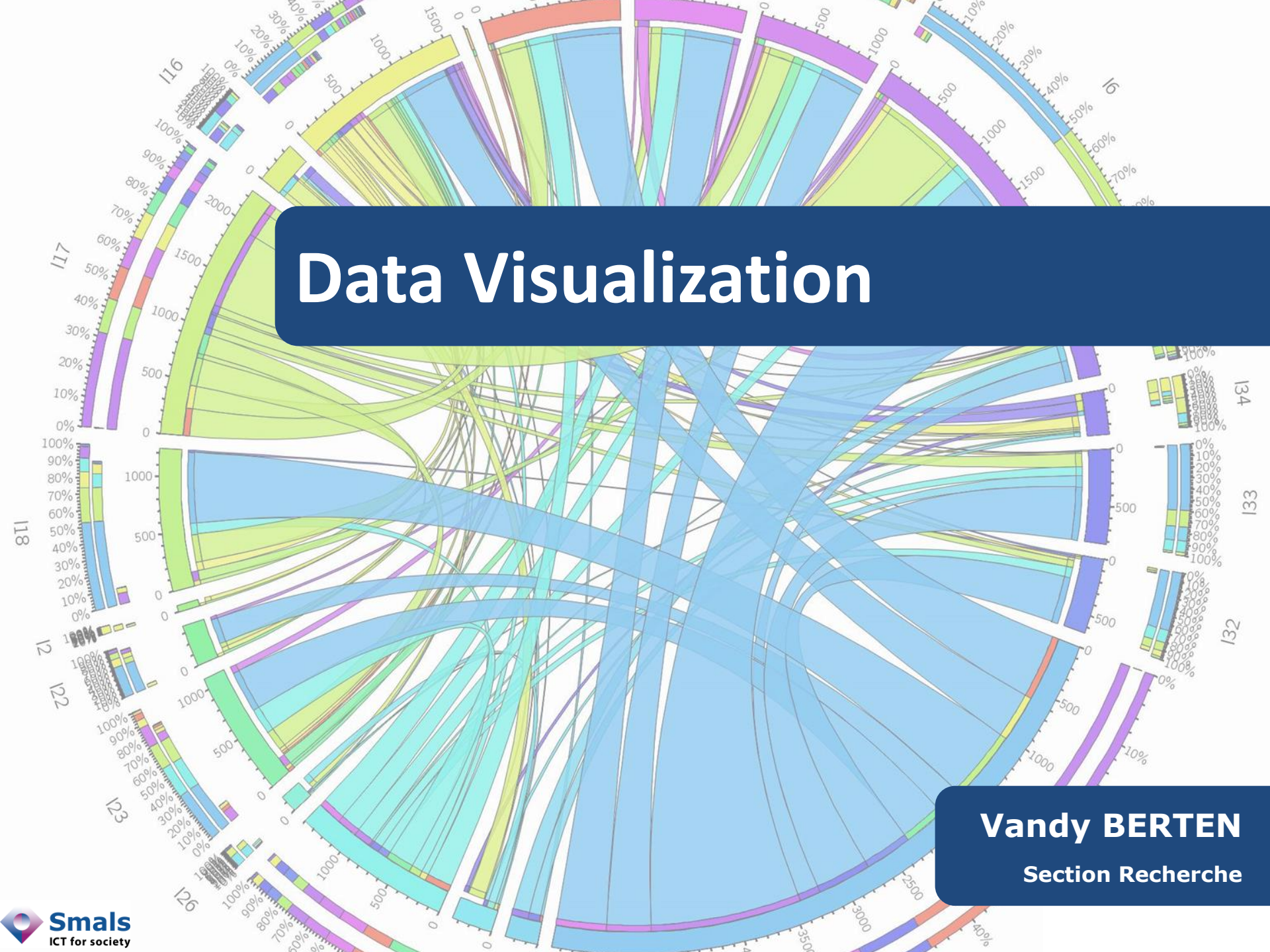


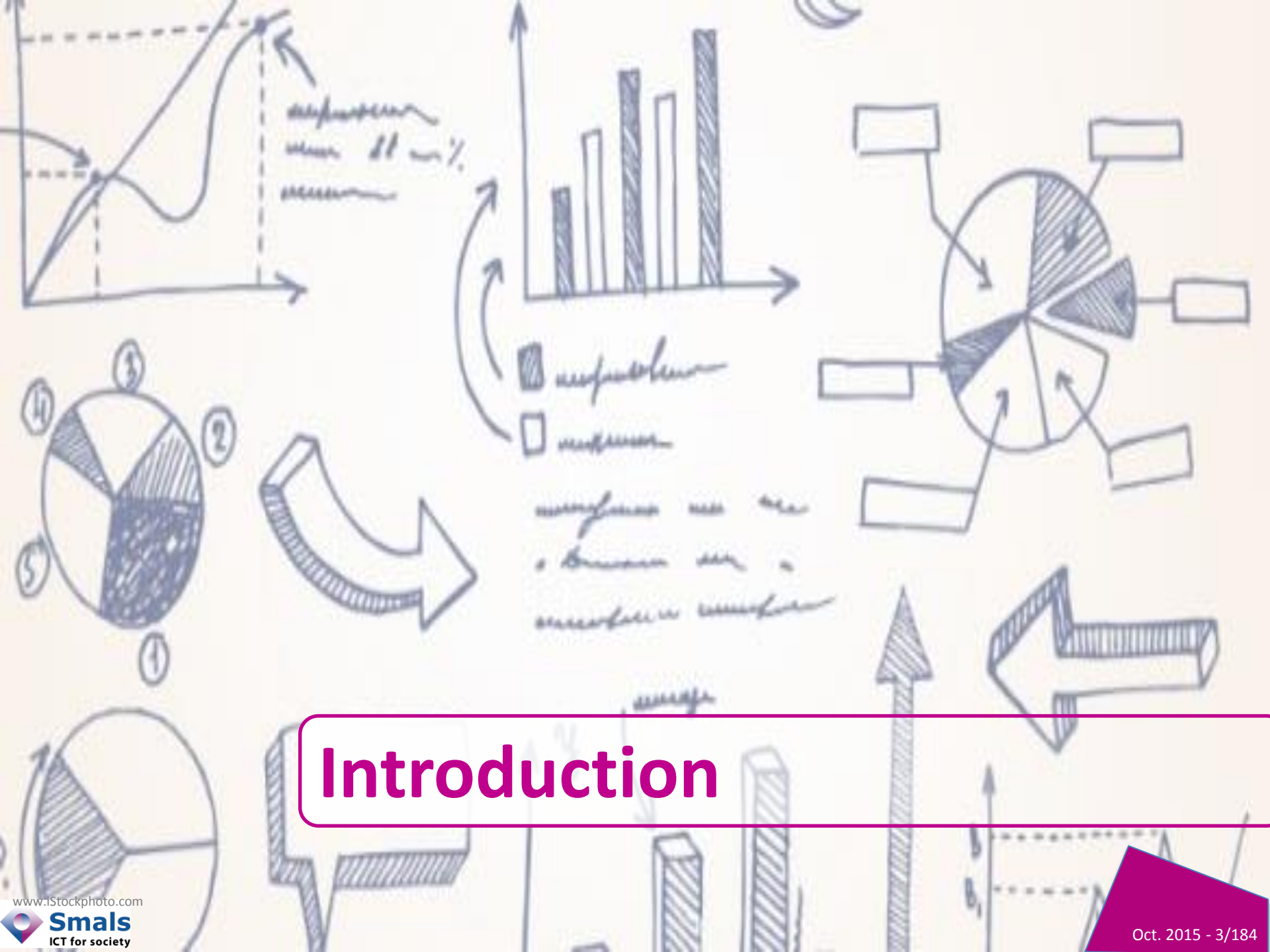
Data Visualization

Vandy BERTEN

Section Recherche

Smals
ICT for society

Table des matières



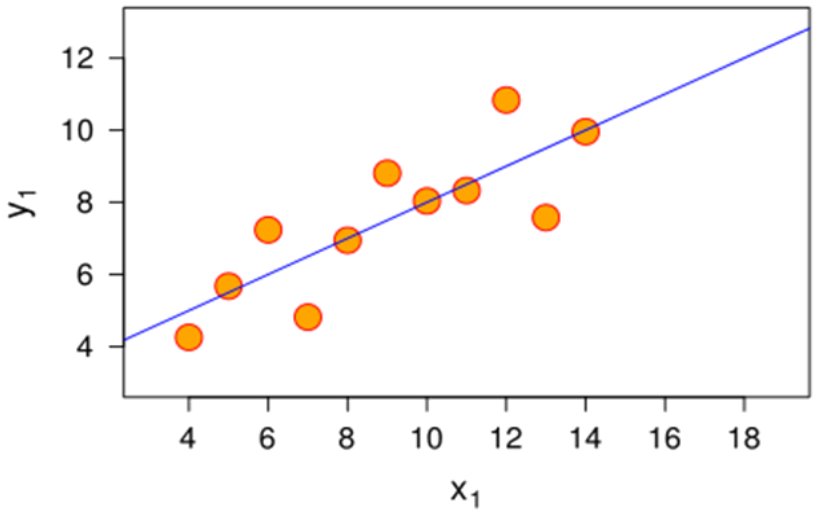
Introduction

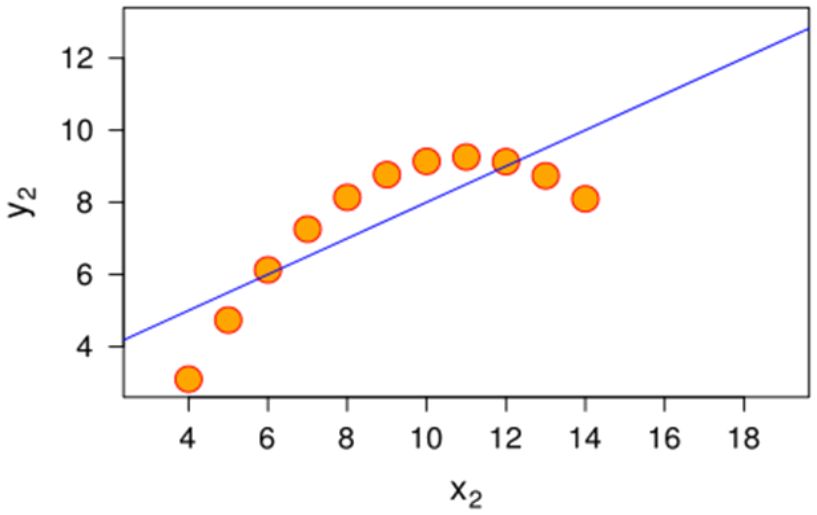
Limites des statistiques (Anscombe)

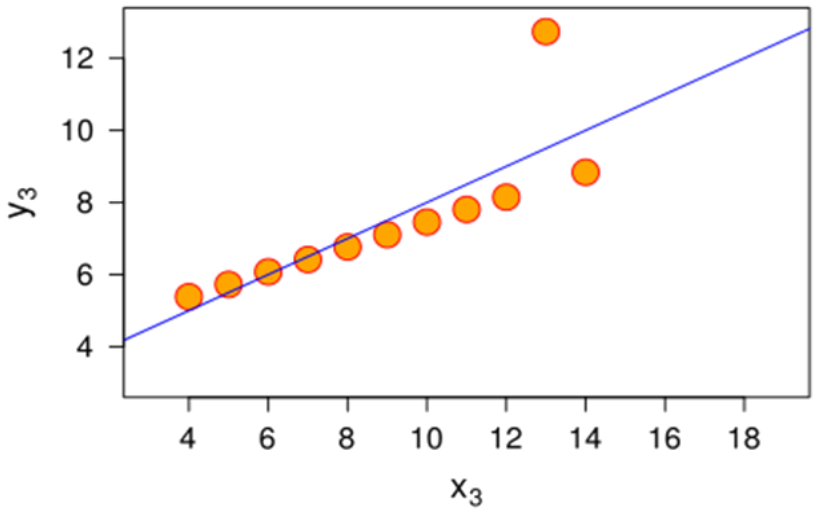
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.				8.0	7.71
9.0	8.81	9.				8.0	8.84
11.0	8.33	11.				8.0	8.47
14.0	9.96	14.				8.0	7.04
6.0	7.24	6.				8.0	5.25
4.0	4.26	4.				19.0	12.50
12.0	10.84	12.				8.0	5.56
7.0	4.82	7.				8.0	7.91
5.0	5.68	5.				8.0	6.89

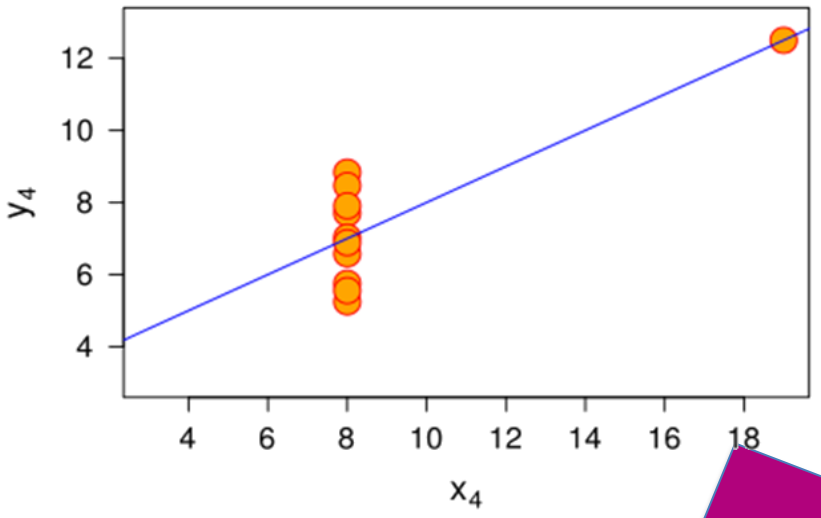
Propriété	Valeur
Moyenne des x	9.0
Variance des x	10.0
Moyenne des y	7.5
Variance des y	3.75
Corrélation entre les x et les y	0.816
Équation de la droite de régression linéaire	y = 3+0.5x
Somme des carrés des erreurs relativement à la moyenne	110.0

Limites des statistiques (Anscombe)









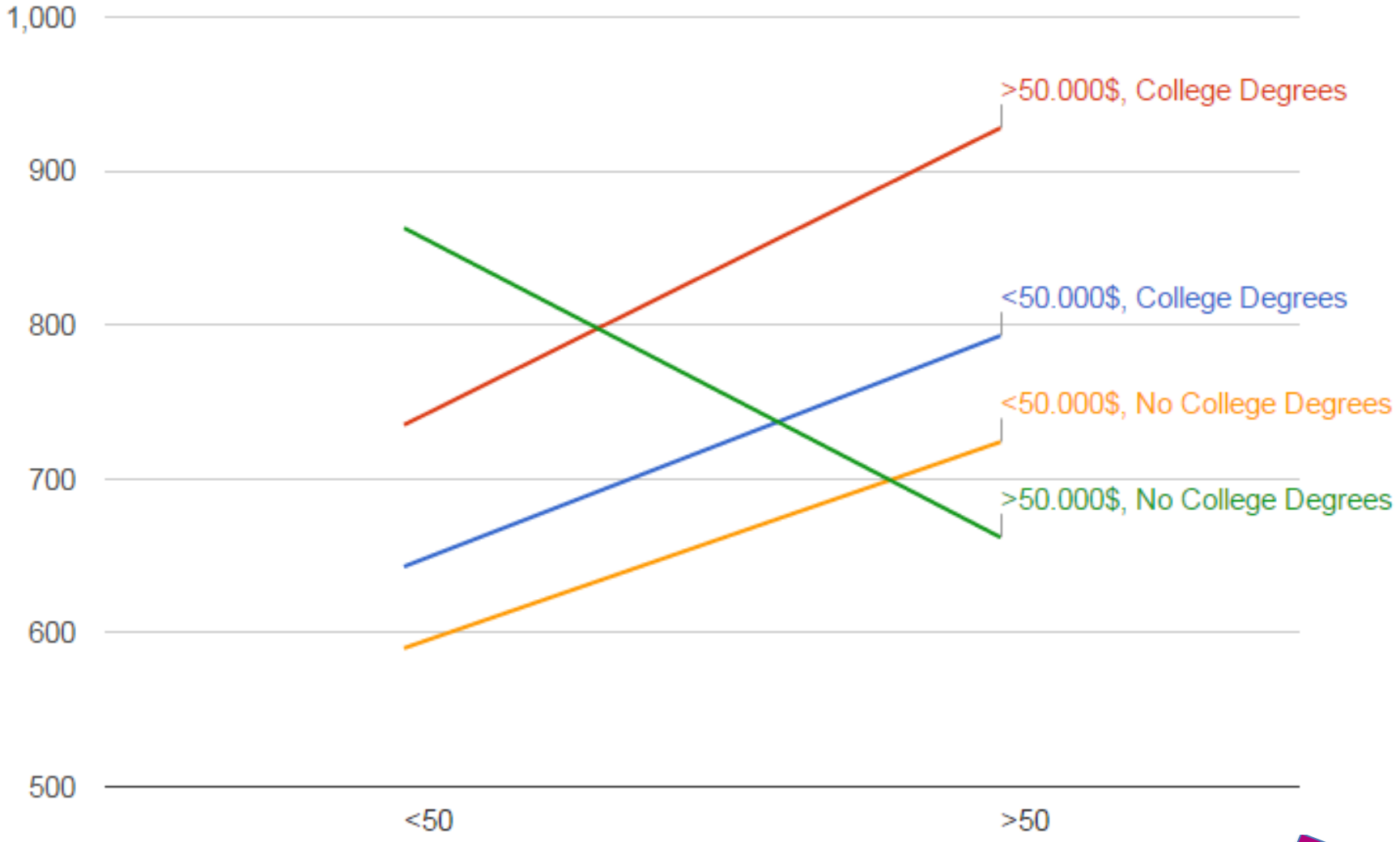
Crédits : Wikipedia

Limites de la cognition

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

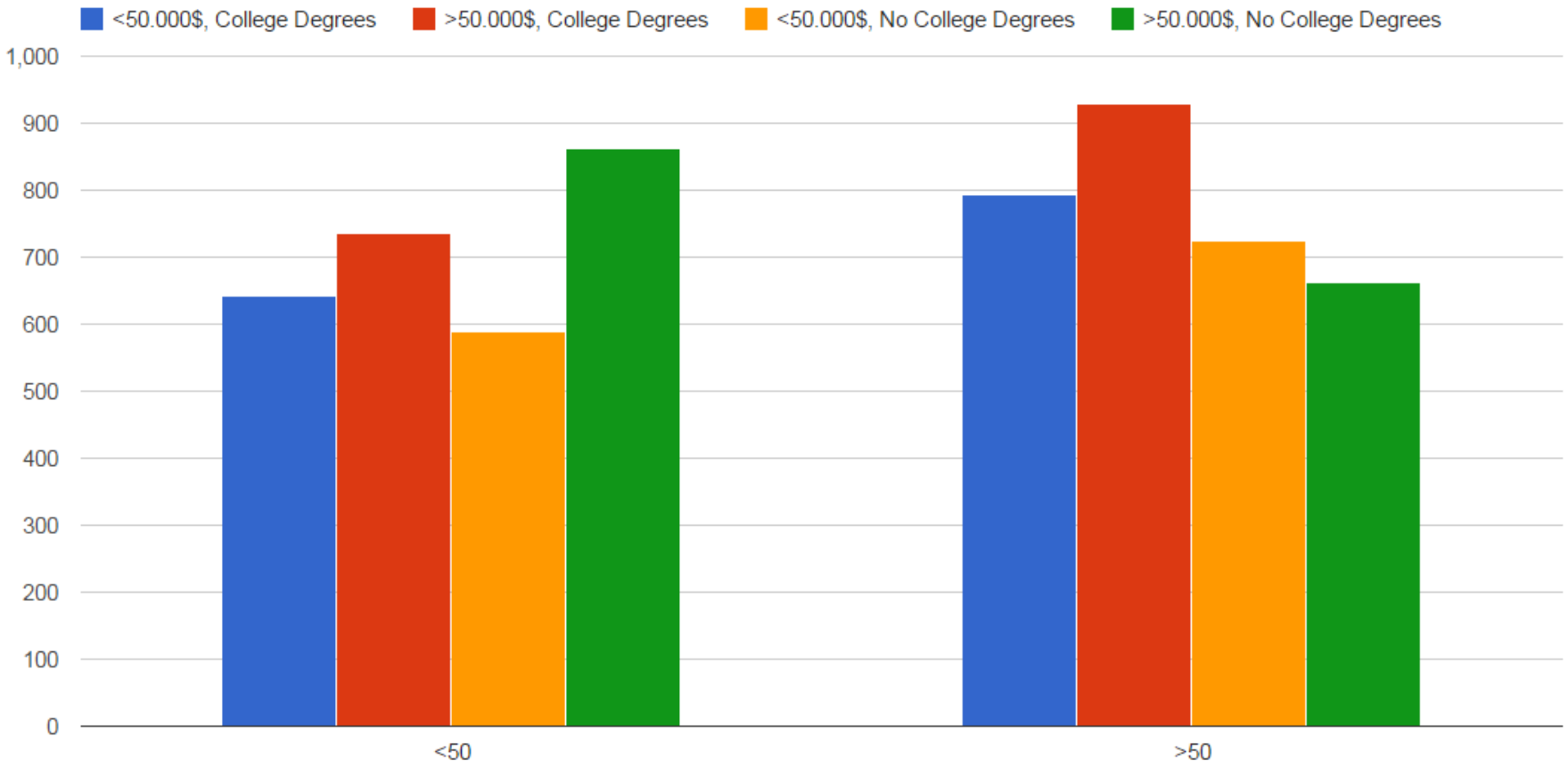
- Est-ce que la satisfaction de tous les employés augmente avec l'âge ?
- Un groupe a-t-il un *pattern* différent des autres ?

Limites de la cognition



Outil : Google Chart

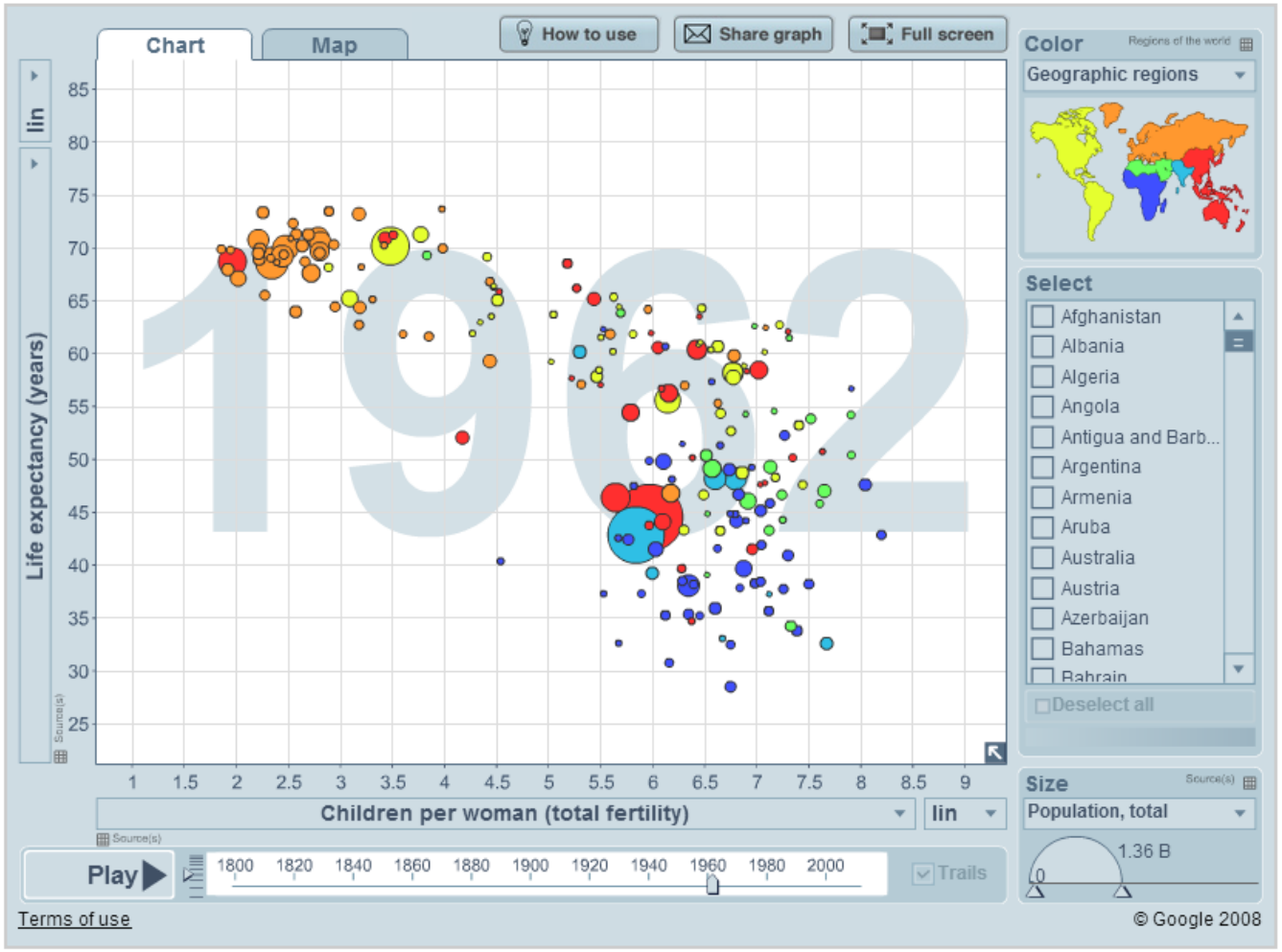
Limites de la cognition



Outil : Google Chart

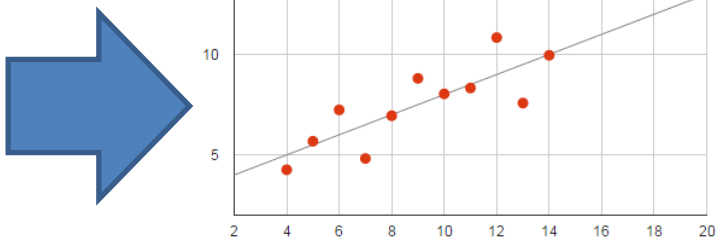
Santé (John Snow, 1854)

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

Science Photo Library/Getty Creative

Data visualization : objectifs

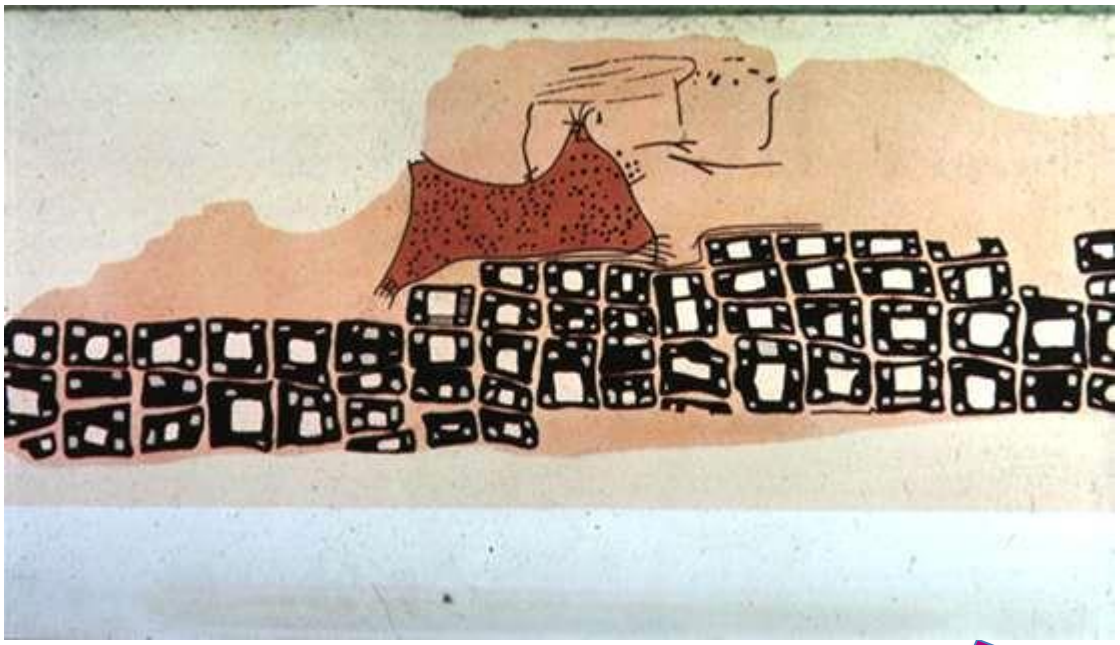
Enregistrer de l'information

Communiquer des informations (visual explanation)

Analyser des données (visual analytics, visual exploration)

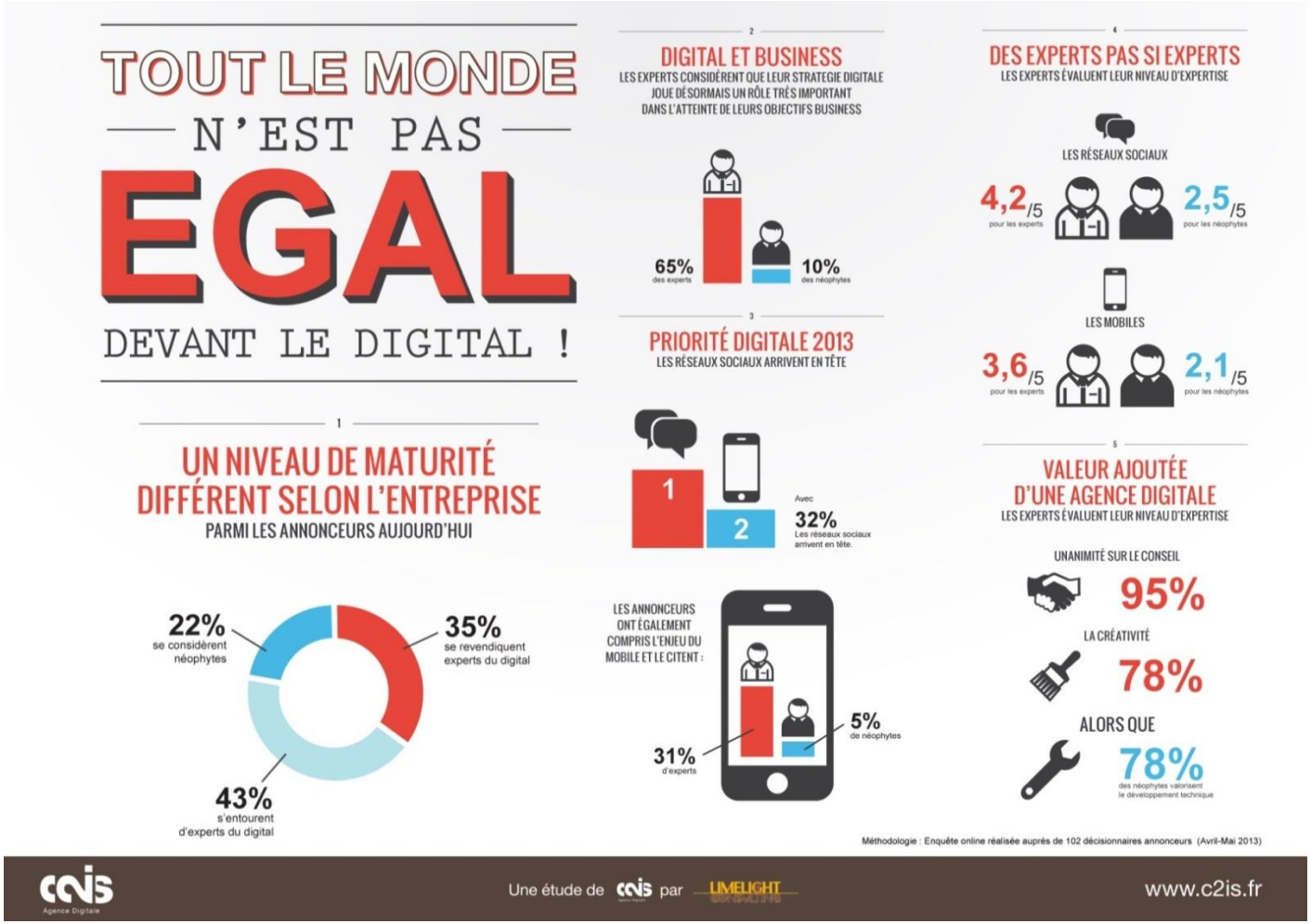
Enregistrer de l'information

Anaximader's Map of the World

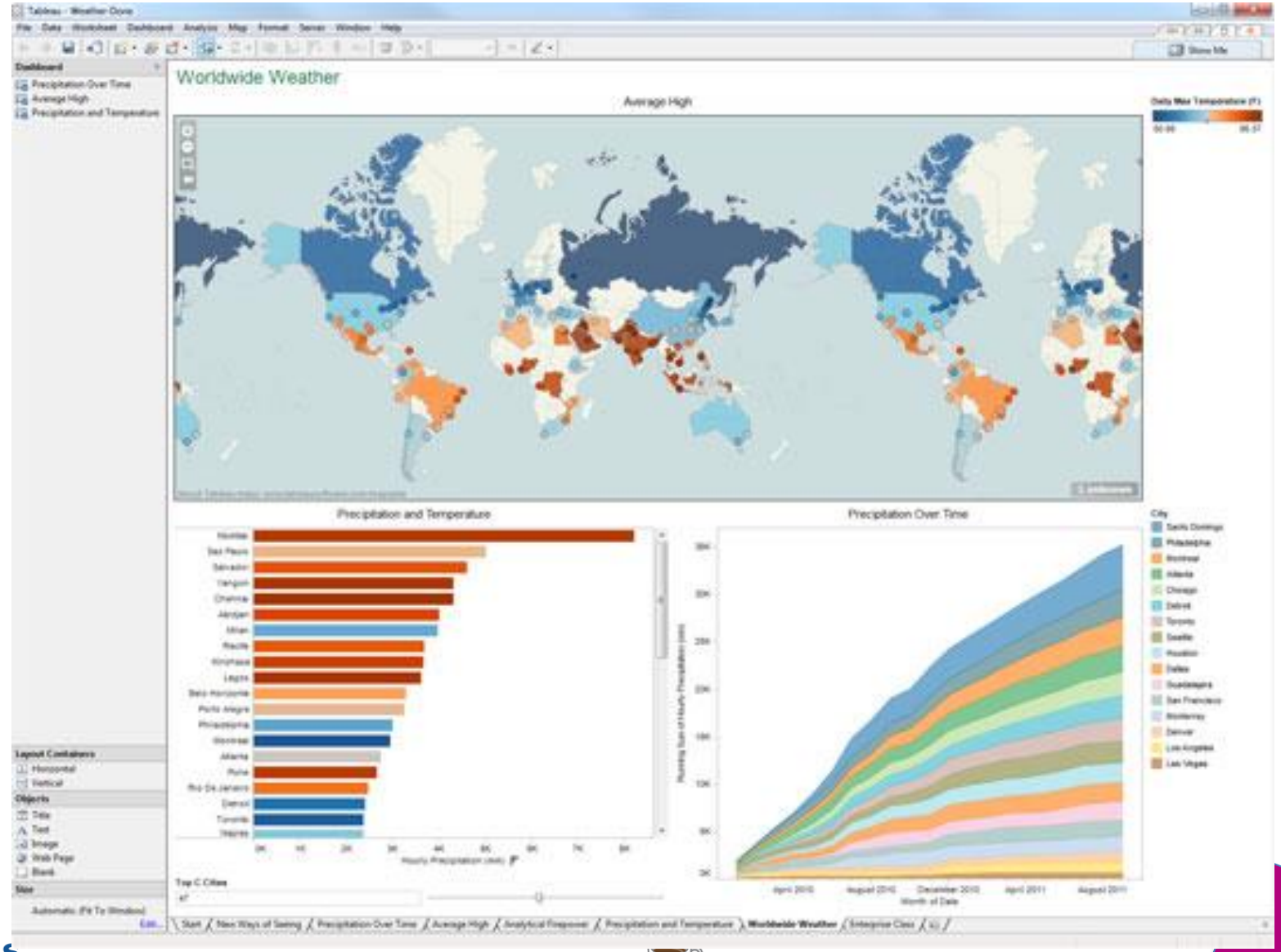


http://www.datavis.ca/milestones/

Communiquer des informations



Analyser des données



Outil : Tableau Software

Les fondamentaux

Outline

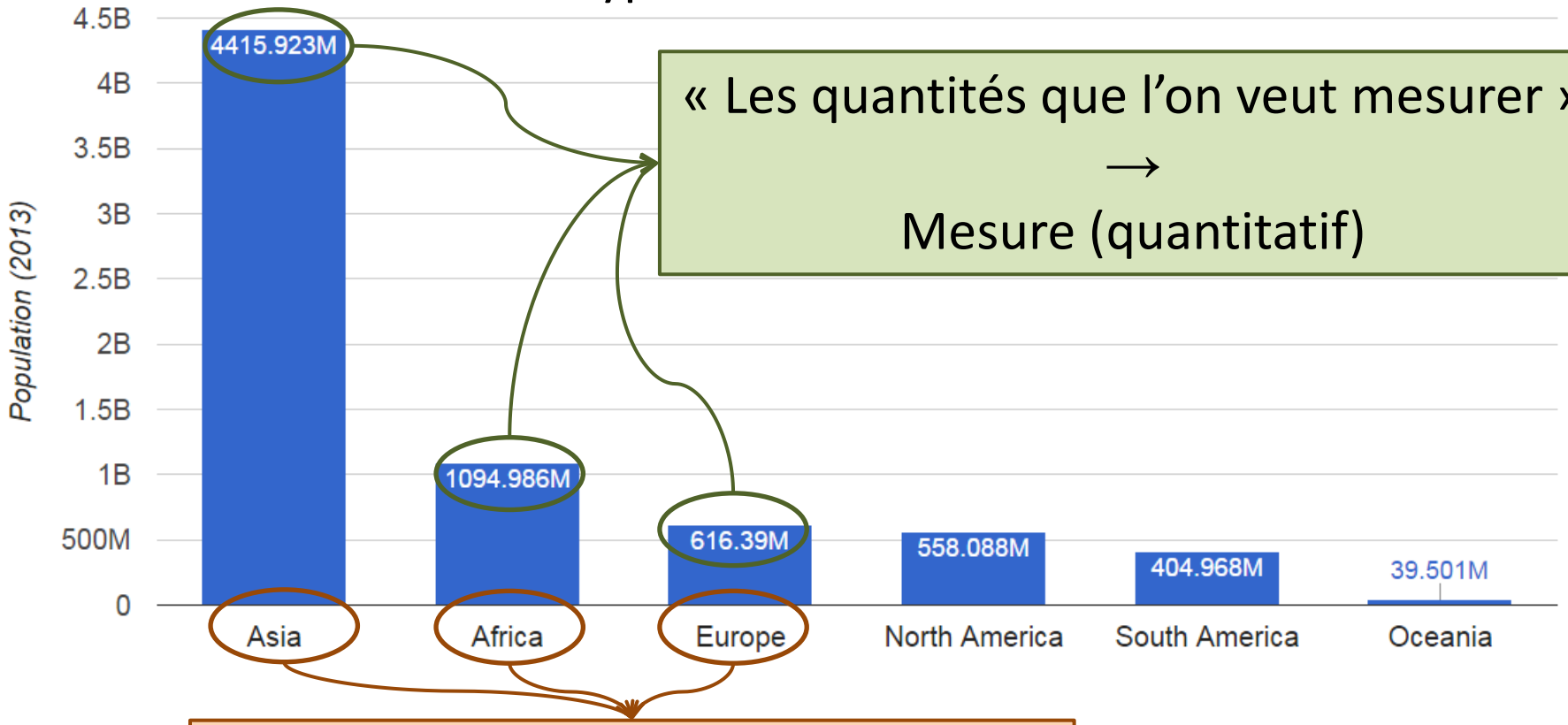
- Données :
  - Classification par fonction (Mesure & dimension)
  - Classification par nature (Quantitatif, qualificatif)
- Graphique :
  - Éléments graphiques (sémiologie)
  - Perception (Gestalt, processus pré-attentif)
- Qualité :
  - Mesures de qualité
  - Exemples

Les fondamentaux

MESURES & DIMENSIONS

Mesures & Dimensions

Deux types d'informations :



« Les quantités que l'on veut mesurer »

→

Mesure (quantitatif)

« Les choses que l'on veut suivre »

→

Dimension (qualitatif/quantitatif)

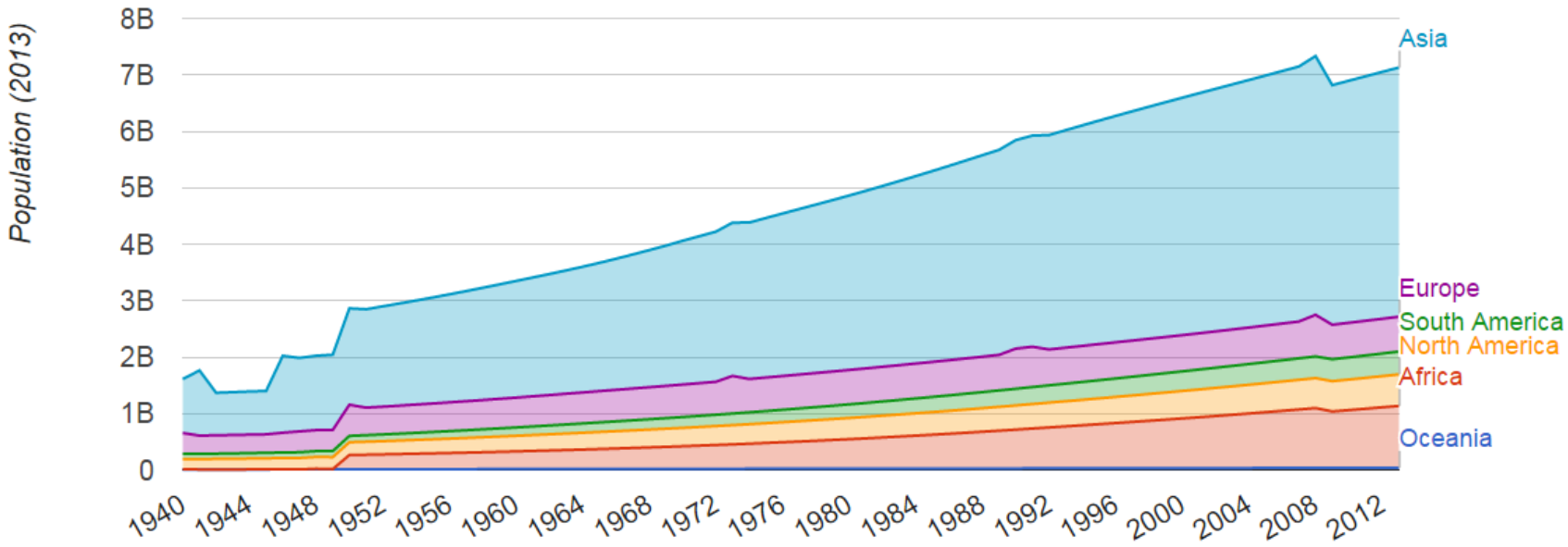
Outil : Google Chart

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)



