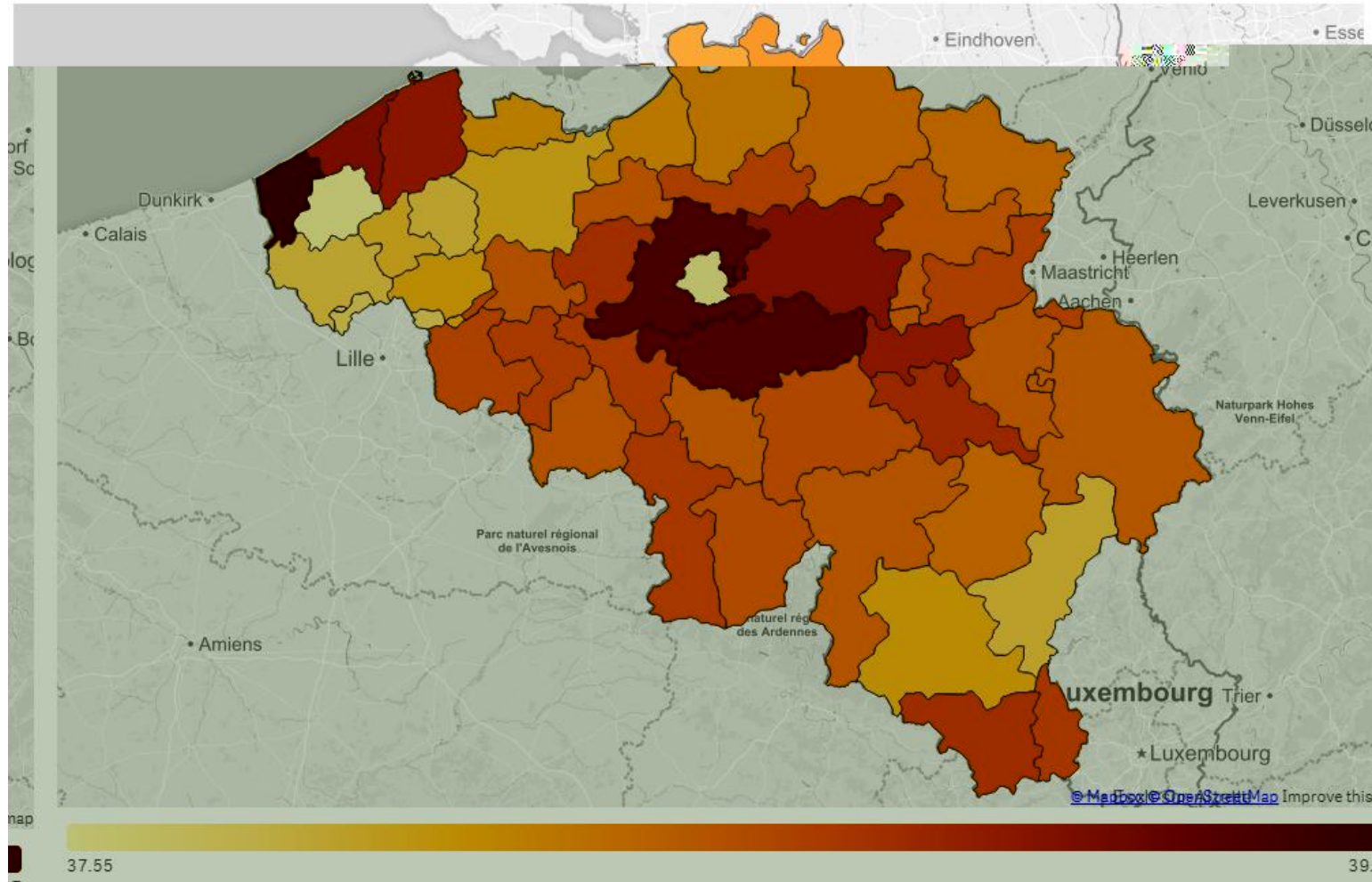
# Point & bubble



- Couleur : dimension (couleur discrète) ou mesure (couleur continue)

