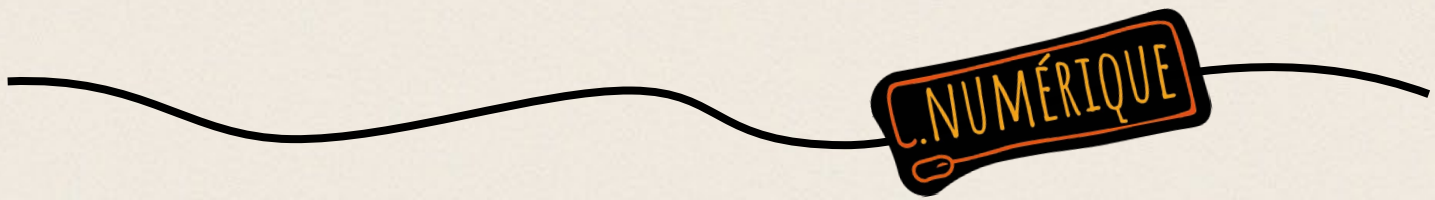
Visualiser les données de mon corpus

Victoria Le Fourner, Florence Perret

Visualiser les données de mon corpus



MESHS Lille Nord de France

Victoria **Le Fourner**

Florence **Perret**

Présentation de l'équipe organisatrice

Victoria Le Fourner
Ingénieure d'études chargée du traitement des données scientifiques
Huma-Num

Contact : victoria.lefourner@univ-lille.fr

Florence Perret
Ingénieure d'études chargée des humanités numériques

Contact : florence.perret@univ-lille.fr

PLAN DE LA FORMATION

1. Introduction à la visualisation de données | 20 min
2. Présentation de Palladio | 20 min
3. Présentation de Carto | 10 min
4. Atelier : préparer ses données | 40 min

Qu'est-ce que la Visualisation de données ?

Définition et rapide historique

La data visualization (visualisation de données, représentation graphique de données) est un ensemble de méthodes visant à structurer et résumer de manière graphique des données recueillies et stockées pour permettre l'accès à une nouvelle compréhension de ce jeu de données.

1662 John Graunt publie les 1ères tables de mortalité recensant les décès à Londres selon l'âge : début de l'histoire des statistiques sociales

1786 William Playfait publie *The Commercial and Political Atlas*. Première représentation moderne de données et premier diagramme en bâtons.

1858 Florence Nightingale créée le *Diagramme des causes de mortalité au sein de l'armée en Orient* et améliore les diagrammes circulaires de William Playfair (1801)

1869 Charles Minard réalise sa *Carte figurative des pertes successives en hommes de l'armée française dans la campagne de Russie 1812-1813*