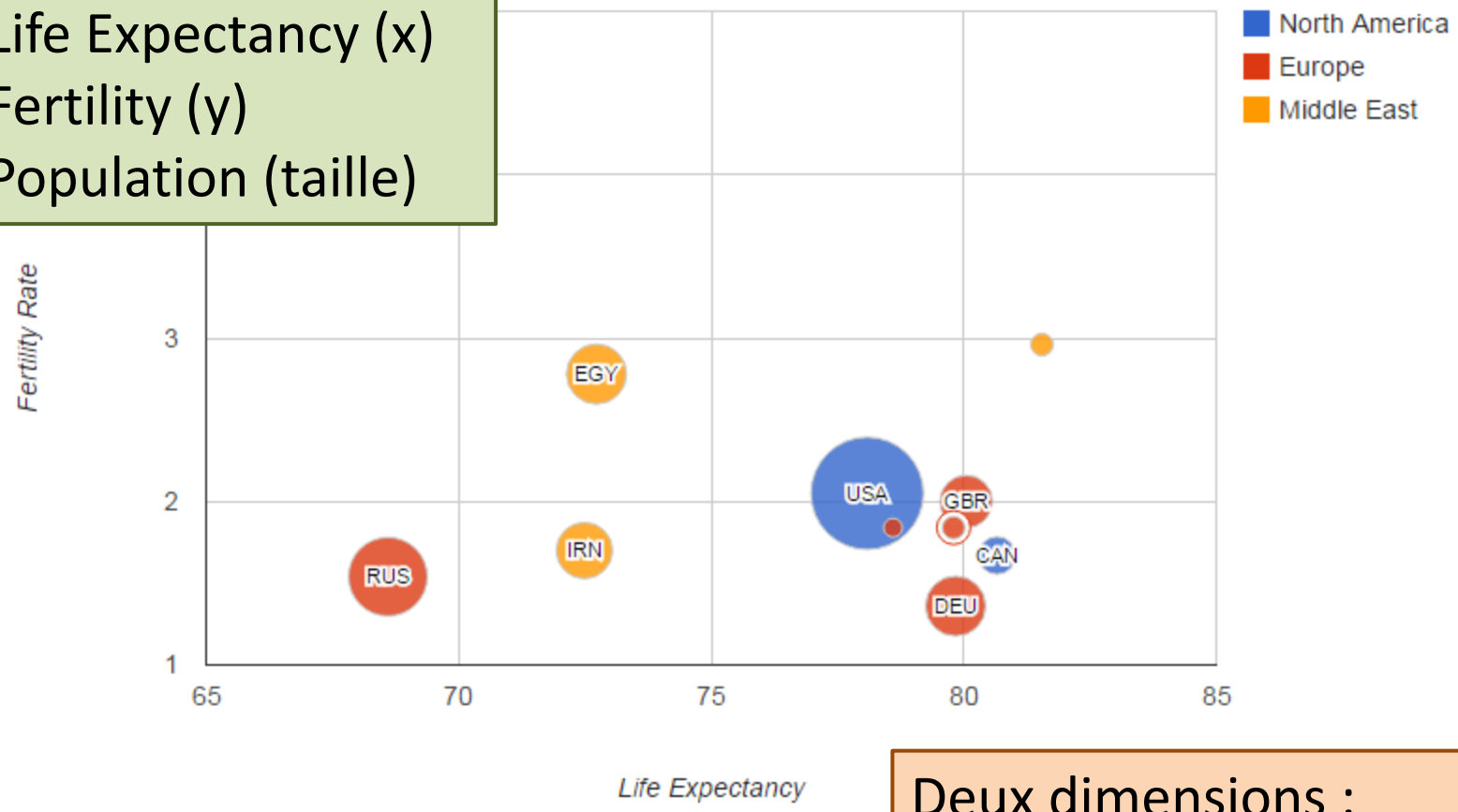
Deux dimensions (année, continent)

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)

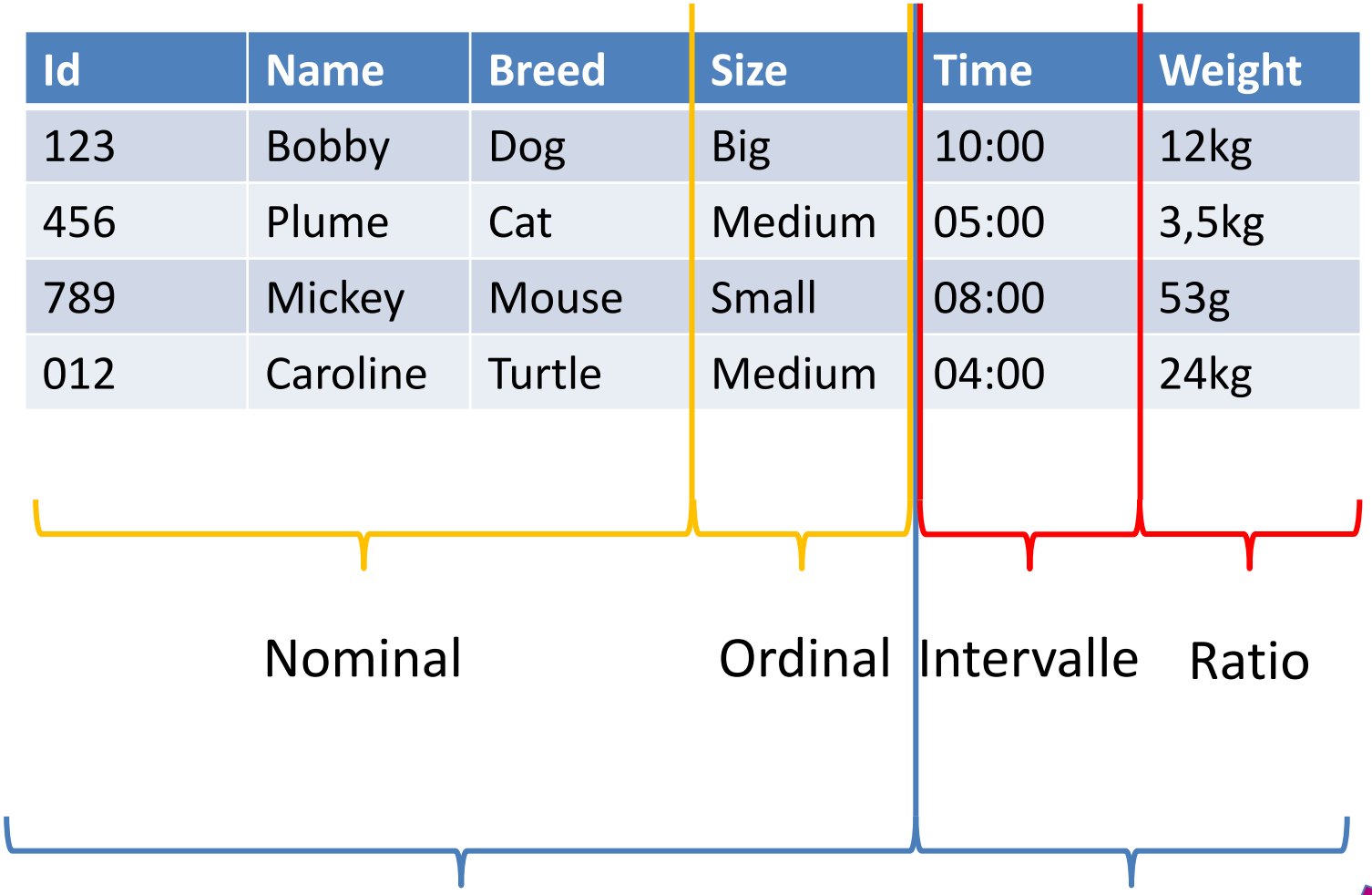
Outil : Google Chart

Les fondamentaux

SCALE OF MEASUREMENT

Scale of measurement (Stevens)

Id	Name	Breed	Size	Time	Weight
123	Bobby	Dog	Big	10:00	12kg
456	Plume	Cat	Medium	05:00	3,5kg
789	Mickey	Mouse	Small	08:00	53g
012	Caroline	Turtle	Medium	04:00	24kg



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

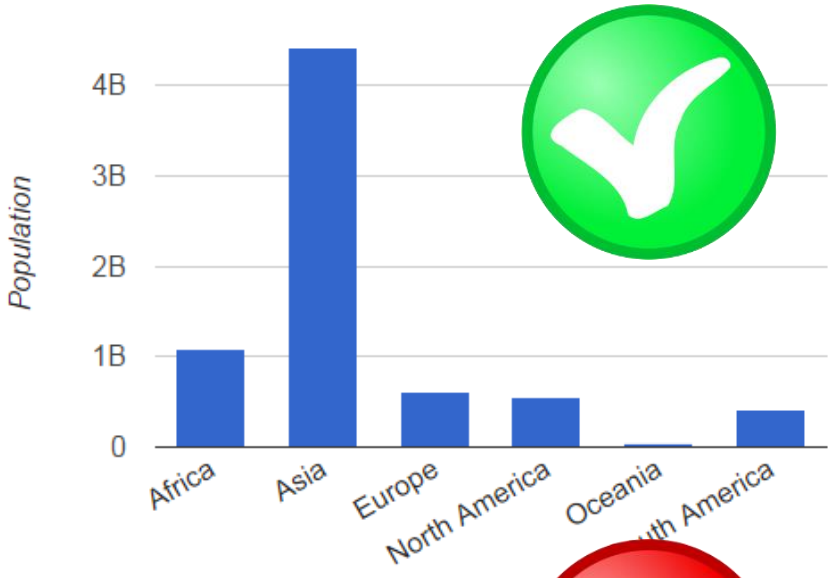
25 secondes

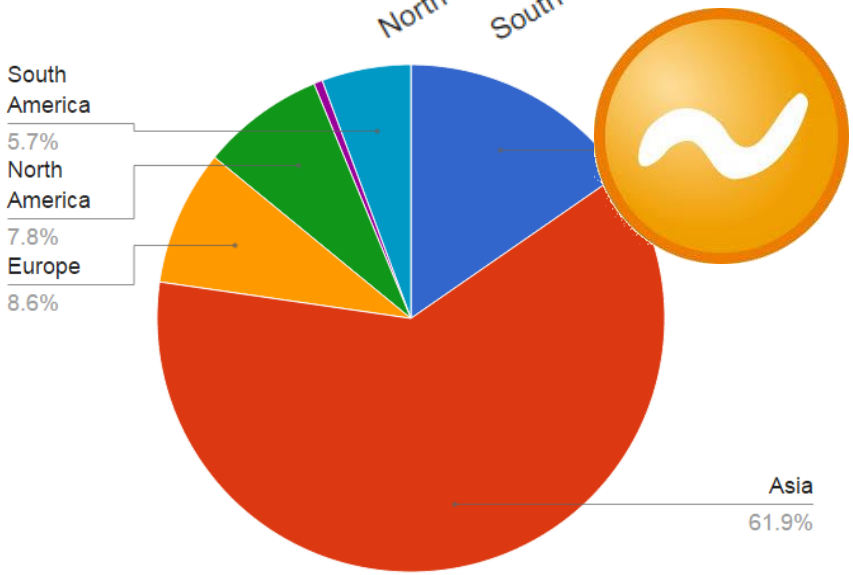
105€

- , +, /

=, ≠, <, >

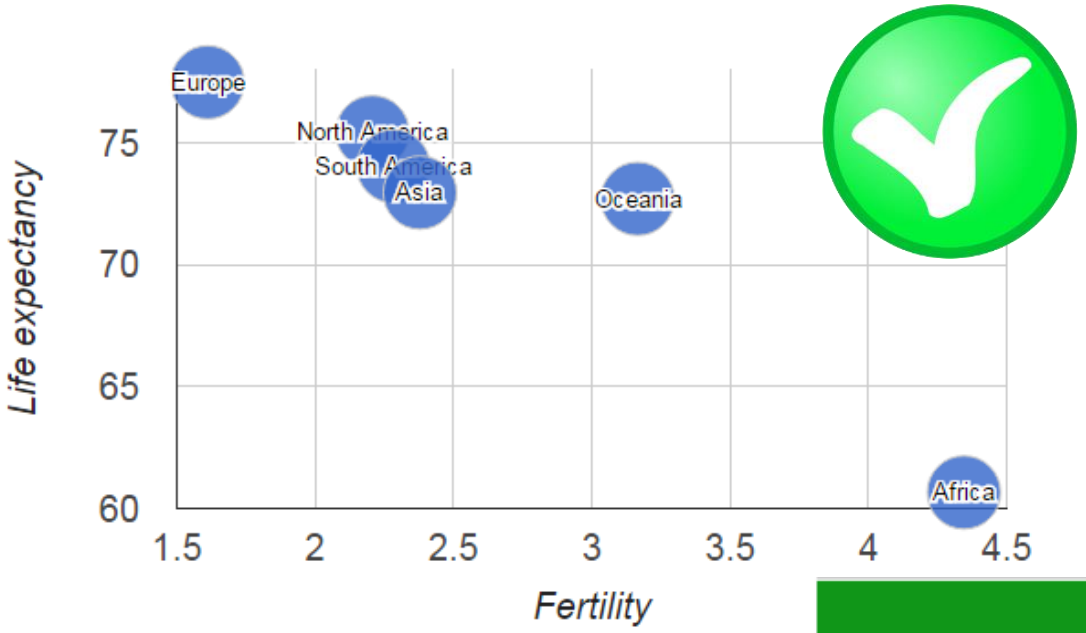
Nominal : Dimension

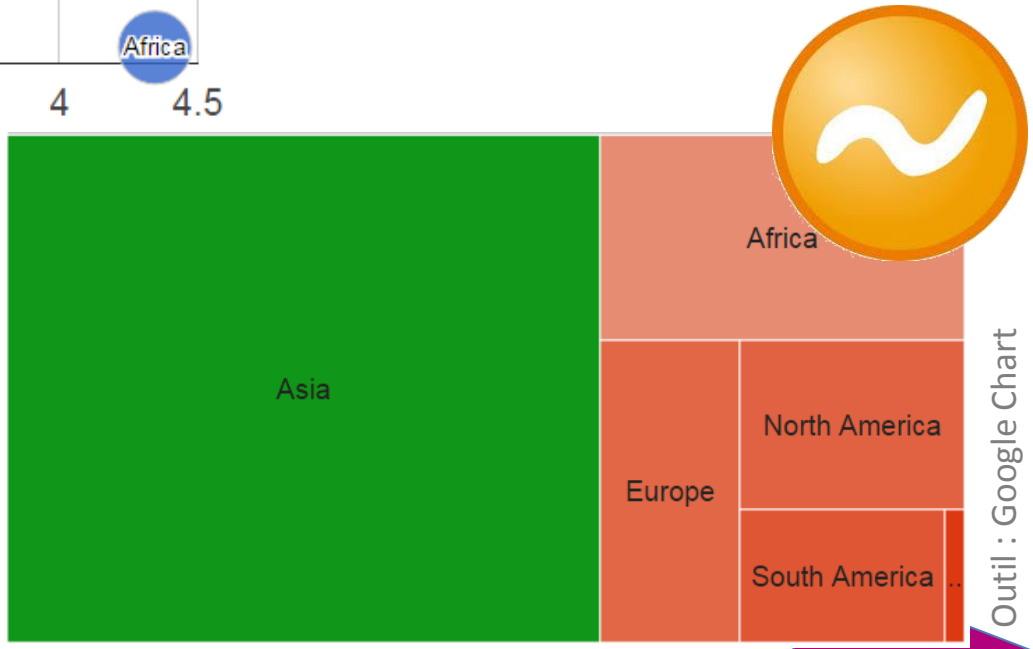




Outil : Google Chart

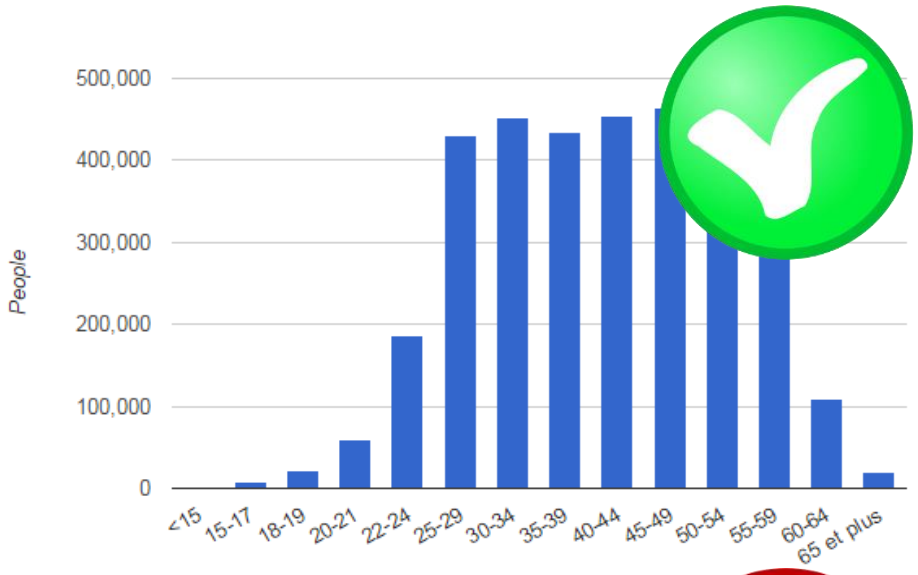
Nominal : Dimension

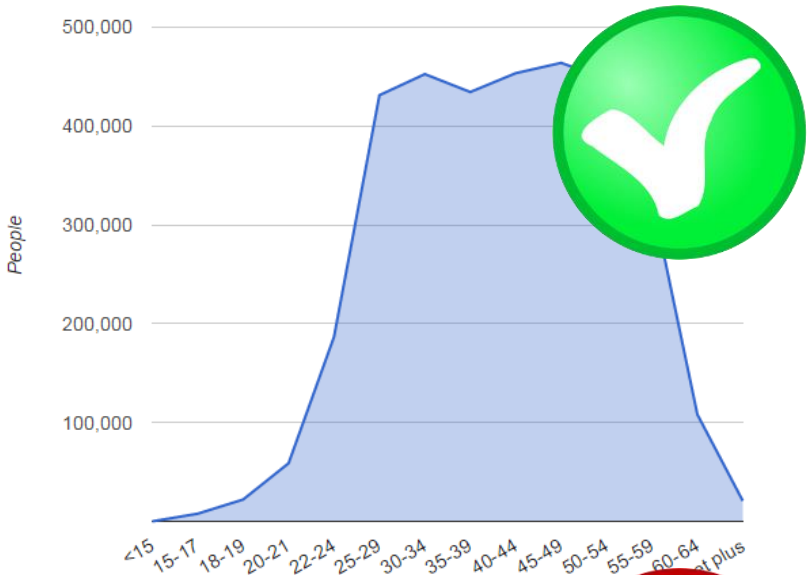


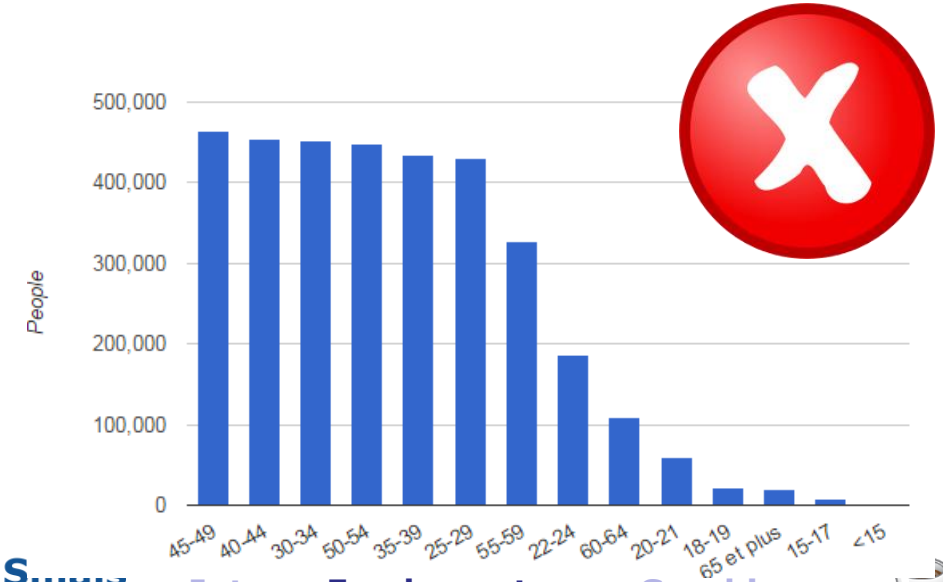


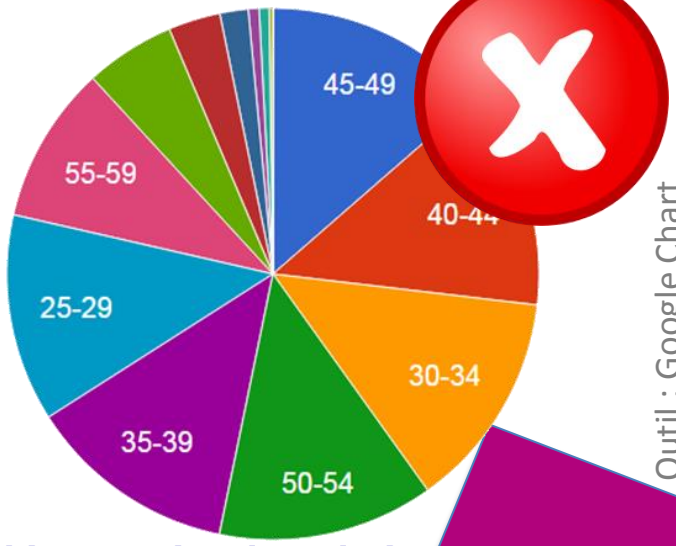
Outil : Google Chart

Ordinal : Dimension



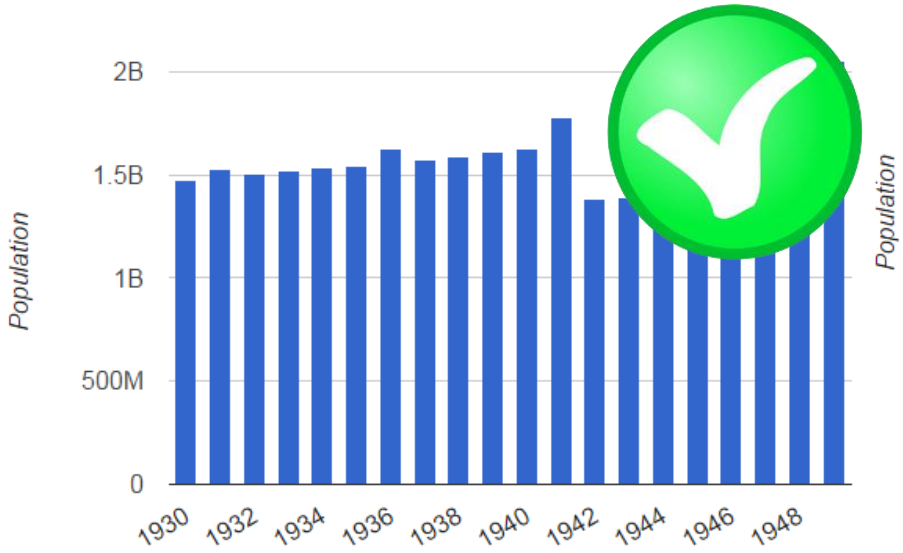


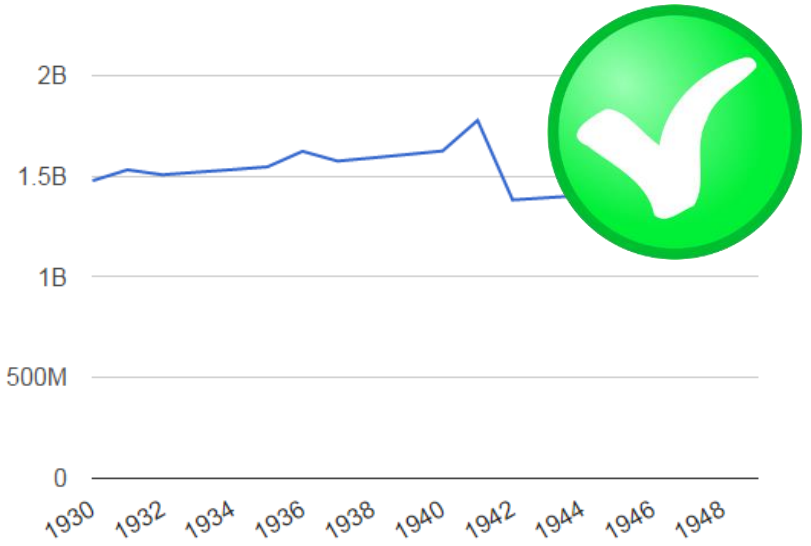


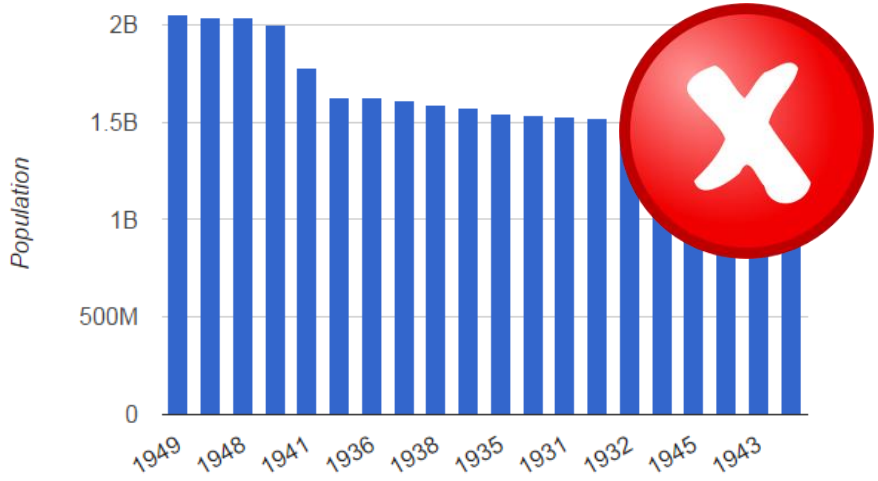


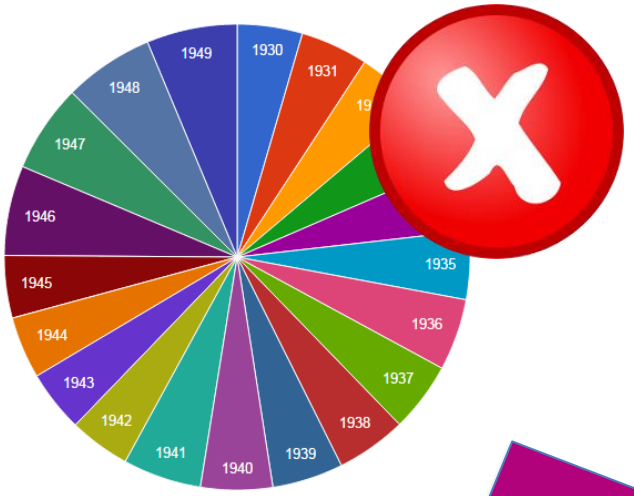
Outil : Google Chart

Intervalle : dimension









Outil : Google Chart

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

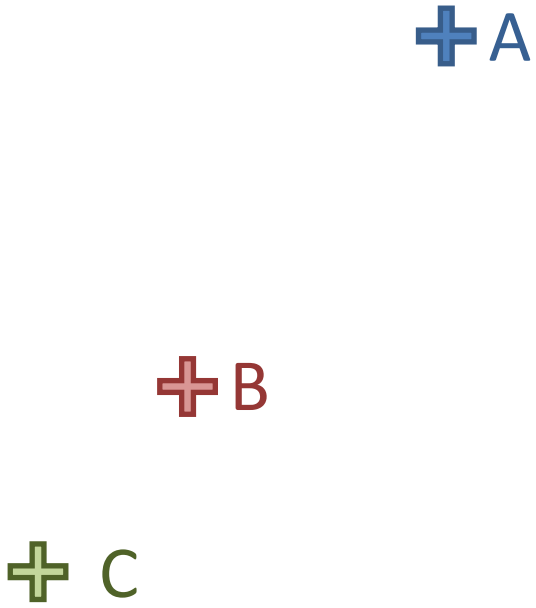
SÉMIOLOGIE GRAPHIQUE

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

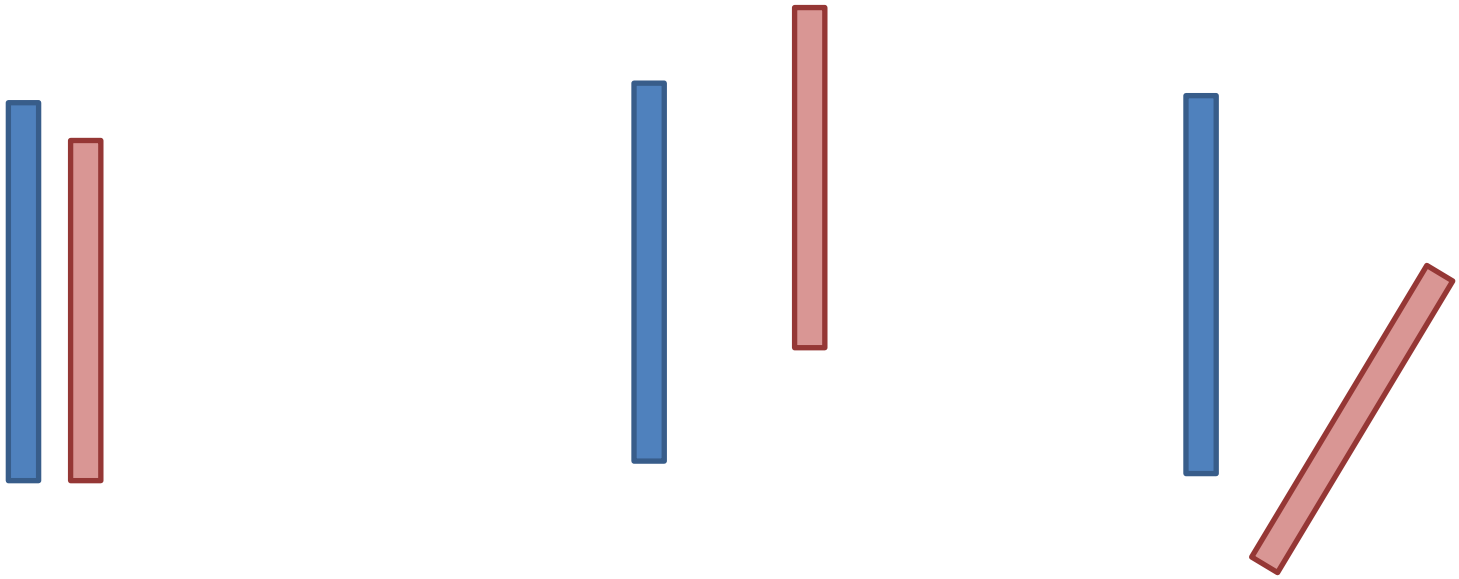
Sémiologie graphique : Position

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



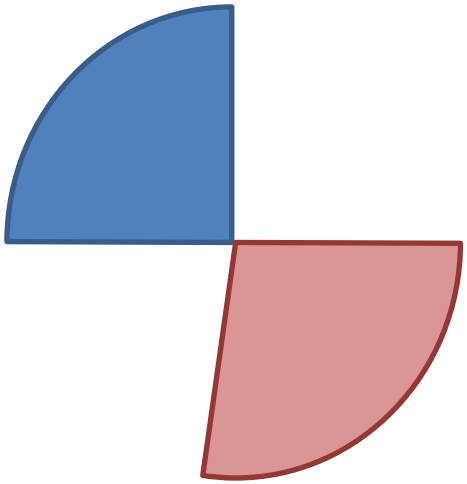
Sémiologie graphique : Longueur

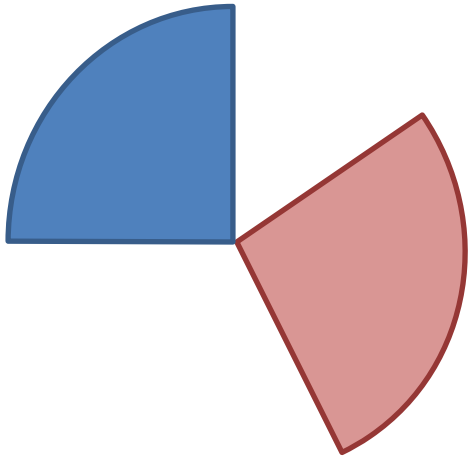
- Quelle est la plus longue barre (ratio : 9/10)?



Sémiologie graphique : Angle

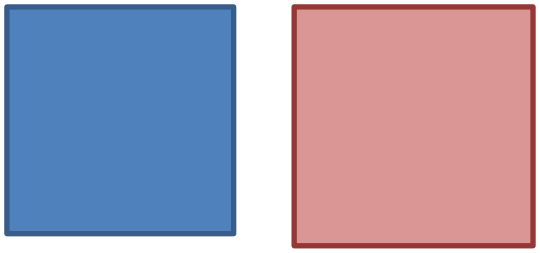
- Quelle tranche est la plus grande ?

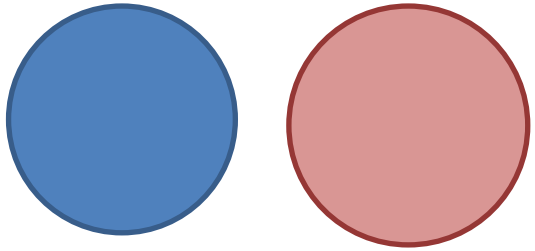


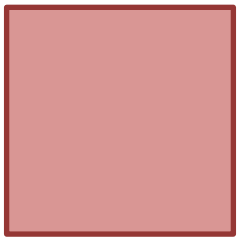


Sémiologie graphique : Surface

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





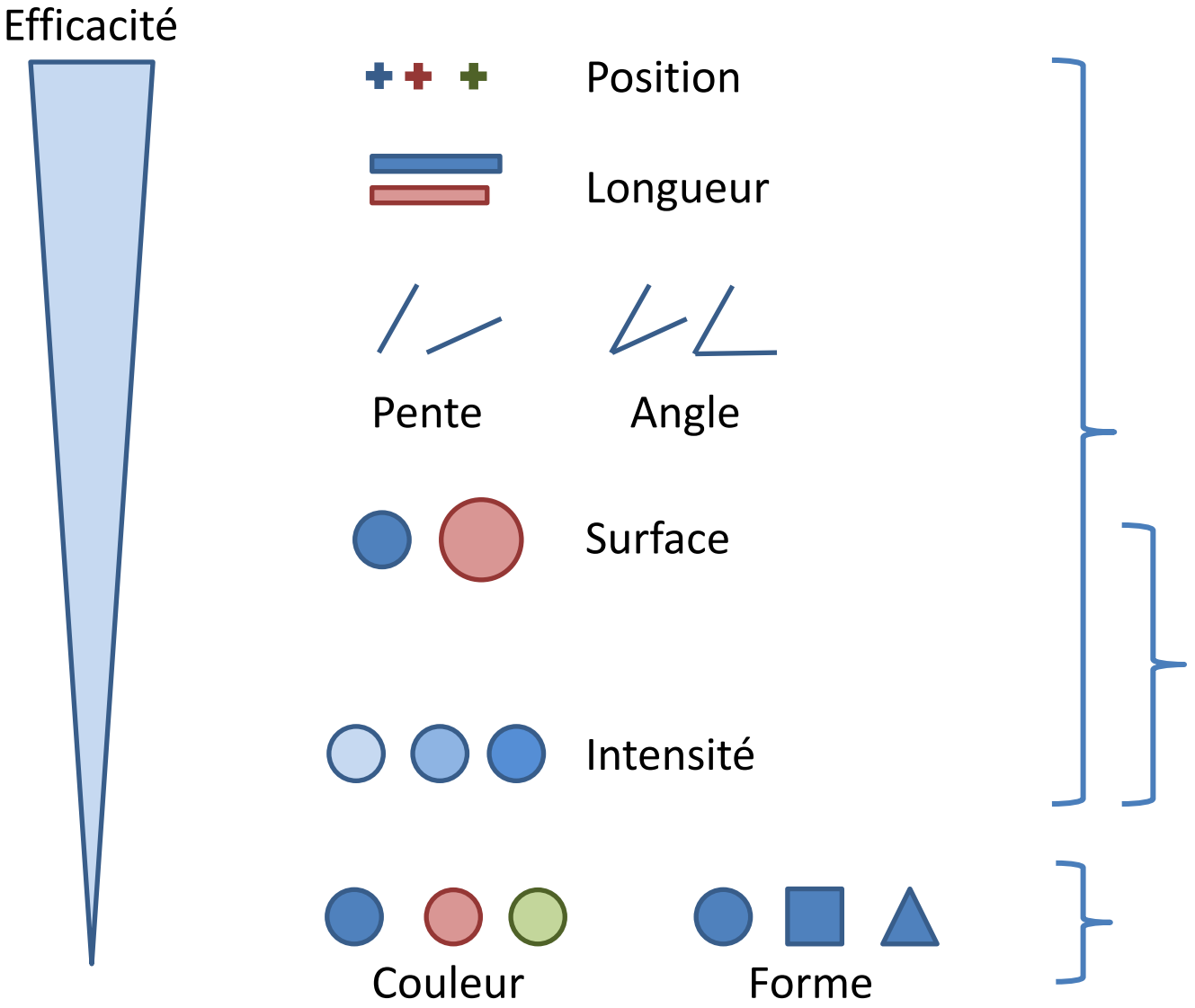


Sémiologie graphique : Couleur

0 10

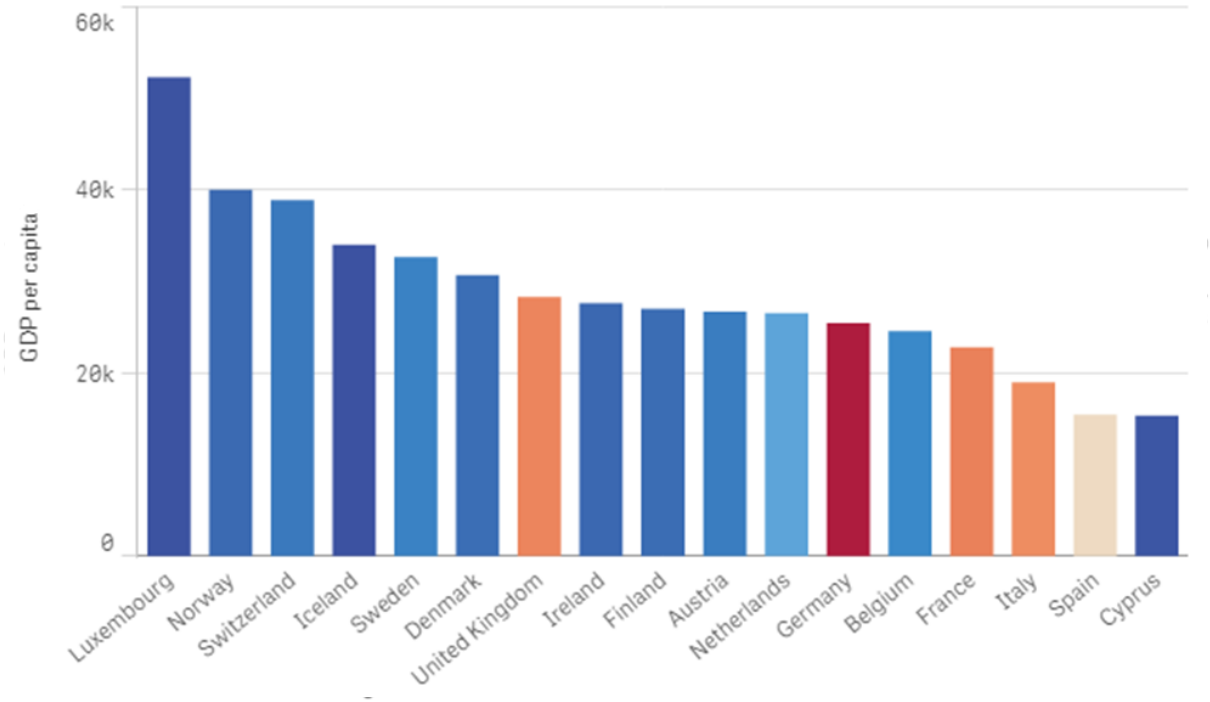
Combien représentent ces boules :

Sémiologie graphique



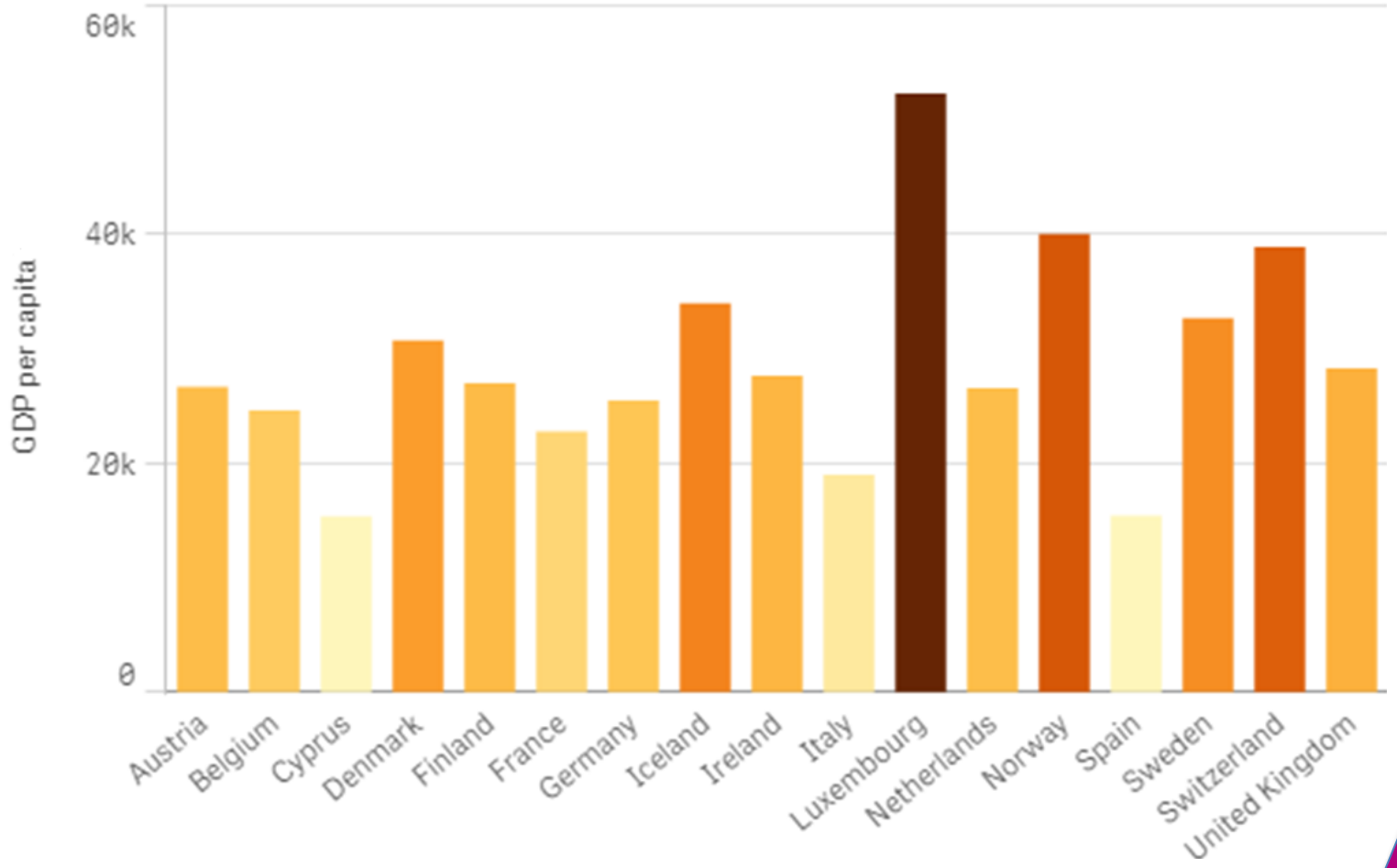
Sémiologie graphique

- 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

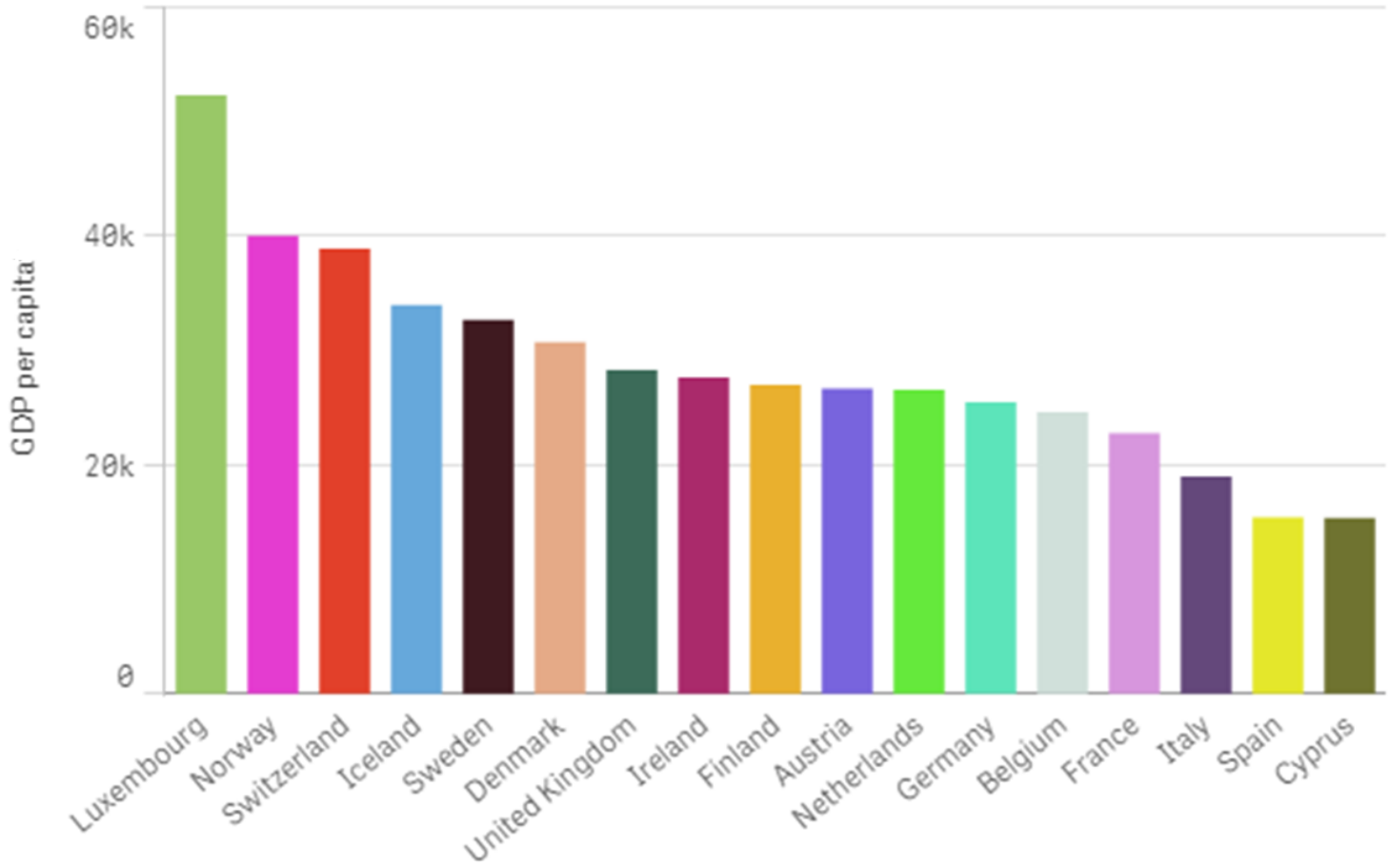
- Éventuellement, duplication des canaux :