# Carte choroplèthe

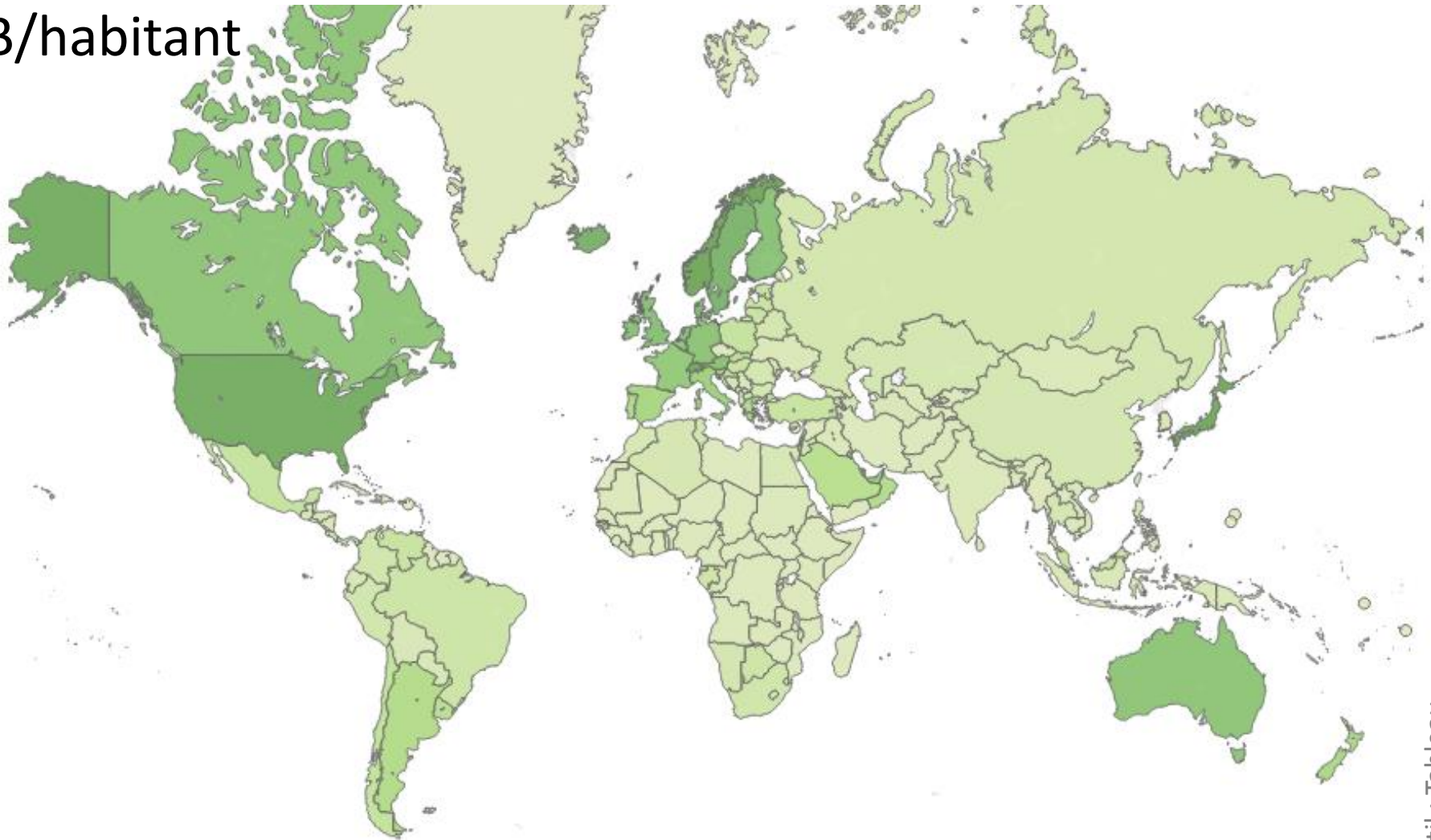
Âge moyen des travailleurs



Outil : QlikSense

# Carte choroplèthe

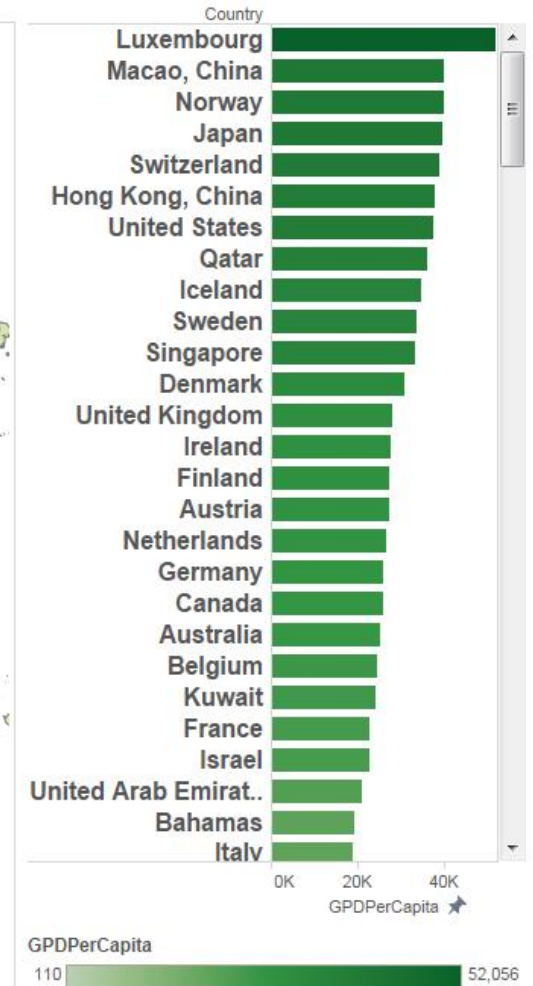
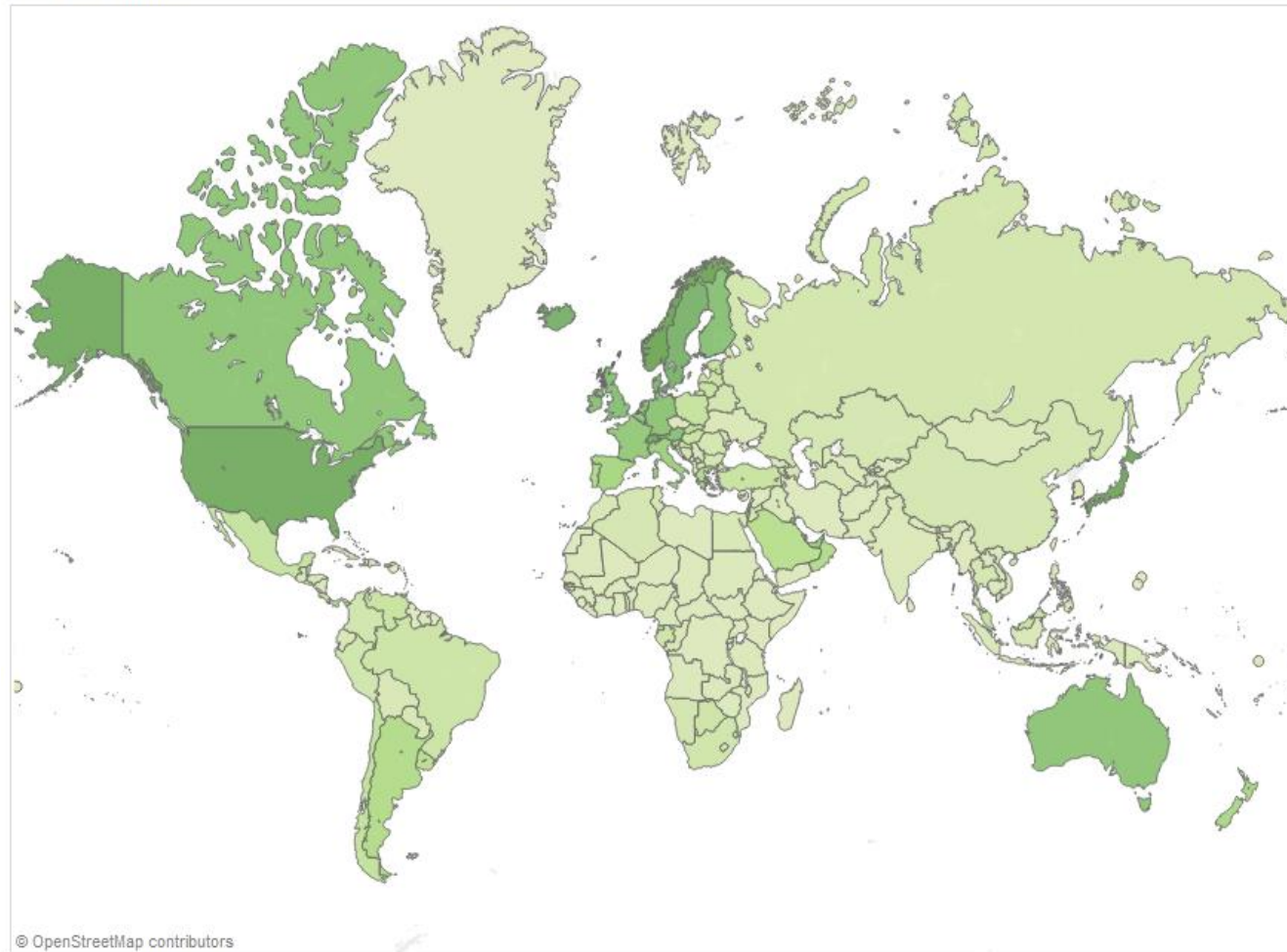
PIB/habitant



Outil : Tableau

# Carte choroplèthe

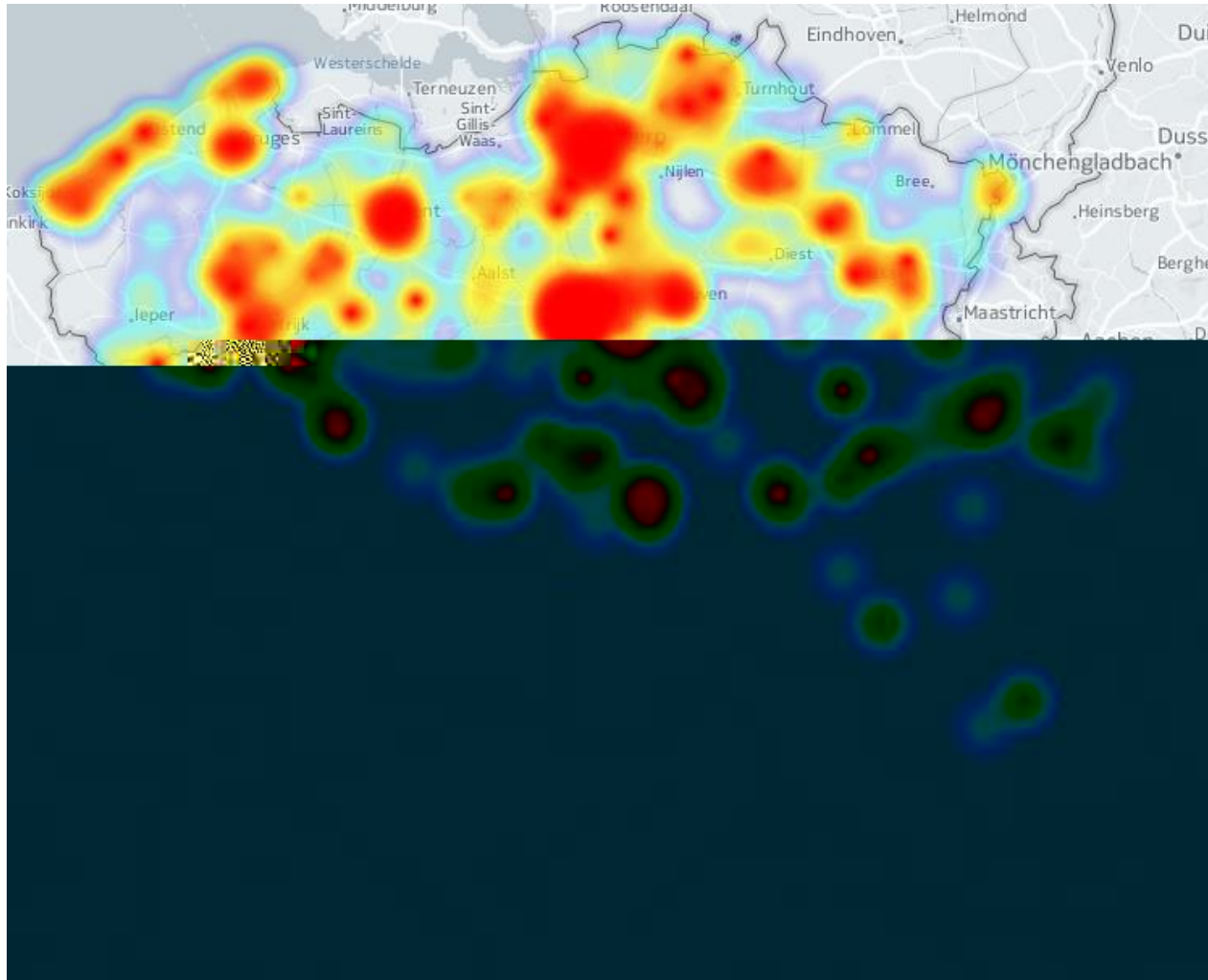
GPD per capita



Outil : Tableau

# Heatmap

Données :