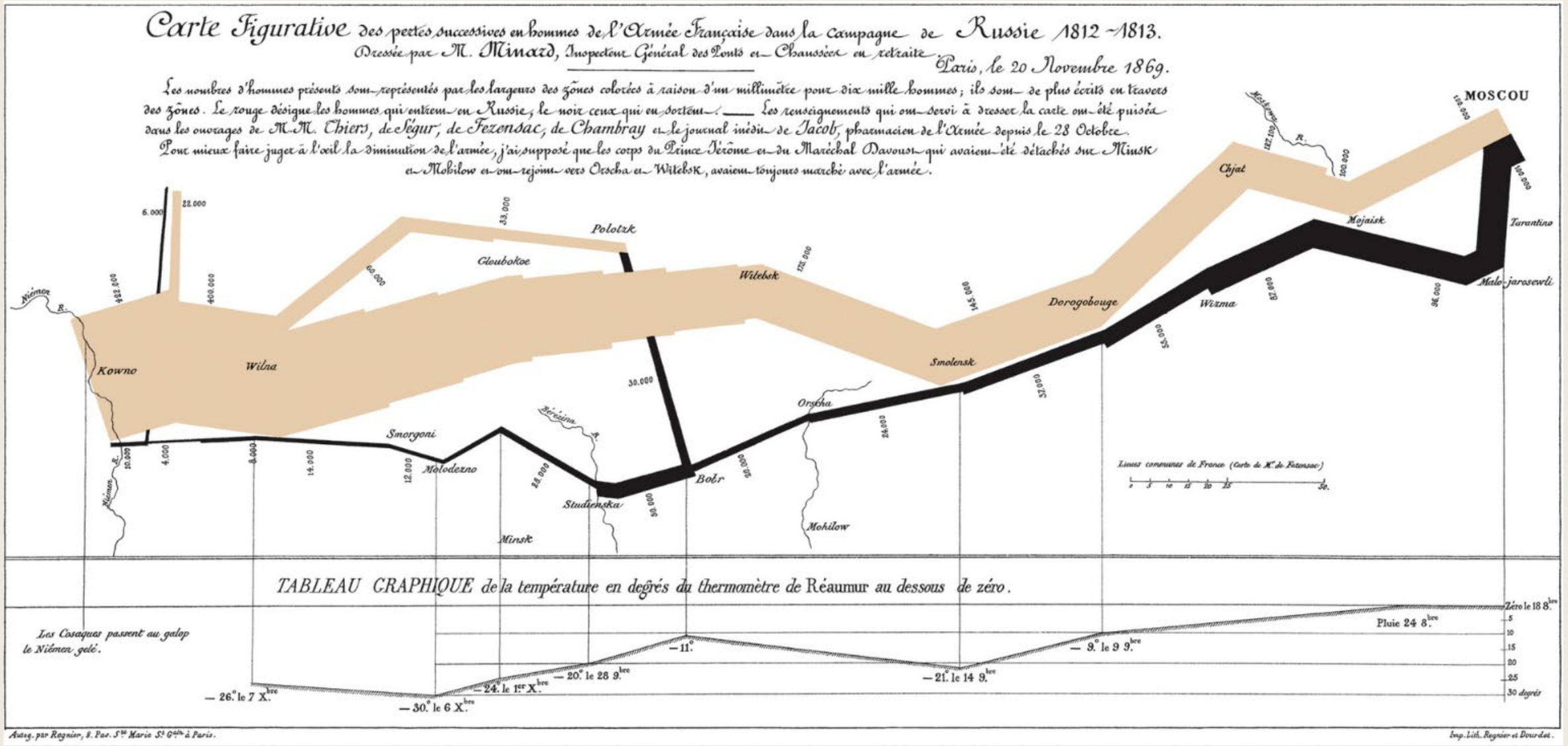
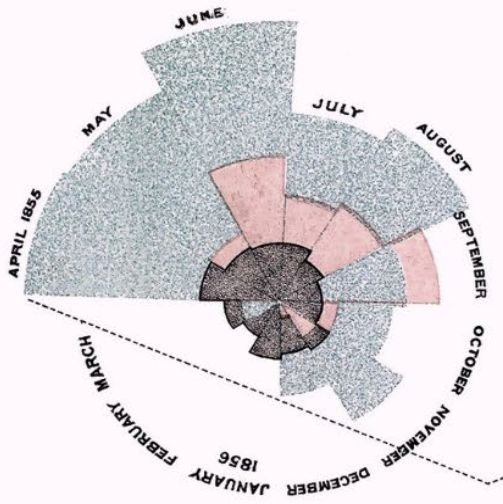