Outil : Qlik Sense

Sémiologie graphique

- À éviter : canal « inutile »



Outil : Qlik Sense

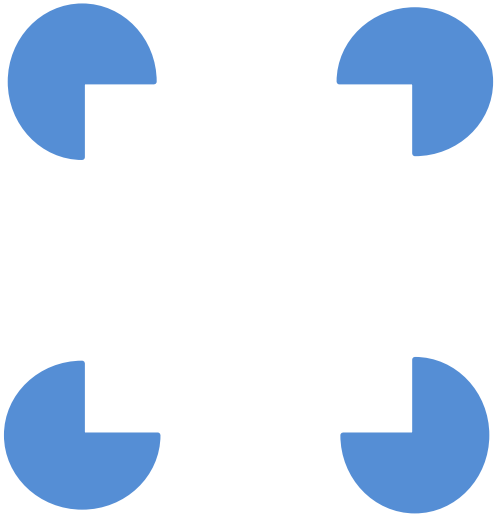
Les fondamentaux

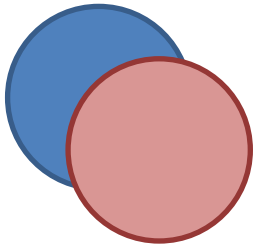
LOIS DE GESTALT

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

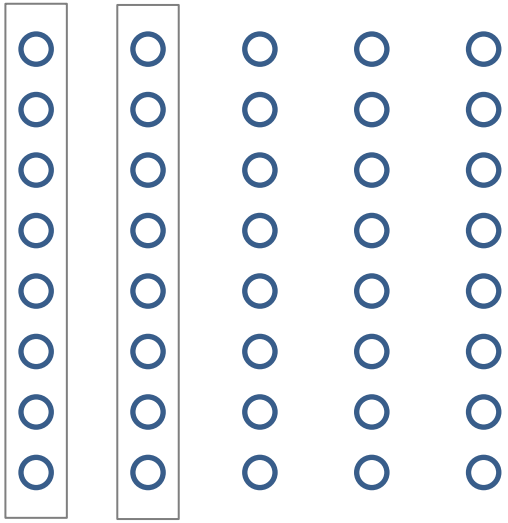


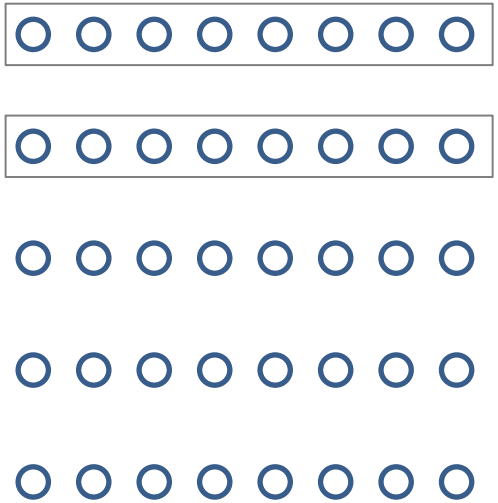


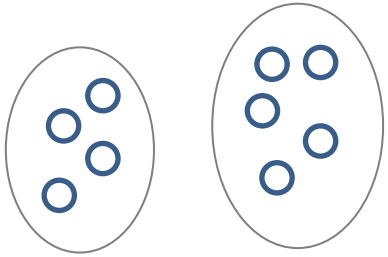
Fermeture : Terminaison des contours manquants

Crédits : Hanspeter Pfister, Harvard

Gestalt Laws : Proximity







Proximité : regroupement des entités proches

Gestalt Laws : Proximity

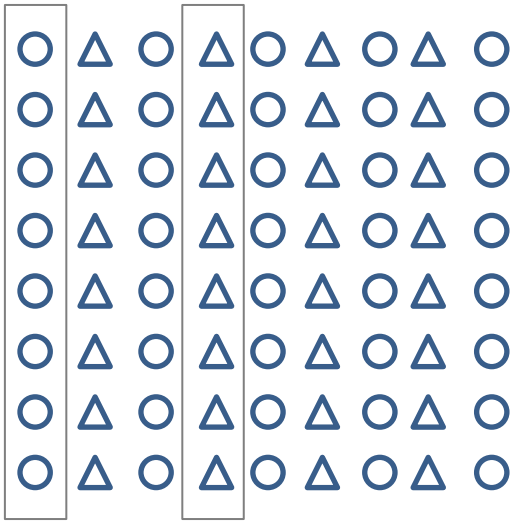
Mise en évidence des colonnes :

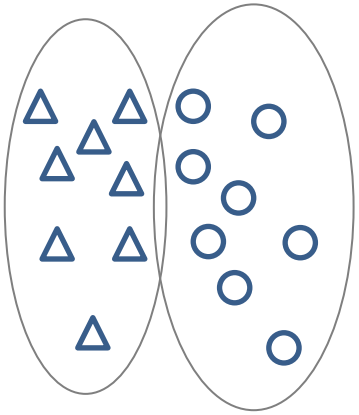
1	5	3	1	7
2	5	4	7	4
5	7	3	5	8
6	8	4	3	2
7	7	9	7	1

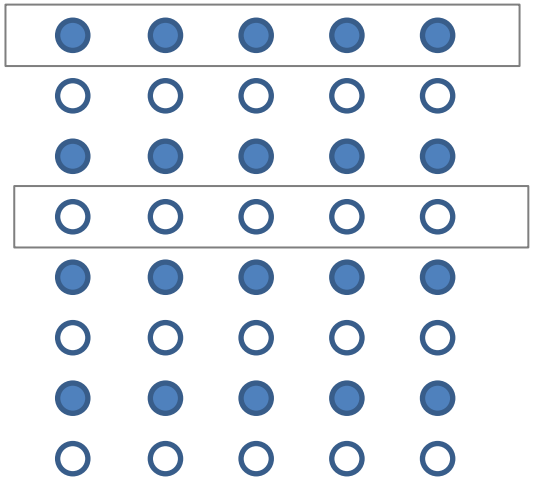
Mise en évidence des lignes :

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

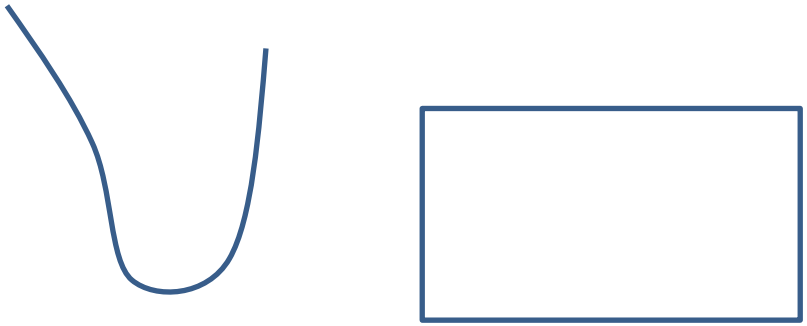
1	5	3	1	7
2	5	4	7	4
5	7	3	5	8
6	8	4	3	2
7	7	9	7	1

Mise en évidence des lignes :

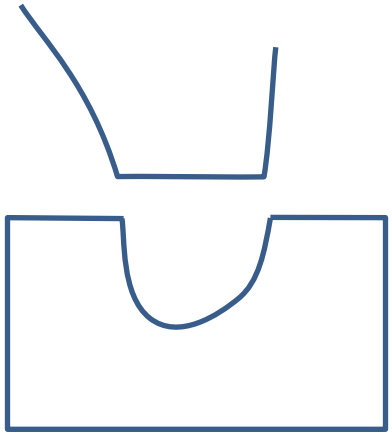
1	5	3	1	7
2	5	4	7	4
5	7	3	5	8
6	8	4	3	2
7	7	9	7	1

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

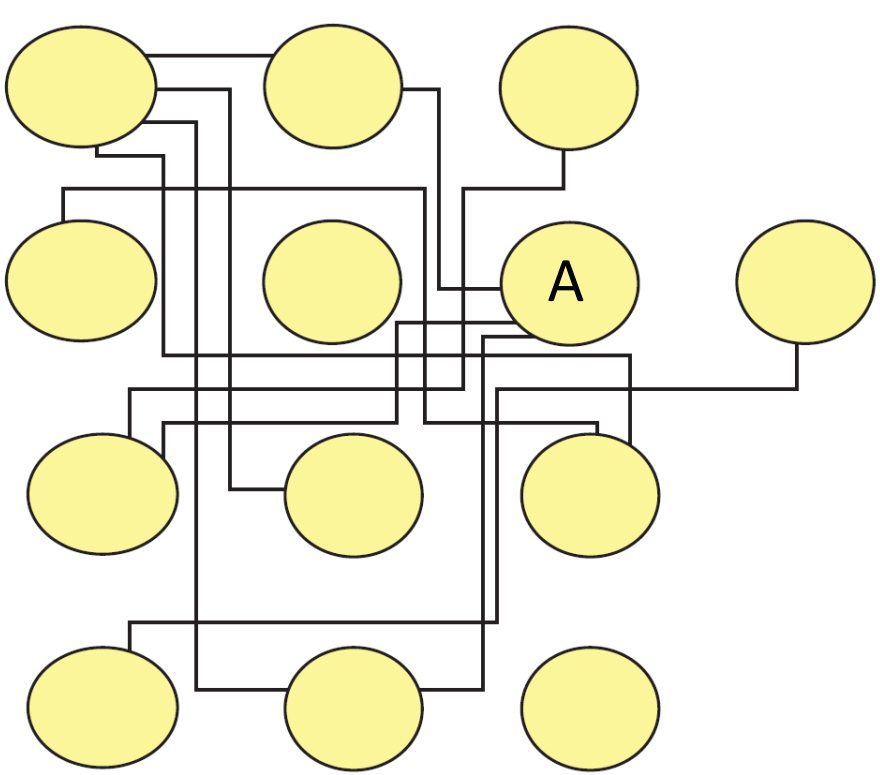


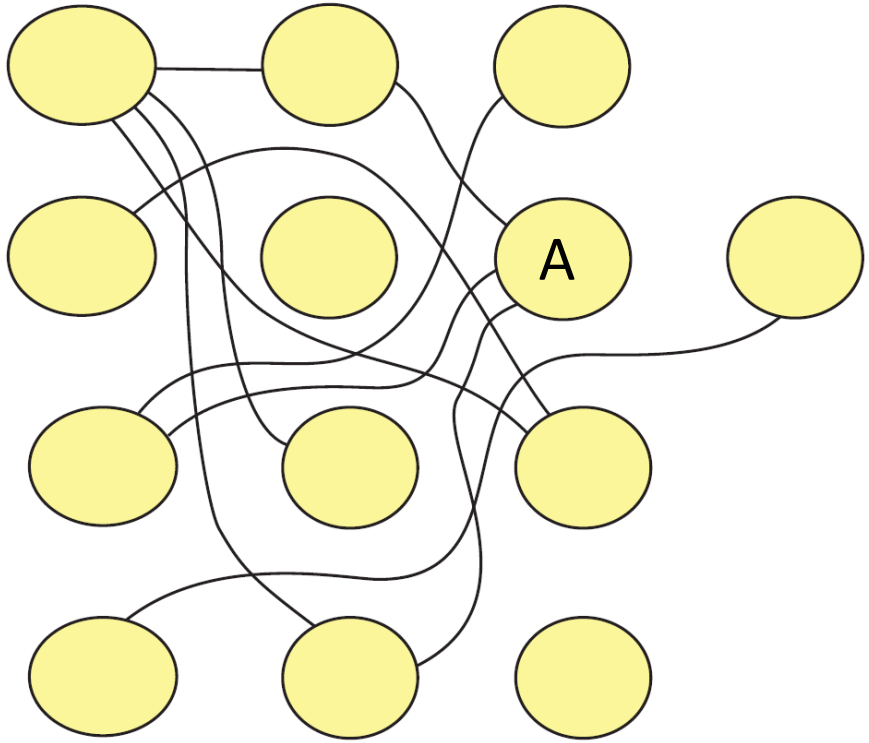
Continuité :
Construction sur base d'éléments lisses et continus



Gestalt Laws : Continuity

- À quoi est connecté A ?





[V]

Les fondamentaux

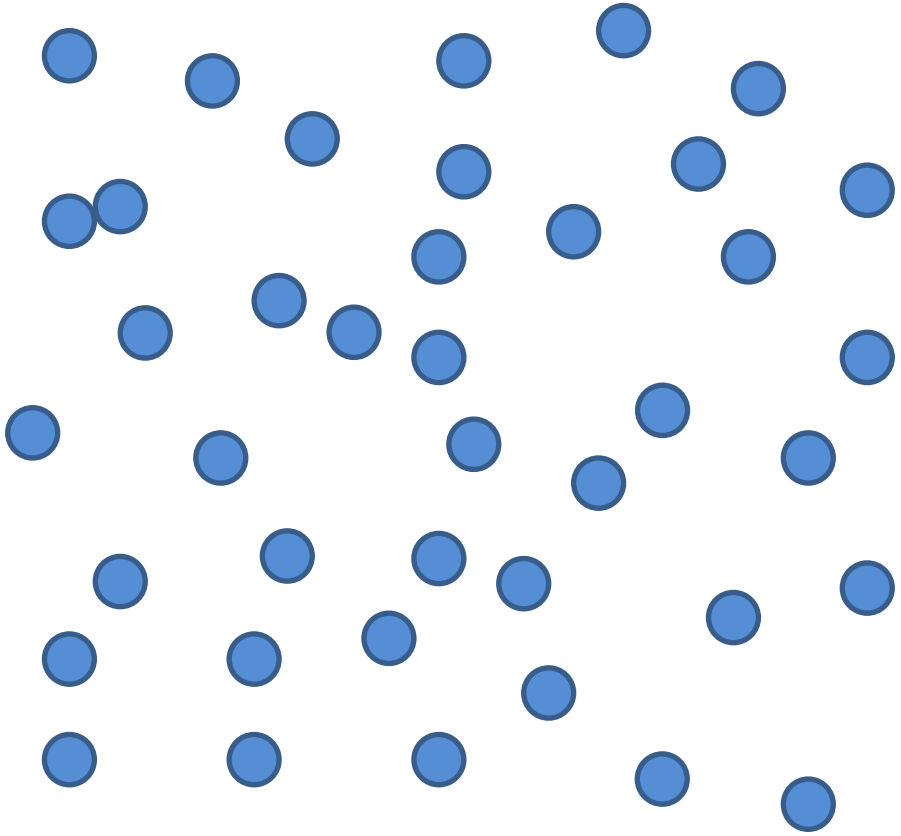
PERCEPTION PRÉ-ATTENTIVE

Perception pré-attentive

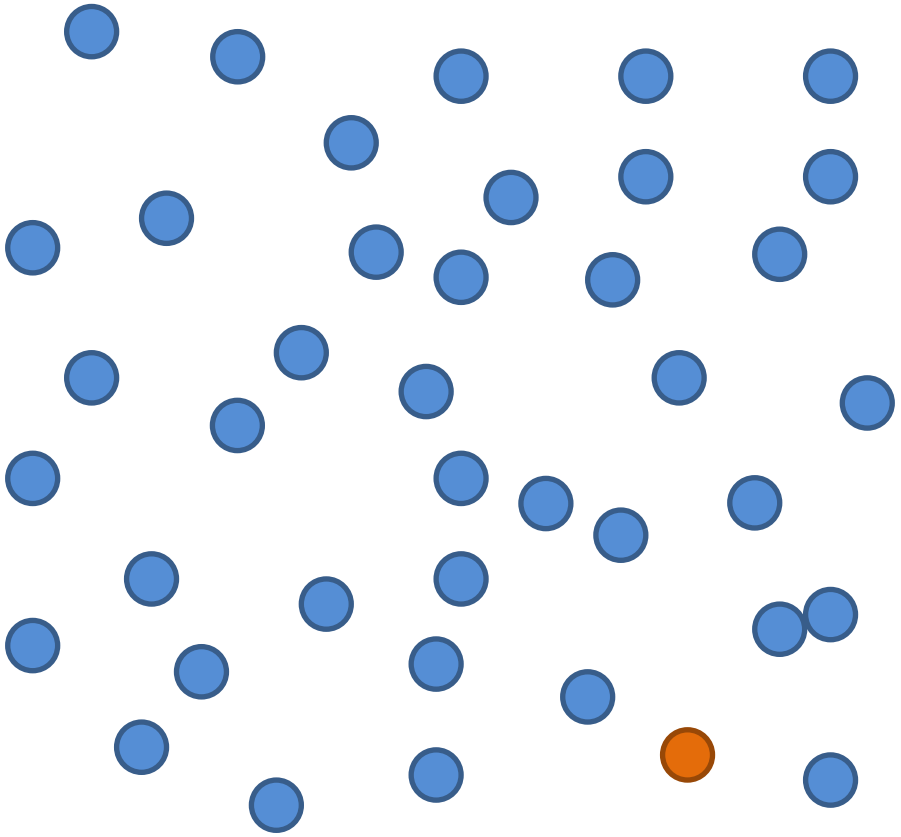
- Certains éléments sont identifiés « *instantanément* »
- Trouvez le 3 dans la série :

6546496874540668654687845487684846864
8525768768468989789755468456862135768
7861687196876868178676786876876714546

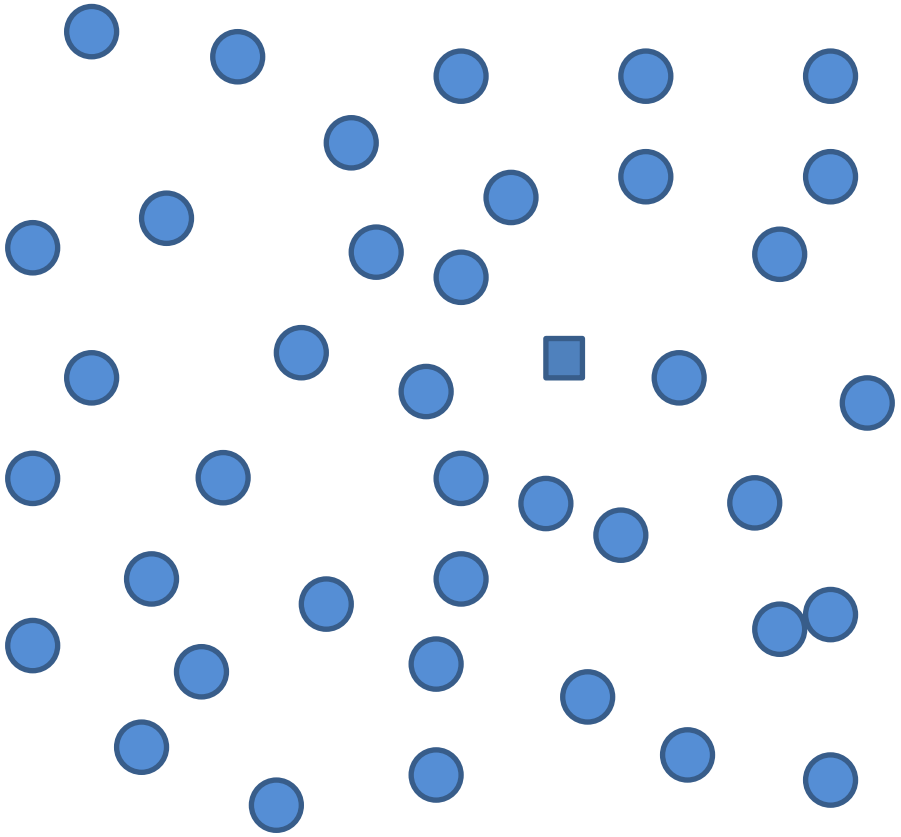
Perception pré-attentive : Couleur



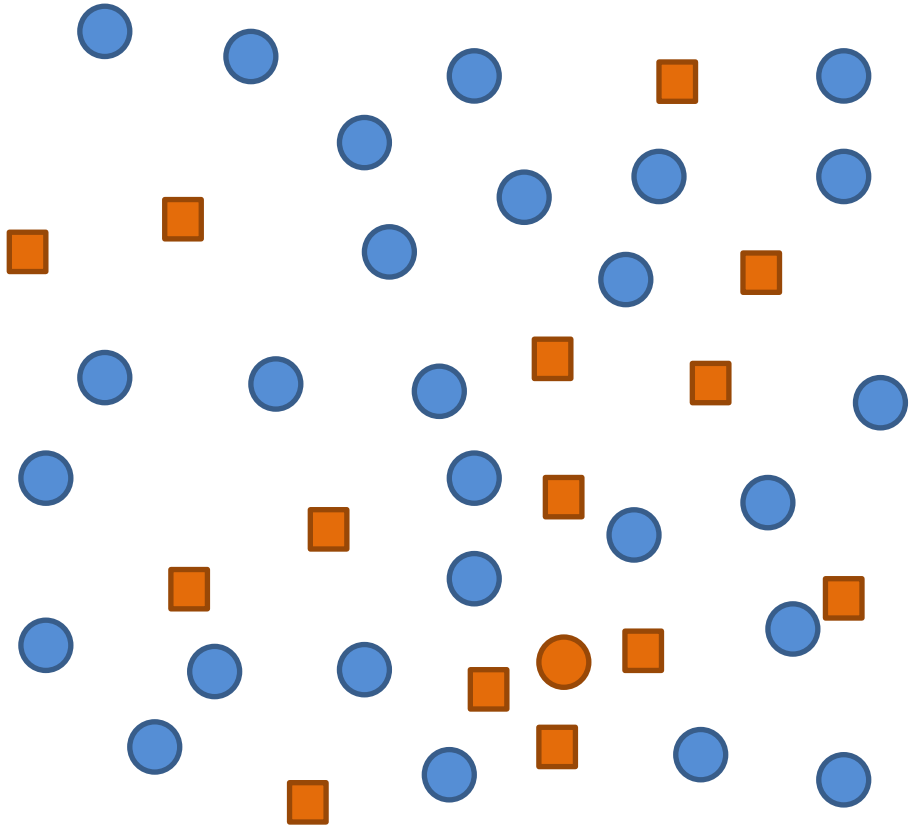
Perception pré-attentive : Couleur



Perception pré-attentive : Forme



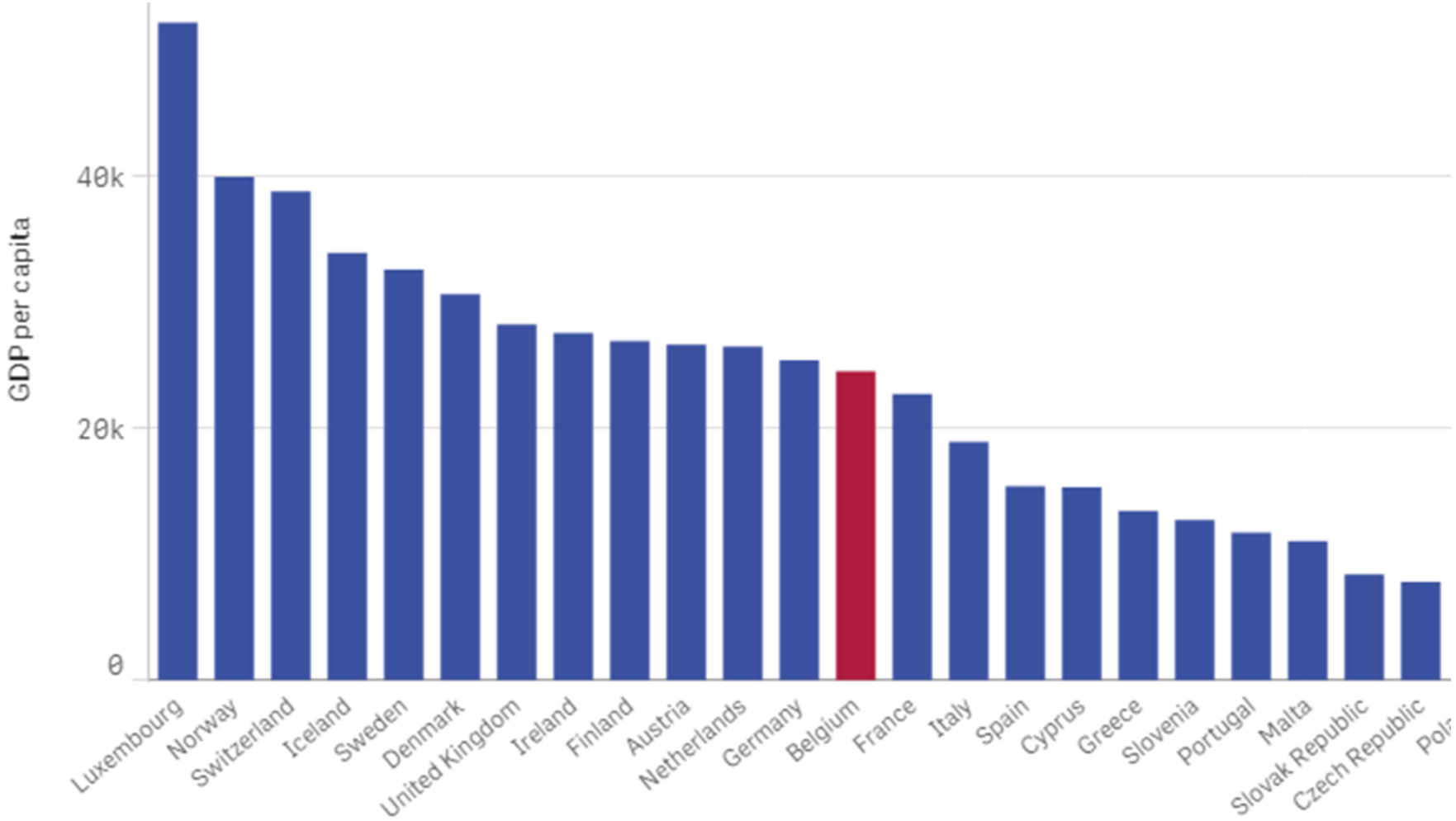
Perception pré-attentive : Combinaison



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

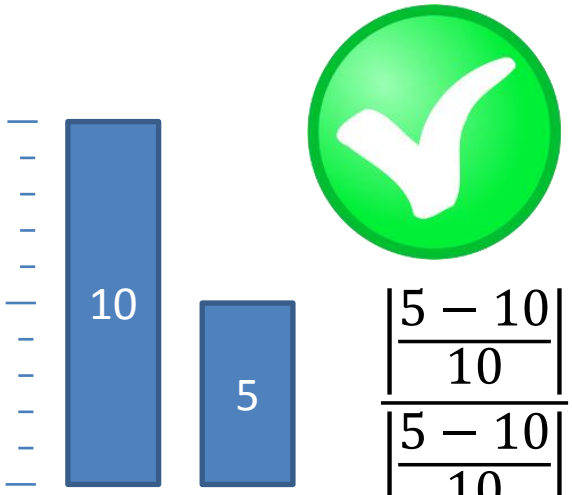
- Data-ink factor :

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

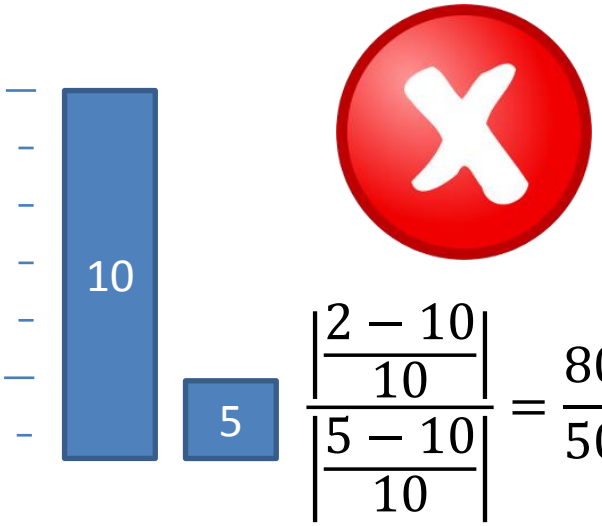
- Data density :

  À quel point l'espace est utilisé efficacement

Lie Factor



$$\frac{\left|\frac{5-10}{10}\right|}{\left|\frac{5-10}{10}\right|} = \frac{50\%}{50\%} = 1$$

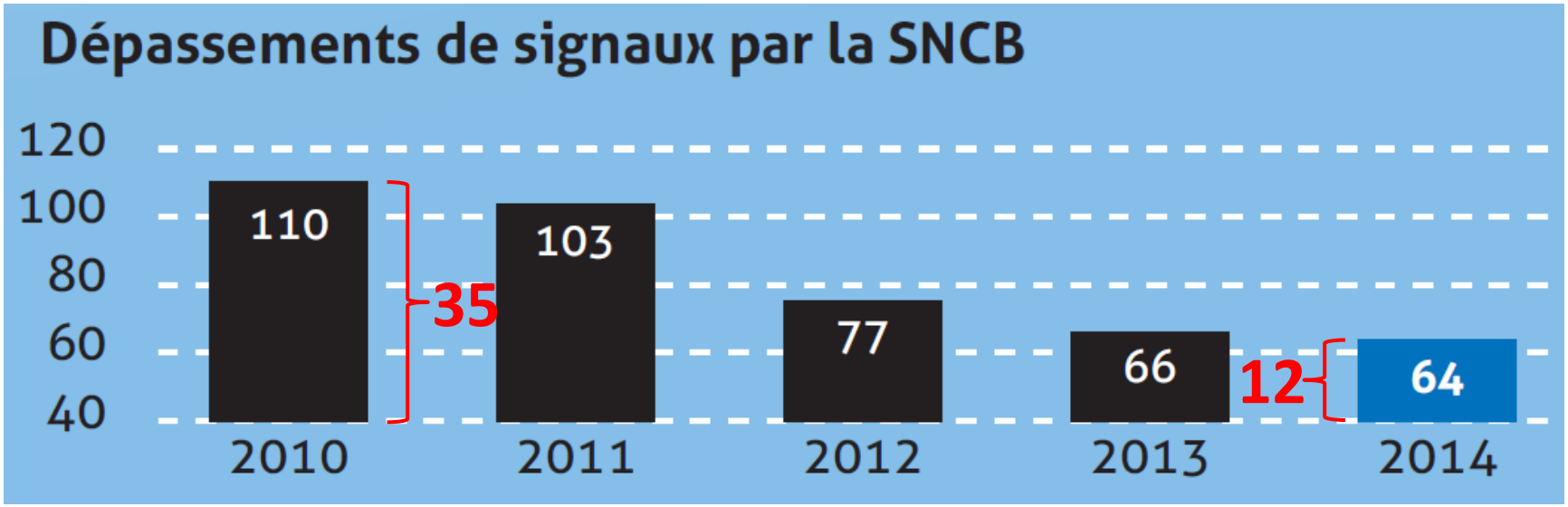


$$\frac{\left|\frac{2-10}{10}\right|}{\left|\frac{5-10}{10}\right|} = \frac{80\%}{50\%} = 1{,}6$$

$$\text{Lie factor} = \frac{\text{size of effect in graphic}}{\text{size of effect in data}}$$

$$\text{Size of effect} = \left|\frac{\text{2d value} - \text{1st value}}{\text{1st value}}\right|$$

Lie Factor : Exemple



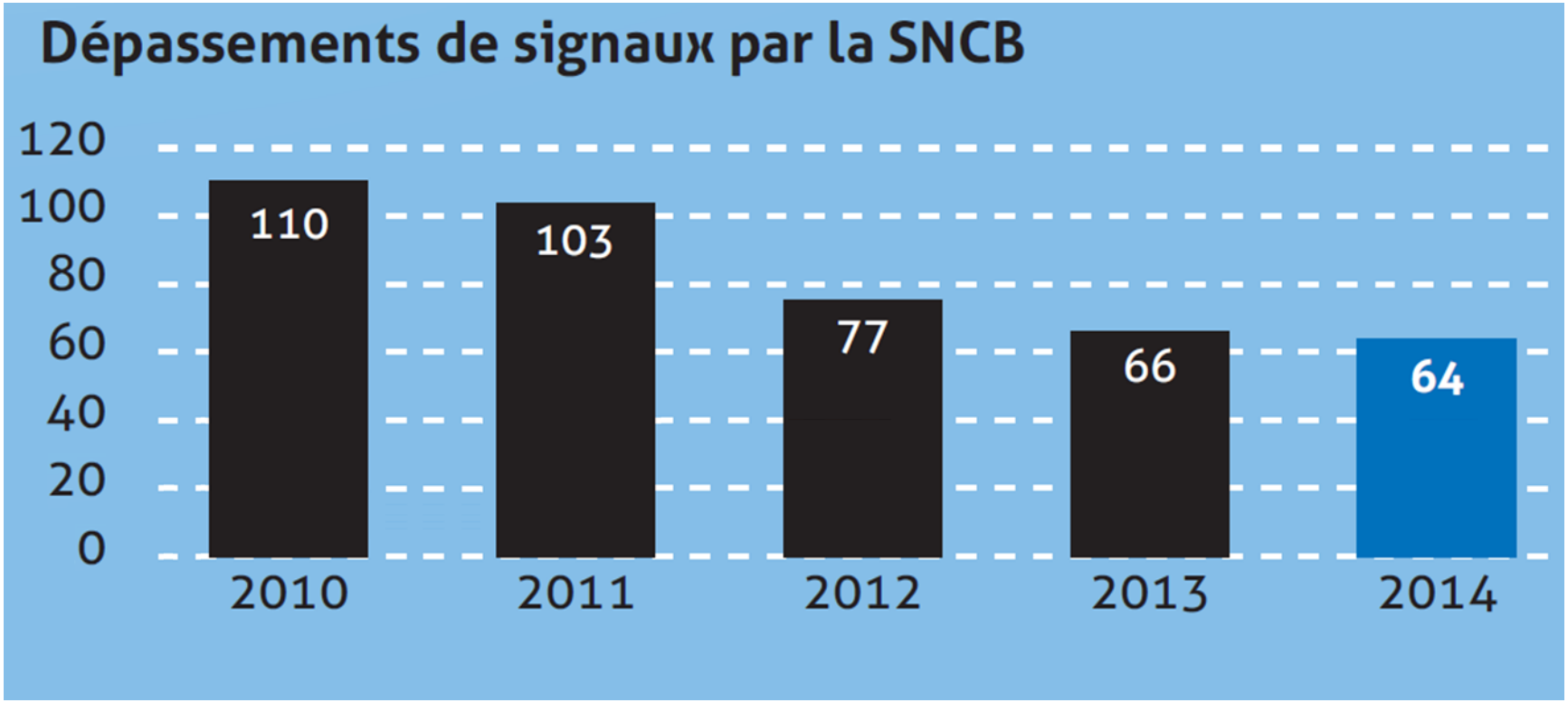
$$\text{Data} = \left|\frac{64-110}{110}\right| = 41{,}8\%$$

$$\text{Graphic} = \left|\frac{12-35}{35}\right| = 65{,}7\%$$

$$\text{Lie Factor} = \frac{65{,}7\%}{41{,}8\%} = 1{,}57$$

Sncb.be

Lie Factor : Exemple

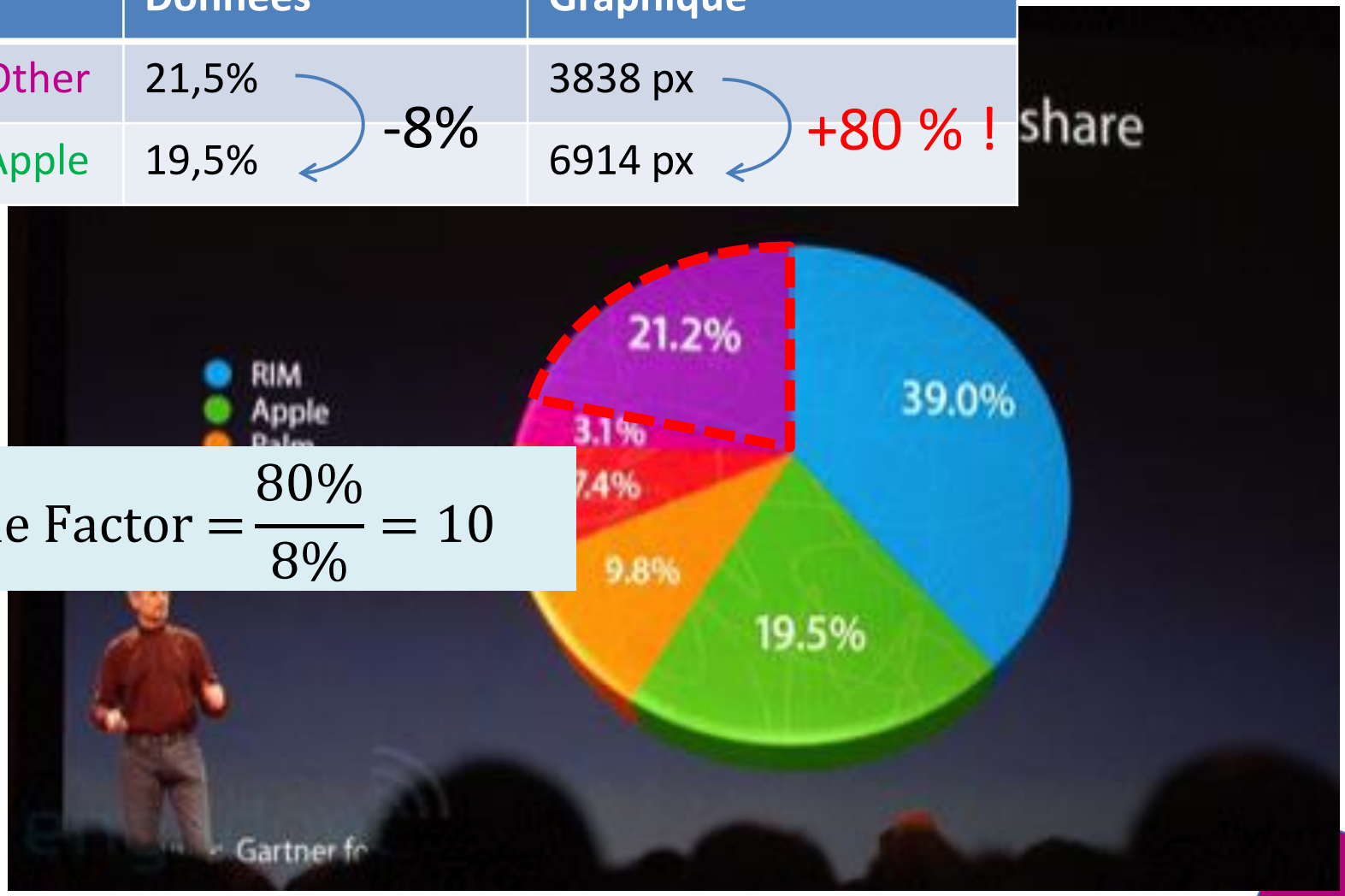


Lie factor : Exemple

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

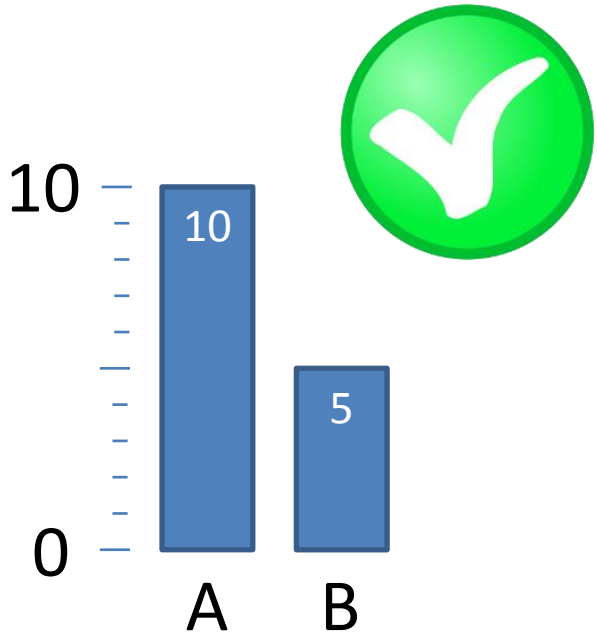
-8%

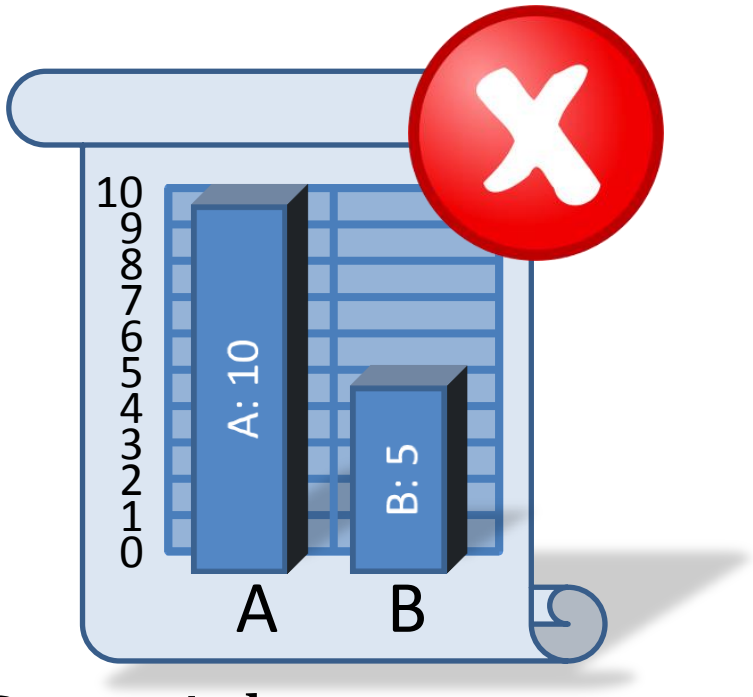
+80 % !



$$\text{Lie Factor} = \frac{80\%}{8\%} = 10$$

Data-ink ratio



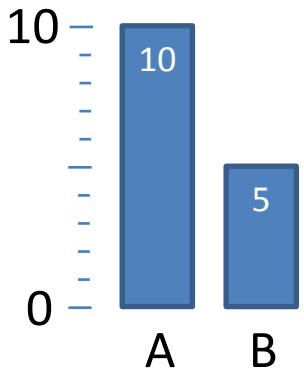


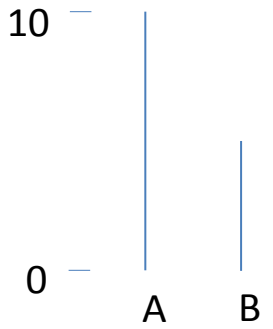
$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink in the graphic}}$$

$$= 1.0 - \text{proportion of a graphic that can be erased without loss of data-information}$$

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 !