Outil : CartoDB

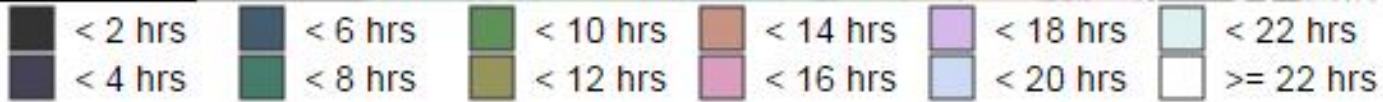
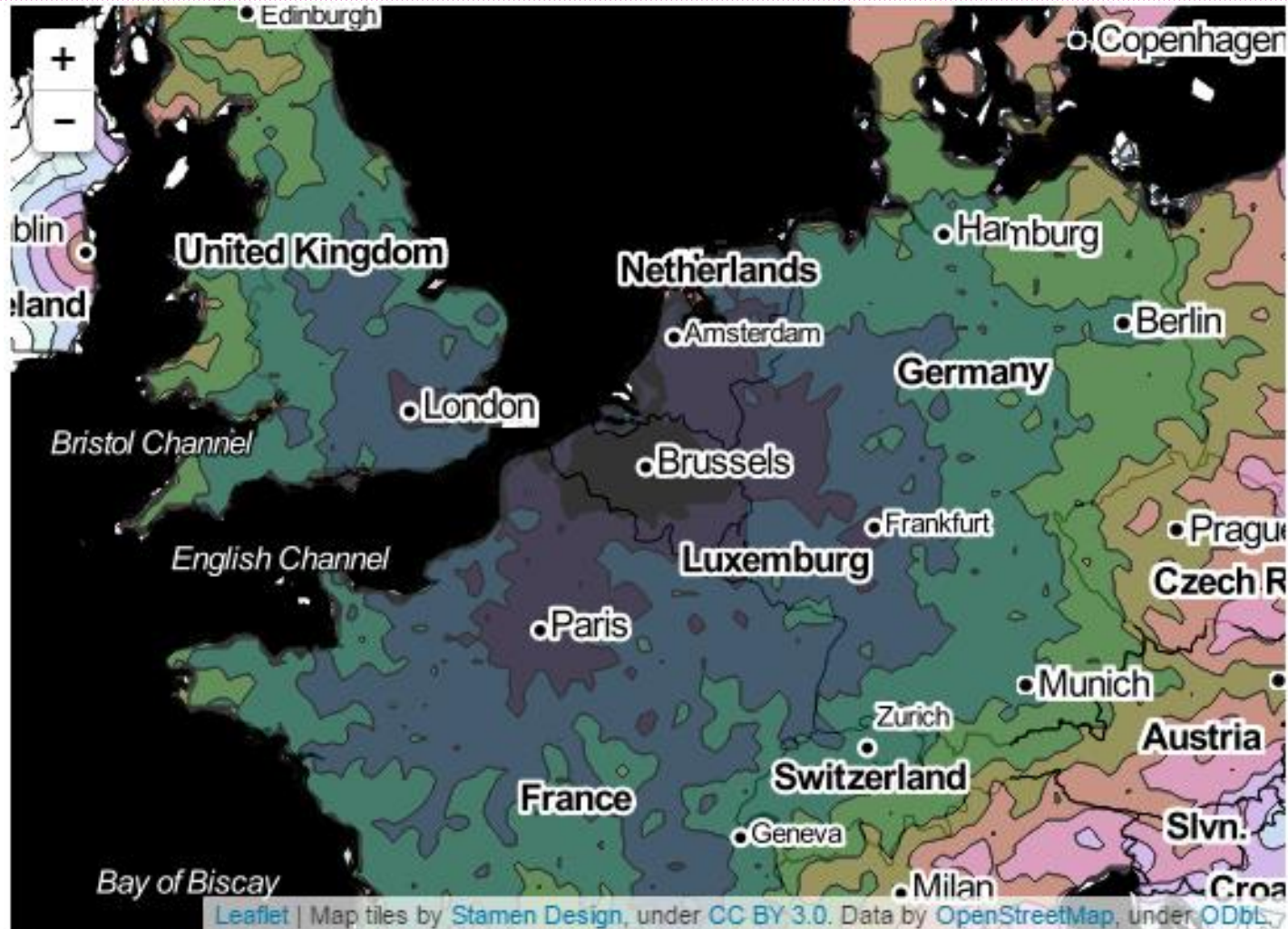
# Carte isoplèthe

---

- Carte isoplèthe : relie par une courbe tous les points possédant une **même caractéristique**
- Isochrone : même **temps de trajet**
- Isotherme : même **température**
- Isobare : même **pression atmosphérique**
- Isohypse : même **altitude**



# Carte isochrone



[http://emptypipes.org/supp/isochrone\\_brussels/](http://emptypipes.org/supp/isochrone_brussels/)

Visualisation géographique

# GÉOCODAGE

# Géocodage

---

- Géocodage : transformation d'une **adresse** en **coordonnées géographiques**
- Utilité :
  - **Visualisation**
  - Détection de **doublons/fraudes/...**
  - Détection d'**erreur**/data quality
- Nombreuses API gratuites (limitées) et commerciales, faciles d'utilisation

# Types de données

---

- **Coordonnées** :
  - degré-minute-seconde : 50°50'10.1"N, 4°20'17.8"E
  - décimales : 50.8361263, 4.3382716
  - OK pour carte **bubble**, **heatmap**
- **Adresse** : Avenue Fonsny 20, 1060 Saint-Gilles
  - Souvent des problèmes de *data quality*
  - Nécessite un « **géocodage** », difficile « **on-site** »
  - Pas d'adresse pour une borne d'autoroute, une forêt...
  - OK pour carte **bubble**, **heatmap**
- **Entité** : Bruxelles, Belgique, 1160, Hal-Vilvorde...
  - Table « entité – frontière – centre » facile à connaître
  - OK pour carte **bubble** (centre) **choroplèthe** (frontières)

# Géocodage : Google Maps

- Avenue Fonsny 20, 1060 Bruxelles :
- <http://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=Avenue%20Fonsny%2020,%201060%20Bruxelles>

```
{
  "results" : [
    {
      "address_components" : [
        {
          "long_name" : "20",
          "short_name" : "20",
          "types" : [ "street_number" ]
        },
        {
          "long_name" : "Avenue Fonsny",
          "short_name" : "Avenue Fonsny",
          "types" : [ "route" ]
        },
        {
          "long_name" : "Sint-Gillis",
          "short_name" : "Sint-Gillis",
          "types" : [ "locality", "political" ]
        },
        {
          "long_name" : "Brussel",
          "short_name" : "Brussel",
          "types" : [ "administrative_area_level_1",
"political" ]
        }
      ],

```

```
      {
        "long_name" : "Belgique",
        "short_name" : "BE",
        "types" : [ "country", "political" ]
      },
      {
        "long_name" : "1060",
        "short_name" : "1060",
        "types" : [ "postal_code" ]
      }
    ],
    "formatted_address" : "Avenue Fonsny 20, 1060 Sint-Gillis, Belgique",
    "geometry" : {
      "bounds" : {
        "northeast" : {
          "lat" : 50.83613219999999,
          "lng" : 4.3385958
        },
        "southwest" : {
          "lat" : 50.83612489999999,
          "lng" : 4.338581500000001
        }
      }
    }
  ],

```

```
    "location" : {
      "lat" : 50.83612489999999,
      "lng" : 4.3385958
    },
    "location_type" : "RANGE_INTERPOLATED",
    "viewport" : {
      "northeast" : {
        "lat" : 50.83747753029149,
        "lng" : 4.339937630291503
      },
      "southwest" : {
        "lat" : 50.83477956970849,
        "lng" : 4.337239669708499
      }
    },
    "partial_match" : true,
    "place_id" :
"EitBdmVudWUgRm9uc255IDlwLCAxMDYwIF
NpbmQtR2lsbGlzLCBCZWxnaCOr",
    "types" : [ "street_address" ]
  },
  "status" : "OK"
}
```

# Géocodage : Google Maps

## Adresse standardisée

```
{ "long_name" : "20", "types" : [ "street_number" ] },  
{ "long_name" : "Avenue Fonsny", "types" : [ "route" ] },  
{ "long_name" : "Avenue Fonsny", "types" : [ "route" ] },
```

## Géométrie

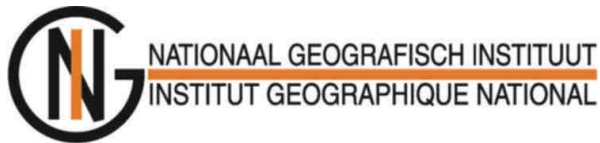
```
{ "long_name" : "Avenue Fonsny", "types" : [ "route" ] },  
  "types" : [ "route" ],  
{ "long_name" : "Avenue Fonsny", "types" : [ "route" ] },  
{ "long_name" : "Avenue Fonsny", "types" : [ "route" ] },  
  "formatted_address" : "Avenue Fonsny, 20, 1300 Wavre, Belgique",  
  "geometry" : {  
    "bounds" : {  
      "northeast" : { "lat" : 50.83613219, "lng" : 4.3385958 },  
      "southwest" : { "lat" : 50.83612489, "lng" : 4.3385815 }  
    },  
    "location" : { "lat" : 50.83612489, "lng" : 4.3385958 }
```

## Précision

```
"location_type" : "RANGE_INTERPOLATED"  
"status" : "OK"
```

# Géocodage

---



# Géocodage : comparaison

---

- Comparaison de services : difficile !
- Littérature : dataset d'adresses/coordonnées **connues** -> difficile à établir, forcément tronqué
- Alternative : comparer les réponses de 7 API
- Voir s'il y a consensus :



Mauvaise  
réponse ?