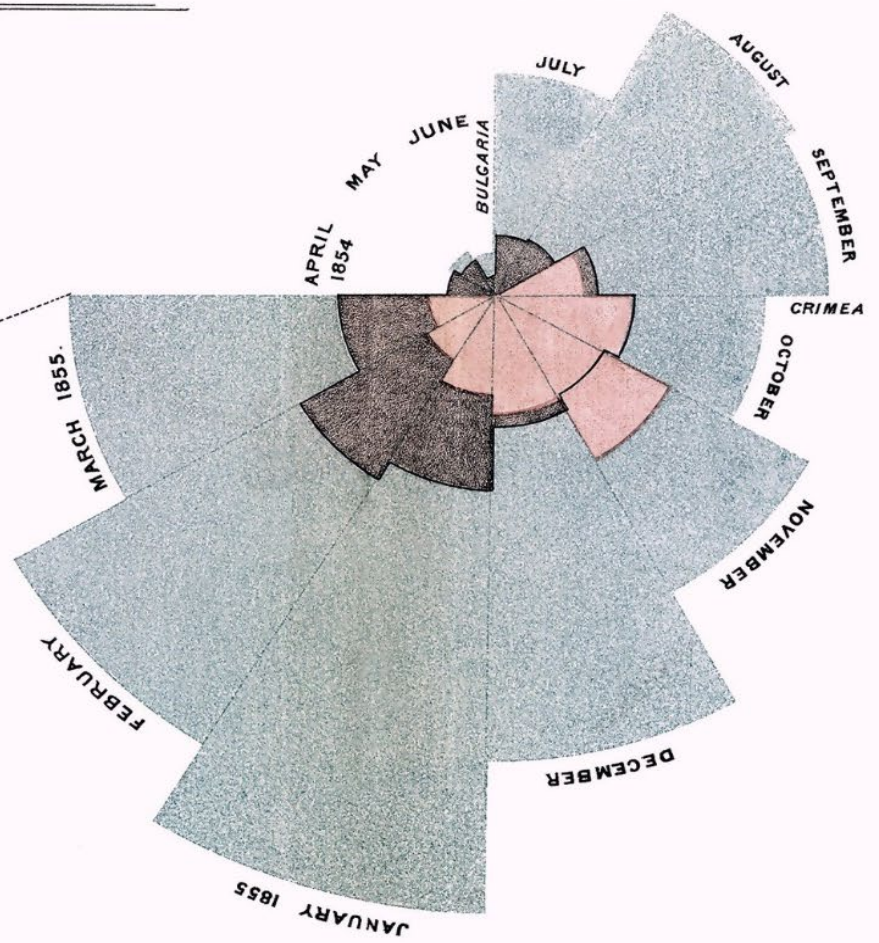