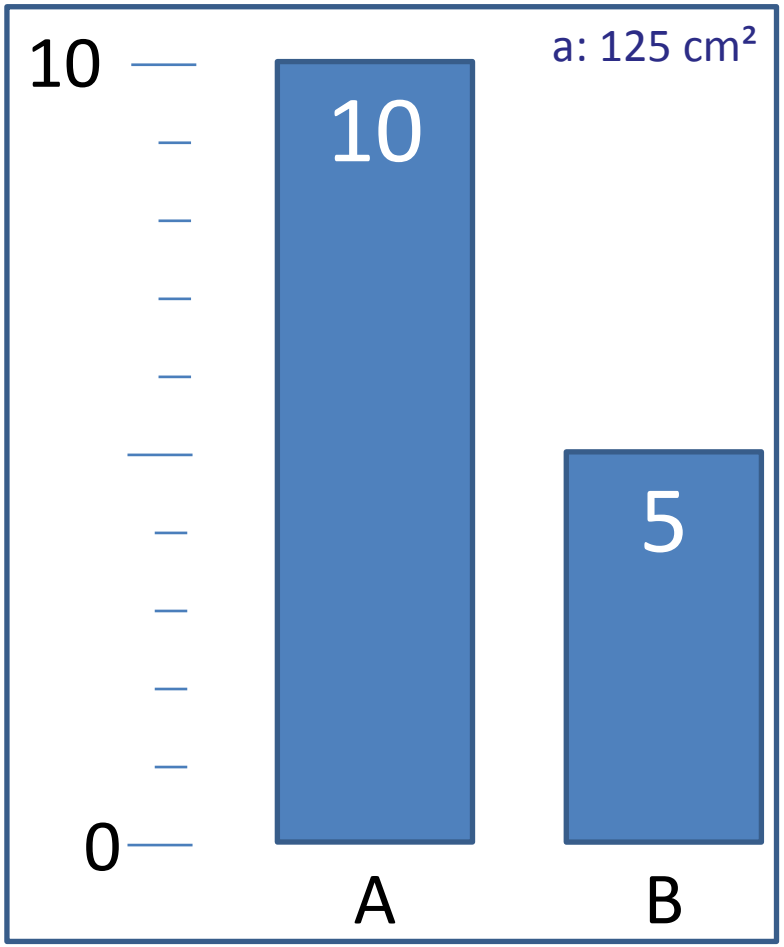


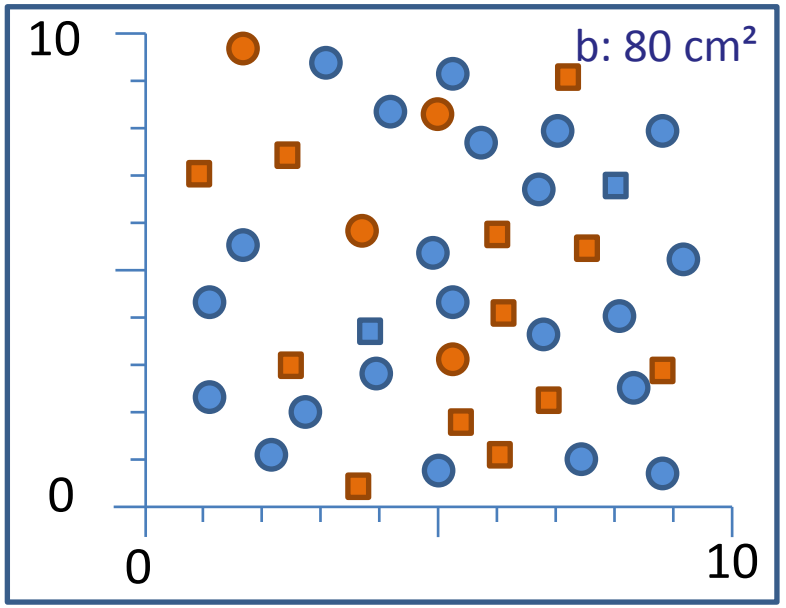


« Above else, show the data »

E. Tufte

Data density





$$\text{Data density} = \frac{\#\text{ entries in the data matrix}}{\text{Area of data graphic}}$$

Data density

a: 125 cm²

Pt	val
A	10
B	5

b: 80 cm²

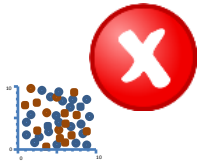
Pt	x	y	coul	form
1	10	6	B	R
2	5	2	O	R
3	8	5	O	C
…	…	…	…	…
40	1	1	B	R

$$\mathrm{DD}_a = \frac{2}{125} = 0{,}016\ d/cm^2$$

$$\mathrm{DD}_b = \frac{40 \times 4}{80} = 2\ d/cm^2$$

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 ?

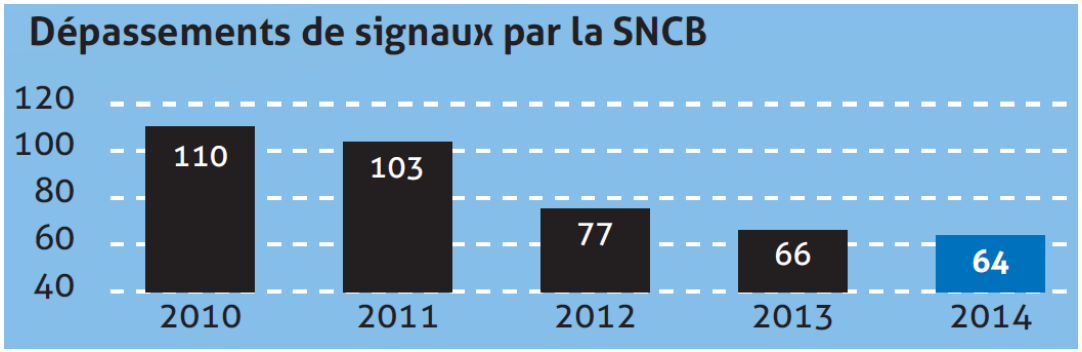
Mesures de qualité : en pratique

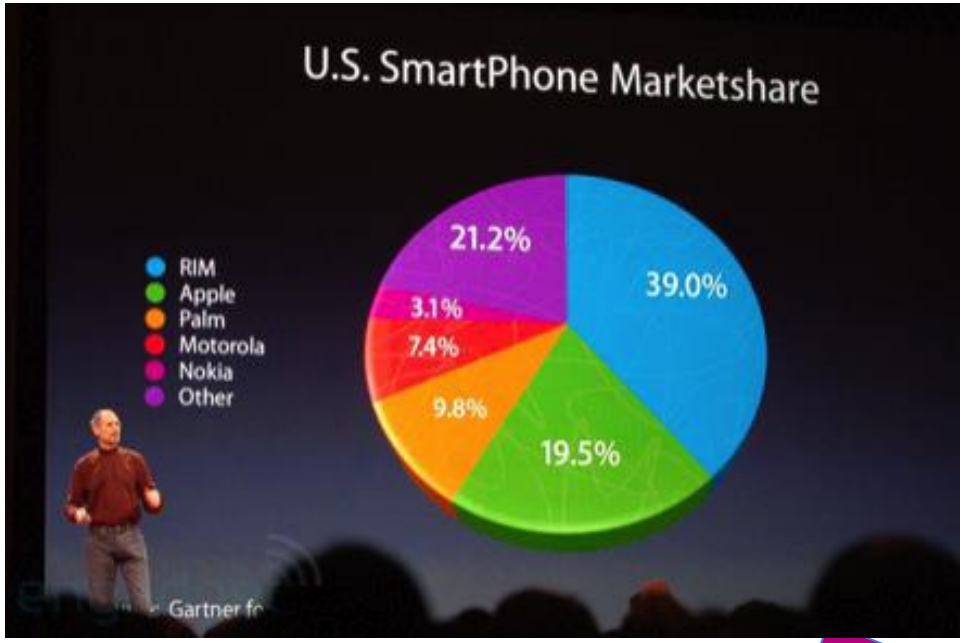
- Viser un « lie factor » de 1
- Minimiser le « data-ink » ratio avec raison :
  - Supprimer la « décoration » (3D, ombre, fond…)
  - Limiter la redondance
- Maximiser la « data-density » sans perdre en lisibilité

Les fondamentaux

ERREURS & MANIPULATIONS

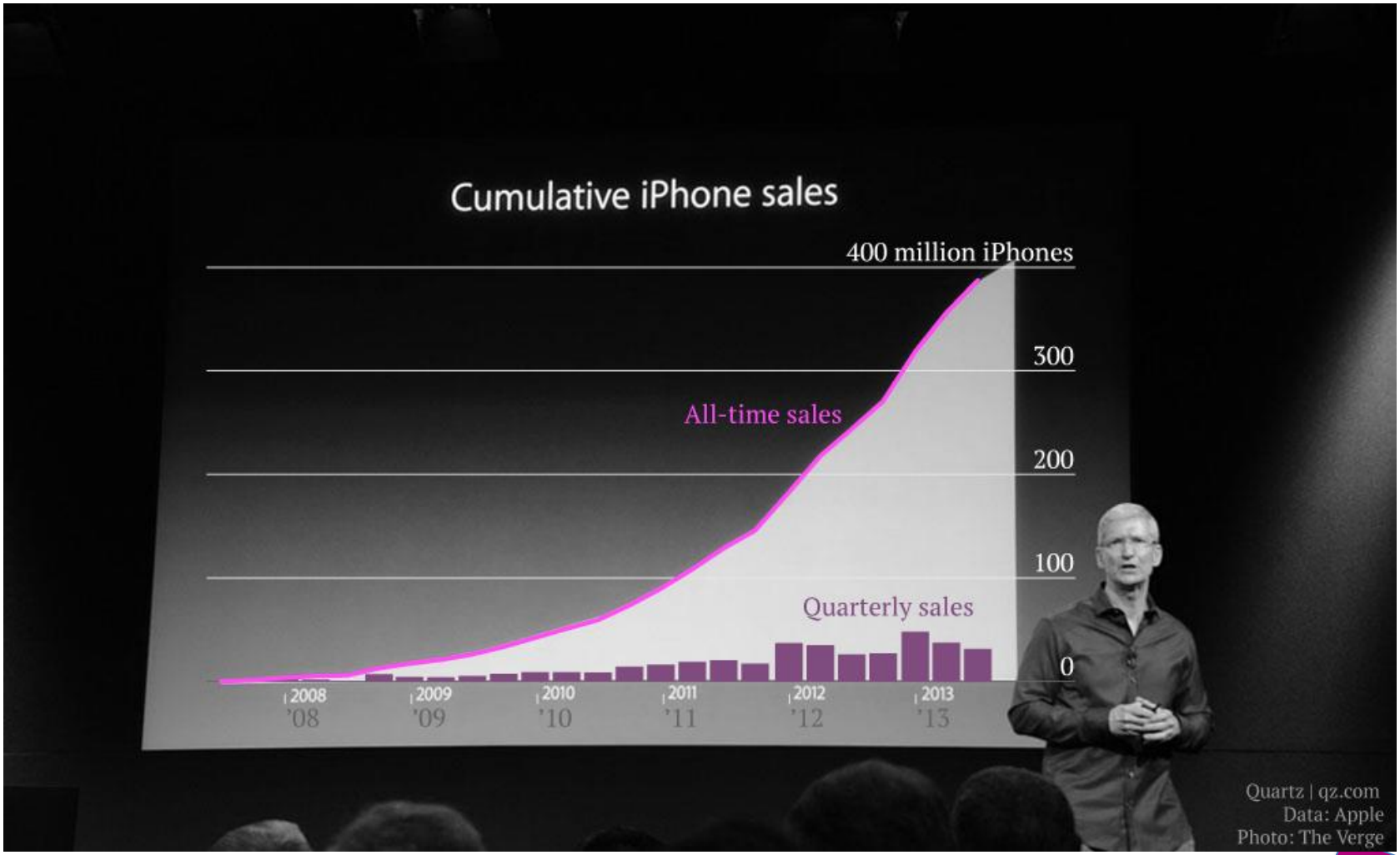
Tronquer un axe, utiliser la perspective



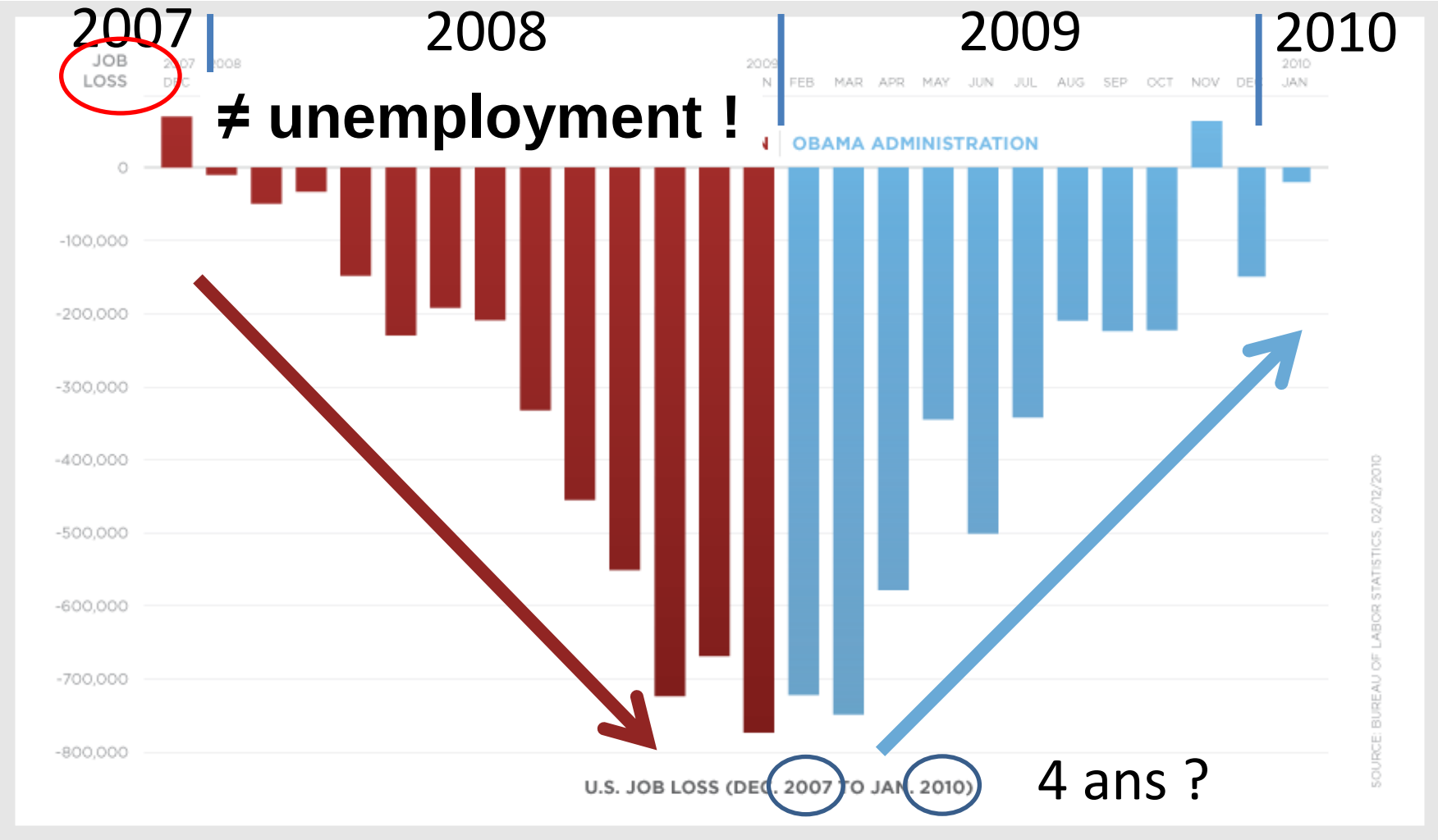


Valeurs cumulatives

Valeurs cumulatives

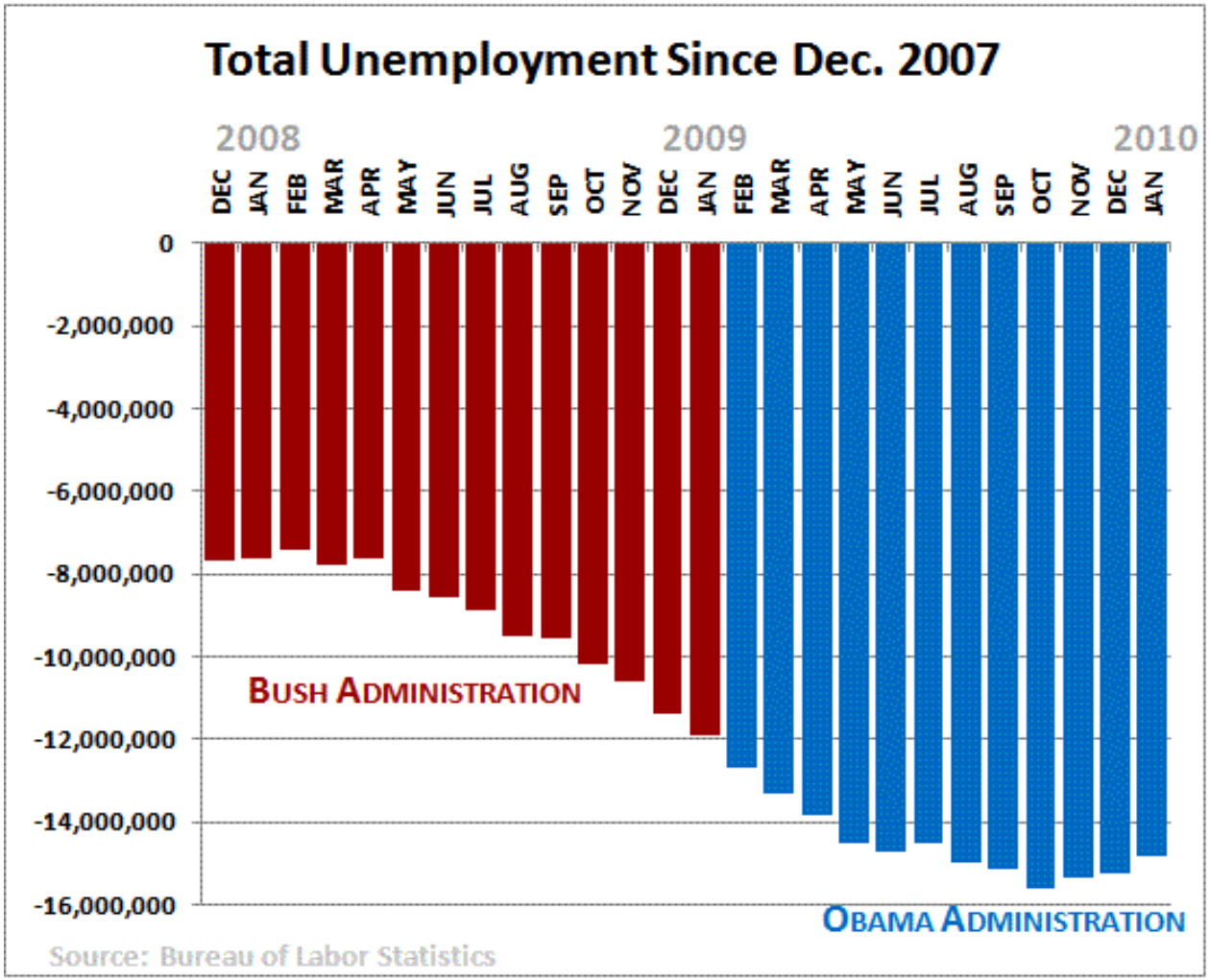


Le contraire

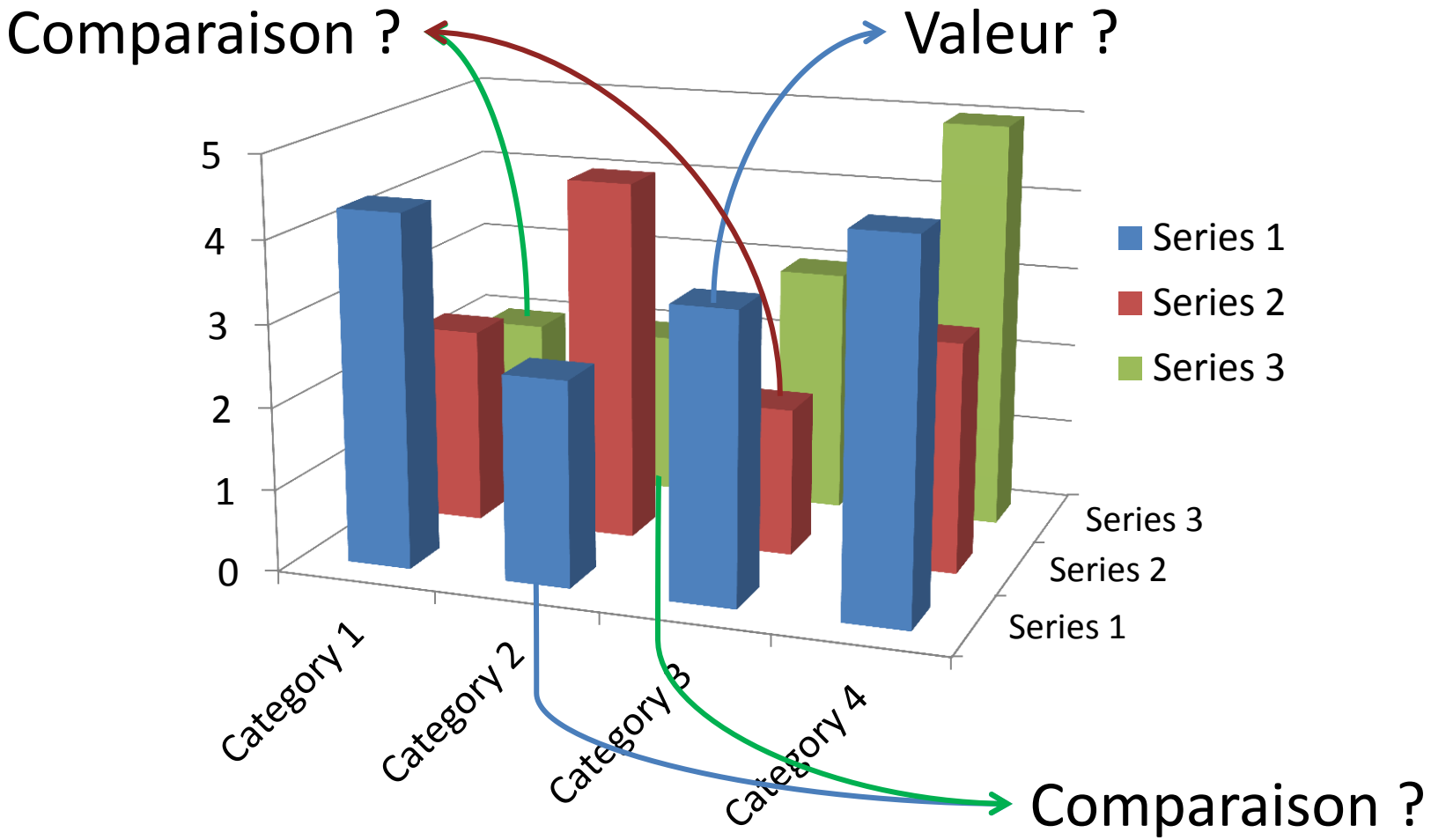


Luminance : **Bush : 57%** **Obama: 84%**

Le contraire



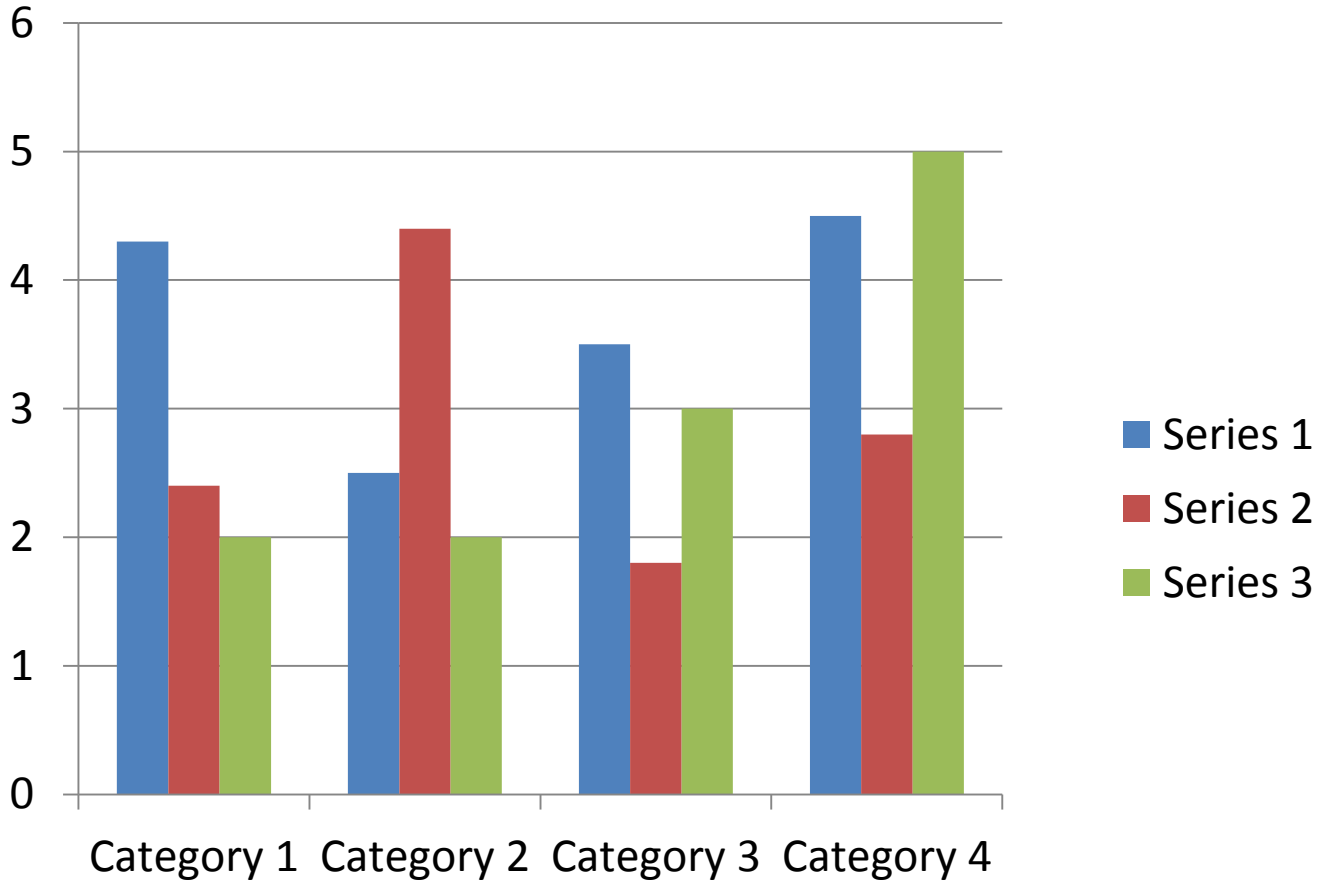
Perspectives & 3D



Excel, « Histogramme 3D » par défaut

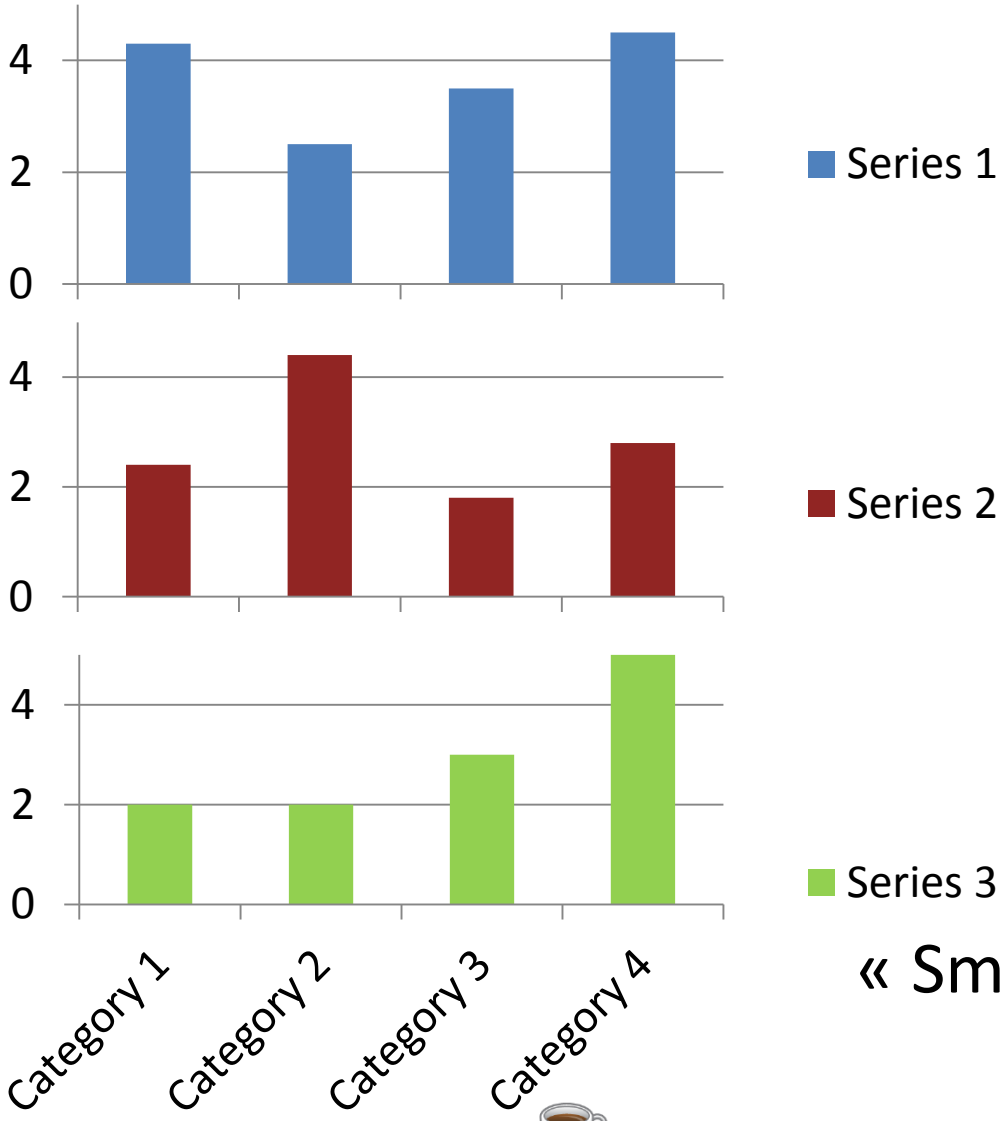
Outil : Excel

Perspectives & 3D



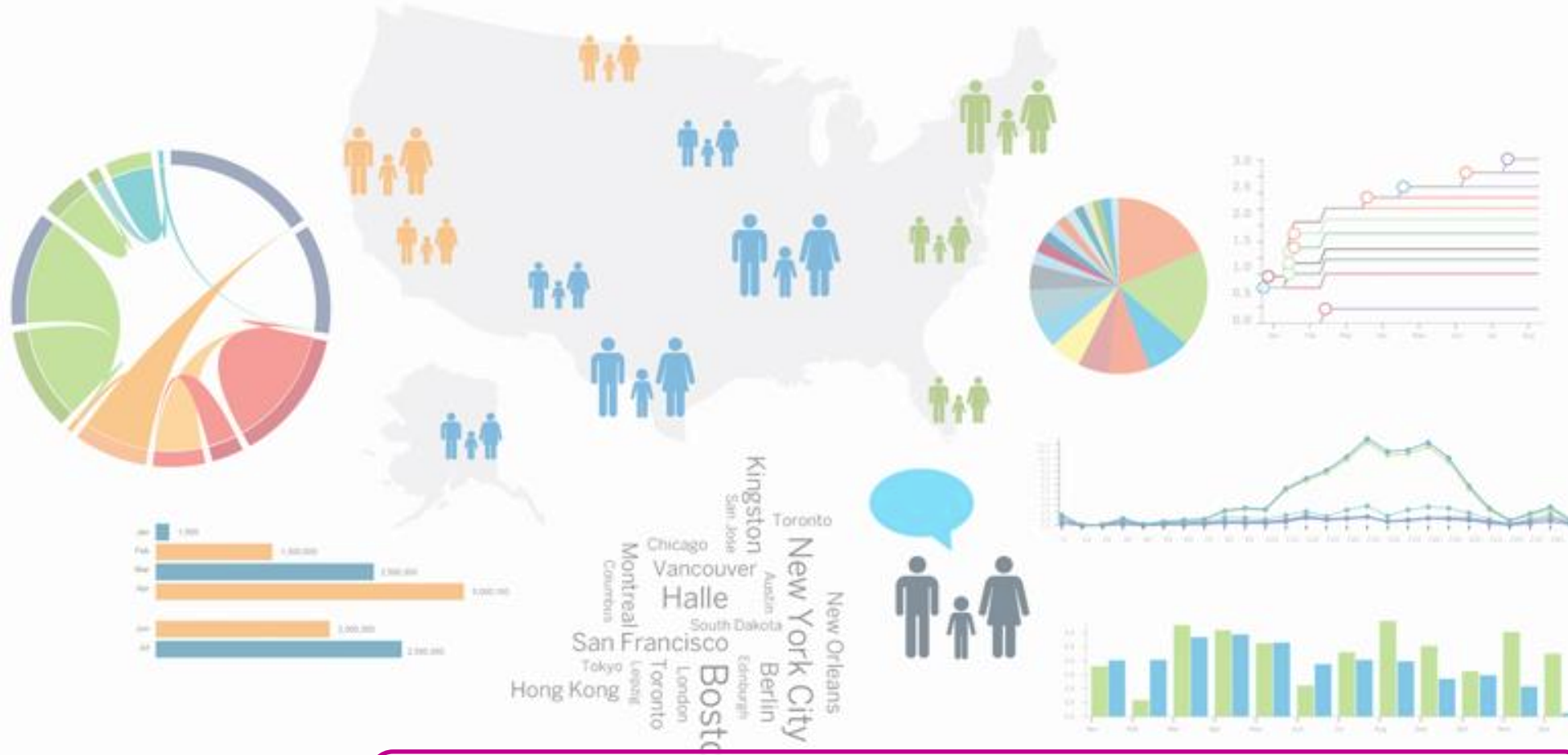
Outil : Excel

Perspectives & 3D



« Small multiples »

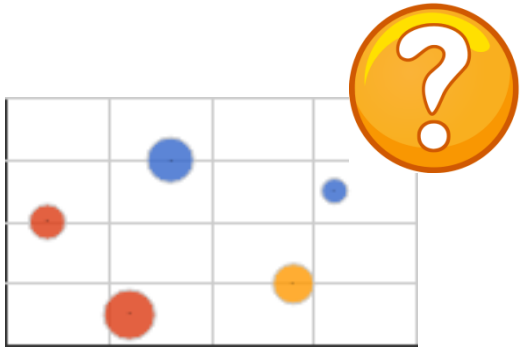
Outil : Excel



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

Outil : Google Charts

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 ?

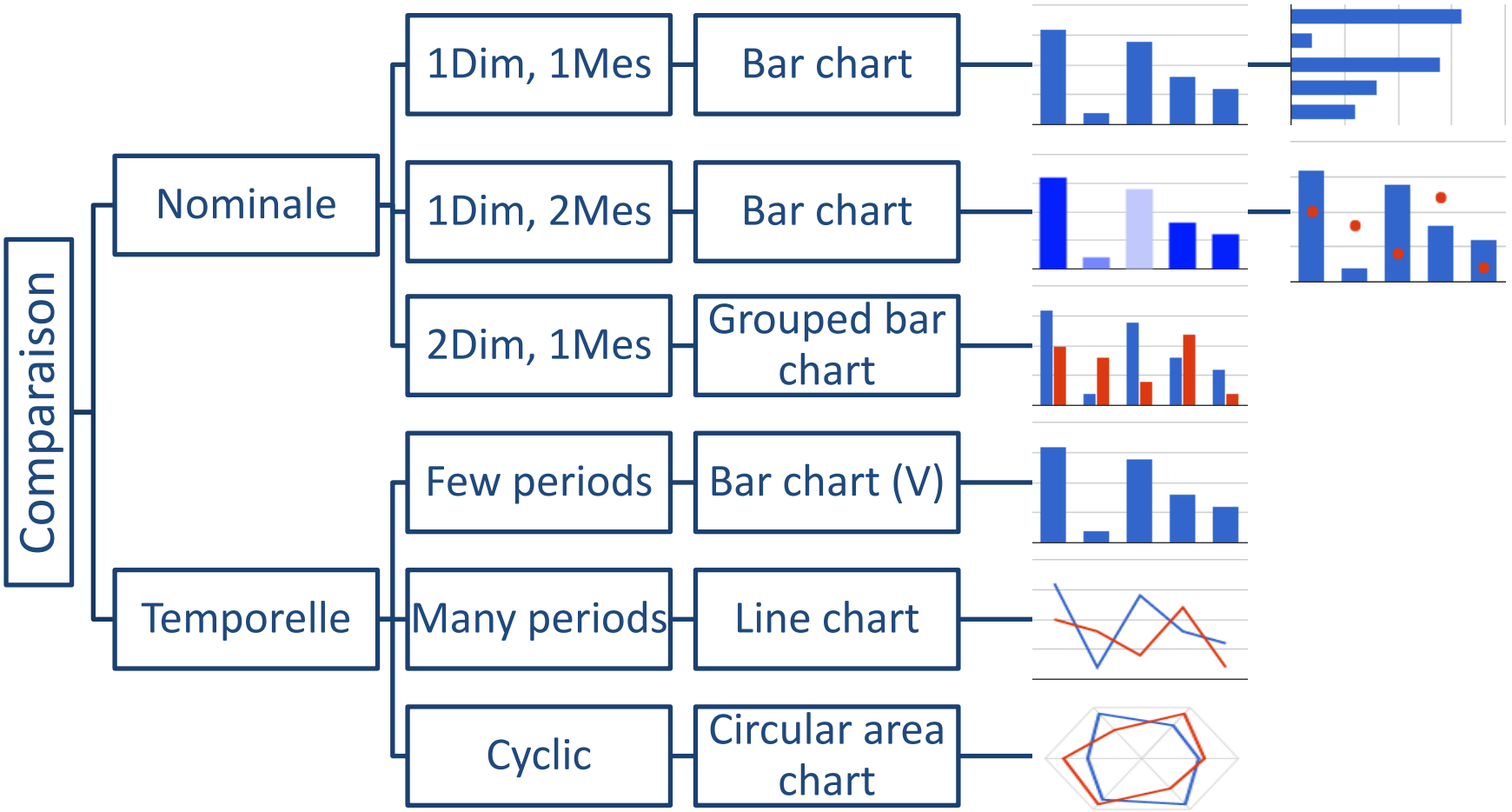
Corrélation
Quel est le lien entre l'âge et répartition des genres ? La nationalité et le secteur ?

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

Choix des graphiques

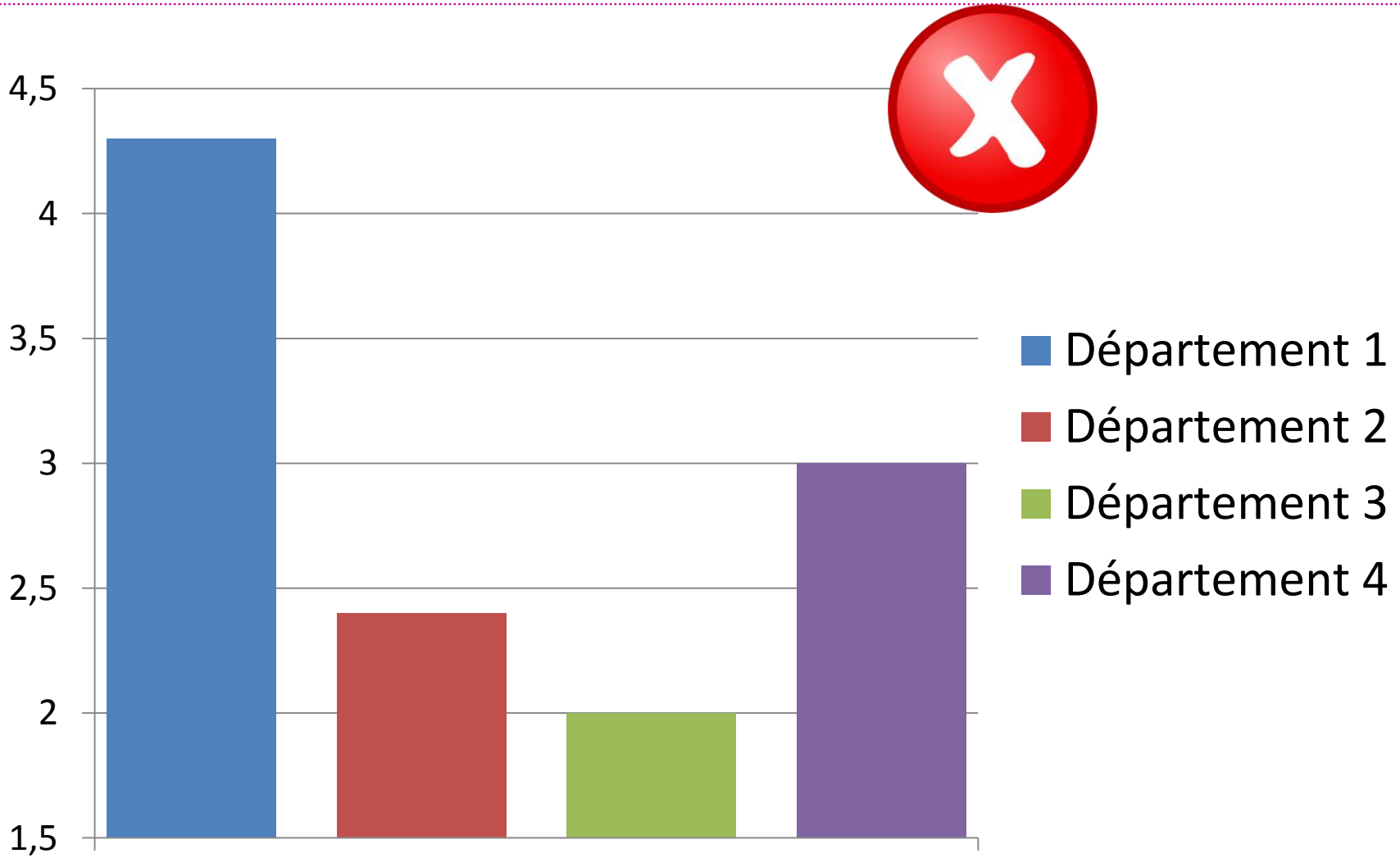
COMPARAISON

Comparaison



Outil : Google Charts, HighCharts

Bar chart



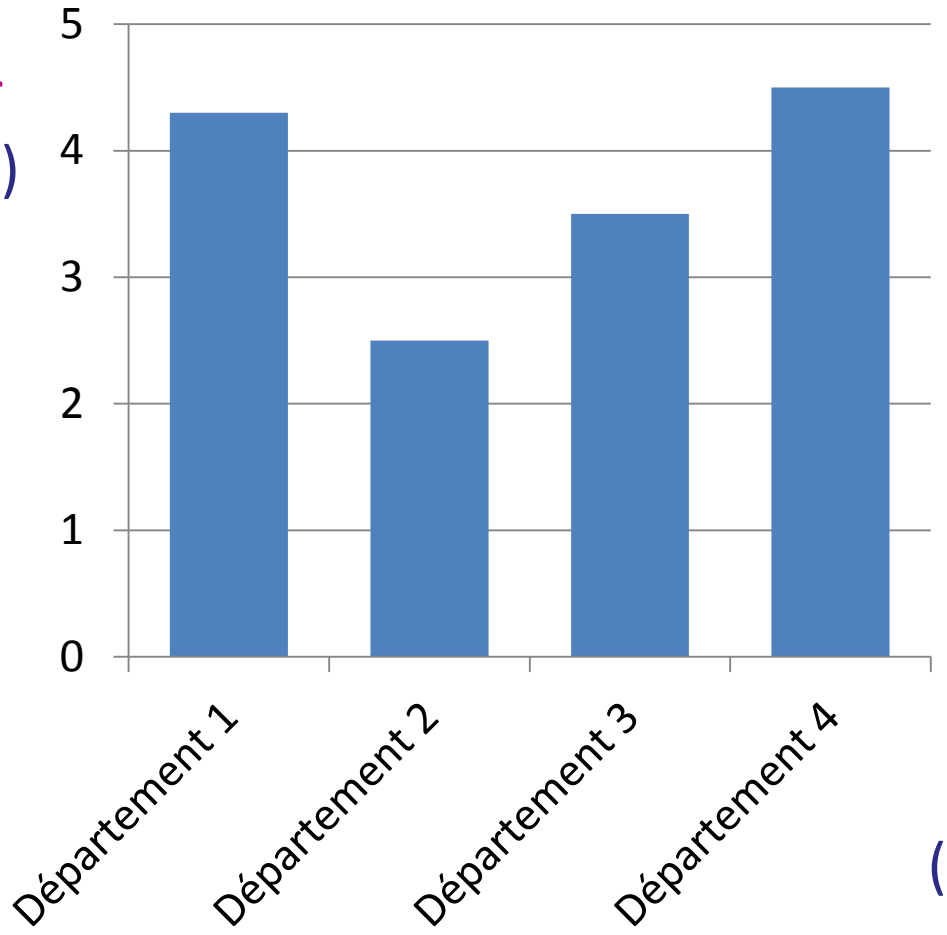
Outil : Excel

Bar chart

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

Var. quantitative ratio

Tronquer = tromper !