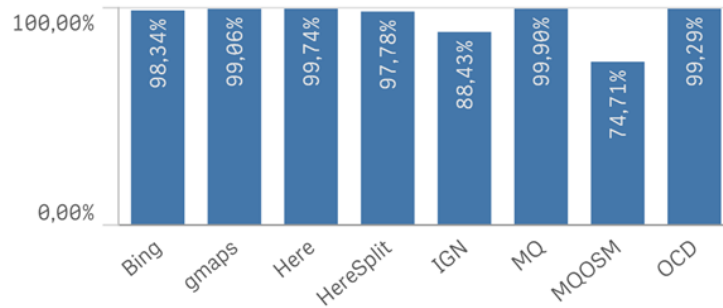


- Dataset : 10,000 adresses aléatoires de la CBE (KBO), Belgique (98,4%) + étranger

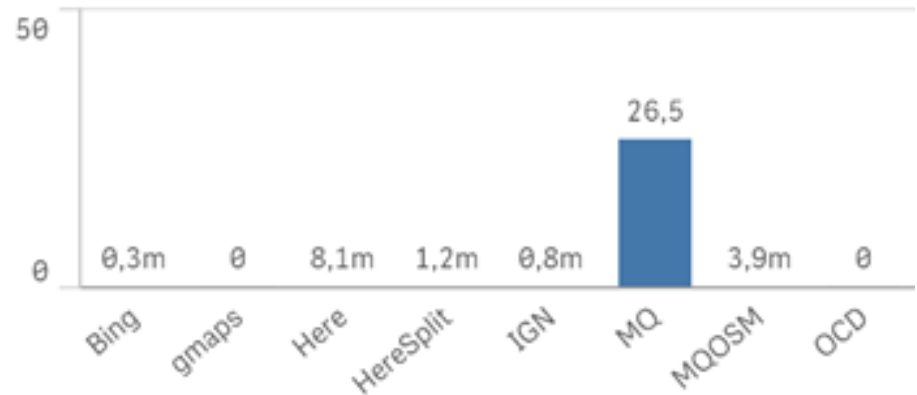


# Géocoding : comparaison

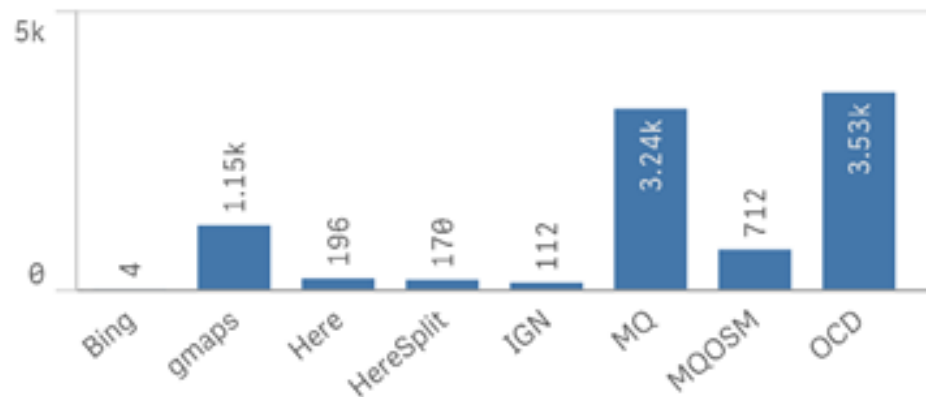
## Match rate



## Avg dist to median



## #dist to median > 0.01



Visualisation géographique

# OUTILS

# Visualisation (gratuits)

---

- Google Fusion Table (Google Drive)
  - Points map/heatmap + géocodage
  - Choroplèthes
  - Permet de fusionner (« join ») des tables
- CartoDB
  - Géocodage d'adresses limité (100/mois)
  - Géocodage de villes non limité
  - De nombreux types de cartes, fusion de tables, filtres SQL...
- QlikView/QlikSense : point & choroplèthe
- Tableau public
  - Géocodage des villes
  - Bloqué par le firewall de Smals

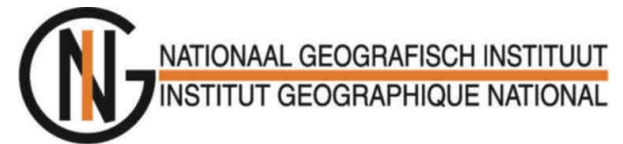
# Géocodage

---

- API :
  - Google Maps (2500/j)
  - OpenStreetMap (usage policy: « no heavy use »)
  - Bing (Microsoft, 30.000/j)
  - Here (Nokia, 100.000/m)
  - IGN (Beta)
- Web interface (batch) :
  - Google Fusion Table (~2/sec)
  - CartoDB (100/mois, 1/sec)
  - BatchGeo (par 250, ~0.5/sec)
  - EasyMapMaker (~1/s), cut&paste

# Chez Smals

- DDT/Checkin@Work :
  - ESRI/ArcGIS : manipulation, interface,
  - IGN : Visualisation
  - Here (Nokia) : géocodage
- SMUREG/MEDEGA (SPF SP) :
  - TeleAtlas (TomTom)
  - RealDolmen



We make ICT work for your business



# Démo CartoDB

The screenshot displays the CartoDB web interface. At the top, there's a navigation bar with 'Infosession' and 'Edit metadata...' on the left, and 'DATA VIEW' and 'MAP VIEW' tabs in the center. On the right of the top bar are 'Edit -' and 'PUBLISH' buttons. Below the navigation bar is a toolbar with 'Add Element', 'Preview Map', and 'Export Image' options. The main area is a map of Belgium with a heatmap overlay showing data density. A search bar is visible in the top right of the map area. On the right side, a 'Map layer wizard' panel is open, showing a 'cartodb' layer. The wizard includes a 'Map layer wizard' title, a 'SQL' icon, and four visualization options: 'TORQUE', 'HEATMAP', 'TORQUE CAT', and 'INTENSITY'. Below these options are settings for 'Marker Fill' (value 10, color yellow, range 0.9) and 'Marker Stroke' (value 1, color white, range 1). At the bottom of the map, there are 'Change basemap' and 'Options' buttons, and a URL: <https://vandy.cartodb.com/viz/1089c51a-6850-11e5-9a40-0e3ff518bd15/map#left>. The footer of the map area includes '© OpenStreetMap contributors © CartoDB, CartoDB attribution'.