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. ___ Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow en ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
Niémen R.
Kowno
Wilna
Smorgoni
Molodezno
Minsk
Studienska
Bérézina R.
Polotzk
Gloubokoe
Witebsk
Bobr
Orscha
Mohilow
Smolensk
Dorogobouge
Wiazma
Chjat
Mojaisk
MOSCOU
Tarantino
Malo-jaroswetz
Kalouga R.
Lieues communes de France (Carte de Mr de Fezensac)
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
— 26.° le 7 X.bre
— 30.° le 6 X.bre
— 24. le 1er X.bre
— 20.° le 28 9.bre
— 11.°
— 21. le 14 9.bre
— 9.° le 9 9.bre
Pluie 24 8.bre
Zéro le 18 8.bre
30 degrés
Autog. par Regnier, 8. Pas. St Marie St Gde à Paris.
Imp. Lith. Regnier et Dourdet.

DIAGRAM of the CAUSES of MORTALITY IN THE ARMY IN THE EAST.

2.
APRIL 1855 to MARCH 1856.



1.
APRIL 1854 to MARCH 1855.



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

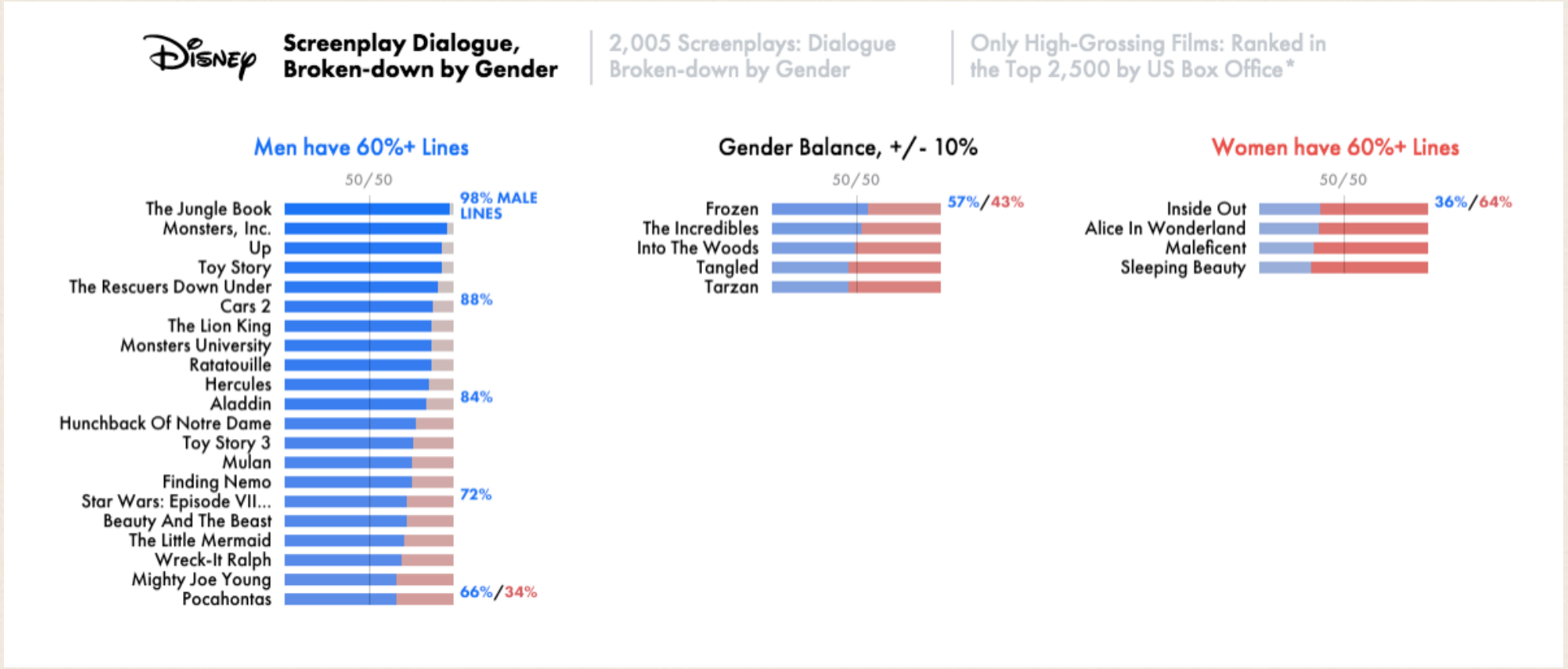
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Interacting With History
300 years of information graphic milestones

Chronologie interactive des travaux de visualisation de données majeurs de 1630 à 1904.
Réalisée par le studio d'infographie Info We Trust en reprenant le forme du Britannia Atlas de John Ogilby (1675)