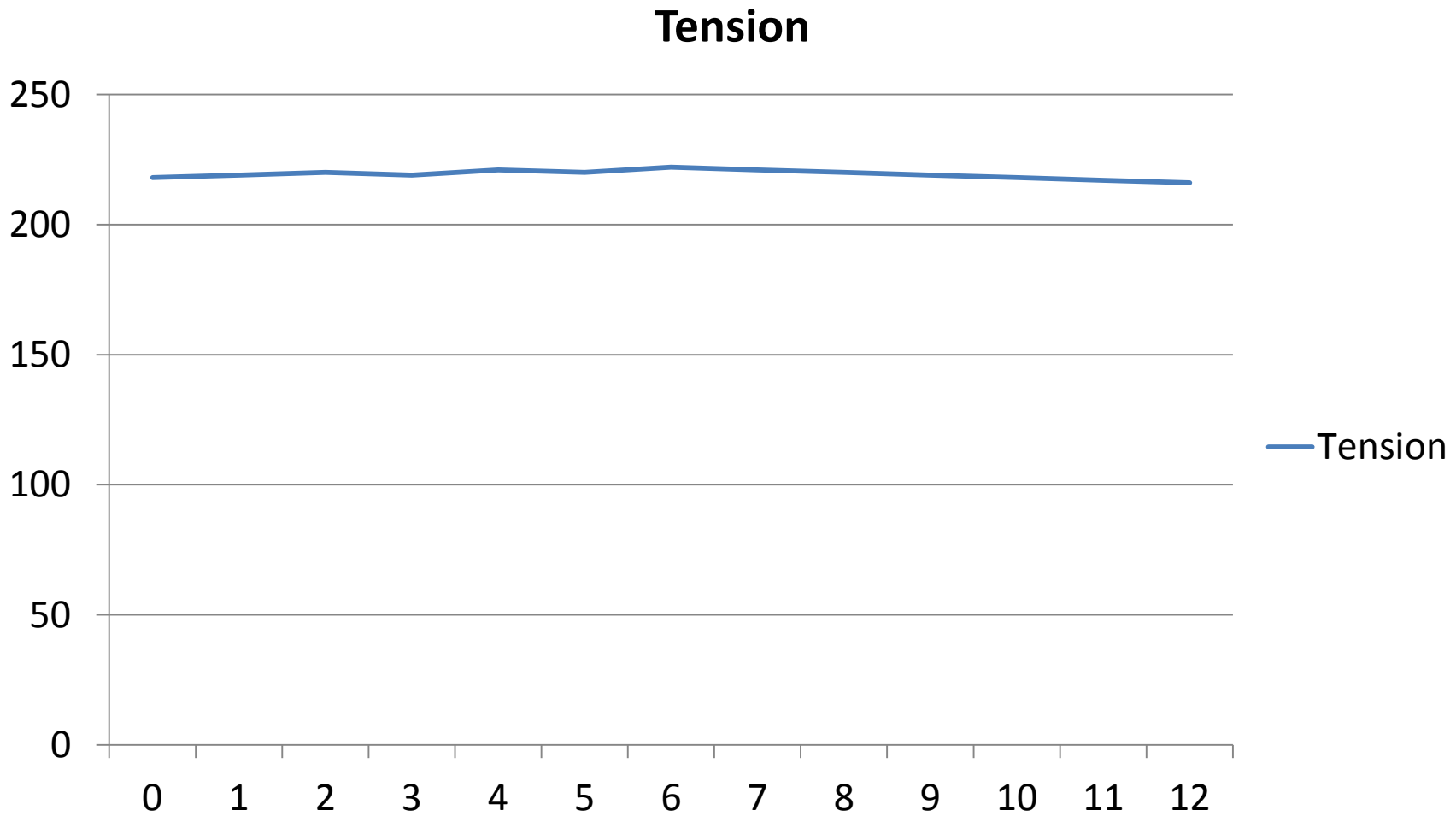
Variable qualitative (nominale ou ordinale)

Outil : Excel

Line chart

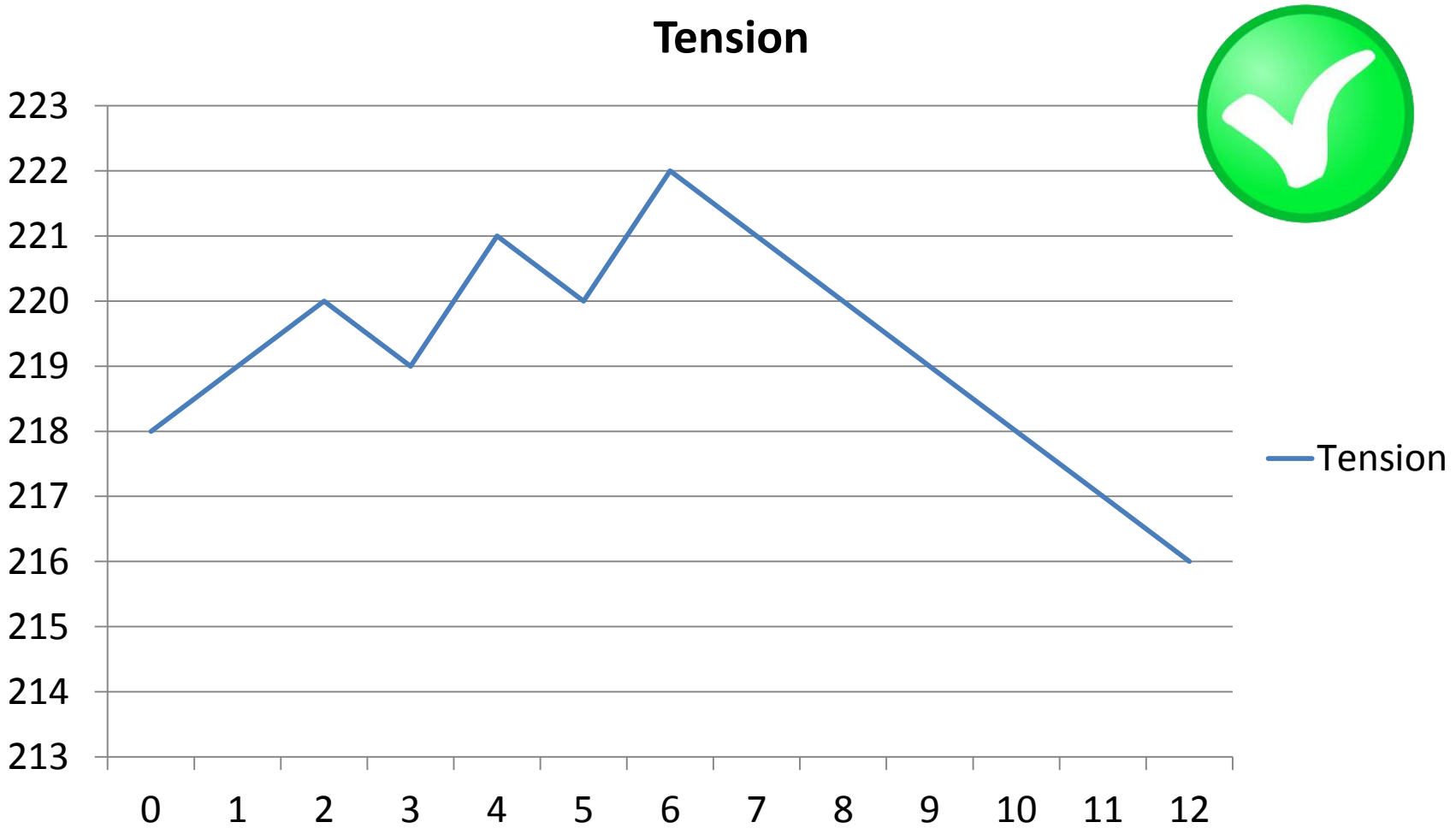
- Axe vertical :
  - Ce que l'œil perçoit, c'est la position des points
  - L'origine de l'axe n'a pas (peu) d'importance, peut être arbitraire… en étant « honnête »
- Axe horizontal : notion de continuité
  - Les valeurs intermédiaires ont un sens
  - Pas de variable nominale (rouge, vert, bleu…)
 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

Line chart

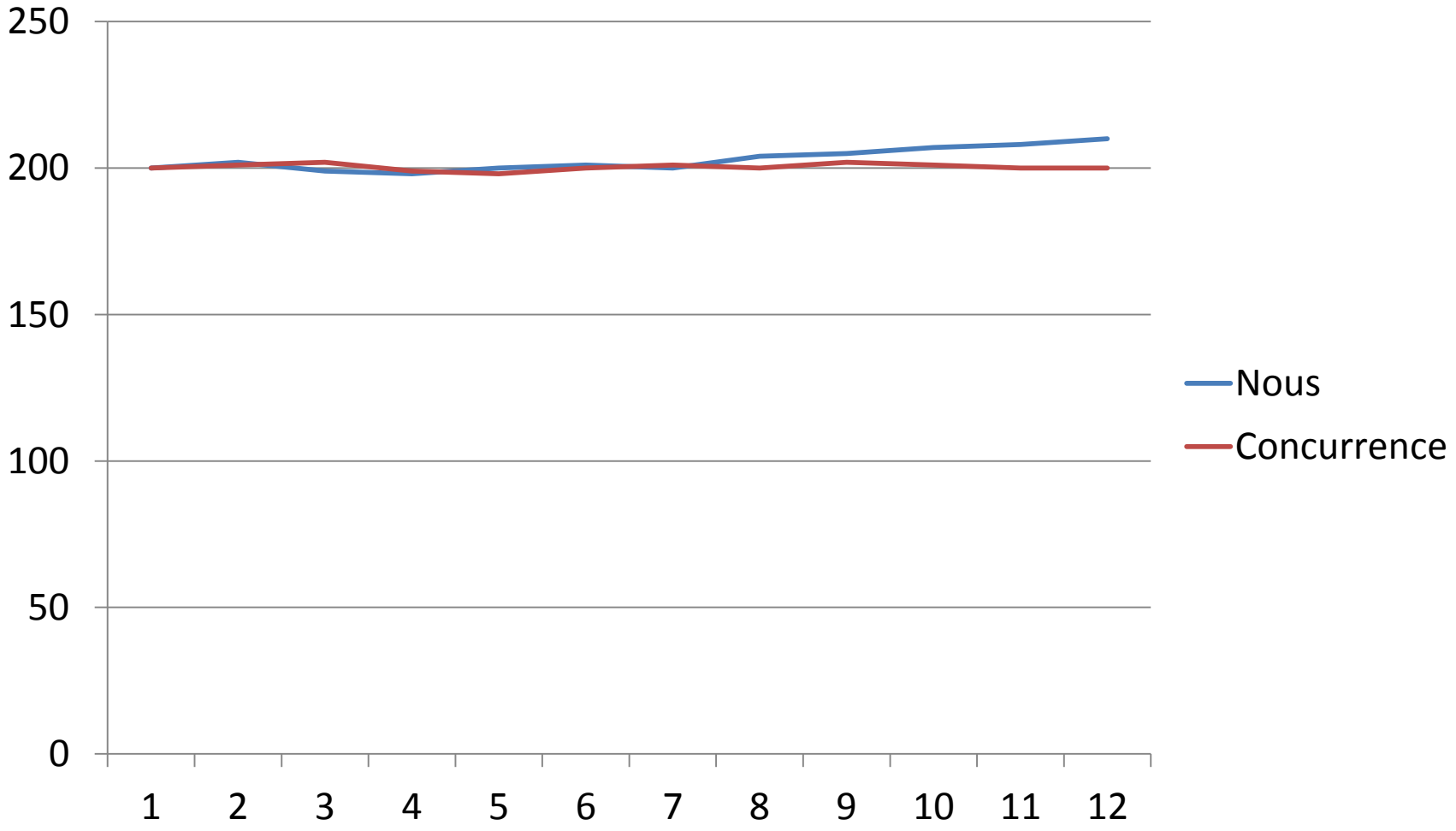


Outil : Excel

Line chart

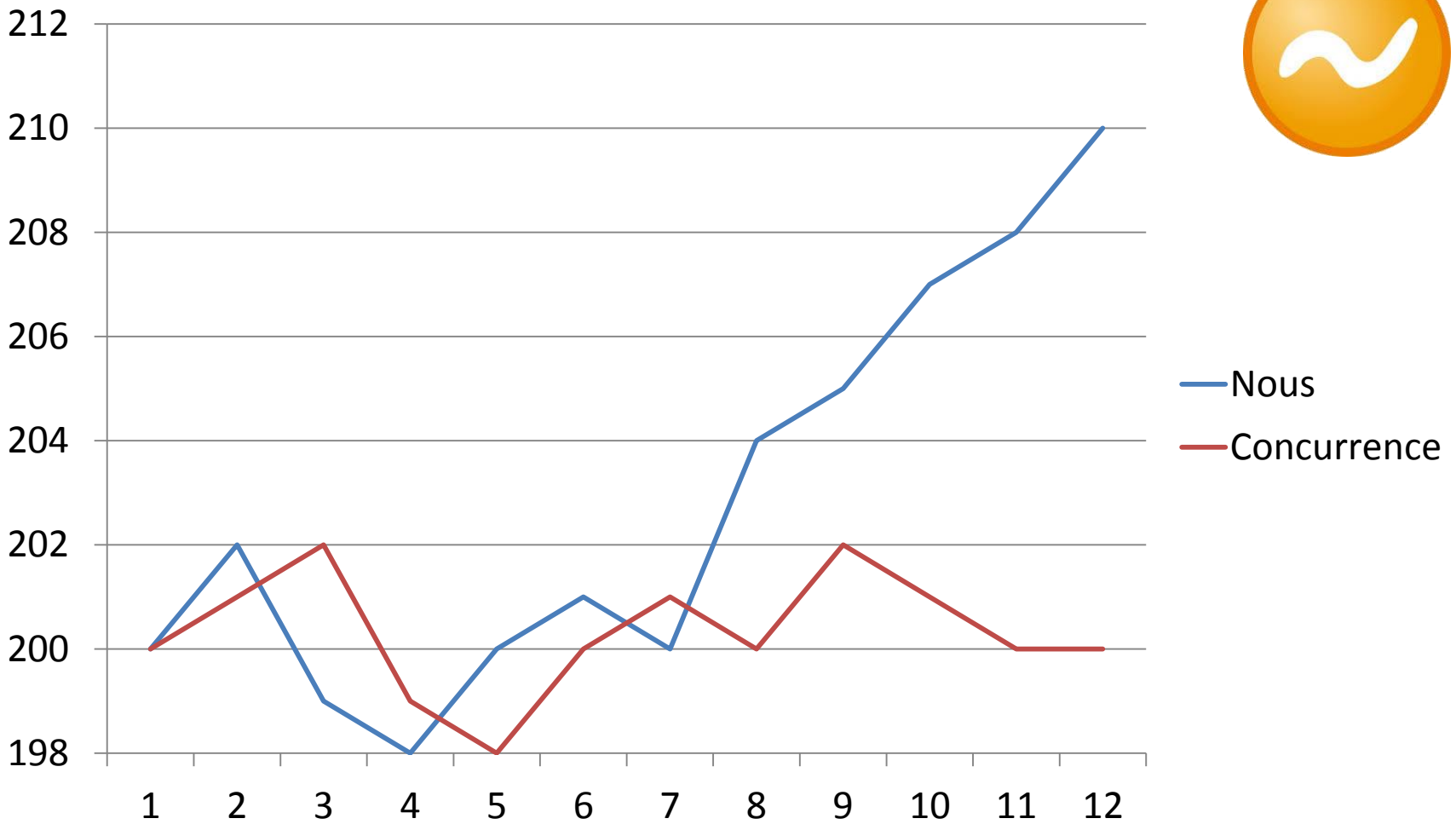

Tension
223
222
221
220
219
218
217
216
215
214
213
0
1
2
3
4
5
6
7
8
9
10
11
12
Tension

Line Chart



Outil : Excel

Line Chart

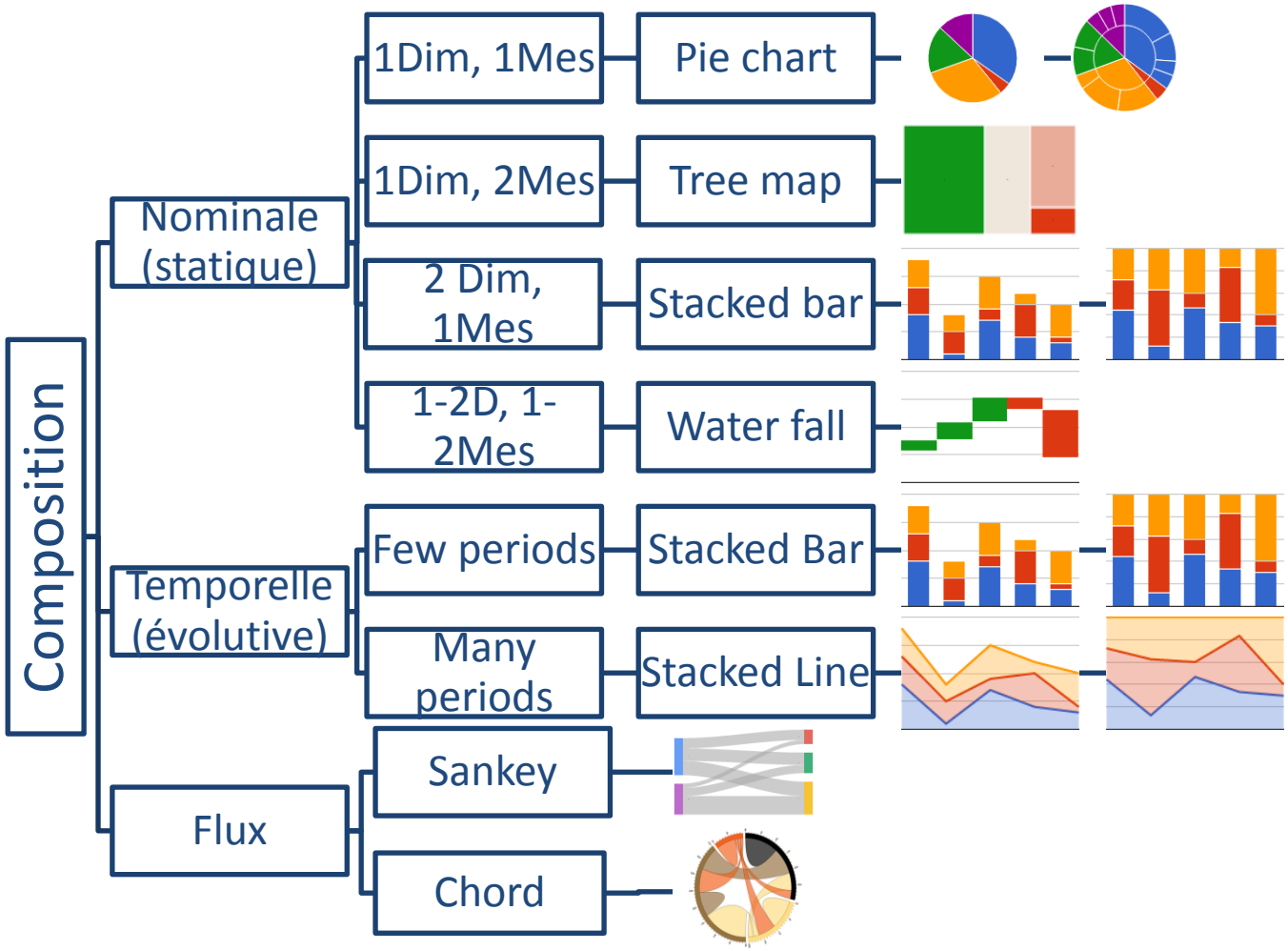


Outil : Excel

Choix des graphiques

COMPOSITION

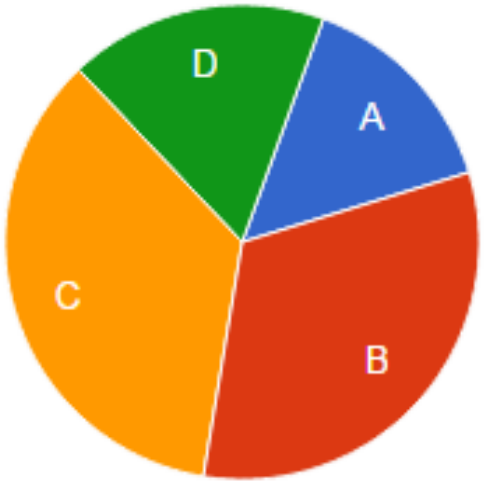
Composition

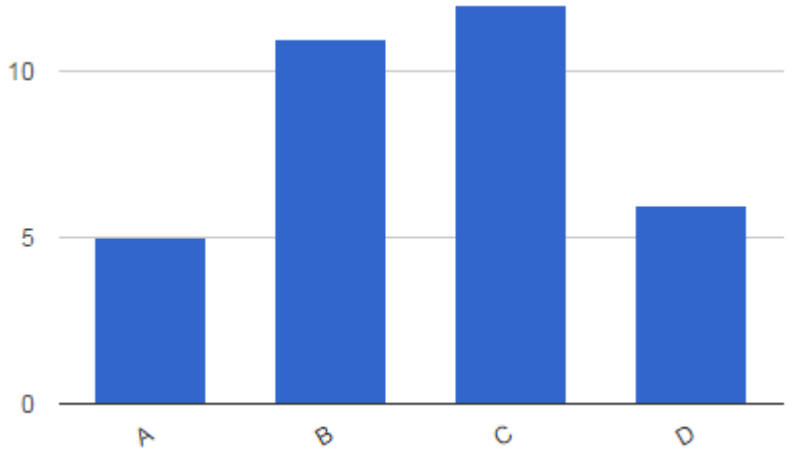


Outil : Google Charts, HighCharts

Pie Chart

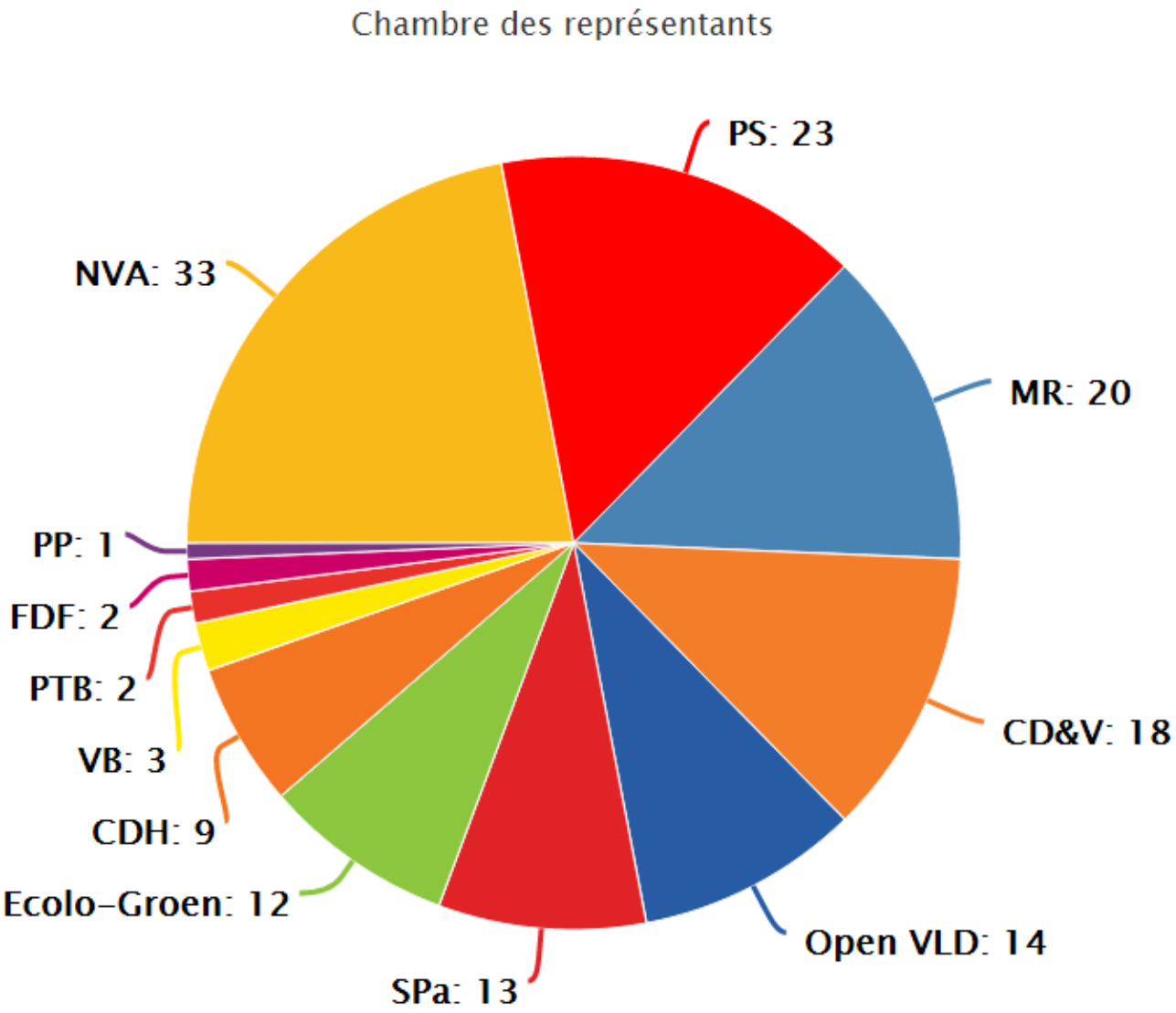
- Avantages :
  - Intuitif, facile à comprendre
  - Donne une bonne vue d'ensemble de la composition : D+A ≈ B ≈ C
- Inconvénients :
  - Peu précis ! A > D ? B > C ? A vs B
  - Peu adapté au-delà de 3-4 tranches





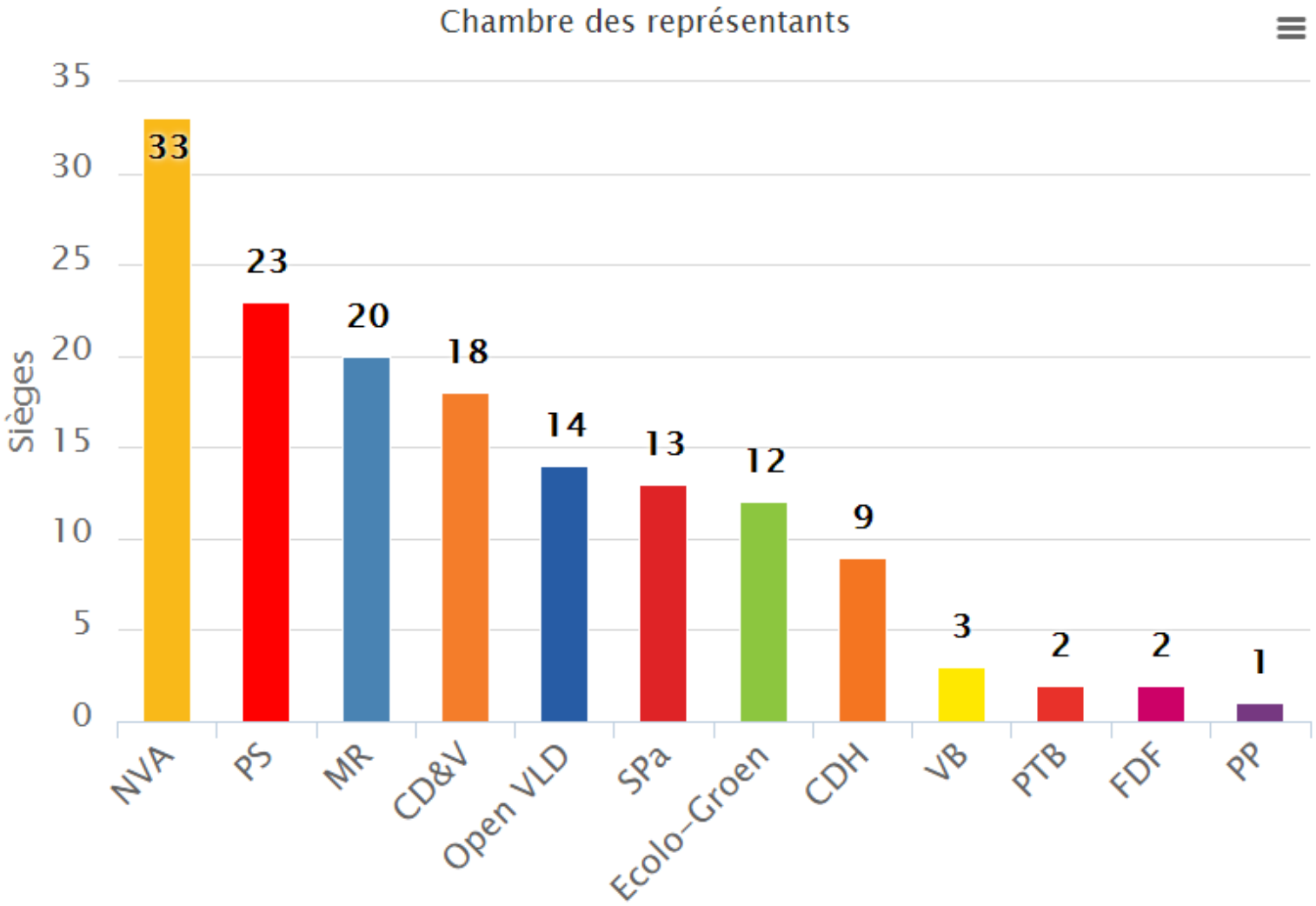
Outil : Google Charts

Pie Chart



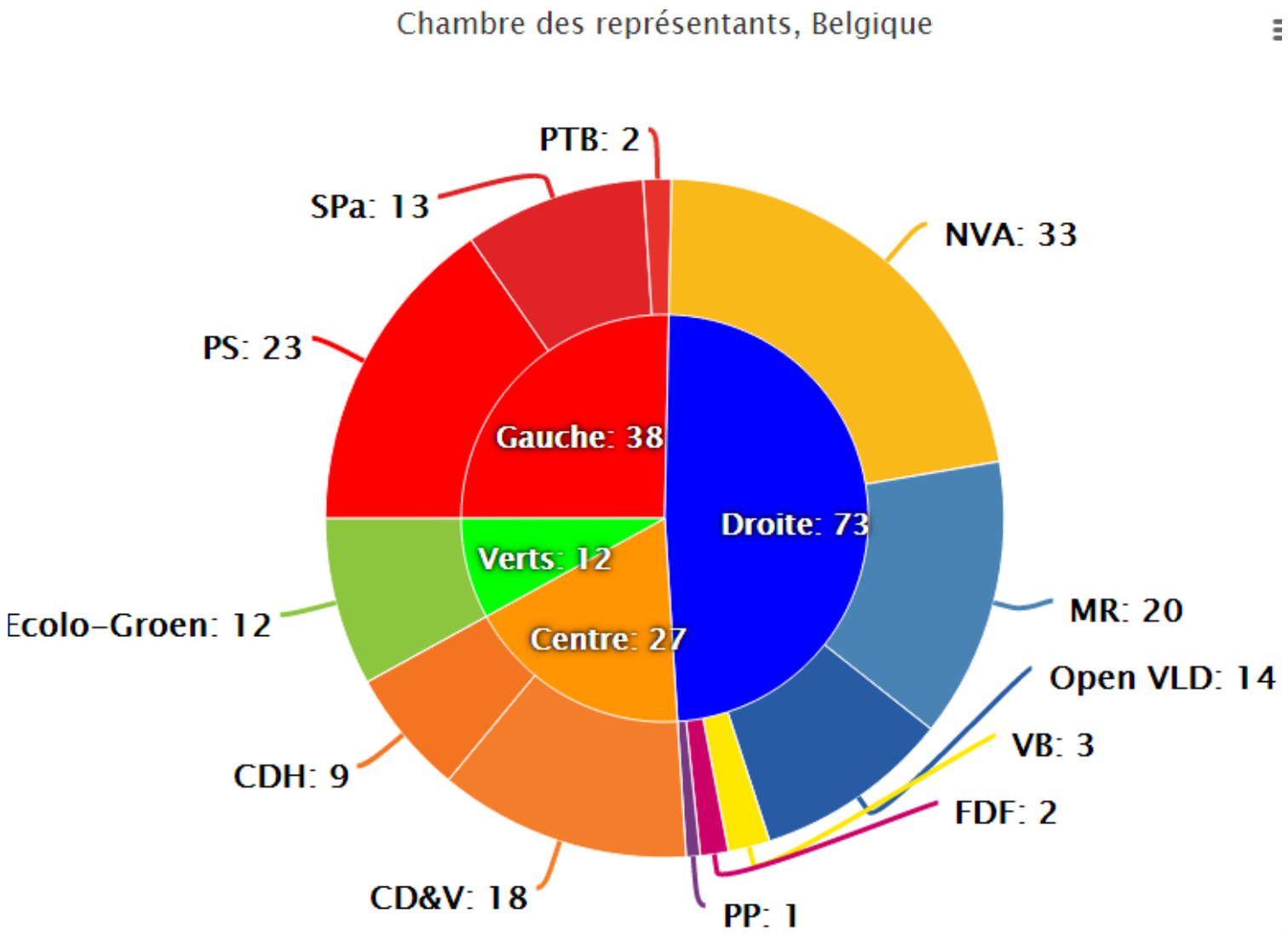
Outil : HighCharts

Pie Chart vs Bar Chart



Outil : HighCharts

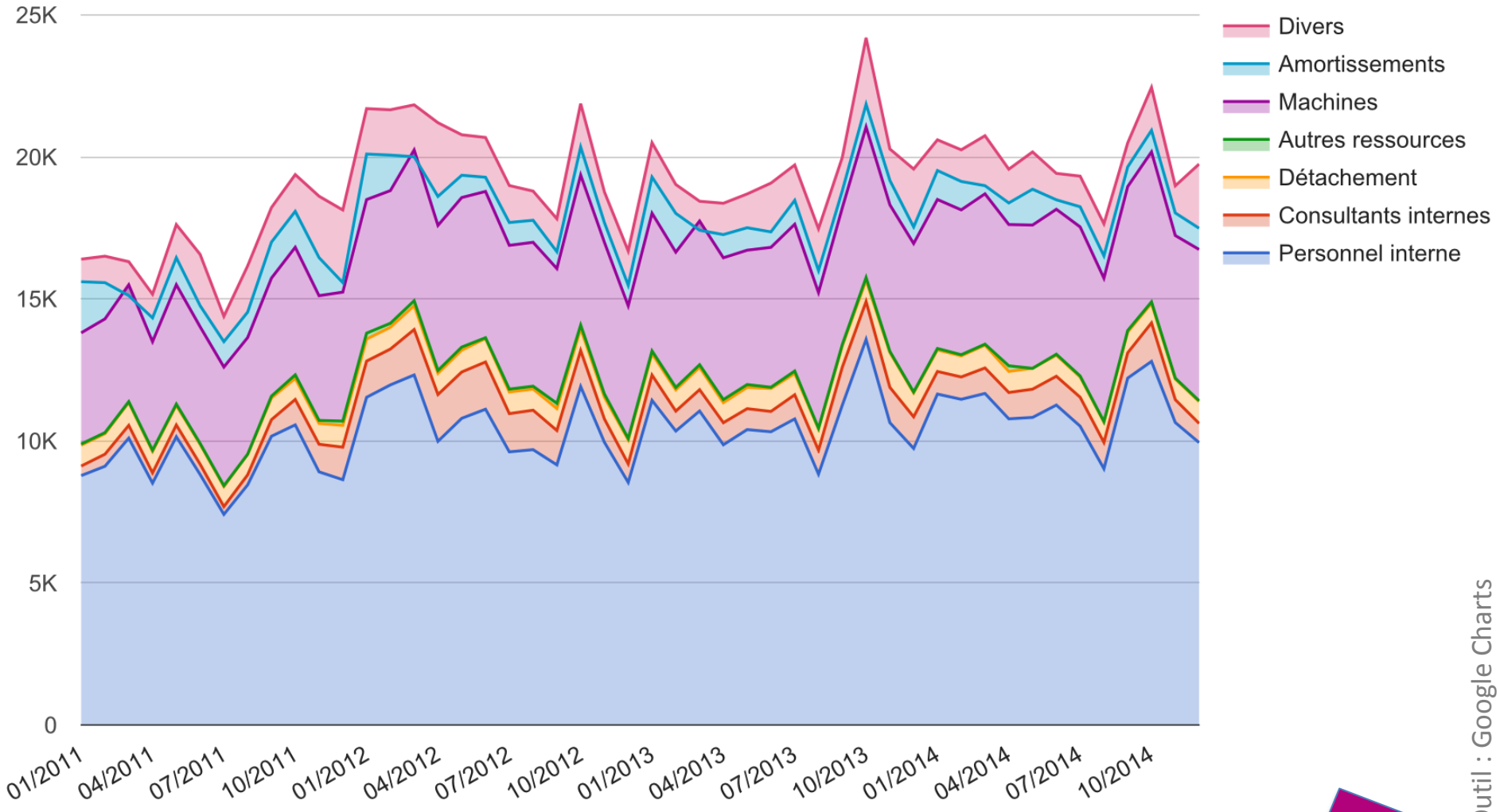
Sun Burst



Outil : HighCharts

Stacked line chart

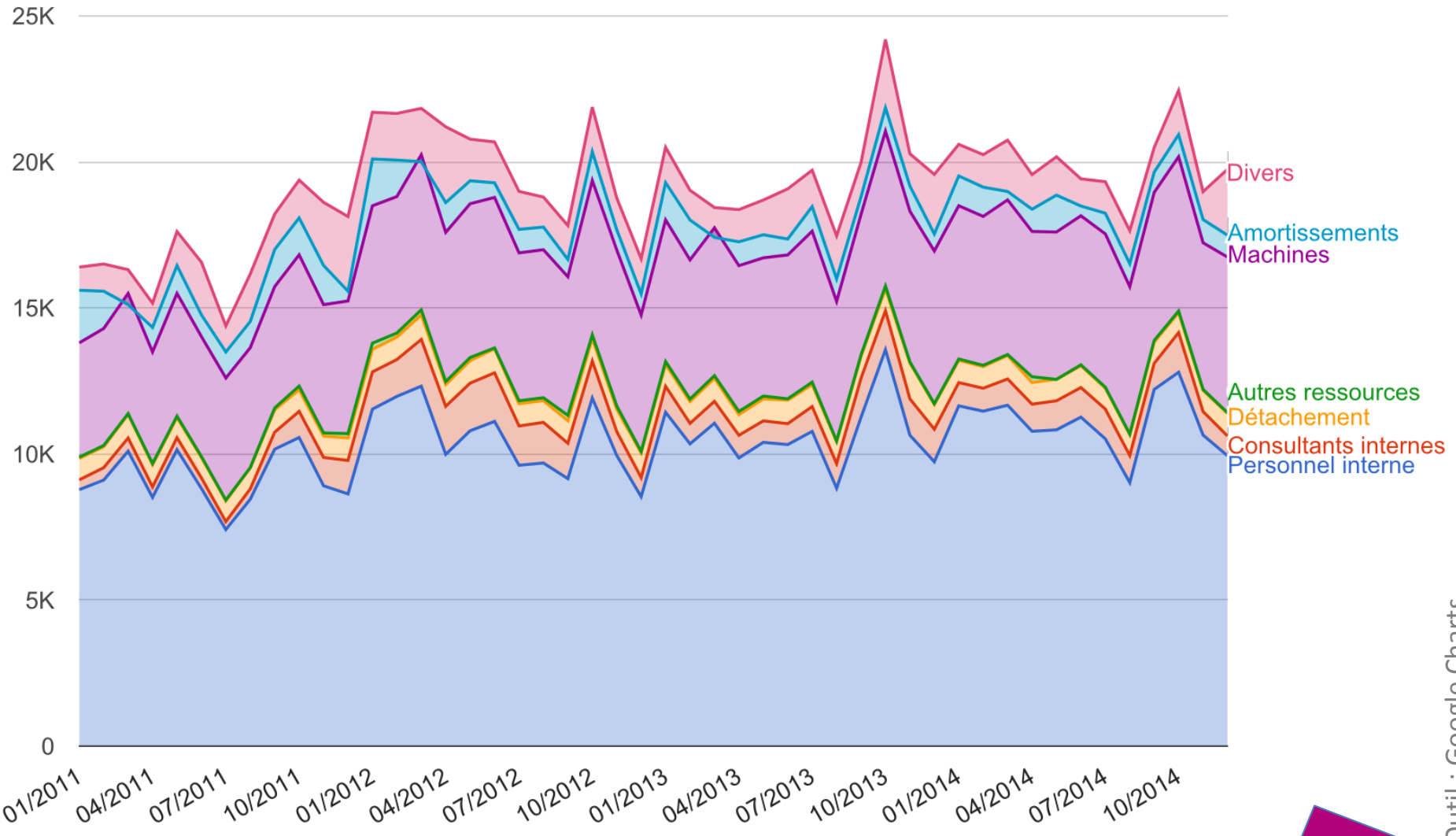
Données : Smals ICT for society



Outil : Google Charts

Stacked line chart

Données : Smals ICT for society

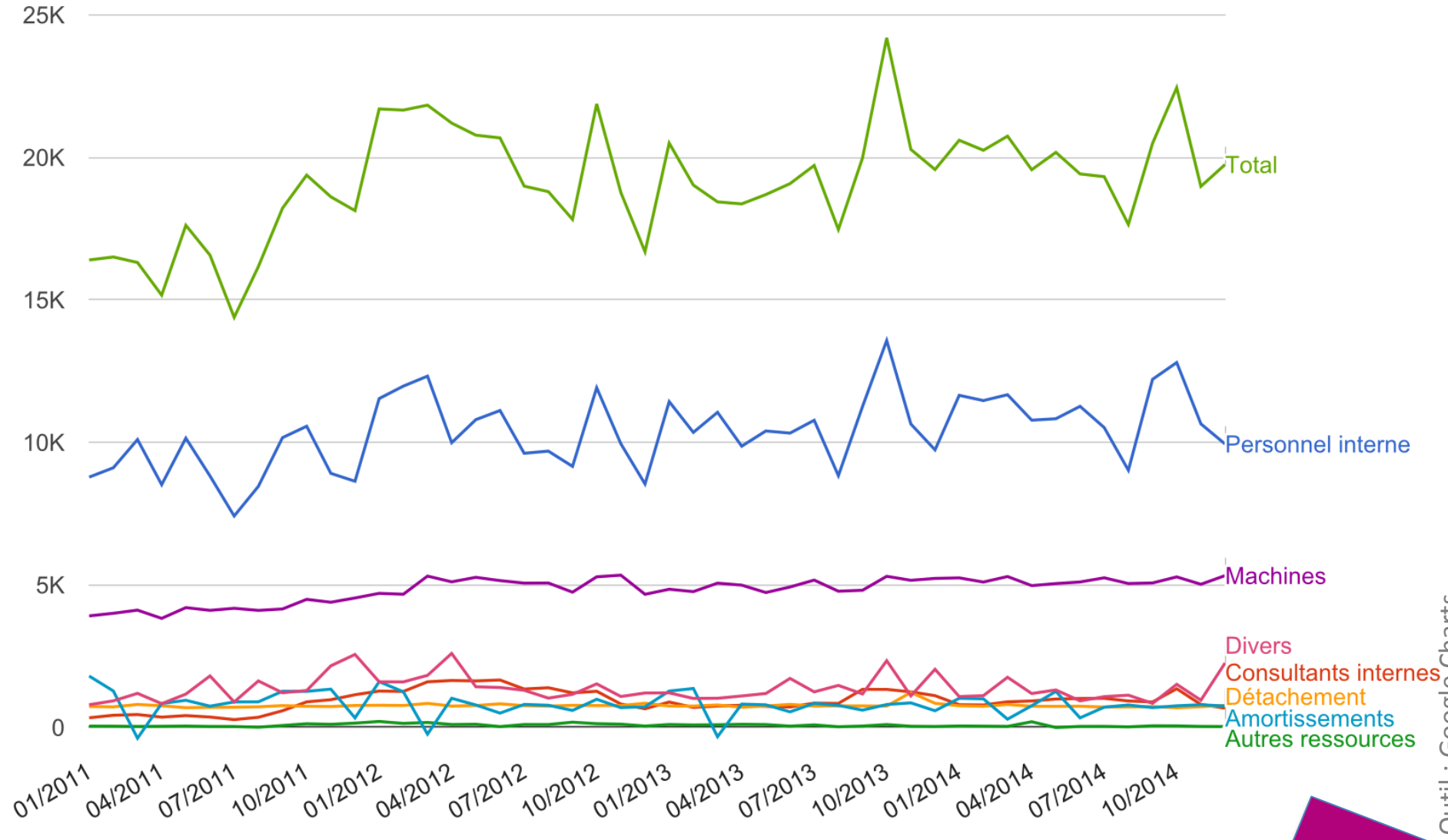


Outil : Google Charts

Données :