### Sales Variation

YTD \$36.17M LYTD \$36.80M

Sales Variation

-1.71%

### Margin Variation

YTD \$15.85M LYTD \$15.52M

Sales Margin Variation

2.13%

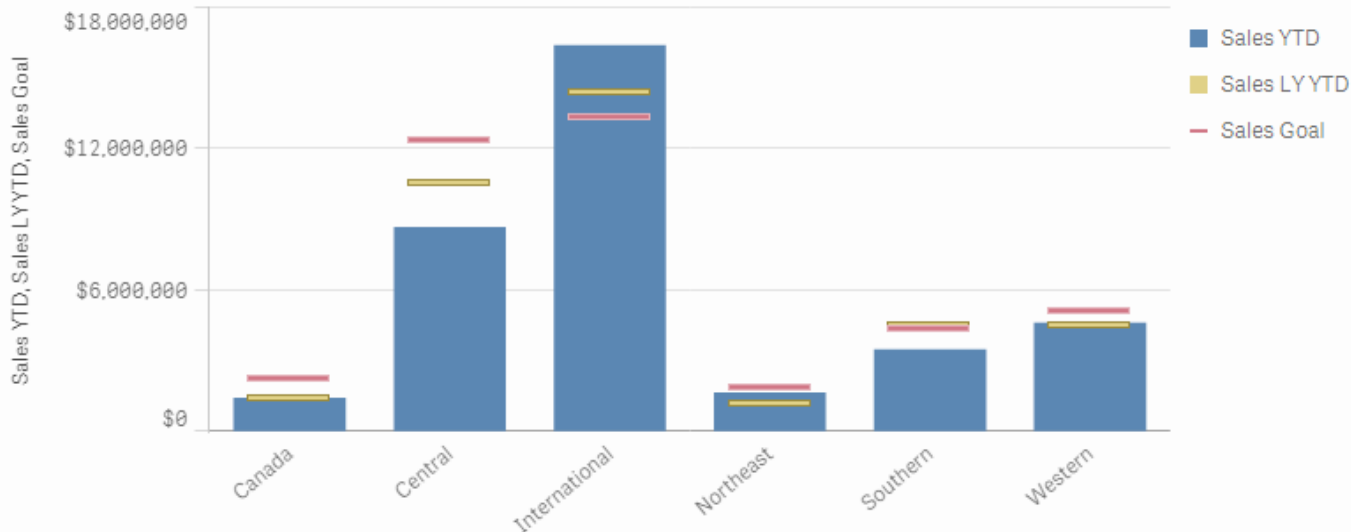
### Sales vs Goal YTD

Budget YTD \$39.44M LYTD \$36.17M

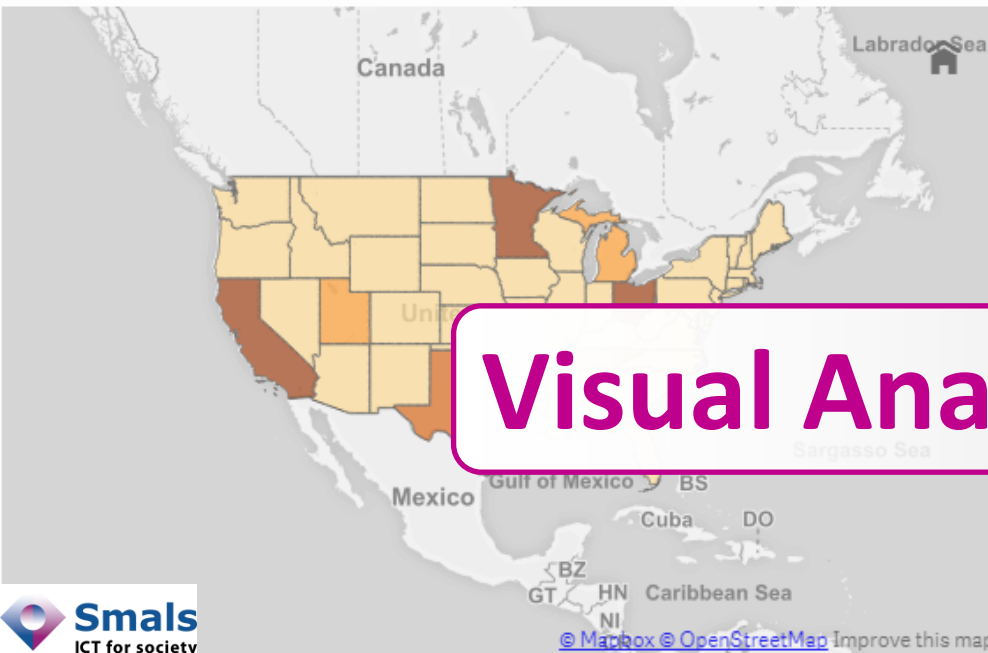
Sales vs Goal

-8.29%

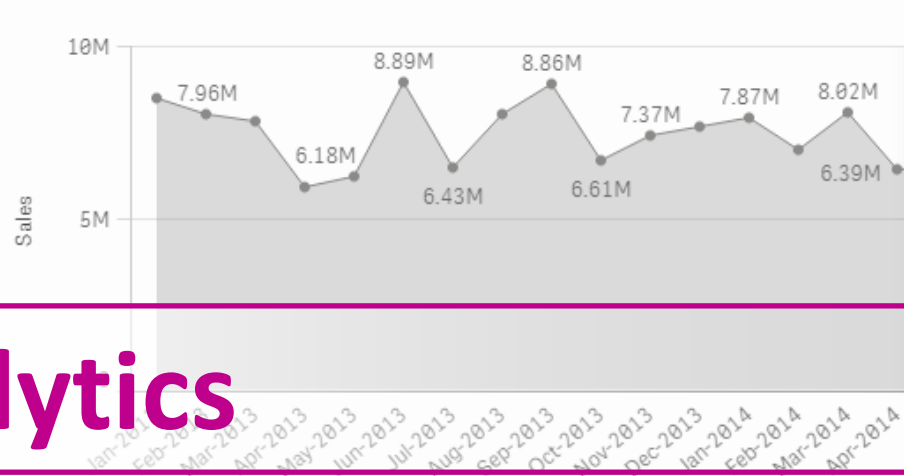
### Sales: YTD vs LYTD vs YTD Goals



### Sales by State



### Sales Evolution



# Visual Analytics

Visual Analytics

# DÉFINITIONS

# Visual Analytics : définition

---

« Visual analytics is the science of **analytical reasoning** facilitated by **interactive visual interfaces** »\*

Approche combinant :

- **Visualisation** (dataviz, sciviz, comp. graphics...)
- **Interaction** (H-C interaction, cognitive psychology, perception...)
- **Data analysis** (Information retrieval, data mining, information/geospatial/scientific/statistical analytics...)

# Objectifs

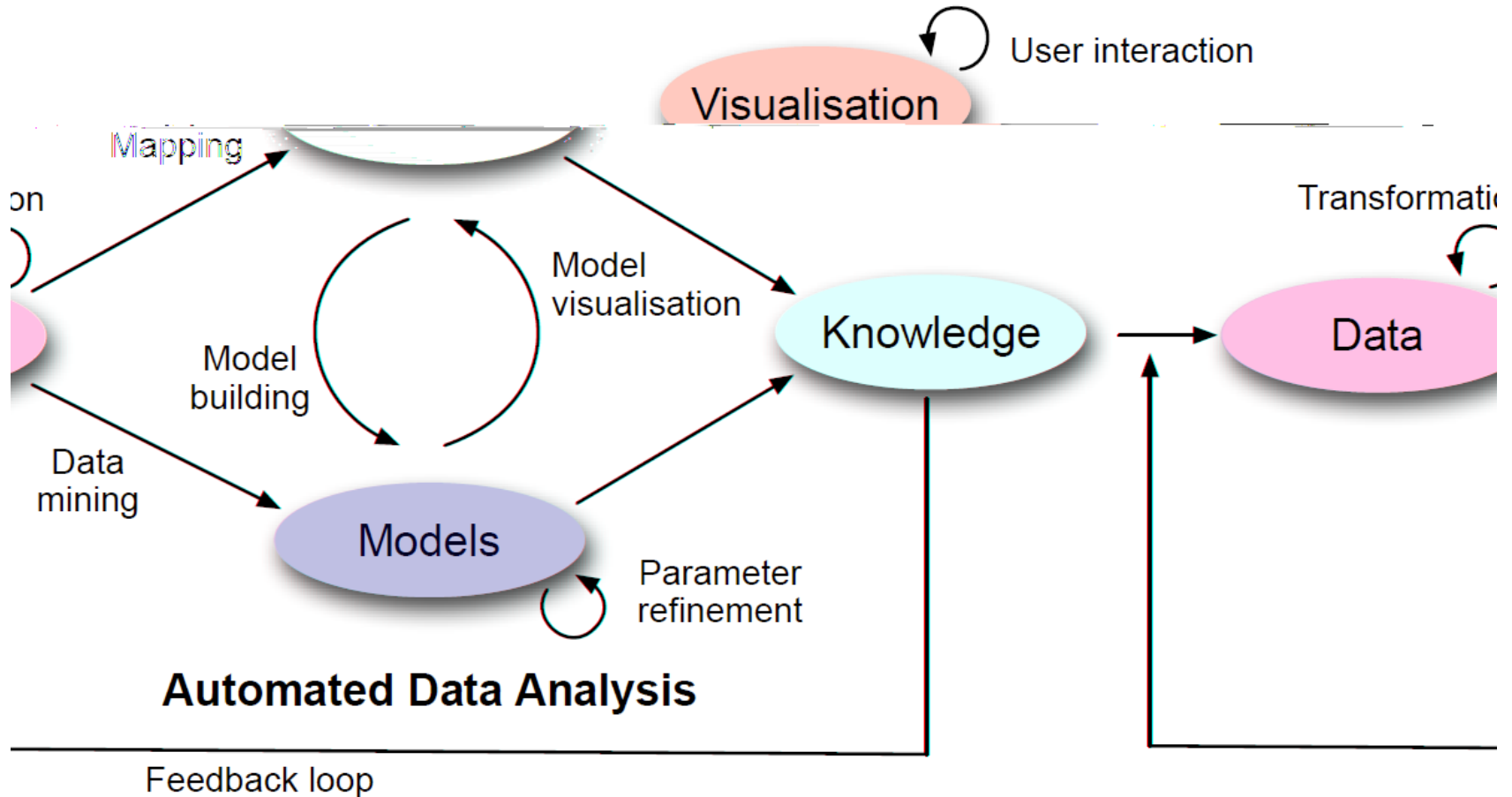
---

- Synthétiser l'information et en tirer une intuition (*insight*) de données massives, dynamiques, ambiguës et souvent contradictoires
- Détecter l'attendu, découvrir l'inattendu
- Fournir des évaluations en temps voulu, défendable et compréhensible
- Communiquer ces évaluations efficacement pour action