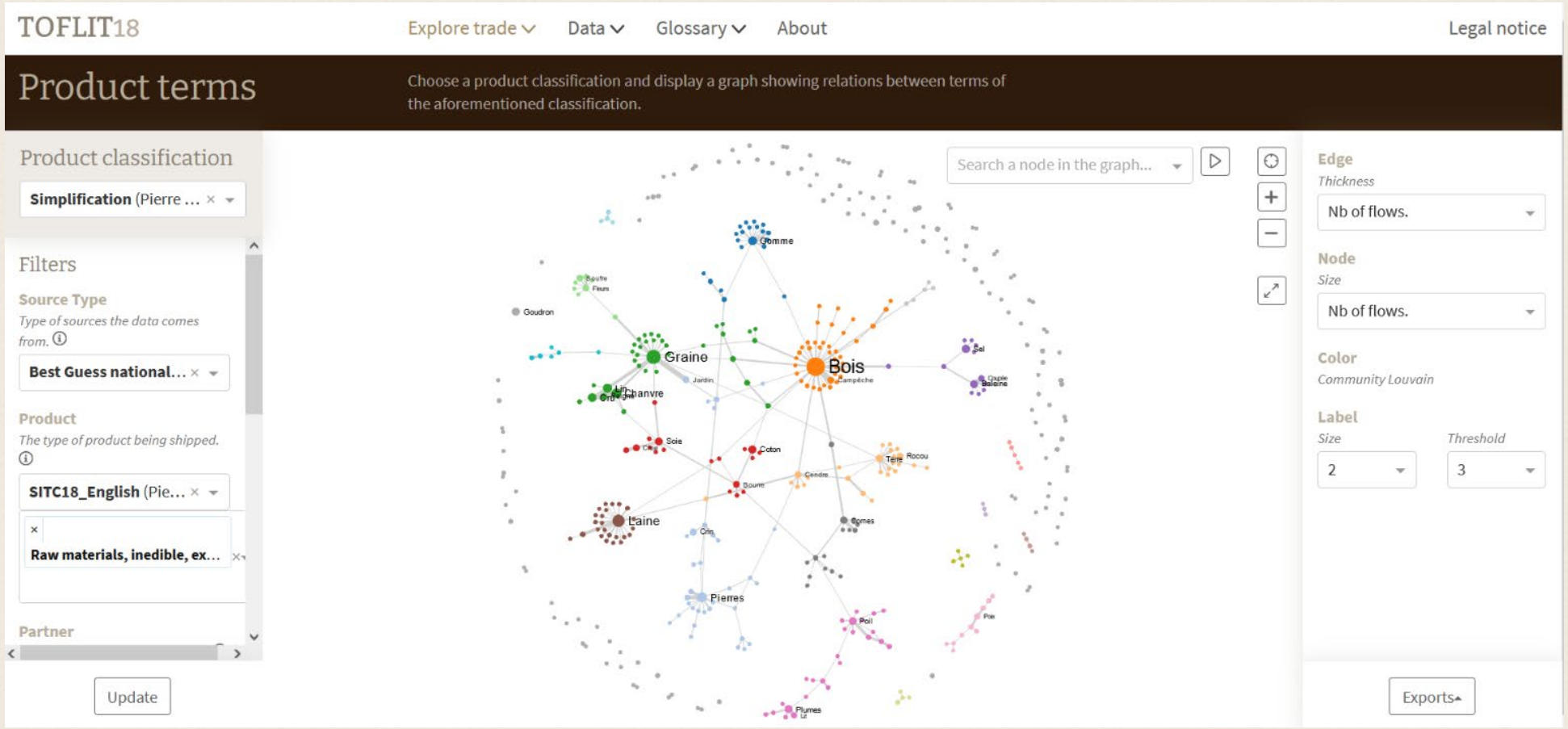
Quelques exemples contemporains

- Les inégalités de genre dans les dialogues des films Disney (The Pudding – 2016)