Smals
ICT for society

Stacked line chart



Outil : Google Charts

Données :

Smals
ICT for society

Stacked line chart

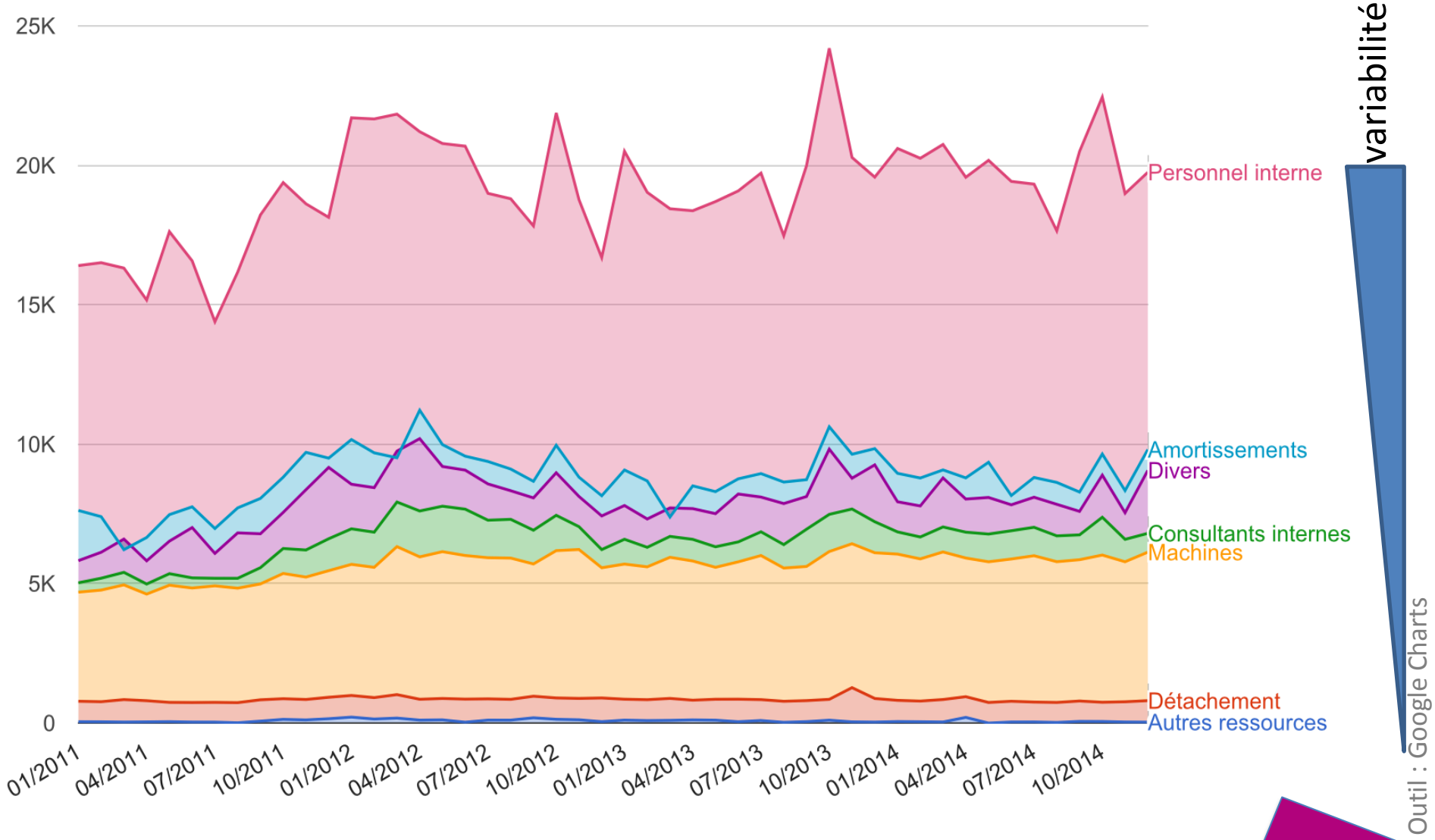
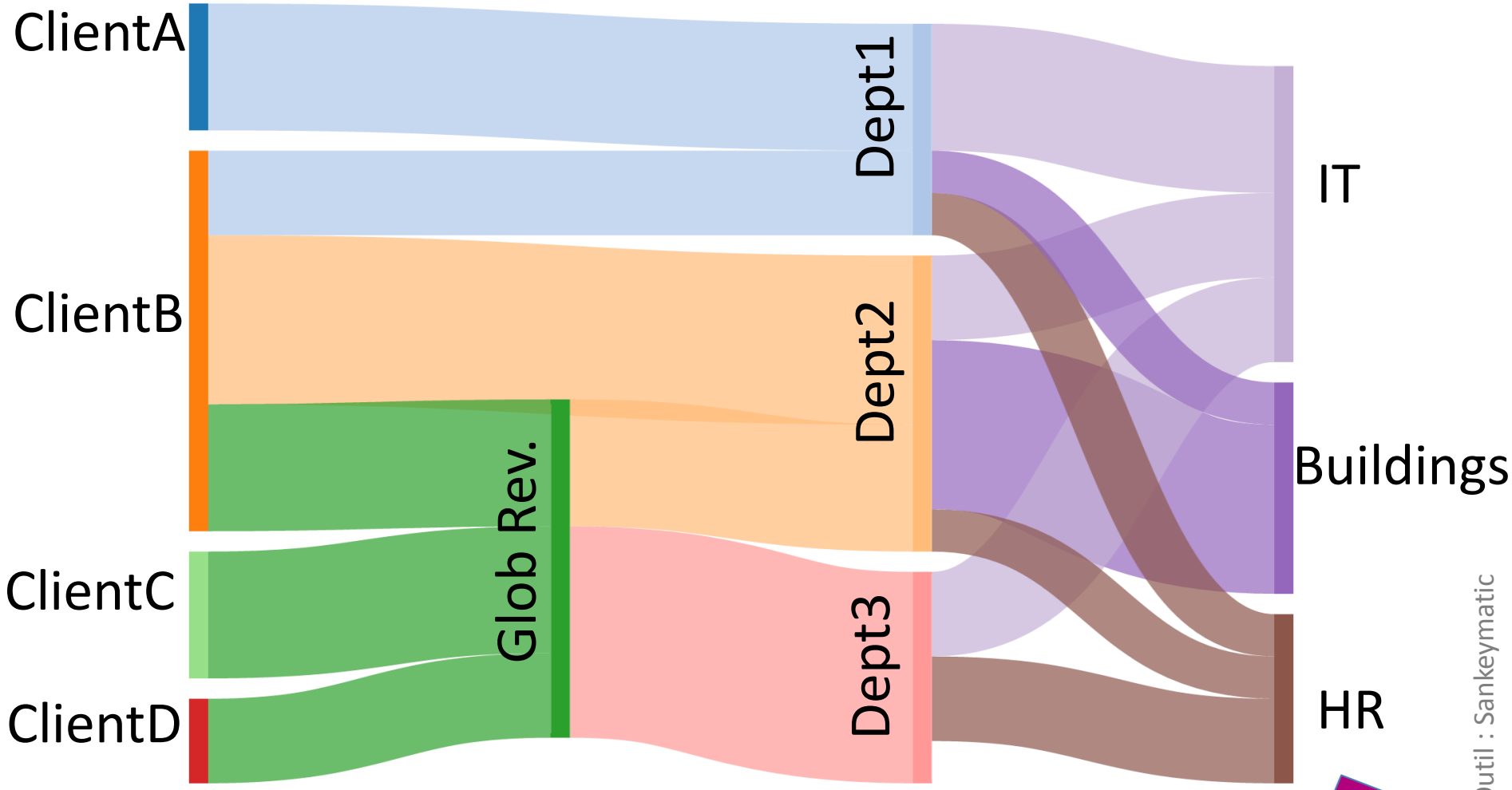


Diagramme de Sankey



Outil : Sankeymatic

Diagramme de Chord

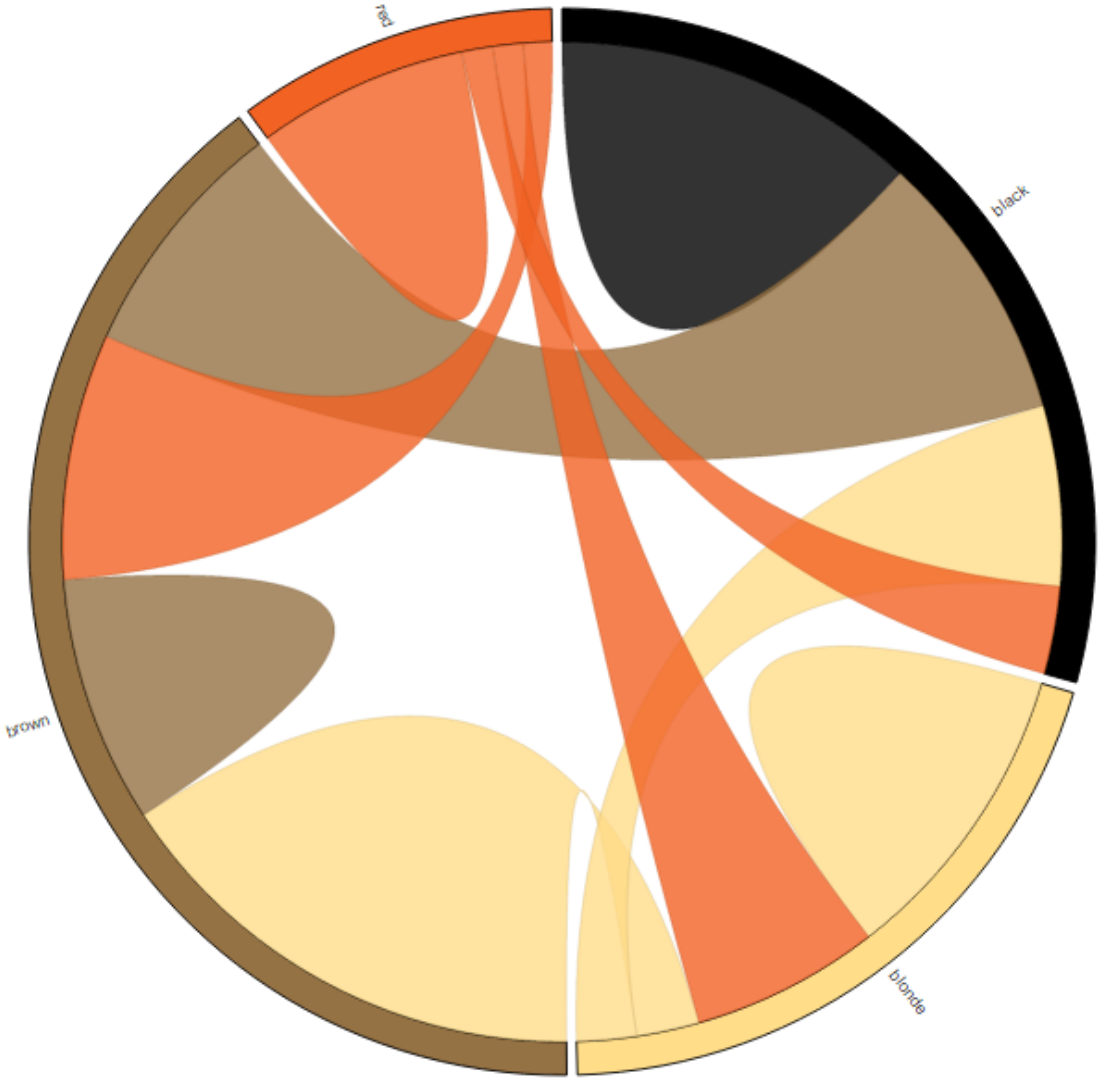
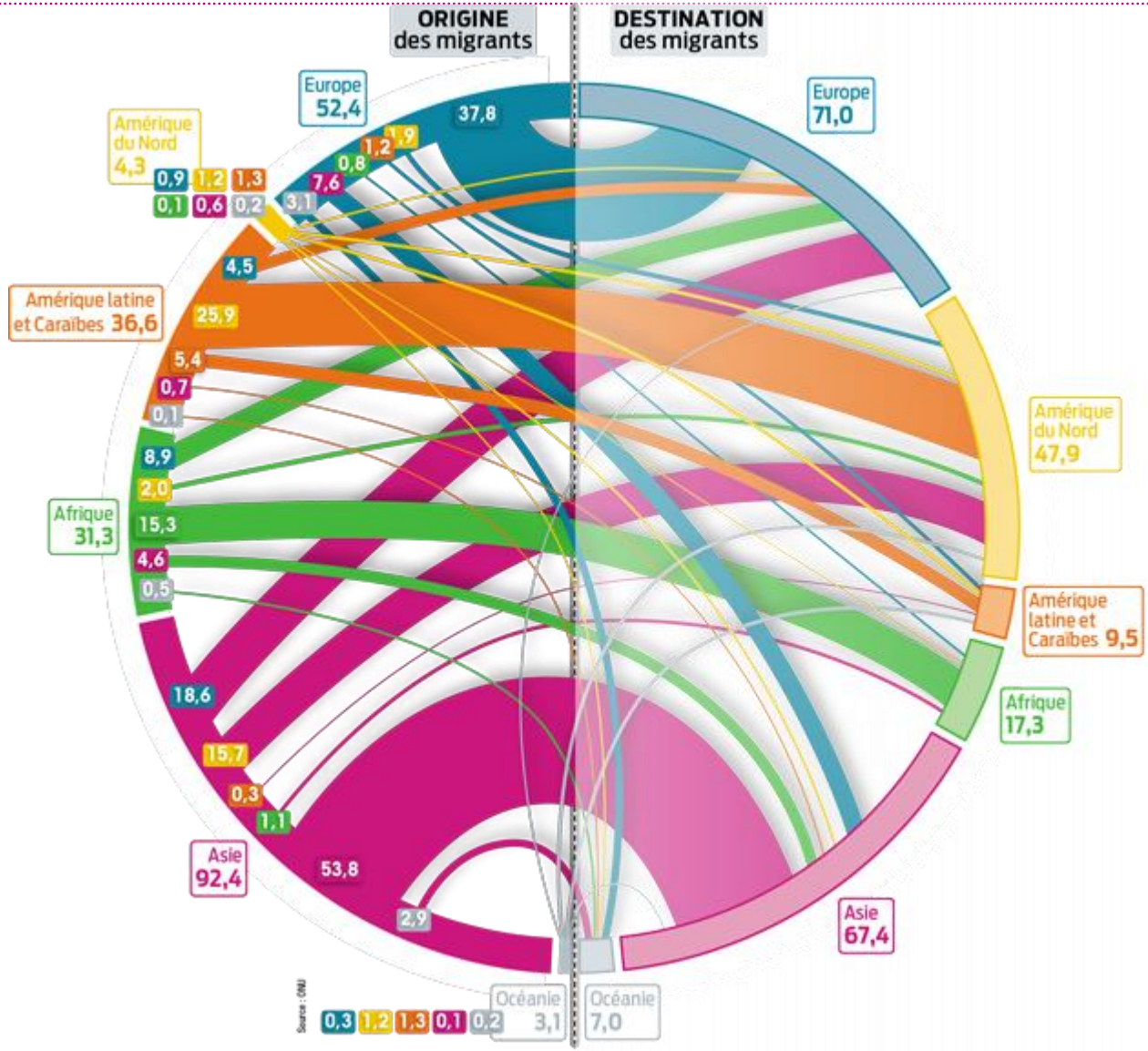


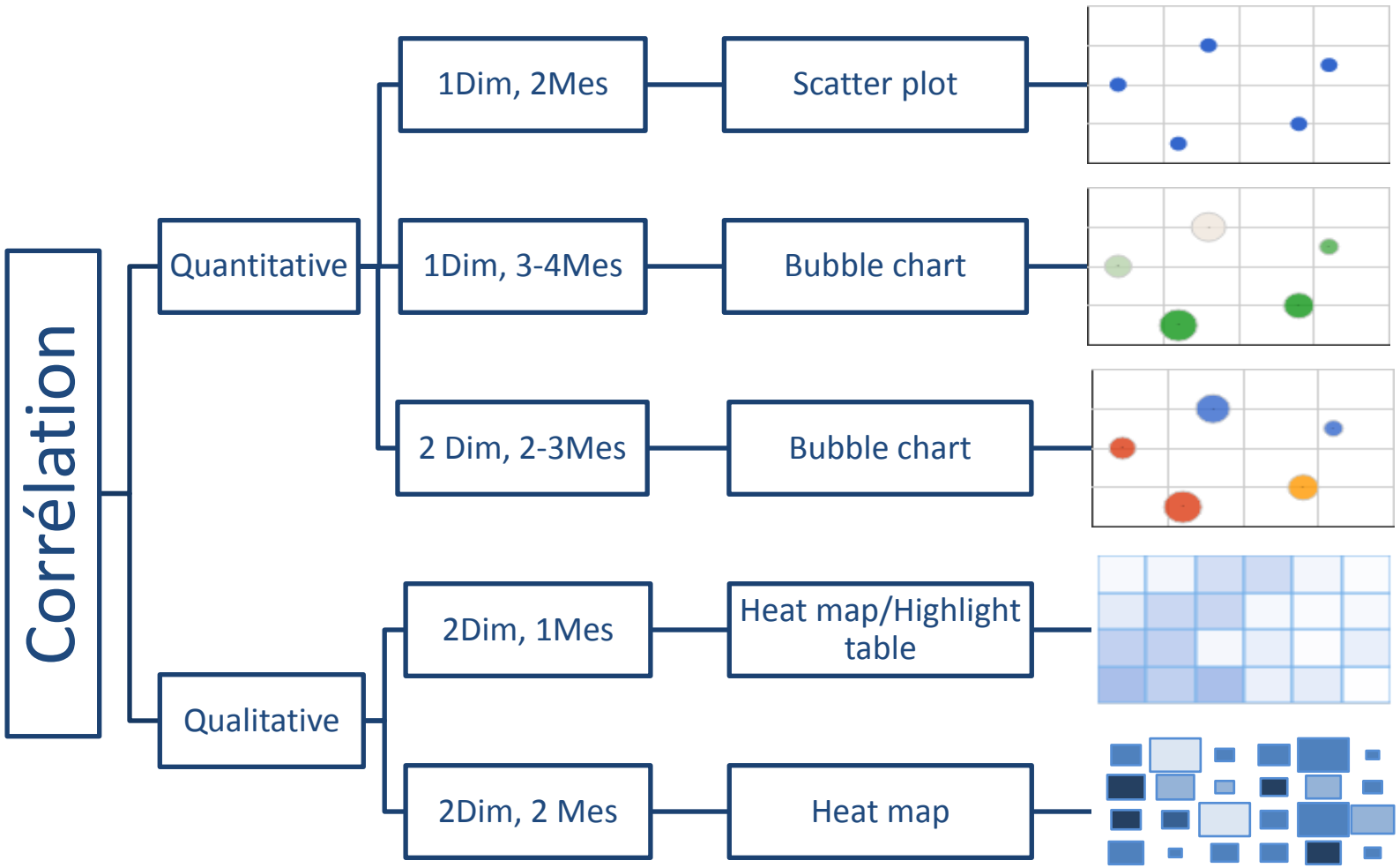
Diagramme de Chord : migrations 2013



Choix des graphiques

CORRÉLATION

Corrélation

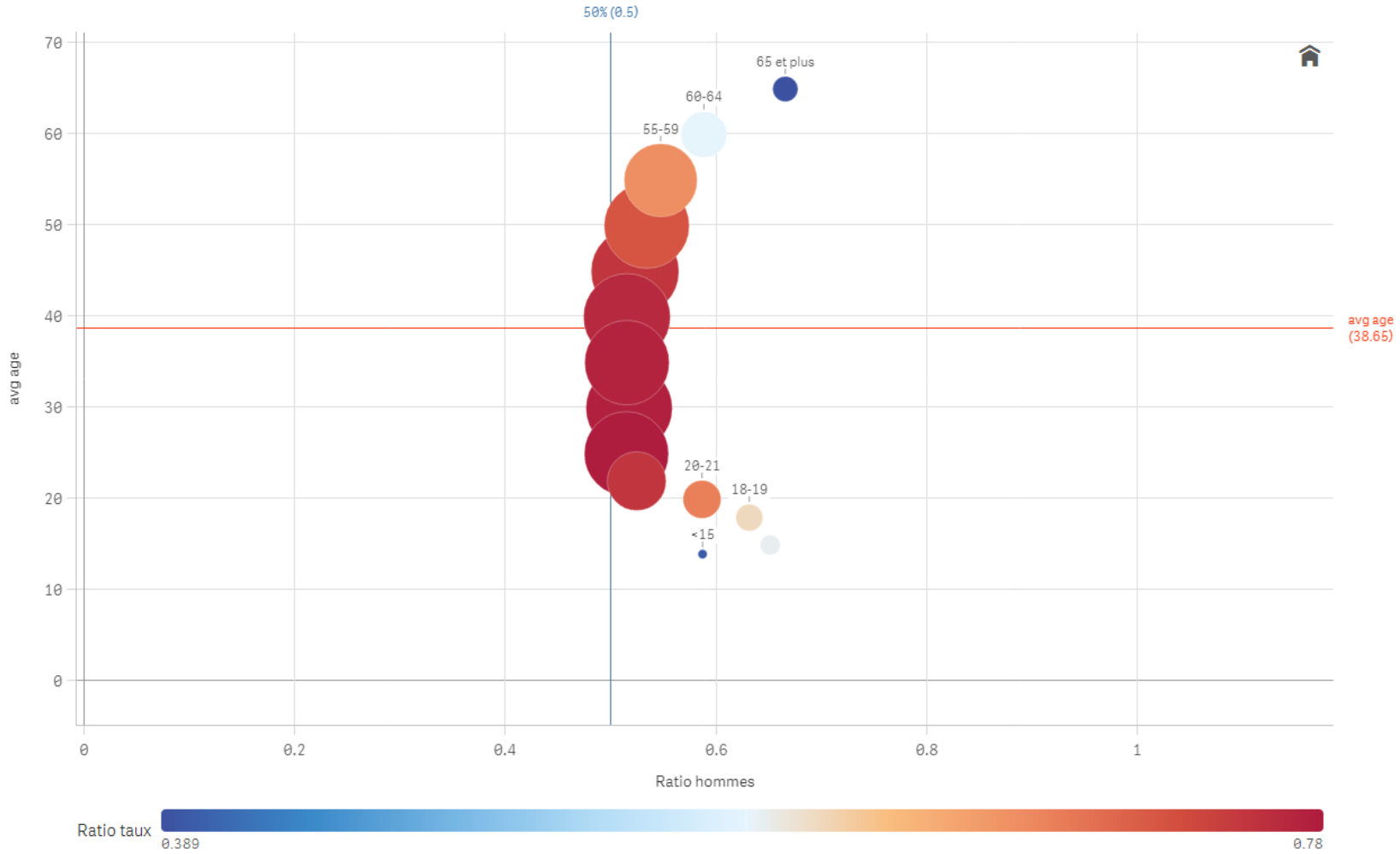


Outil : Google Charts, HighCharts

Bubble Chart

Données (test) :

ONSS

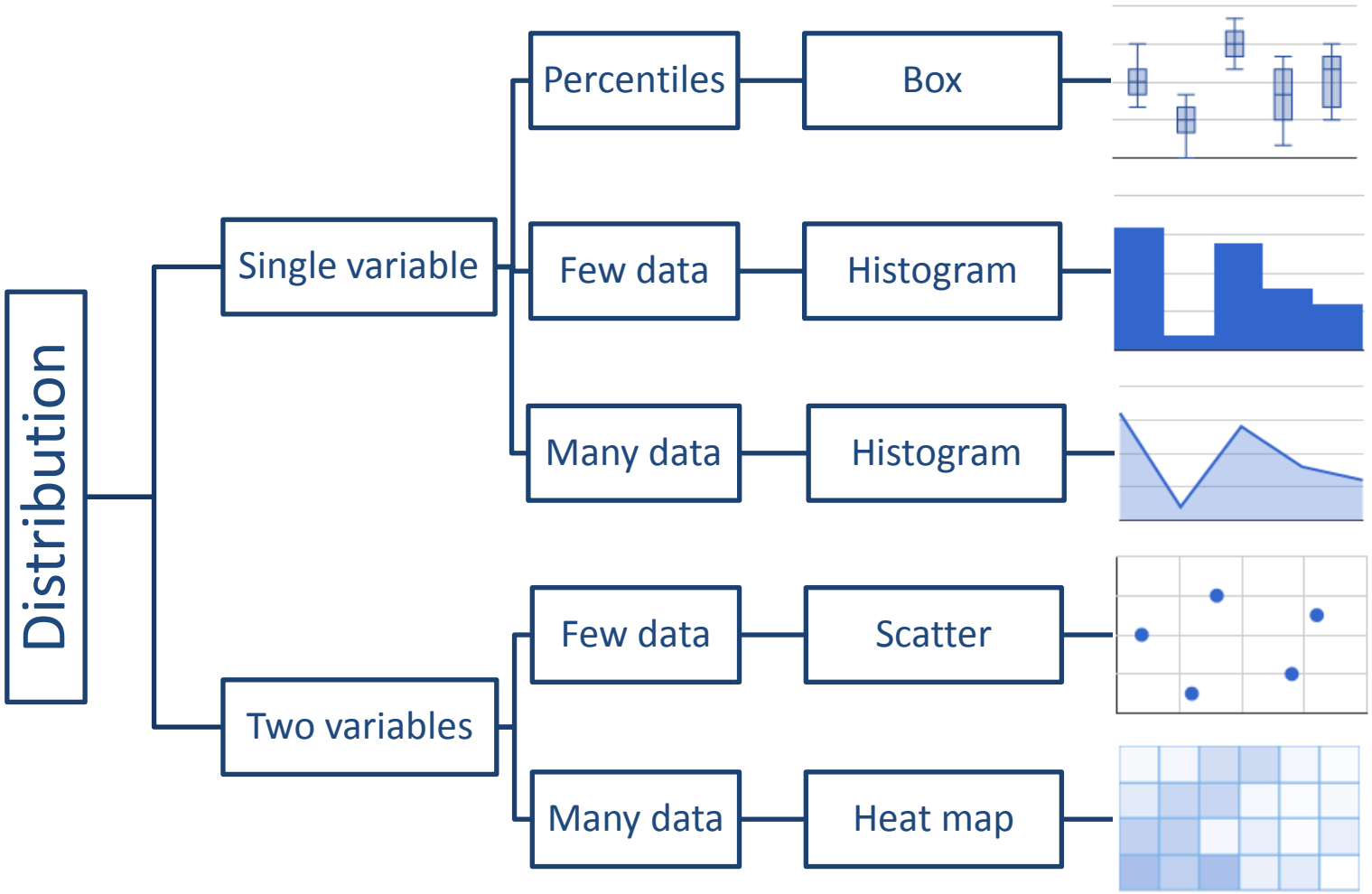


Outil : QlikSense

Choix des graphiques

DISTRIBUTION

Distribution

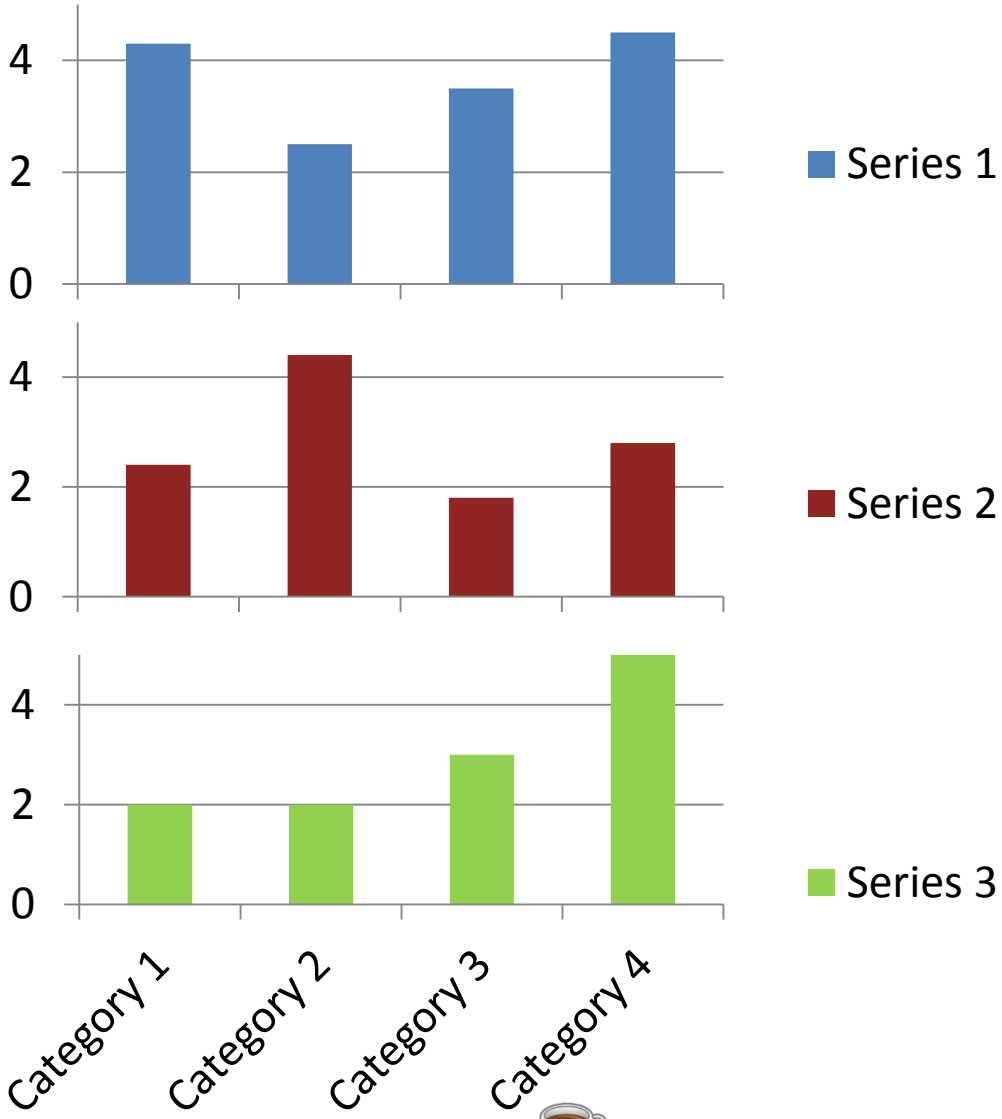


Outil : Google Charts, HighCharts

Choix des graphiques

NOMBREUSES VARIABLES

Small multiple

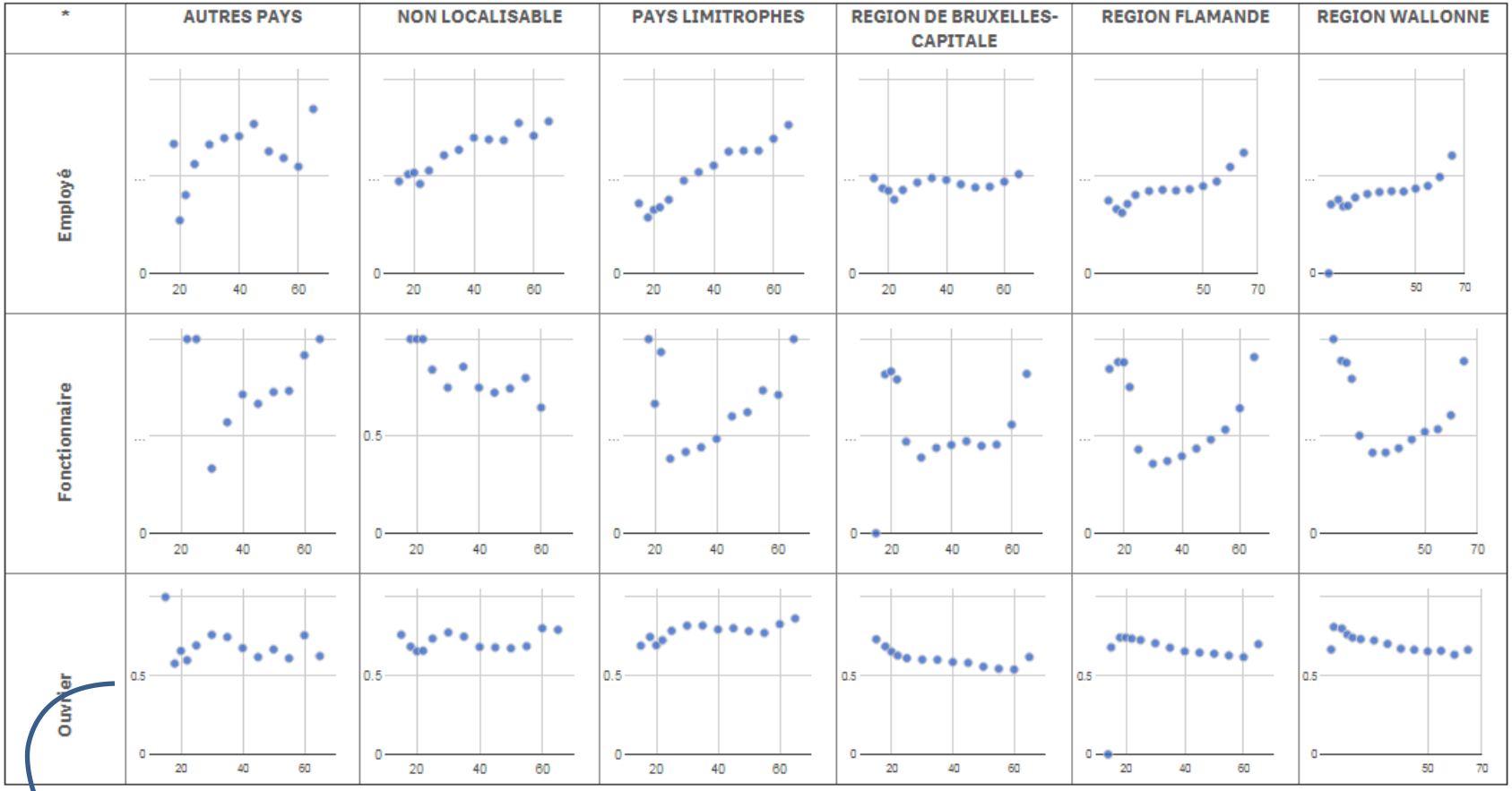


Matrix Plot

Données (test) : ONSS

Valeurs d'une dimension B

Valeurs d'une dimension A



Ratio ♂/♀

Age

Outil : Google Charts, QlikSense

Choix de graphiques

RÉSEAUX

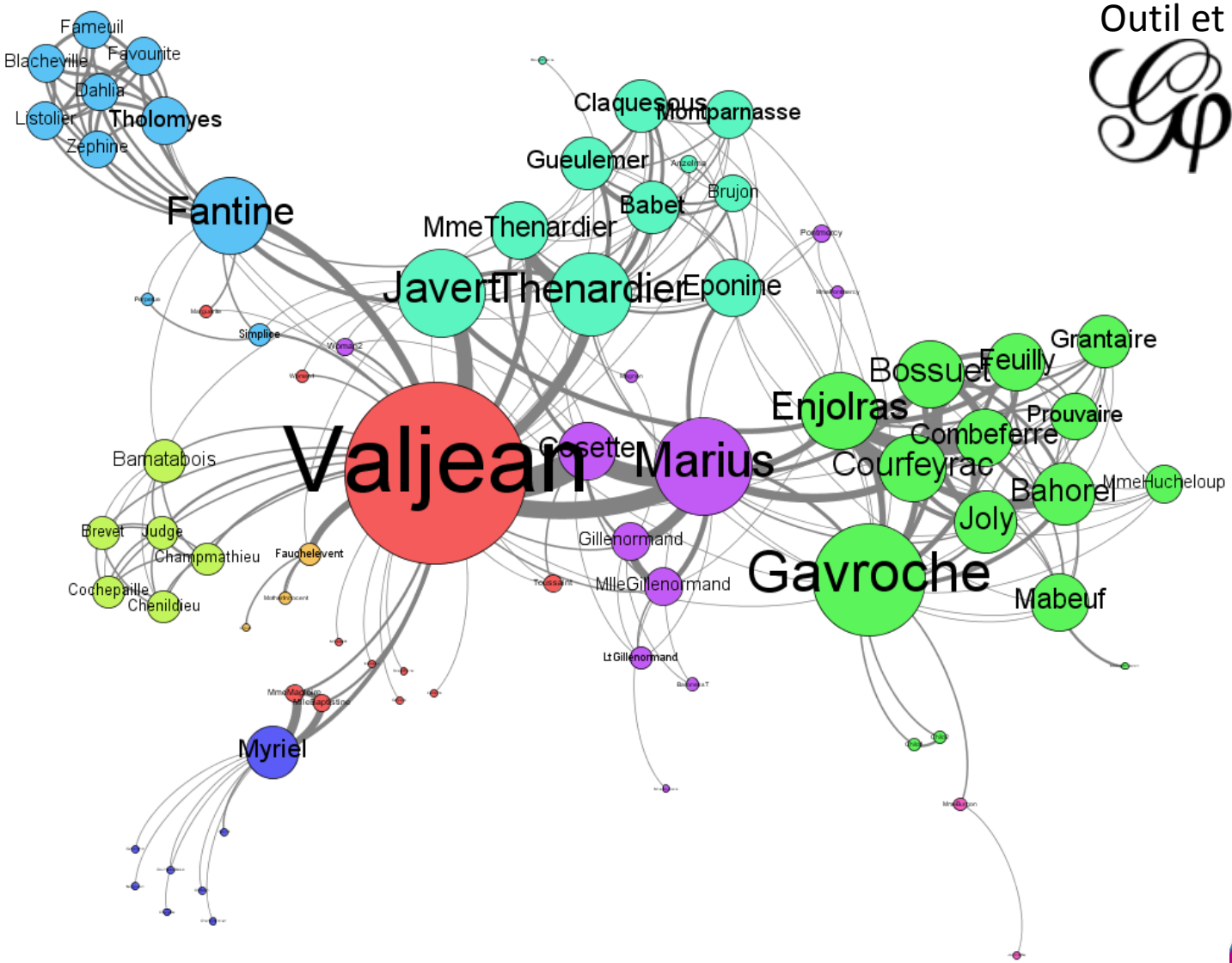
Réseaux

- 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
- …

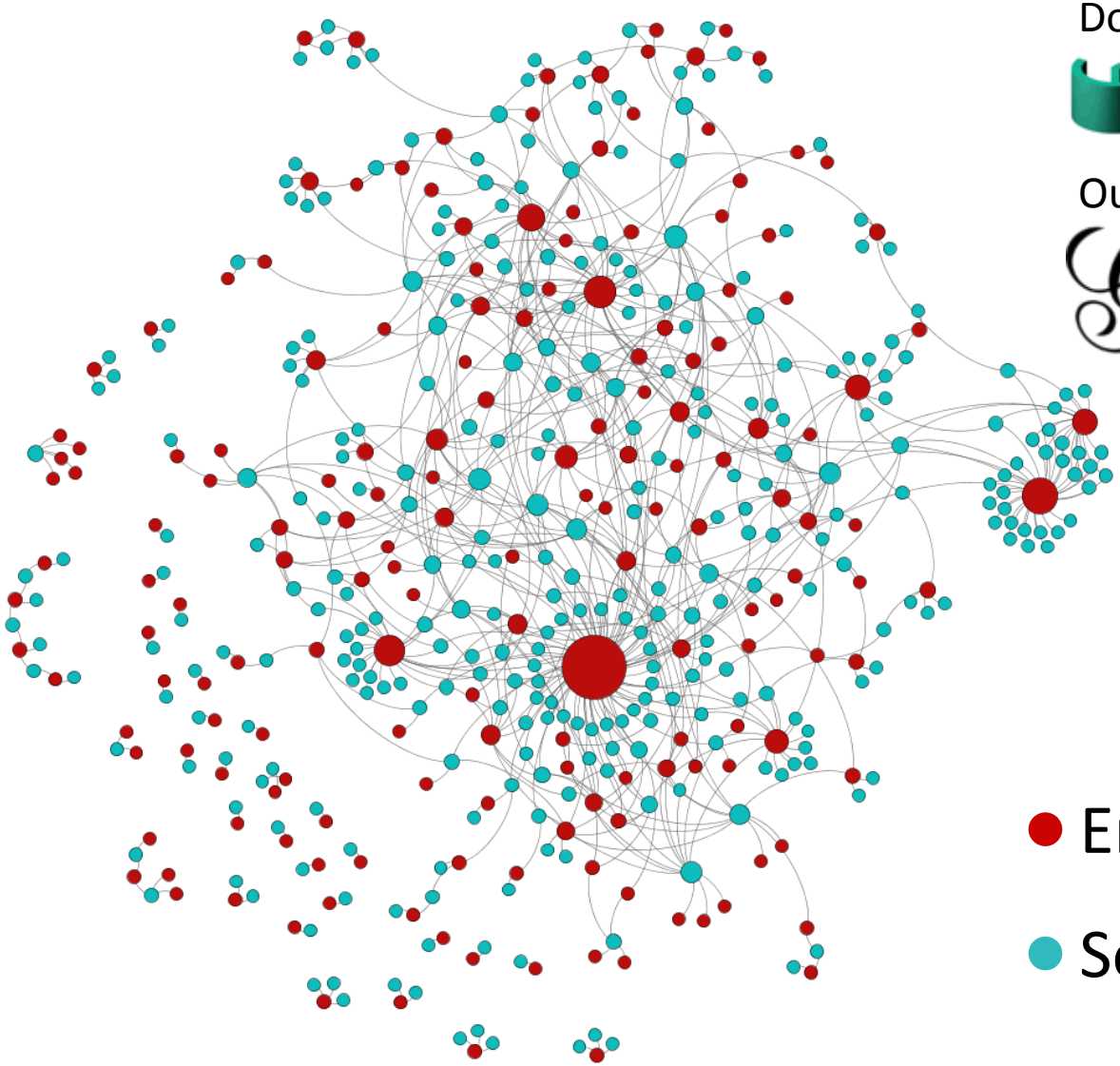
Réseaux : les misérables (V. Hugo)

Outil et données :

Gephi



Réseaux : Entrepreneur – sous-traitant



Données :

ONSS

Outil :

Gephi

- Entrepreneur
- Sous-traitant

Questions ?

Pause !