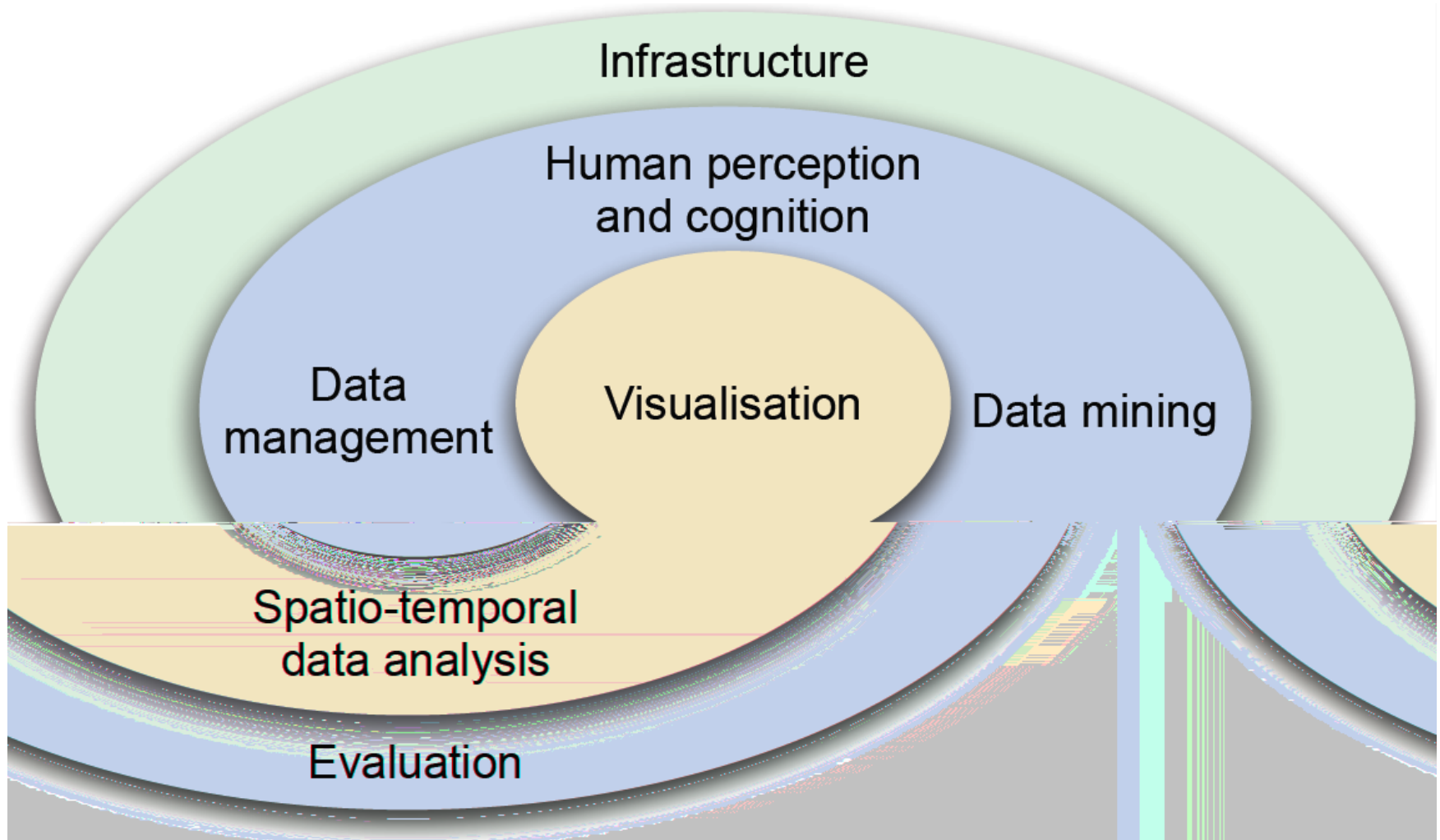


# Visual Analytics Process

## Visual Data Exploration



# Domaines



# Domaines

---

- Visualisation
  - Visualisation **scientifique** (Sciviz) : visualisation multi-D d'**entités physiques** (météo, biologie, chimie, ingénierie...)
  - Visualisation de **données** (Dataviz) : visualisation de **données** « **abstraites** » (démographie, business, mesures, texte...)
  - Visualisations plus complexes que pour le grand public
- Data management
  - **Hétérogénéité** sources (DB, fichiers, page web, streams, texte...) et types (format)
  - **Big Data**
  - **Data quality**

# Domaines

---

- Data Mining
  - Extraction automatique d'information de valeur
  - « Supervised learning » : extraction basée sur des échantillons connus
  - « Unsupervised learning » : extraction sans connaissance préalable
- Spatio-temporal Data Analysis
  - Les données spatiales et les données temporelles requièrent des techniques particulières
  - La combinaison (données spatio-temporelle) augmente la complexité

# Domaines

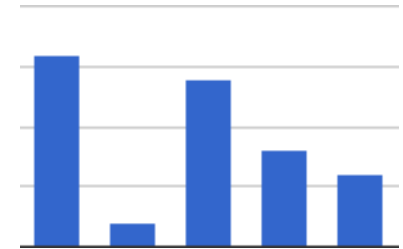
---

- Perception & Cognition
  - Perception : Interprétation de l'environnement (5 sens)
  - Cognition : capacité de comprendre (basée sur l'apprentissage)
- Infrastructure
  - Liaison des différents processus et technologies, souvent incompatibles
- Évaluation

# Données brutes

- Libraires graphiques :

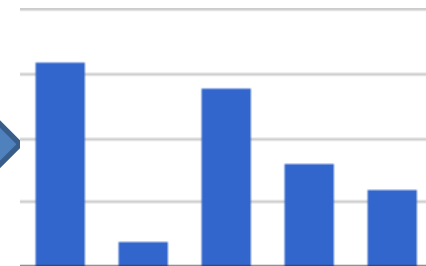
Key	Value
A	8
B	1
C	7
D	4
E	3



- Outils Visual Analytics :

Id	Key
1	X1
2	X2
3	X1
4	X3
5	X2

Key	Attr
X1	A
X2	C
X3	B



# In-Memory

	Live	In-Memory
Principe	Connexion à la source pour chaque affichage	<i>Extract</i> local de ce qui est nécessaire au <i>dashboard</i>
Requêtes	Sur la source	Localement
Mises à jour	Permanententes	Manuelles/programmées
Vitesse	Plus lent	Plus rapide
Problème si	Source pas efficace (fichier CSV...)	Très gros volume (big data)

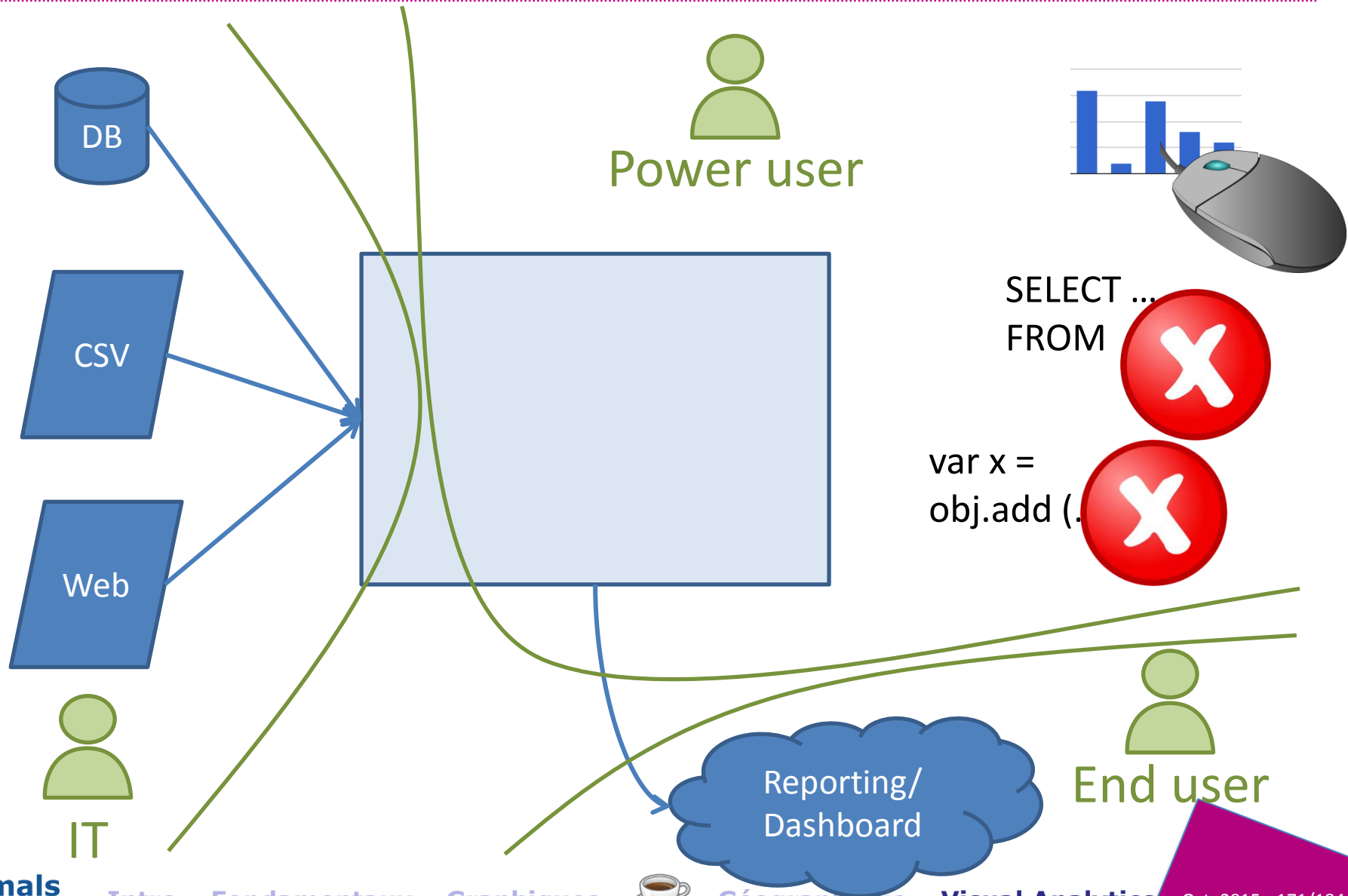
Hybride : certaines sources live, d'autre in-memory