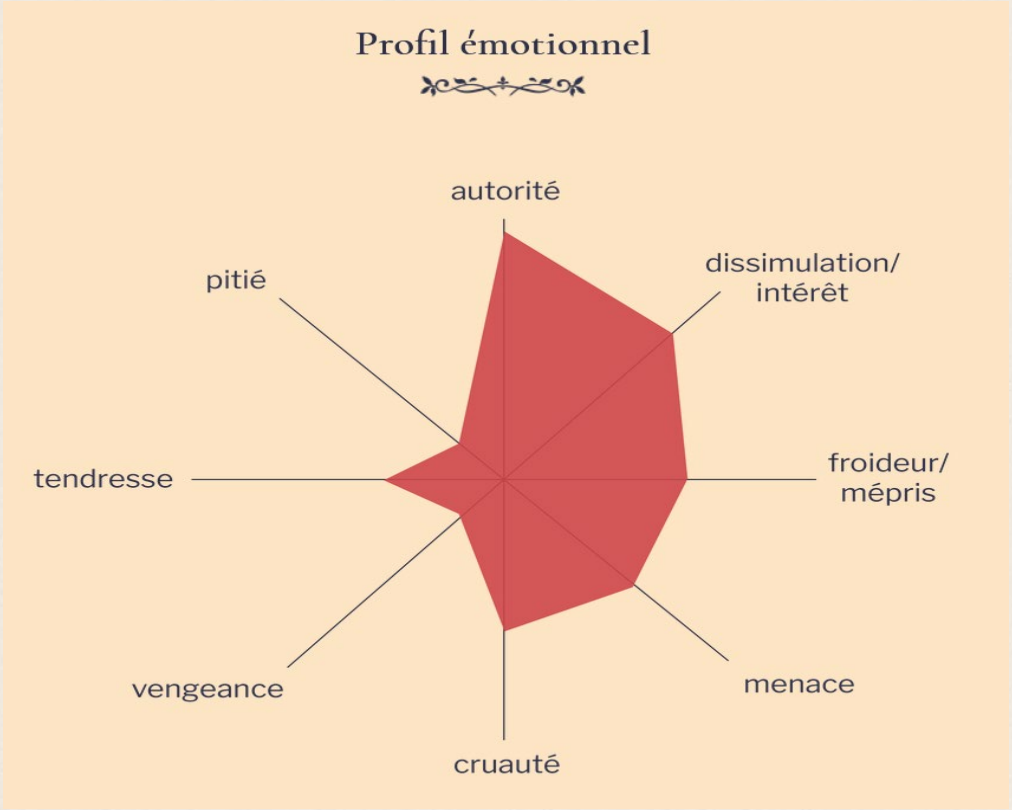
Quelques exemples contemporains

- Le datascape de la base de données TOFLIT18 (ANR – 2014/2017)



Quelques exemples contemporains

- Profils émotionnels des personnages de la pièce *Charles IX* de Marie-Joseph Chénier (1789) – projet de médiation autour du projet des Registres de la Comédie Française



Profil émotionnel de Catherine de Médicis, réalisé par Thibaut Julian, postdoctorant en histoire et littérature française (EHESS) et Julie Machu (étudiante - graphiste)

Pourquoi visualiser ?

Pourquoi visualiser ?

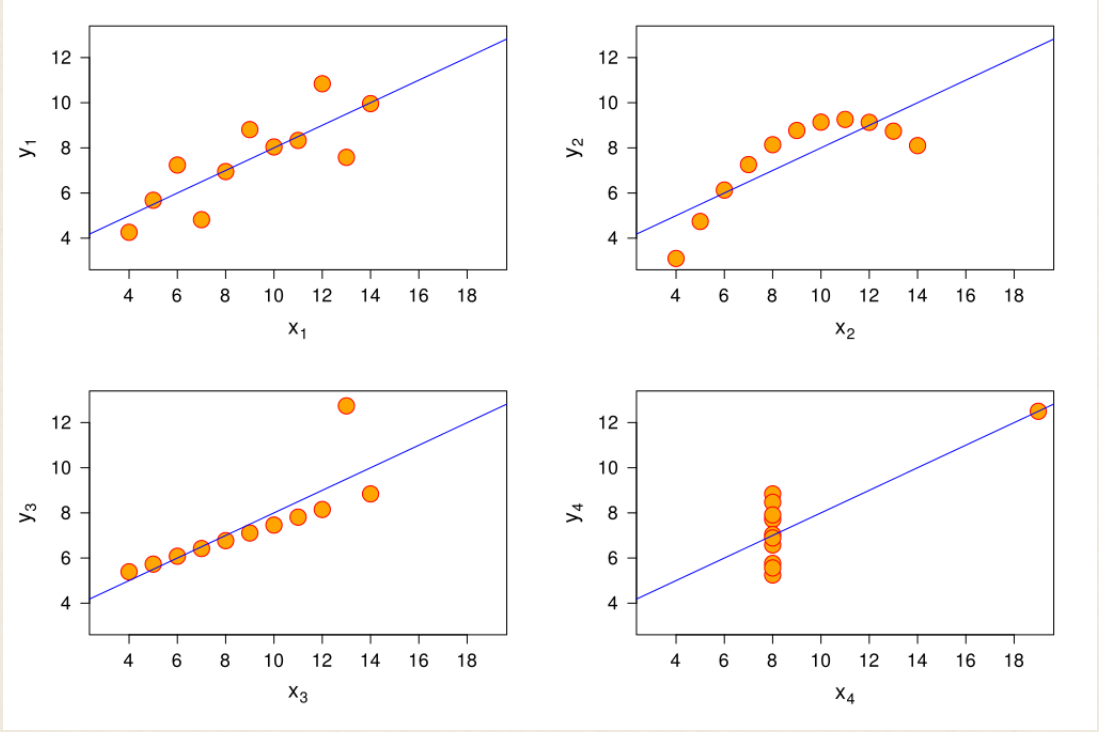
« The greatest value of a picture is when it forces us to notice what we never expected to see. »

Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company.

Pourquoi visualiser ?