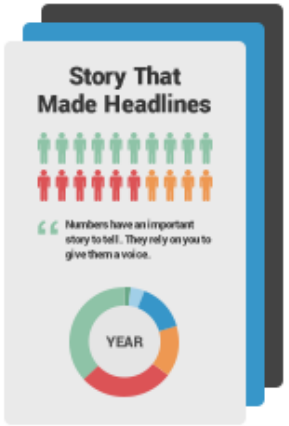
Choix des graphiques

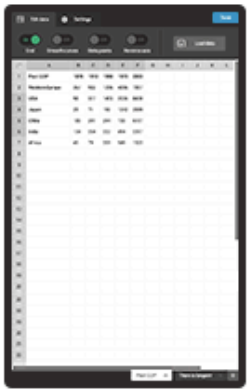
OUTILS

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 »
- Librairies :
  - 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

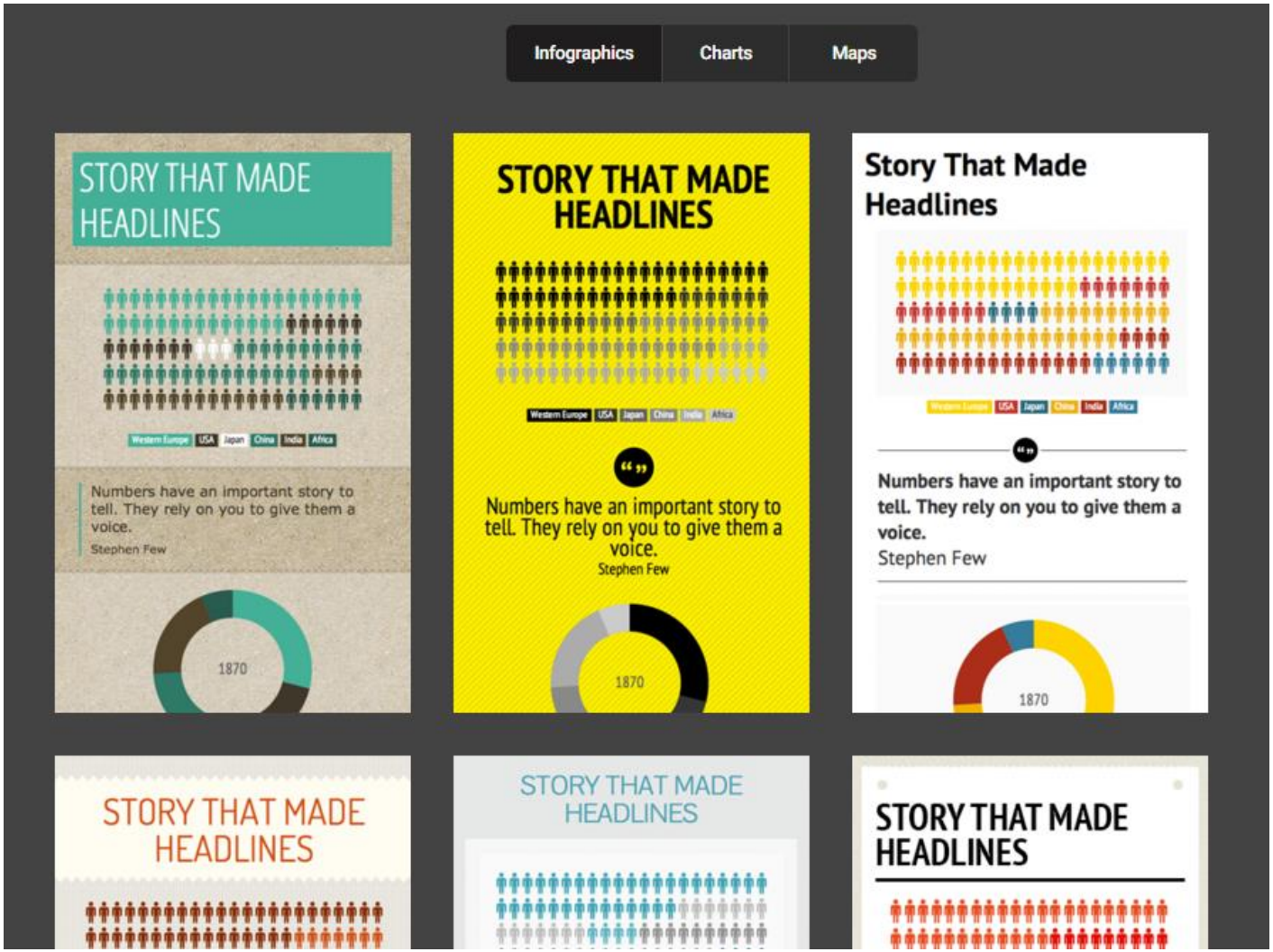
Add charts, maps, videos, images, icons and more

3

Publish & Share

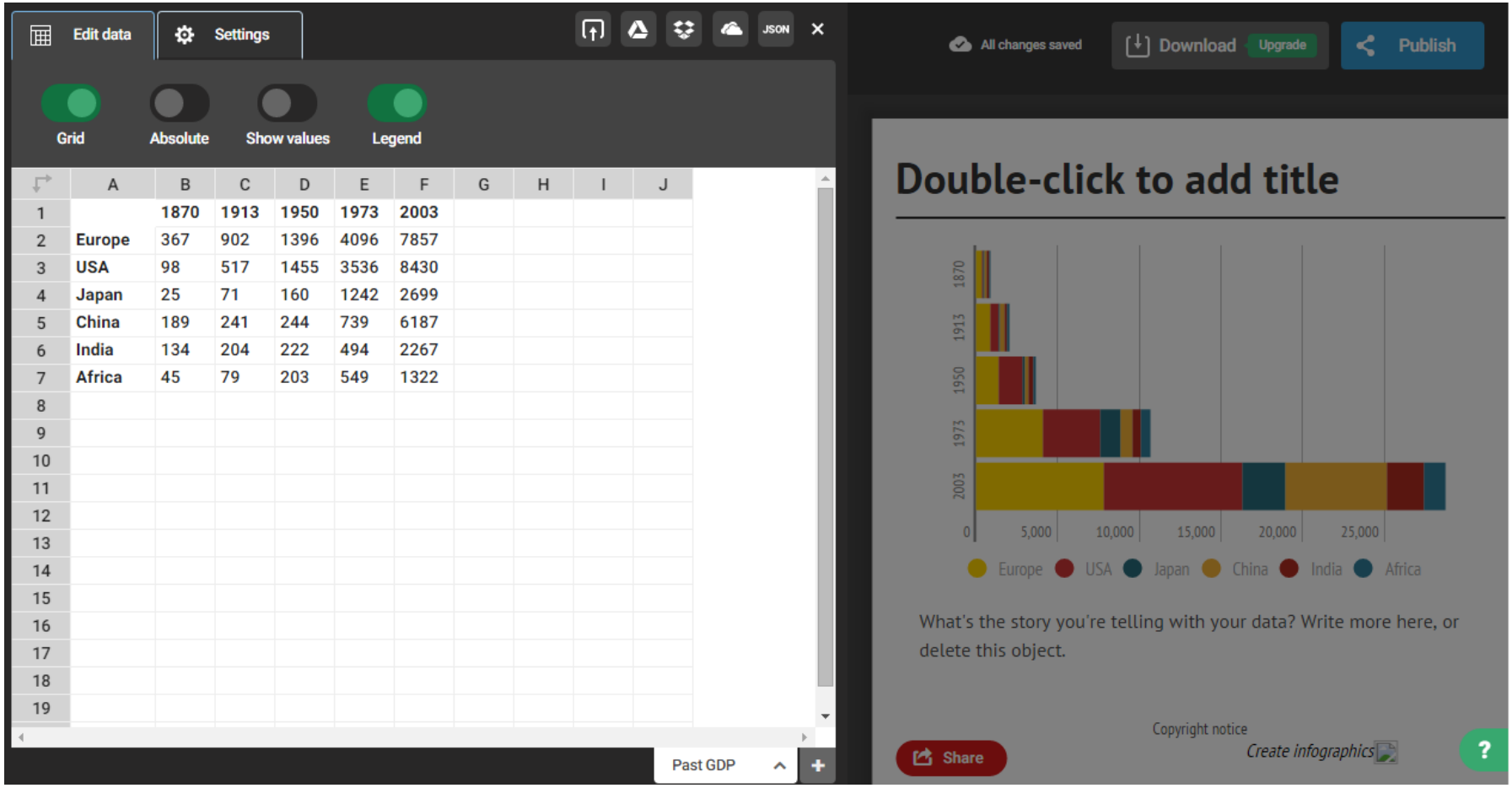
Use our one-click share buttons or easily generate an embed code to post on your blog or website

Outils spécialisés : infogr.am



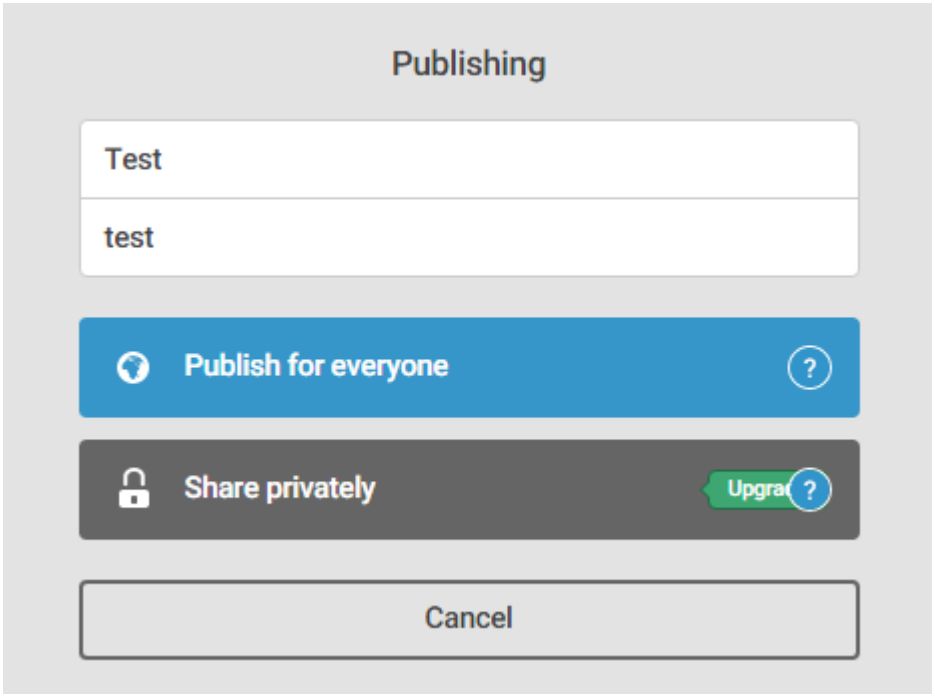
Outils spécialisés : infogr.am

Outils spécialisés : infogr.am



	A	B	C	D	E	F
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

Outils spécialisés : infogr.am



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, 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 évènements
- Beaucoup d'exemples pour tout type de graphique
- Très puissant, mais très bas niveau
- À la base de beaucoup de librairies

Librairies javascript : D3

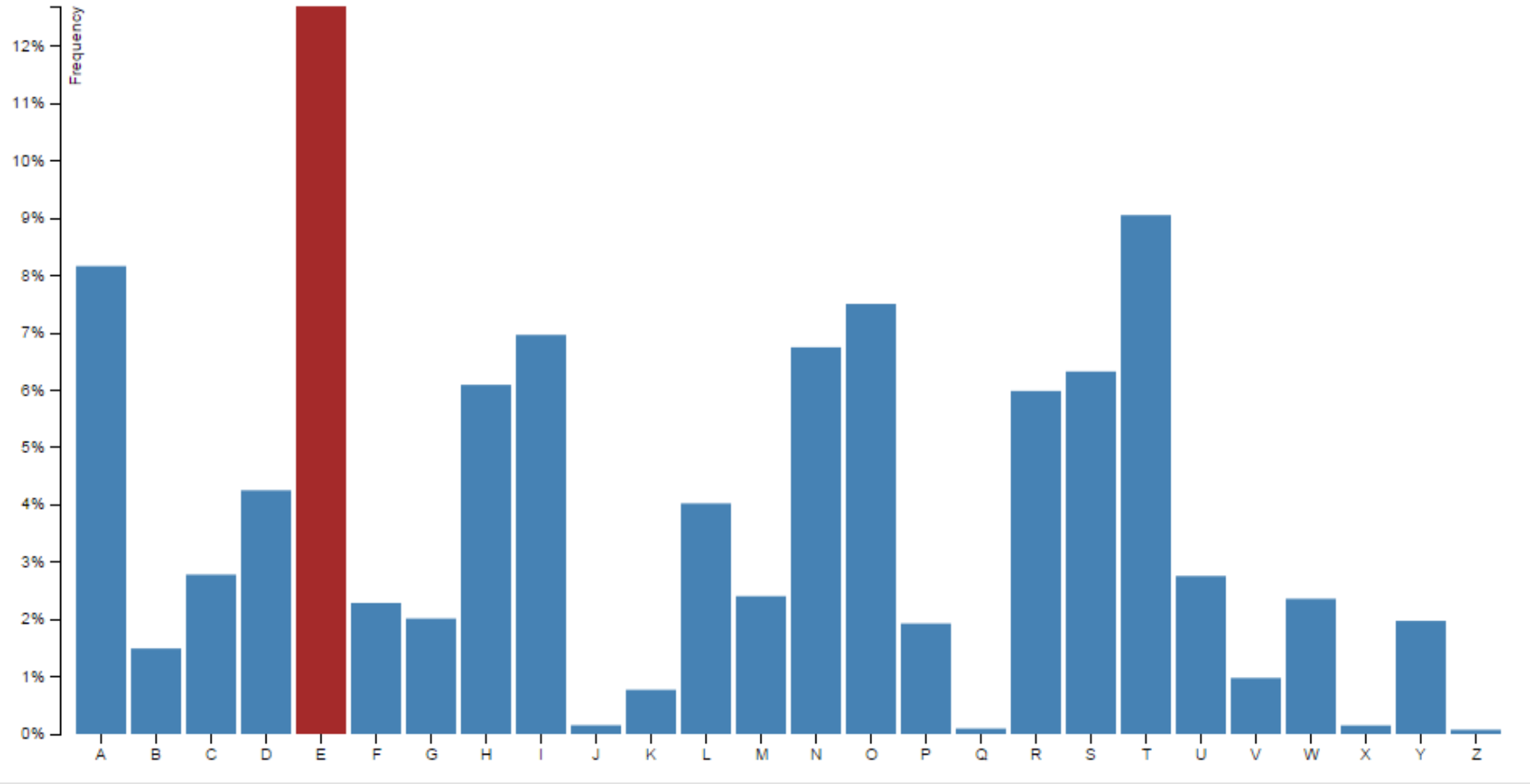
```
<!DOCTYPE html> <meta charset="utf-8">
<style>
.bar { fill: steelblue; }
.bar:hover { fill: brown; }
.axis { font: 10px sans-serif; }
.axis path,
.axis line { fill: none; stroke: #000; shape-rendering: crispEdges; }
.x.axis path { display: none; }
</style>
<body>
<script src="https://cdnjs.cloudflare.com/ajax/libs/d3/3.5.5/d3.min.js"></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
....	

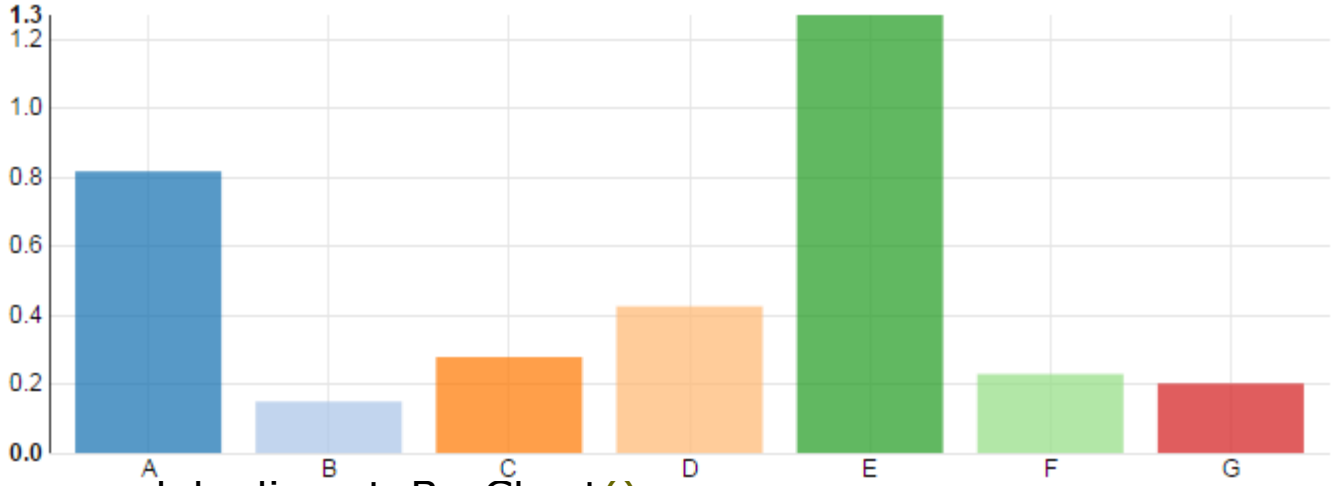
http://bl.ocks.org/mbostock/3885304

Librairies javascript : D3



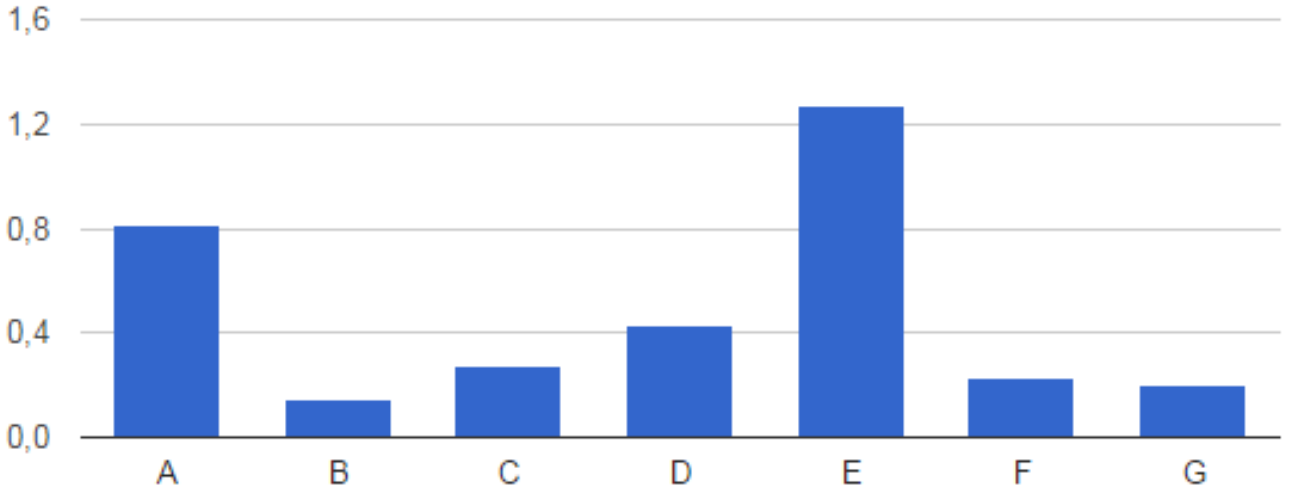
NVD3

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


  nv.addGraph(fur
        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



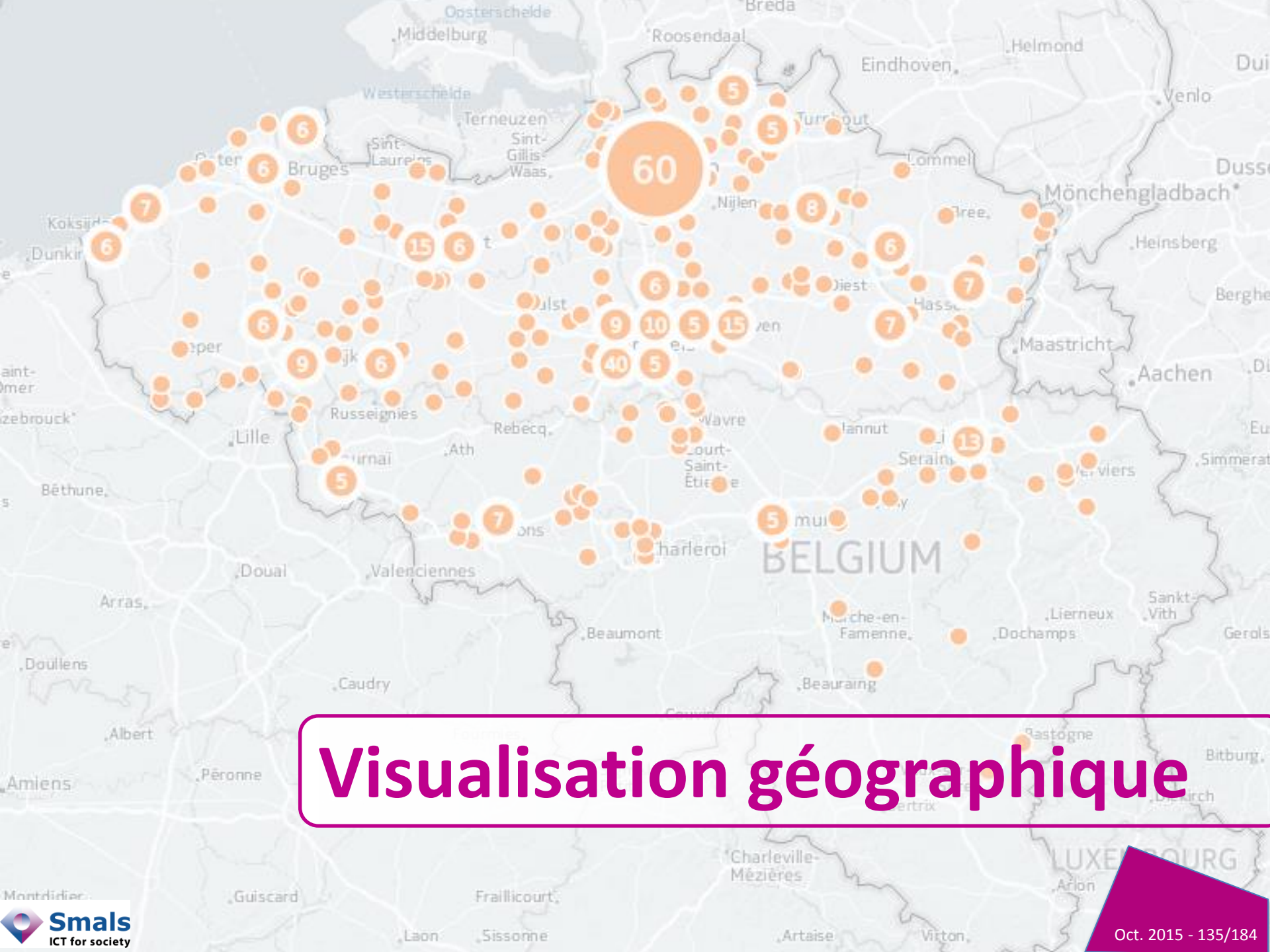

```
<script type="te
google.load("vis
google.setOnLoa

function drawCh
        var data
'Frequency'], ["A
["E",1.2702], ["F ,.2288 ], [ G ,.2015 ]]);

        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 »

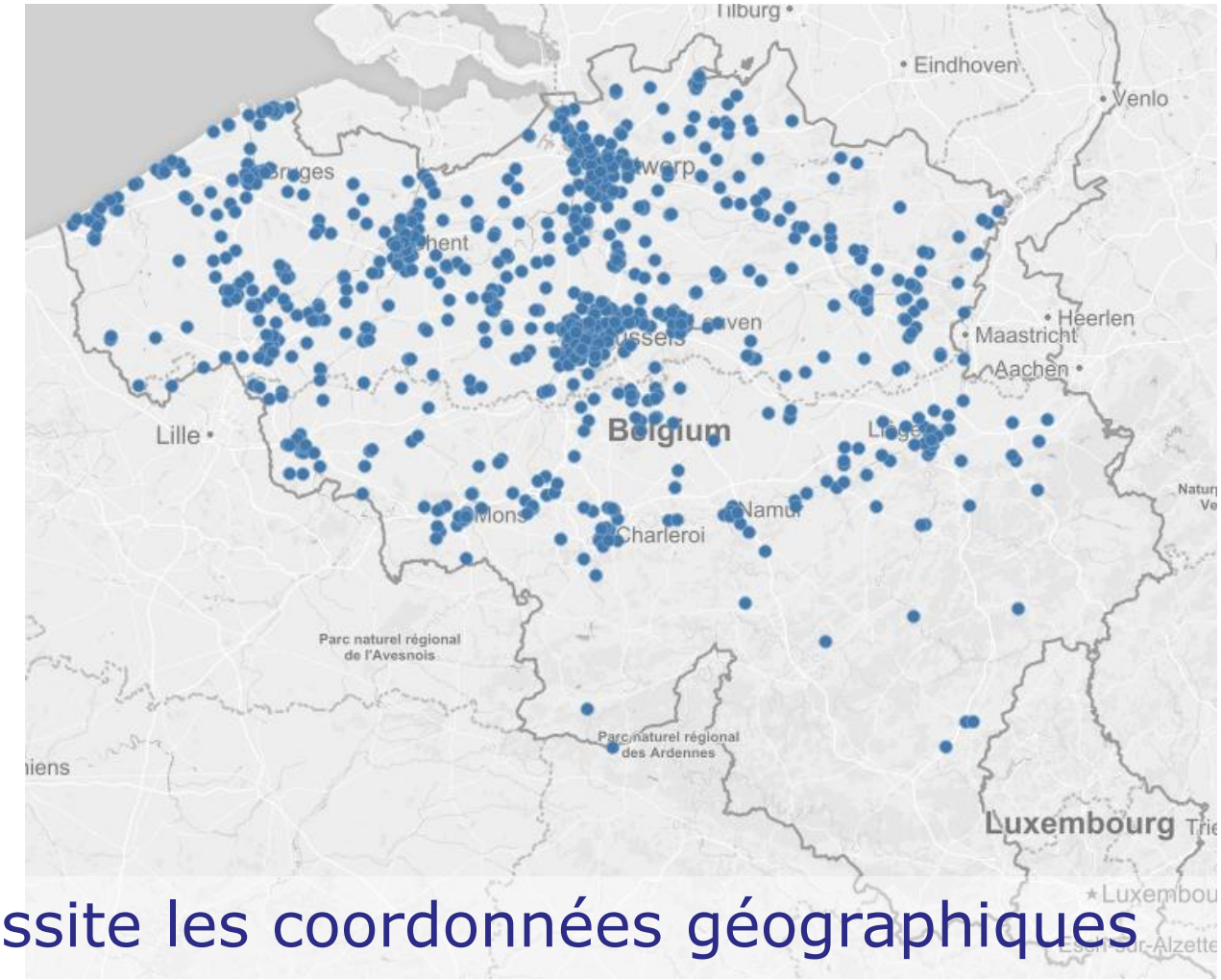

Visualisation géographique

Visualisation géographique

TYPES DE VISUALISATION

Point & bubble

Données : ONSS

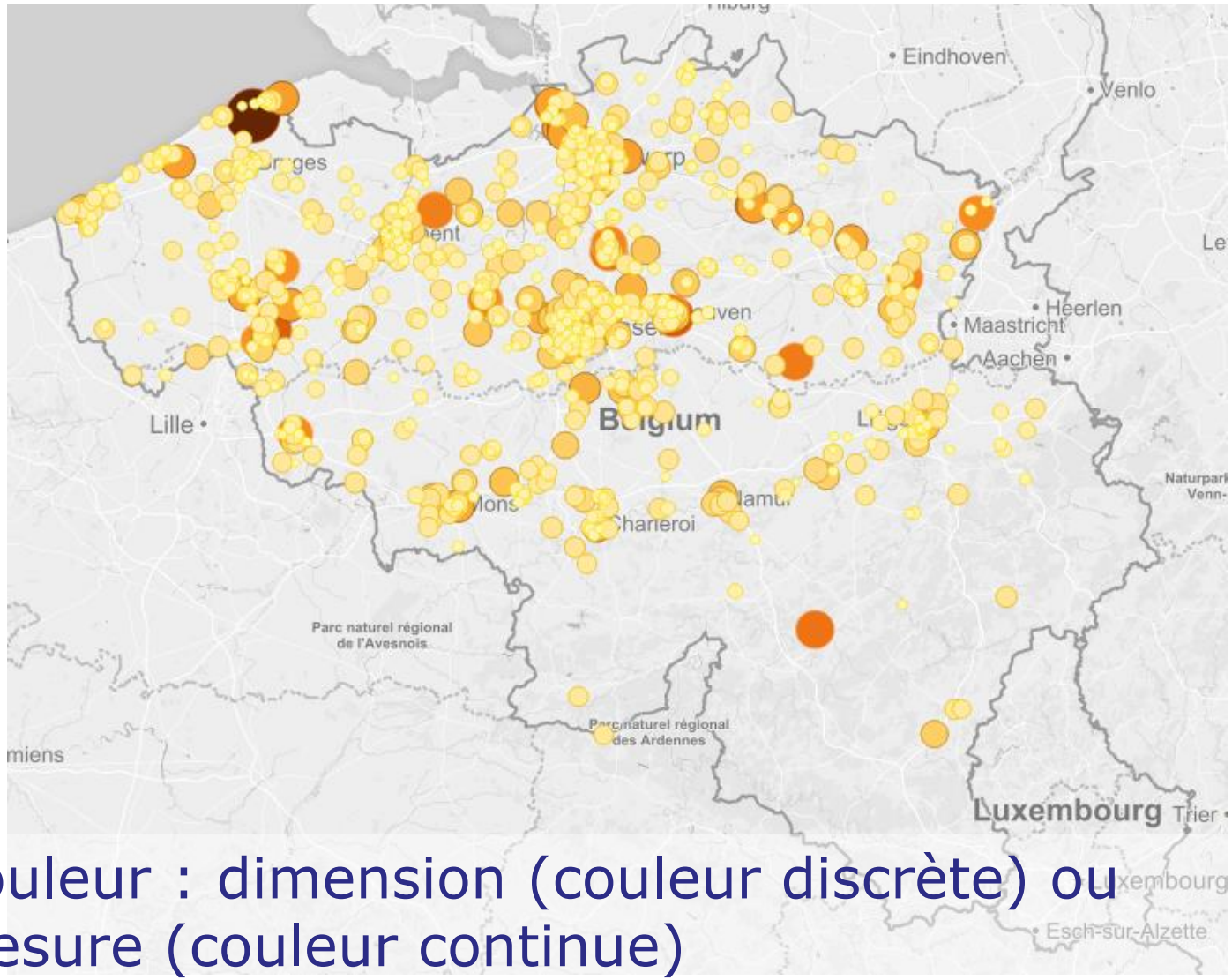


- Nécessite les coordonnées géographiques

Outil : QlikSense

Point & bubble

Données : ONSS



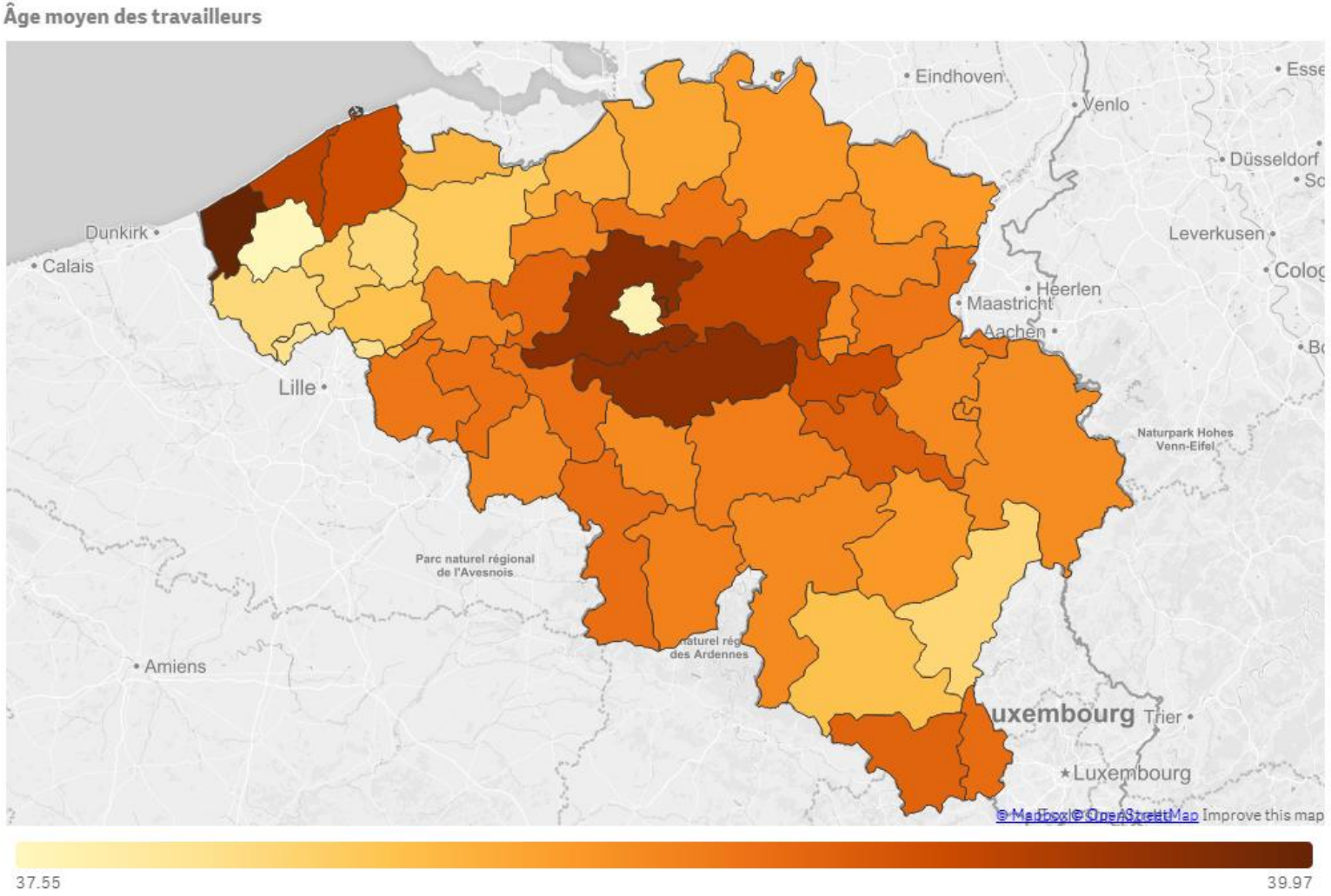
- Couleur : dimension (couleur discrète) ou mesure (couleur continue)

Outil : QlikSense

Carte choroplèthe

Données (test) :

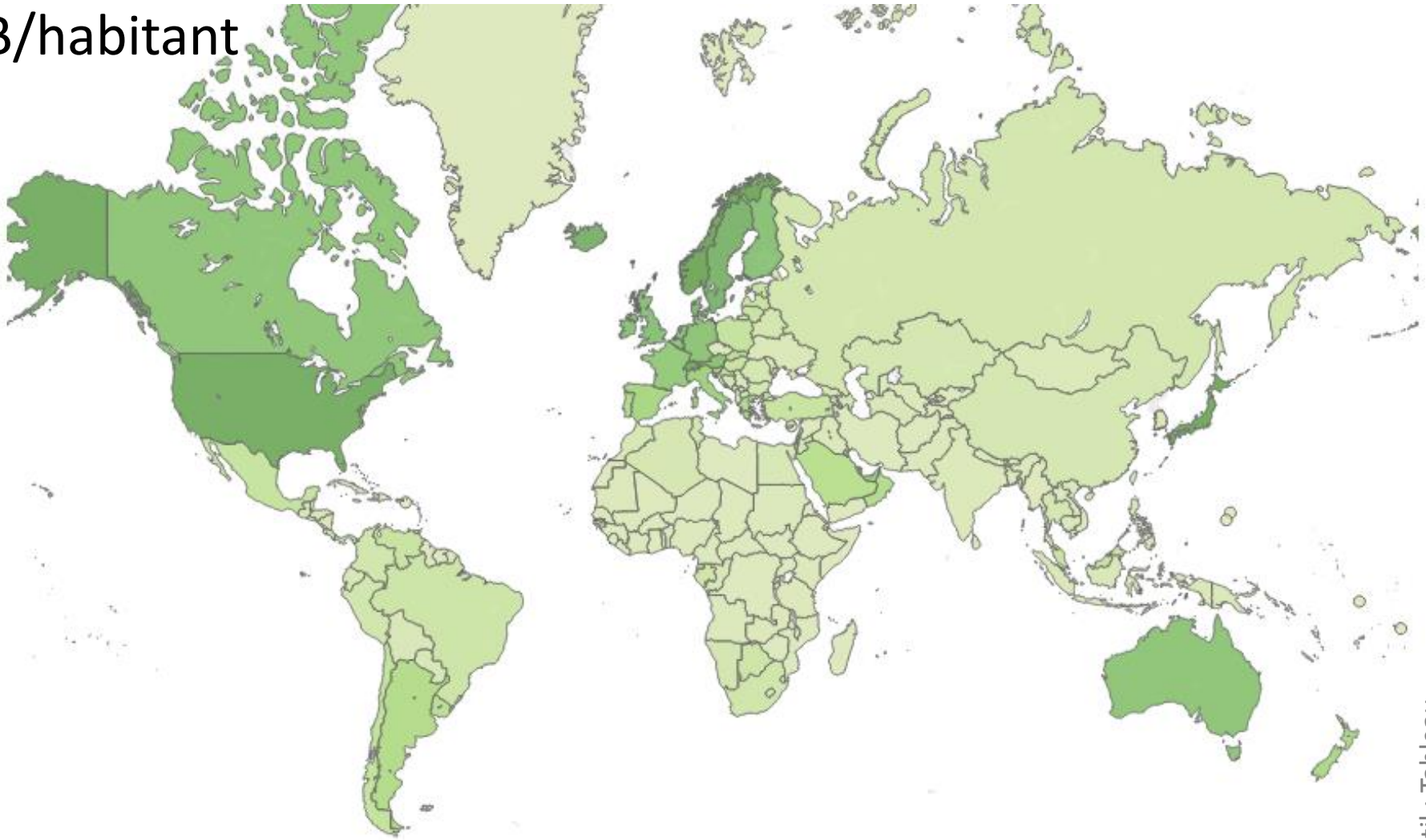
ONSS



Outil : QlikSense

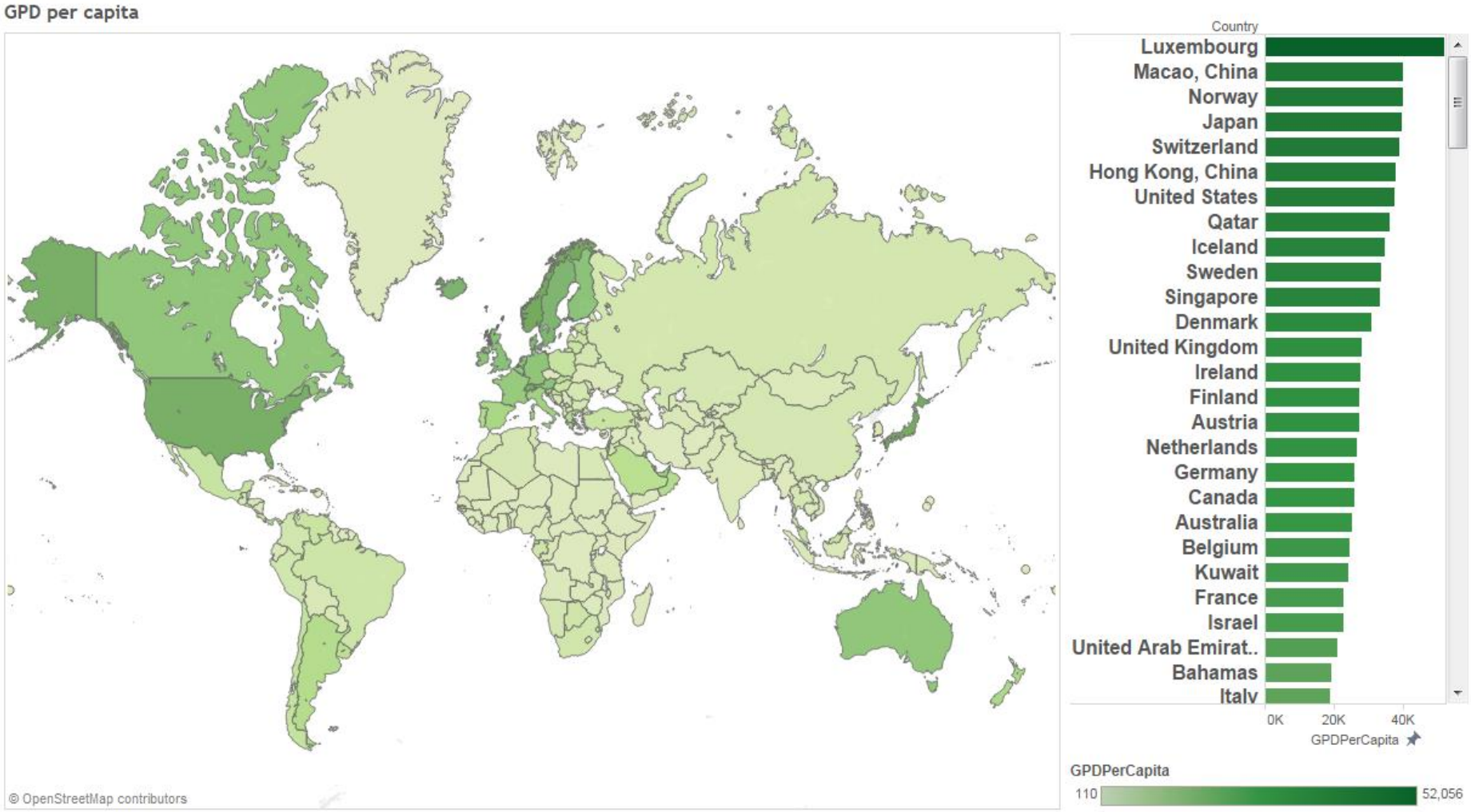
Carte choroplèthe

PIB/habitant



Outil : Tableau

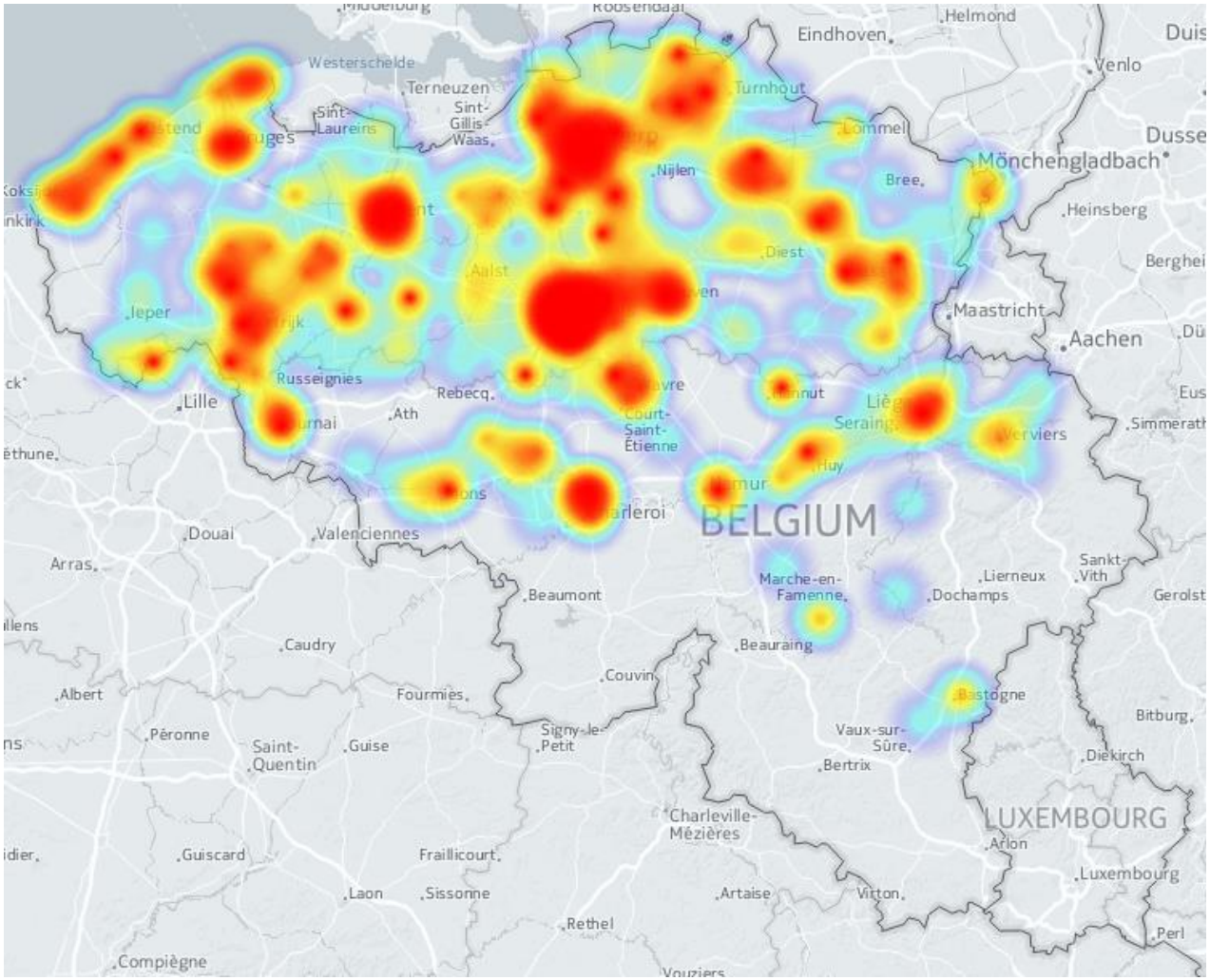
Carte choroplèthe



Outil : Tableau

Heatmap

Données : ONSS

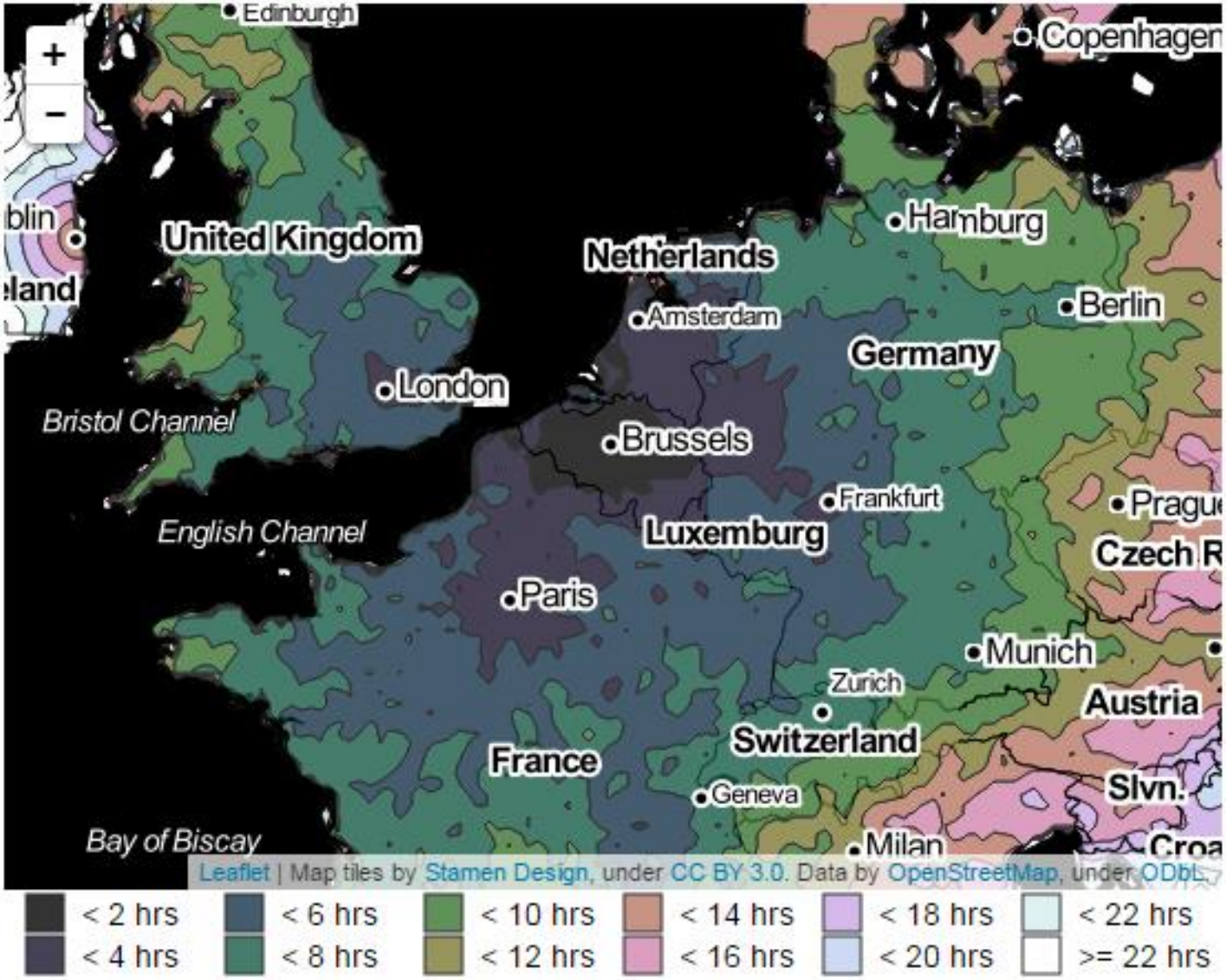


Outil : CartoDB

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/

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 connaitre
  - 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%20Fonsy%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" :
"EitBdmVudWUgRm9uc255IDIwLCAxMDYwIF
NpbnQtR2lsbGlzLCBCZWxnacOr",
         "types" : [ "street_address" ]
      }
   ],
   "status" : "OK"
}
```

Géocodage : Google Maps

Adresse standardisée

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

"formatted_
```