Visual Analytics

# OUTILS

# Outils



# Tableau Software



- Créé par Pixar
- Gartner :
  - Tableau's intuitive, visual-based data discovery capabilities have transformed business users' expectations about what they can discover [...]
- Life ou In-Memory
- Méthodes statistiques, *forecast*, *trendlines*
- Choix des dimensions, suggestion de graphique
- Tableau Public : version gratuite, tout en ligne, données publiques (bloqué par le FW :-s )
- Uniquement Drag&Drop

# Qlik

# Qlik

- Anciennement « Quik » (1993)
- QlikView :
  - Gartner: « ... mature, self-contained, [...] used by IT or more technical users for building intuitive and interactive dashboard applications [...] »
  - Personal Edition: gratuit, pas de partage

## QlikView

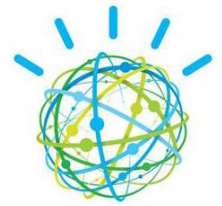
- QlikSense (Gratuit)
  - « Self-service BI »



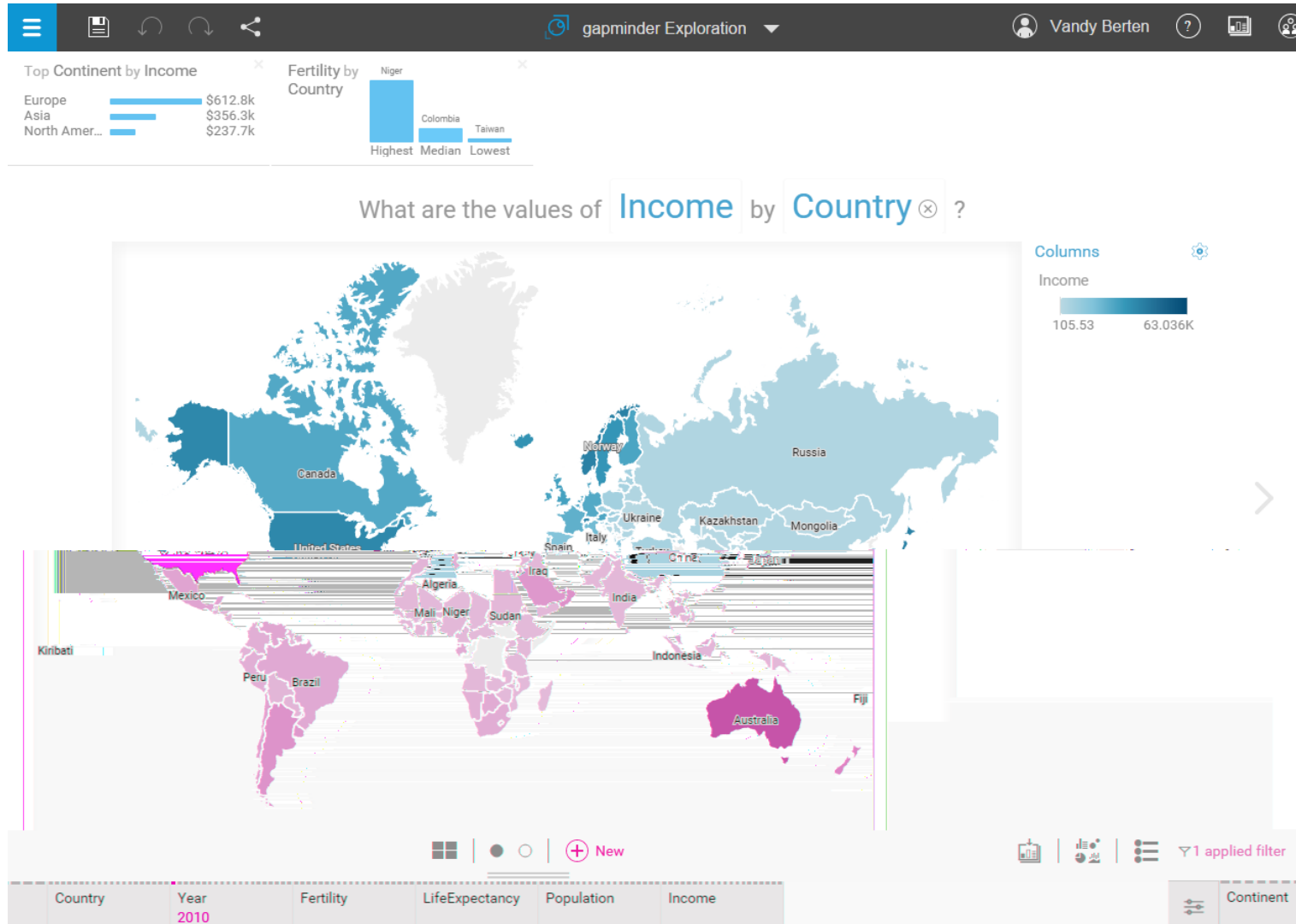
- *In-Memory* (ou hybride « *direct query* ») **Qlik® Sense**
- *Discovery & Reporting*, peu de *data analytics*
- Choix du graphique puis des *dimensions*
- Beaucoup d'extensions (QlikMaps...)
- Drag&Drop ou Script

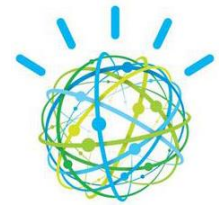
# Use cases Qlik & Tableau

	Qlik	Tableau
Usage gratuit	<ul style="list-style-type: none"> <li>- QlikSense (Partage via « Qlik Cloud » ou par envoi des fichiers)</li> <li>- QlikView Personal Ed. (Pas de partage)</li> </ul>	Tableau Public : Client desktop, données publiques sur un <i>cloud</i>
Usage interne	<ul style="list-style-type: none"> <li>- Named User Licence + partage fichiers</li> <li>- QlikView Server :               <ul style="list-style-type: none"> <li>- Small Business Ed. (<math>\leq 25</math> users)</li> <li>- Enterprise Edition</li> </ul> </li> </ul>	Tableau Server + Tableau Desktop + Tableau Reader (gratuit)
Usage public	QlikView Server EE + Information Access Server	Tableau Server (Lic. CPU)



# IBM Watson Analytics





# IBM Watson Analytics

What do you want to explore next?

what is the population per country



Very relevant

What are the values of **Population by Country**?



Very relevant

What is the breakdown of **Population by Country**?



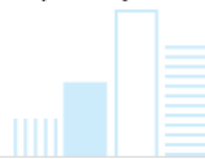
Somewhat relevant

What is the relationship between **Population and Income by Country**?



Somewhat relevant

How do the values of **Population** compare by **Continent**?



Somewhat relevant

What are the values of **Income and Population by Country**?



Somewhat relevant

What are the values of **Fertility and Population by Country**?



Somewhat relevant

What are the values of **Life Expectancy** and **Income by Country**?

Somewhat relevant

What are the values of **Income by Country** and **Life Expectancy**?

Somewhat relevant

What are the values of **Fertility by Country** and **Life Expectancy**?

Somewhat relevant

What are the values of **Life Expectancy by Country** and **Fertility**?