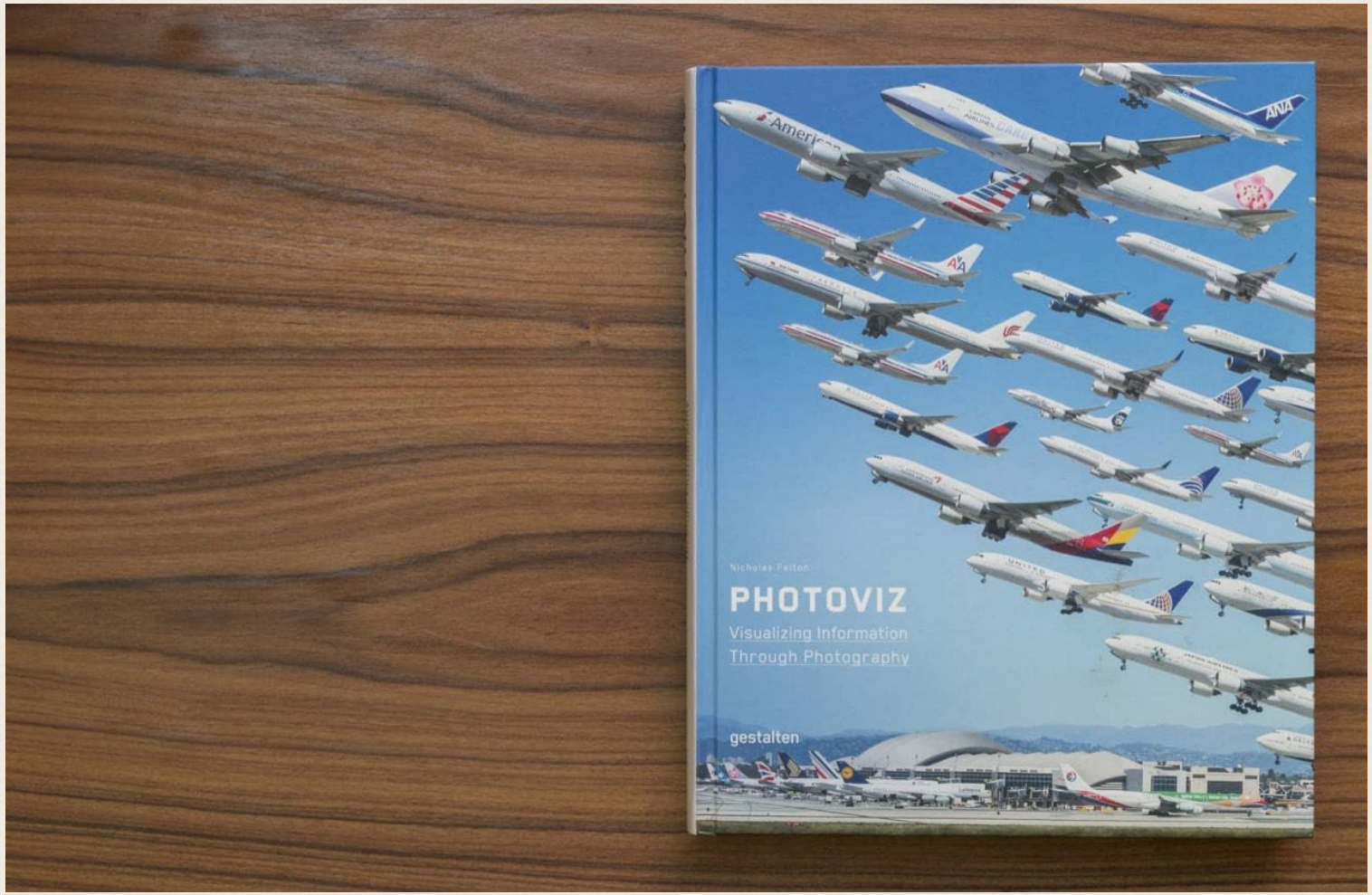
Quartet d'Anscombe

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89



Le **quartet d'Anscombe** (1973) montre bien l'intérêt de l'usage de la visualisation dans le contexte d'une étude statistique.