Géometrie

```
"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


mapquest


bing


here


NATIONAAL GEOGRAFISCH INSTITUUT
INSTITUT GEOGRAPHIQUE NATIONAL


OpenCage Data


Open
Street
Map

Géocodage : comparaison

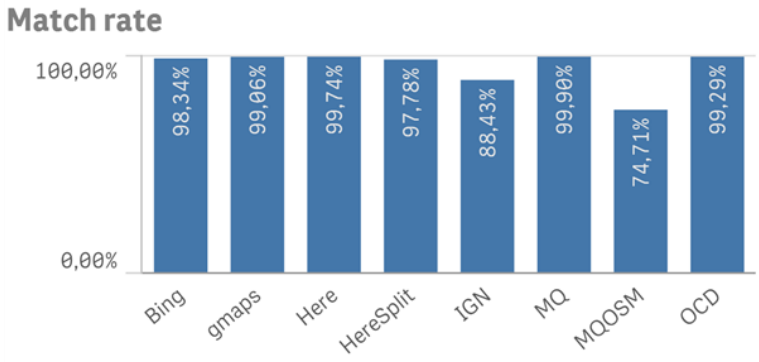
- Comparaison de services : difficile !
- Littérature : dataset d'adresses/coordonnées connues -> difficile à établir, forcément tronqué
- Alternative : comparer les réponses de 7 API
- Voir s'il y a consensus :

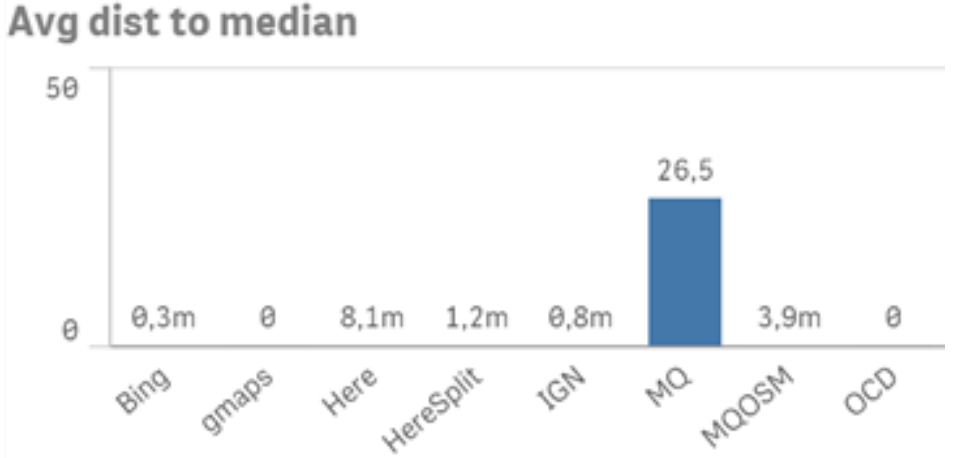
Mauvaise réponse ?

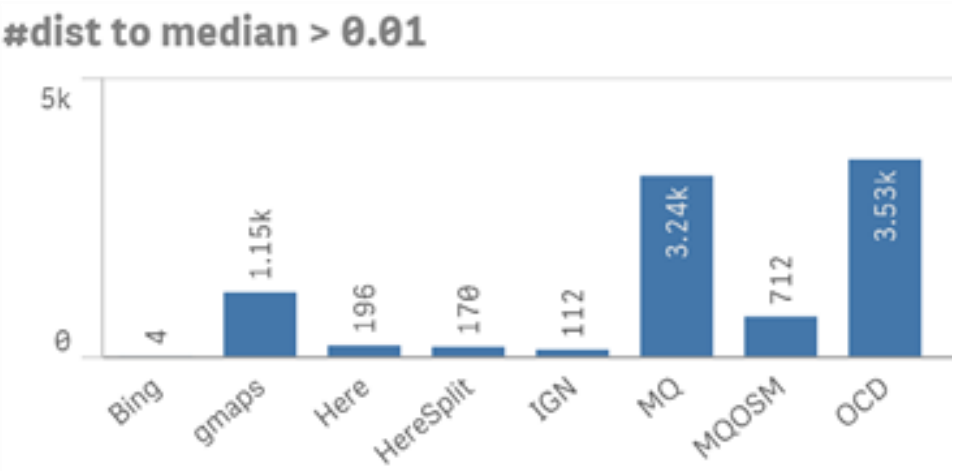
- Dataset : 10,000 adresses aléatoires de la CBE (KBO), Belgique (98,4%) + étranger

Données :

Géocoding : comparaison







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

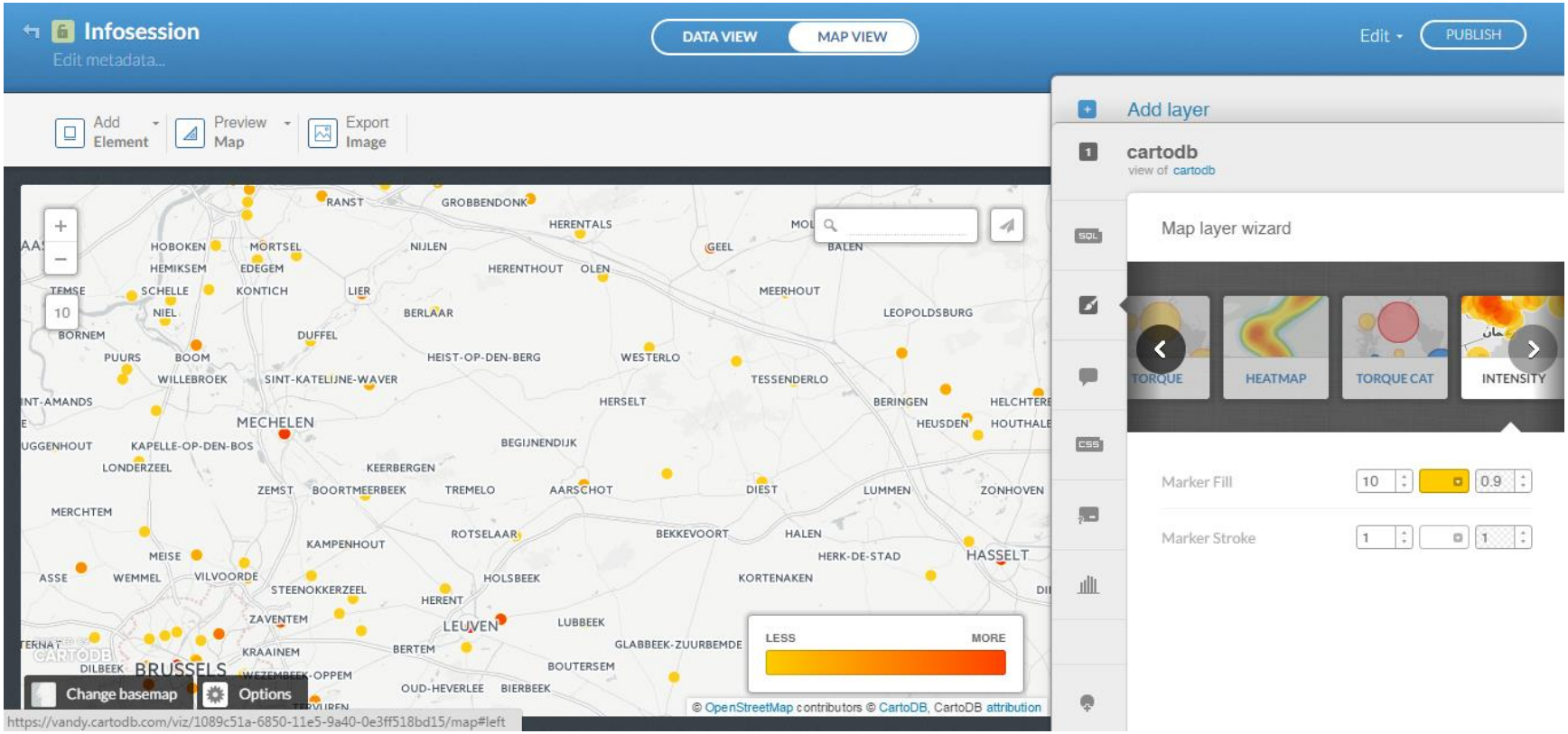
Géocodage

- 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

Démo CartoDB



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%

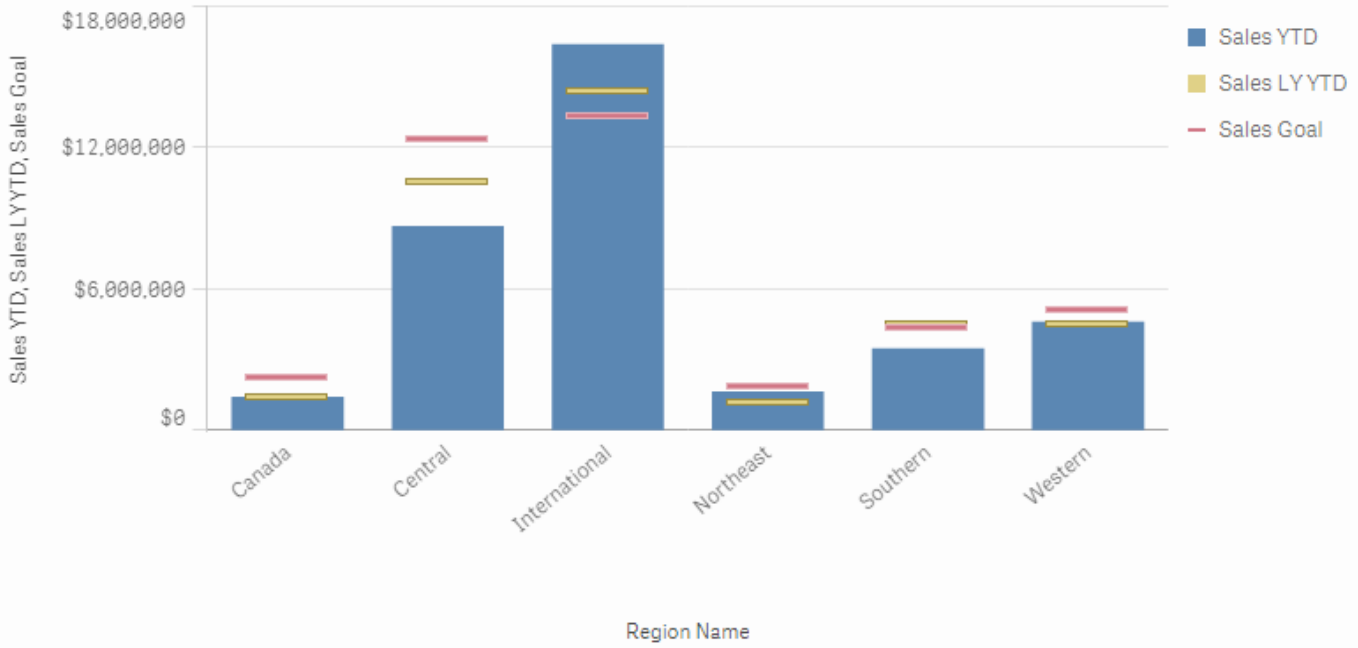
Sales vs Goal YTD

Budget YTD $39.44M LYTD $36.17M

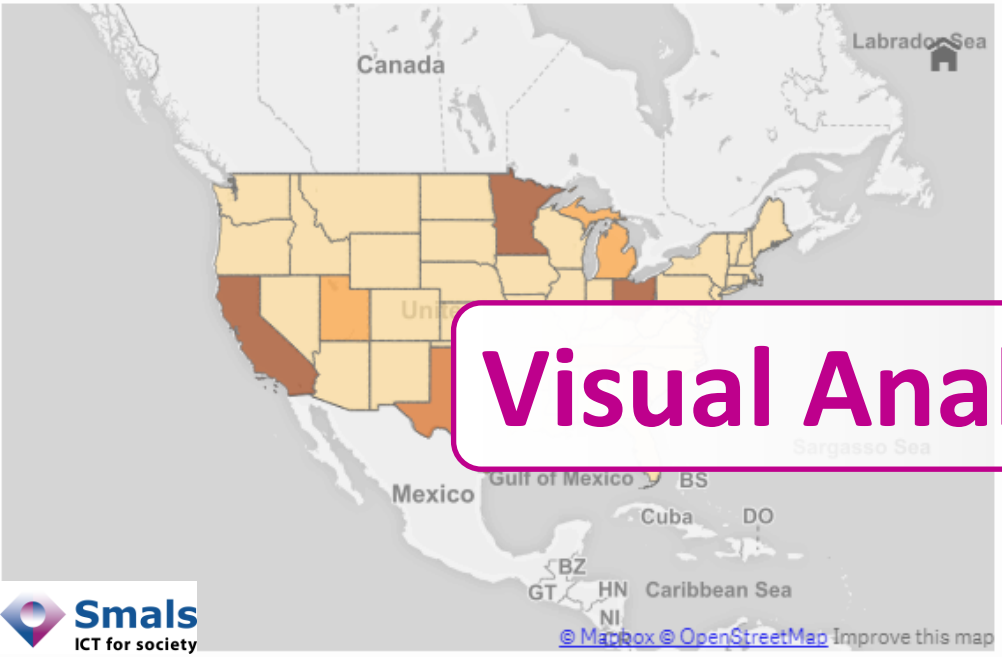
Sales vs Goal

-8.29%

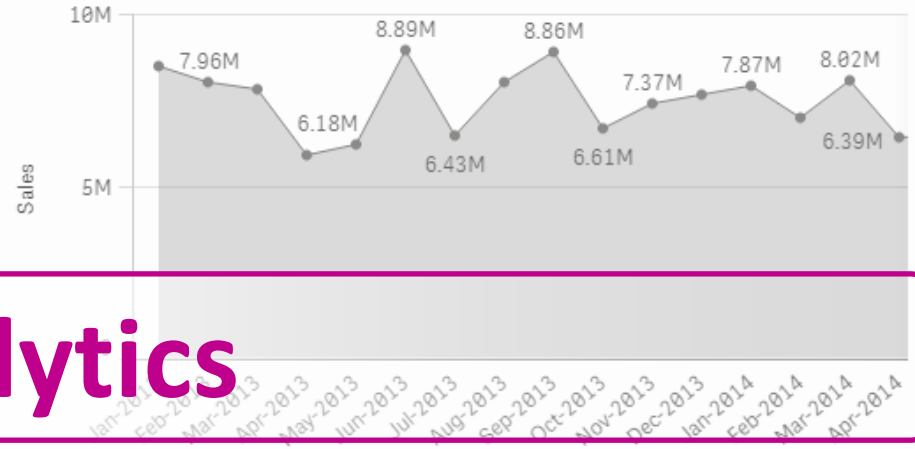
Sales: YTD vs LYTD vs YTD Goals



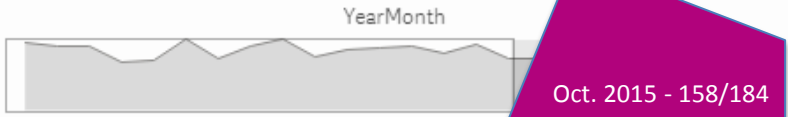
Sales by State



Sales Evolution



Visual Analytics



Smals
ICT for society

Visual Analytics

DÉFINITIONS

Visual Analytics : définition

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

Approche combinant :

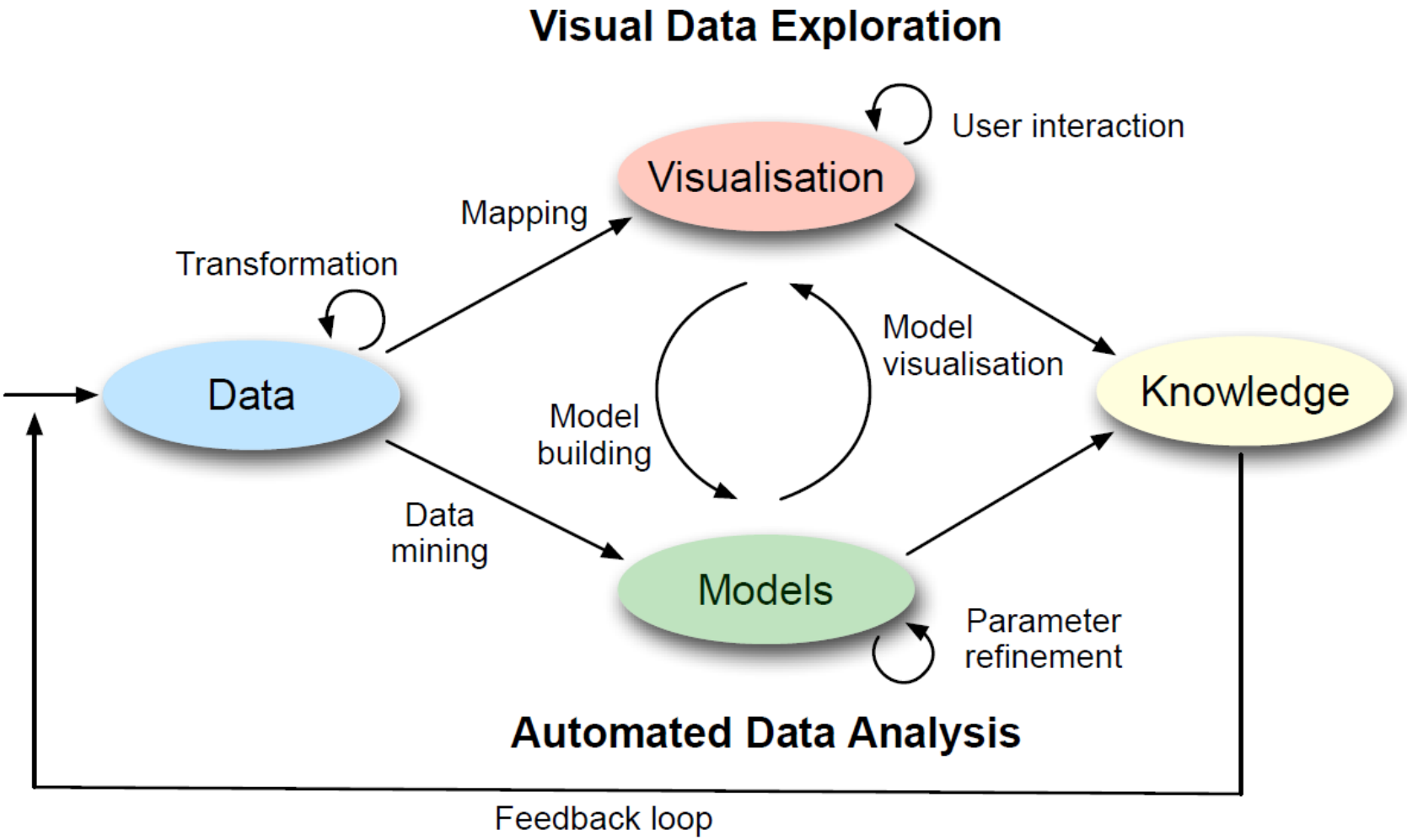
- Visualisation (dataviz, sciviz, comp. graphics…)
- Interaction (H-C interaction, cognitive psychology, perception…)
- Data analysis (Information retrieval, data mining, information/geospatial/scientific/statistical analytics…)

*[ITP]

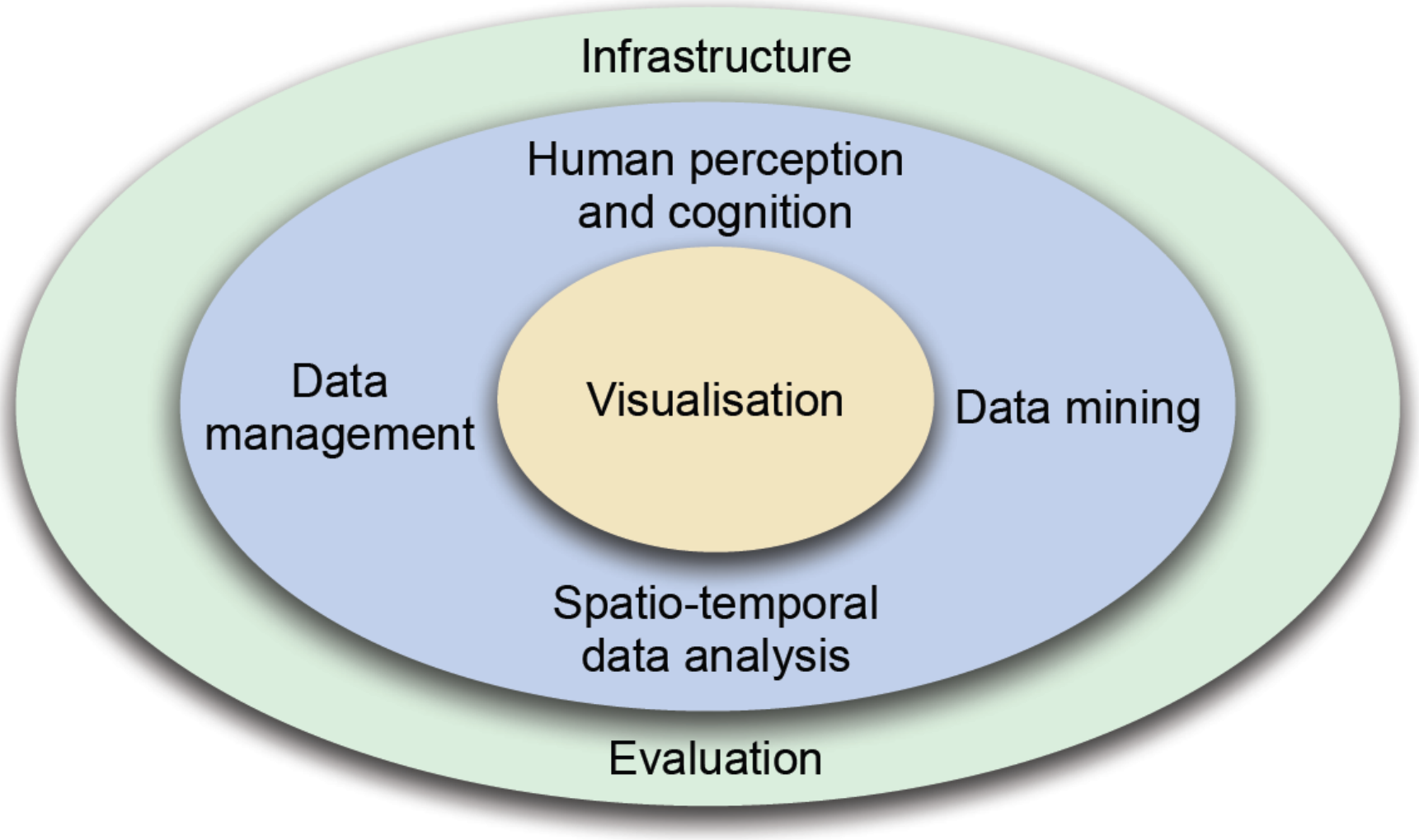
Objectifs

- Synthétiser l'information et en tirer une intuition (*insight)* de données massives, dynamiques, ambigües 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



Domaines


Infrastructure
Human perception and cognition
Data management
Visualisation
Data mining
Spatio-temporal data analysis
Evaluation

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

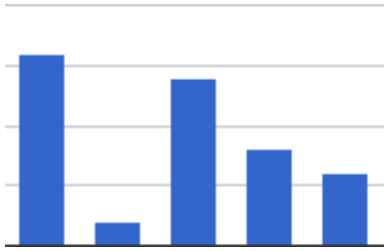
- Perception & Cognition
  - Perception : Interprétation de l'environnement (5 sens)
  - Cognition : capacité de comprendre (basée sur l'apprentissage)
- Infrastructure
  - Liaison des différents processus et technologies, souvent incompatibles
- Évaluation

Données brutes

- Libraires graphiques :

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

Javascript

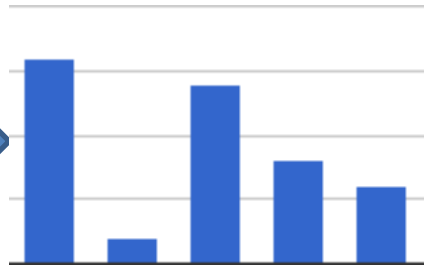


- Outils Visual Analytics :

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

Key	Attr
X1	A
X2	C
X3	B

SELECT Attr, COUNT (Id)
FROM … LEFT JOIN …
GROUP BY Attr



In-Memory

	Live	In-Memory
Principe	Connexion à la source pour chaque affichage	*Extract* local de ce qui est nécessaire au *dashboard*
Requêtes	Sur la source	Localement
Mises à jour	Permanentes	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

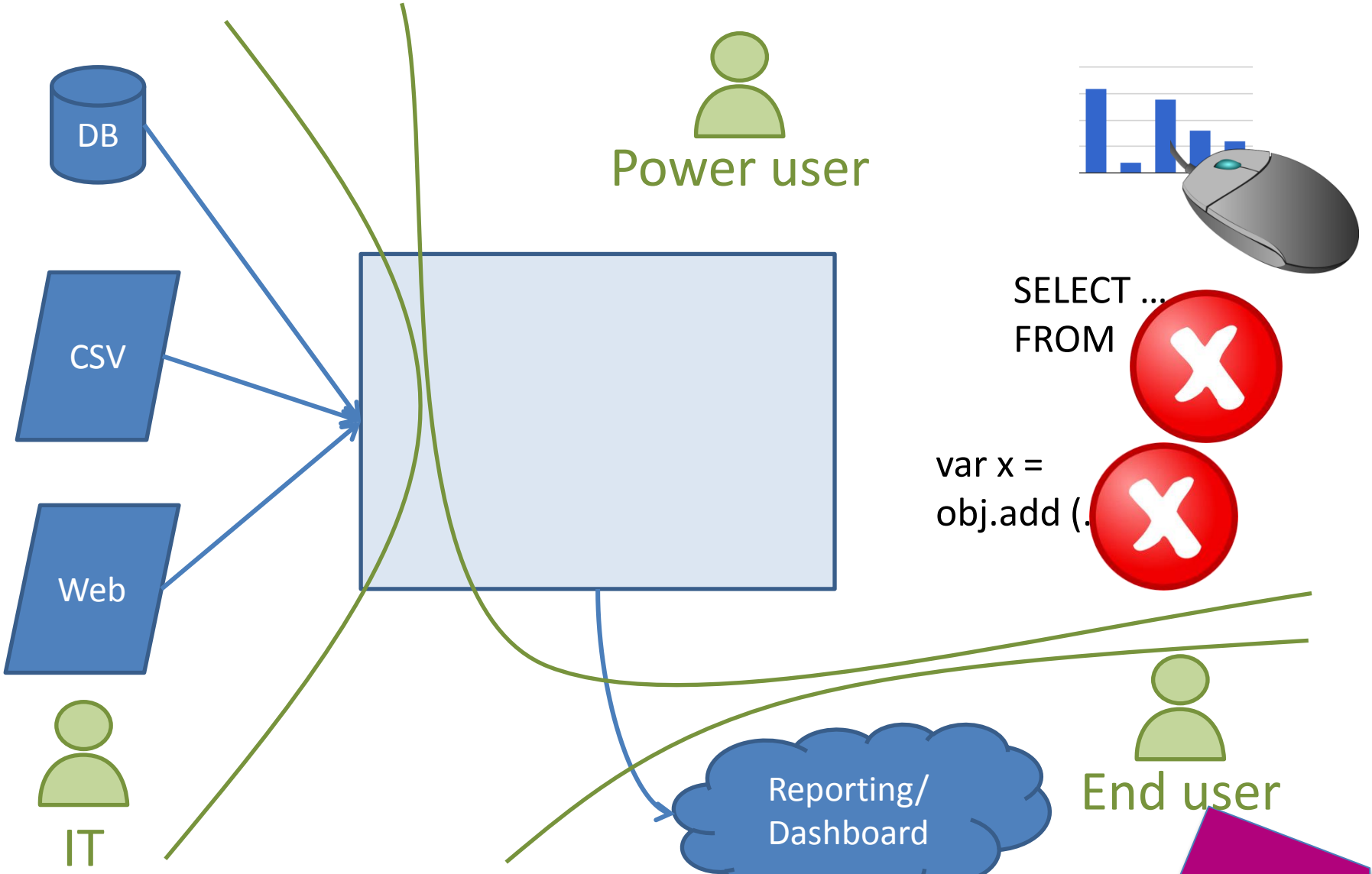

DB
CSV
Web
IT
Power user
SELECT …
FROM
var x =
obj.add (.
Reporting/
Dashboard
End user

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

- 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
- QlikSense (Gratuit)
  - « Self-service BI »
- *In-Memory (*ou hybride *« direct query »)*
- *Discovery & Reporting*, peu de *data analytics*
- Choix du graphique puis des *dimensions*
- Beaucoup d'extensions (QlikMaps…)
- Drag&Drop ou Script

Use cases Qlik & Tableau

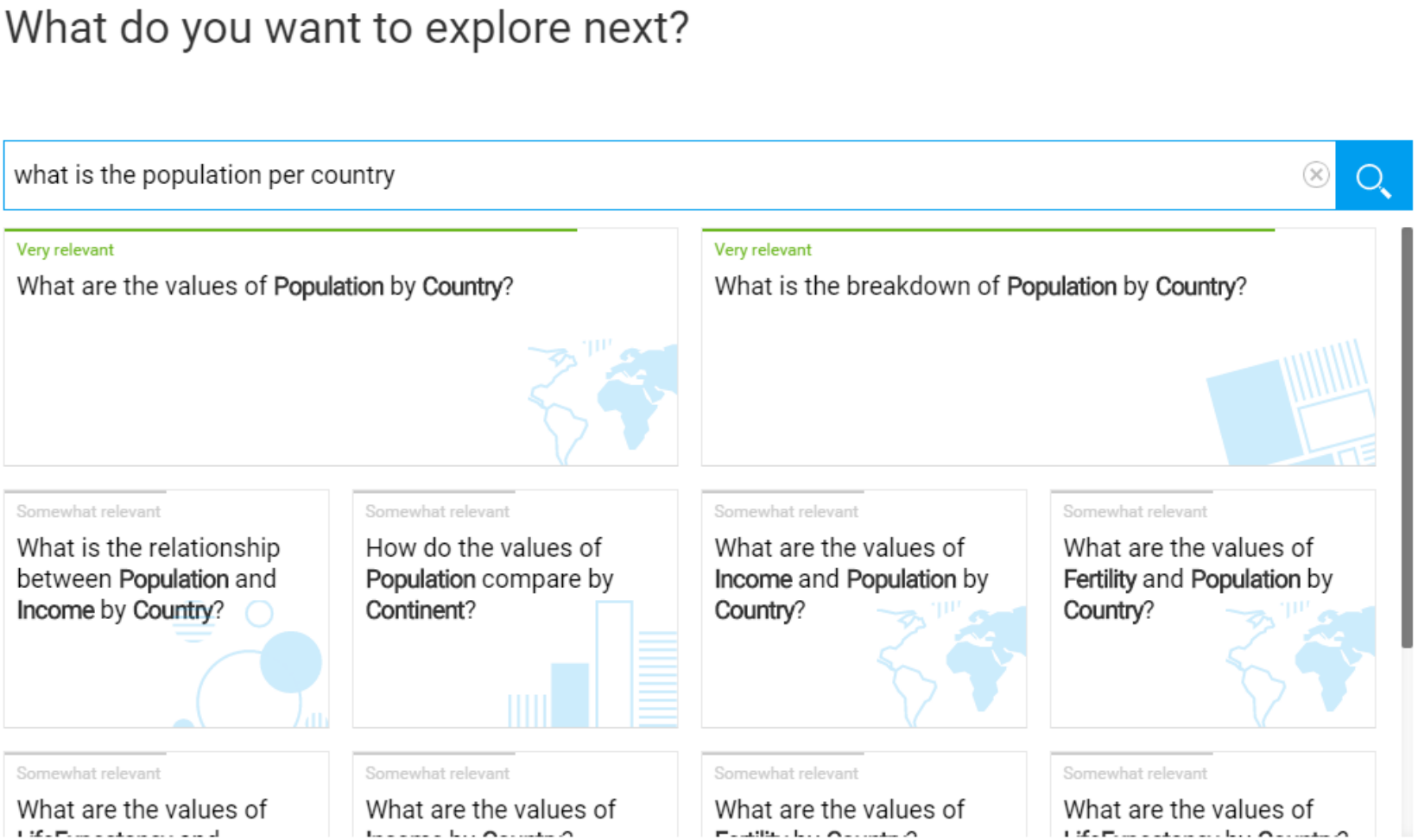
	Qlik	Tableau
Usage gratuit	- QlikSense (Partage via « Qlik Cloud » ou par envoi des fichiers) - QlikView Personal Ed. (Pas de partage)	Tableau Public : Client desktop, données publiques sur un *cloud*

Usage interne	- Named User Licence + partage fichiers
	- QlikView Server :
	  - Small Business Ed. (≤25 users)
	  - Enterprise Edition
		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



Démo Qlik Sense

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

Qlik® Sense

OpenSource

bpost

ONSS

KML_commune
Commune
Contours

KML_arrond
Arrondissement
Contours

KML_prov
Province
Contours

Zip
ZipCode
Commune

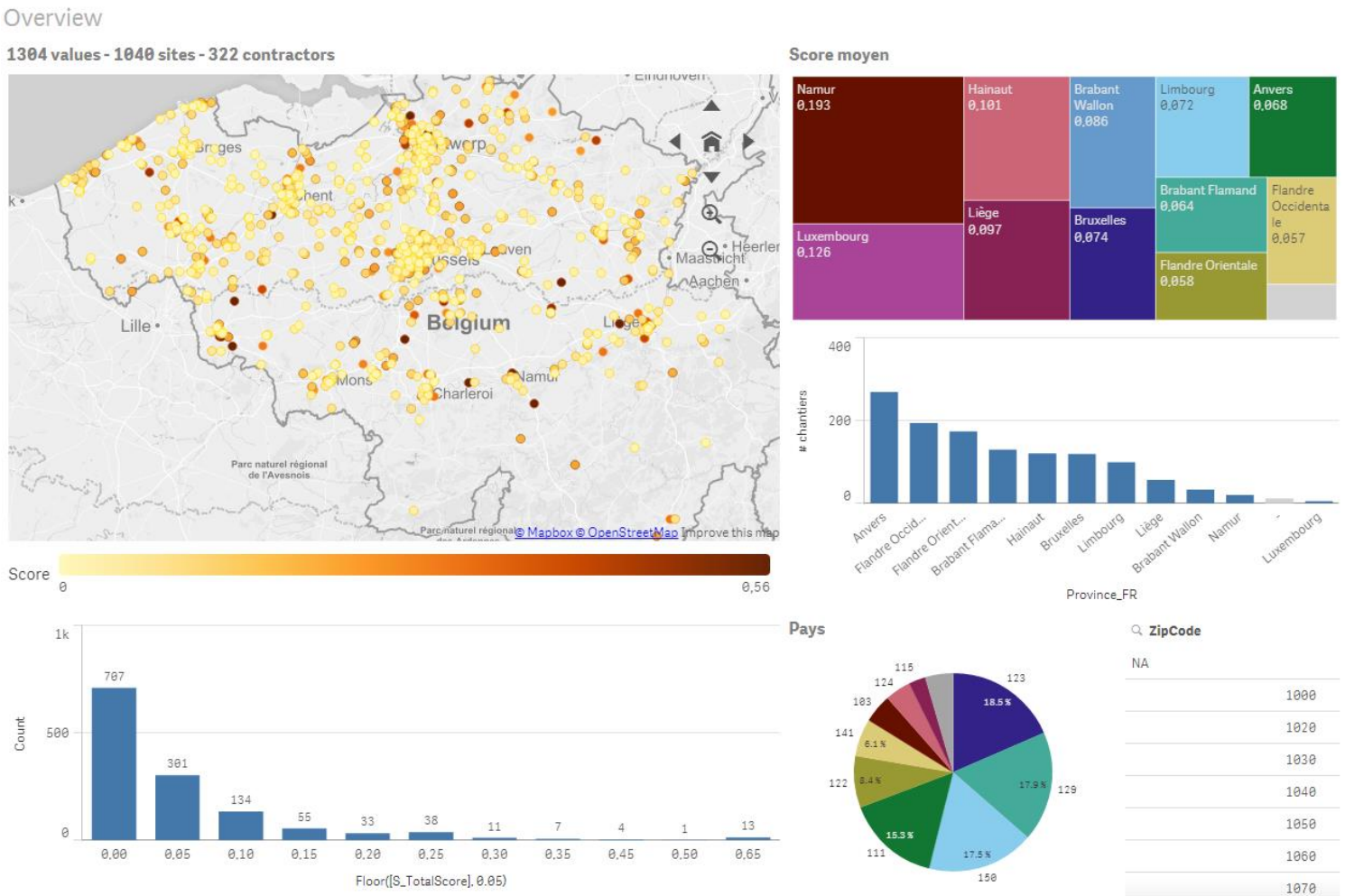
Commune
Commune
Arrondissement

Arrond
Arrondissement
Province

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

Démo

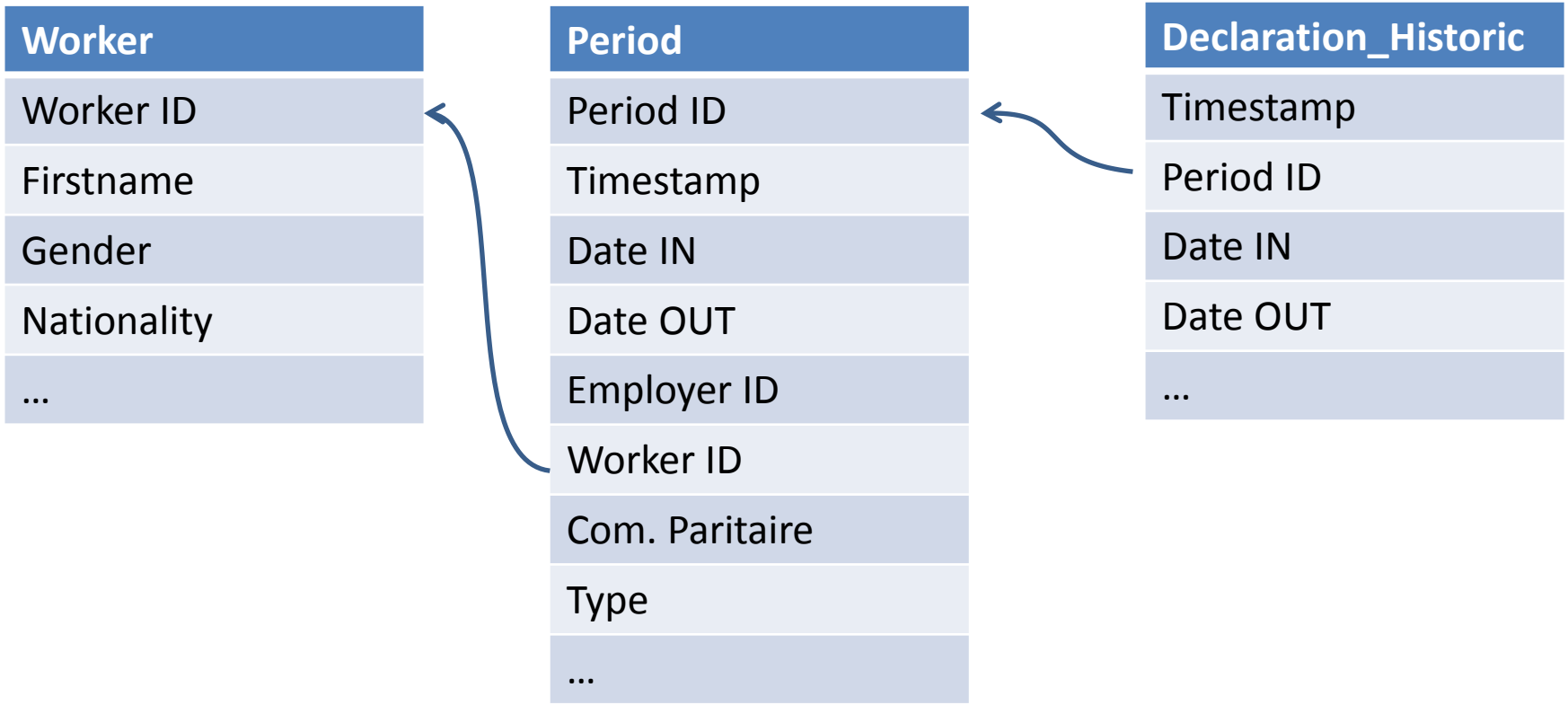
Qlik® Sense



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

More on Smals Research :
Website : www.smals.be
Blog : www.smalsresearch.be
Twitter : @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