# Démo Qlik Sense

- Contexte : données anonymisées de l'ONSS
- On a une liste de chantiers avec pour chacun :
  - L'entrepreneur
  - Les sous-traitants (+ pays)
  - L'adresse
  - Un score de « risque de fraude »
- On géocode les adresses (API Google Maps)
- On croise avec des données de la poste (CP > Commune > Arrondissement > Province > Région) + des données géographiques (contours)



# Démo – schéma de données



## OpenSource



KML_commune
Commune
Contours

KML_arrond
Arrondissement
Contours

KML_prov
Province
Contours

Zip
ZipCode
Commune

Commune
Commune
Arrondissement

Arrond
Arrondissement
Province

Chantiers
Id
ZipCode
Entrepreneur
Sous-traitant
Pays
Adresse
Latitude, Longitude
...

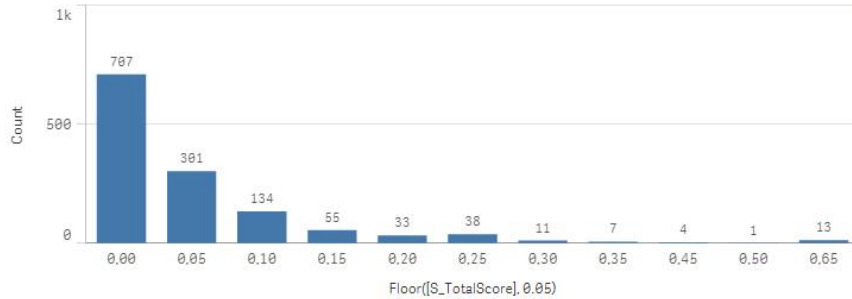
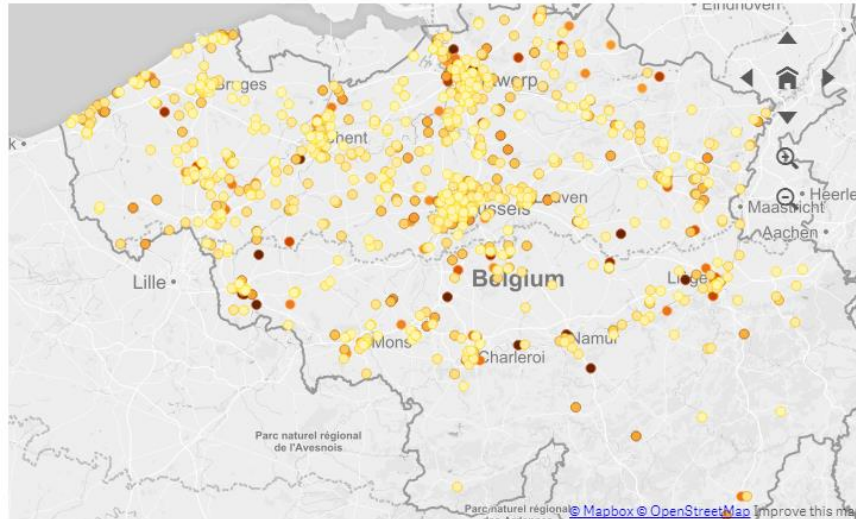




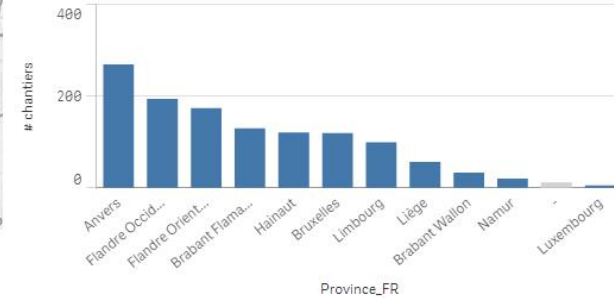
# Démo

## Overview

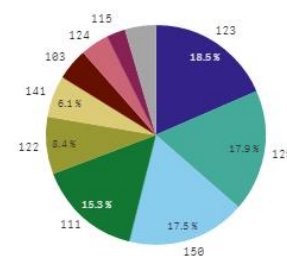
1304 values - 1040 sites - 322 contractors



## Score moyen



## Pays



## ZipCode

ZipCode	Count
NA	0
1000	123
1020	129
1030	150
1040	111
1050	122
1060	141
1070	115

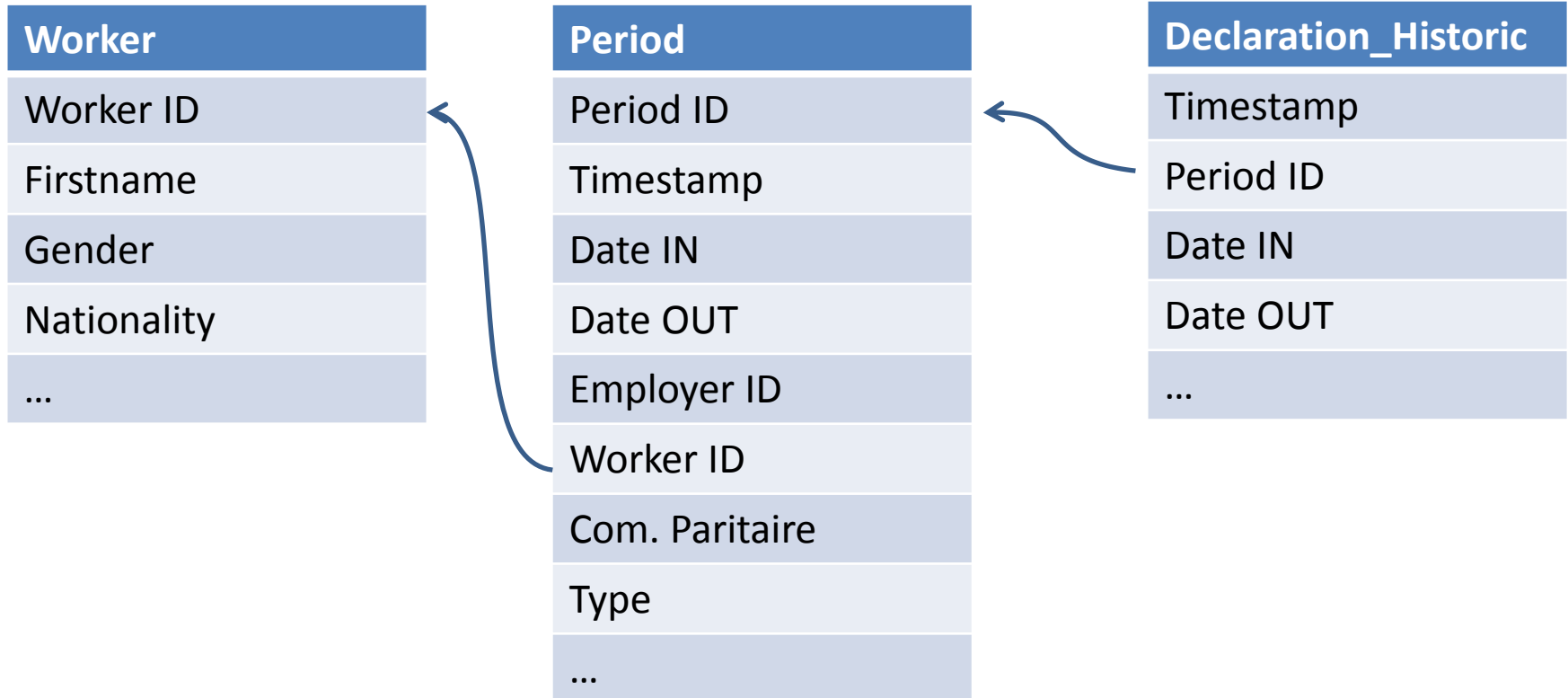


# Démo Tableau

---

- Contexte : déclaration DIMONA
- Données « brutes »
- 3 uses cases :
  - Compréhension des données
  - Détection de problèmes de qualité
  - Détection de cas « suspects »

# Démo Tableau





# Conclusions

# Conclusions

---

Visualisation =  
données →  
information/connaissance

Complémentaire à  
d'autres techniques  
d'analyse (statistiques,  
data  
mining/analytics...)

Très facile de faire une  
mauvaise  
visualisation, même  
avec un bon outil !

Visual Analytics =  
dataviz par le business





Vandy Berten  
02/787.57.32  
[vandy.berten@smals.be](mailto:vandy.berten@smals.be)



More on Smals Research :  
Website : [www.smals.be](http://www.smals.be)  
Blog : [www.smalsresearch.be](http://www.smalsresearch.be)  
Twitter : [@SmalsResearch](https://twitter.com/SmalsResearch)



# Bibliographie

---

- [MIA] « Mastering the Information Age, Solving Problems with Visual Analytics », D. Keim, J. Kohlhammer, G. Ellis and F. Mansmann, Goslar, Germany, 2010 (<http://www.vismaster.eu>)
- [VDQI] « The Visual Display of Quantative Information »
- [SMTN] « Show Me the Numbers »
- [ITP] « Illuminating the Path, The Research and Development Agenda for Visual Analytics », J. J. Thomas, K. A. Cook
- [IV] « Information Visualisation, Perception for Design », C. Ware