Enregistrer l'information



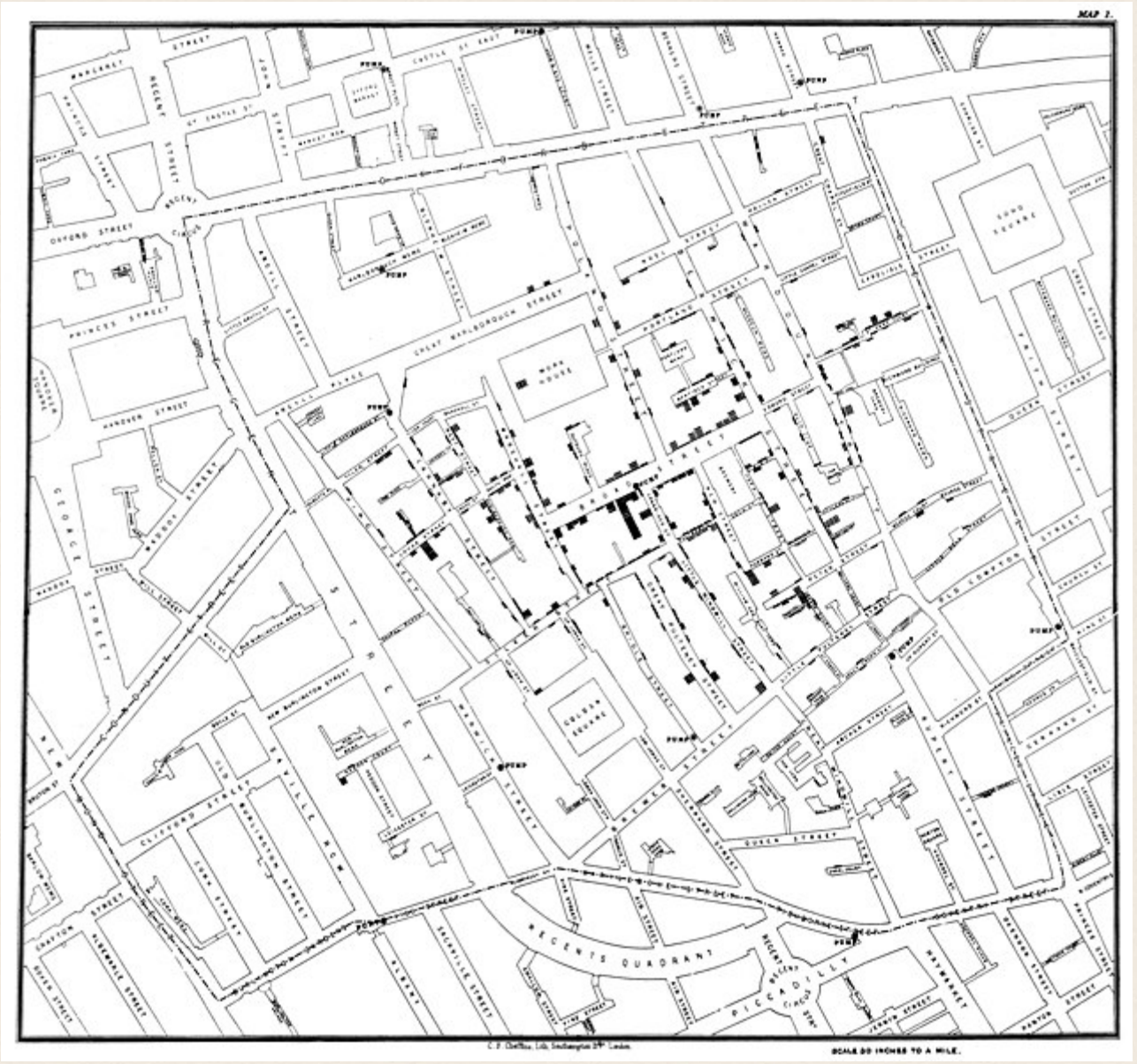
On peut vouloir visualiser pour simplement enregistrer une information de la manière la plus lisible possible

"nearly a day's worth of aircraft movements merged into one visual experience"

Photographie de Mike Kelley, *PhotoViz* (2016)

Analyser l'information

On peut aussi vouloir visualiser pour pouvoir analyser et comprendre

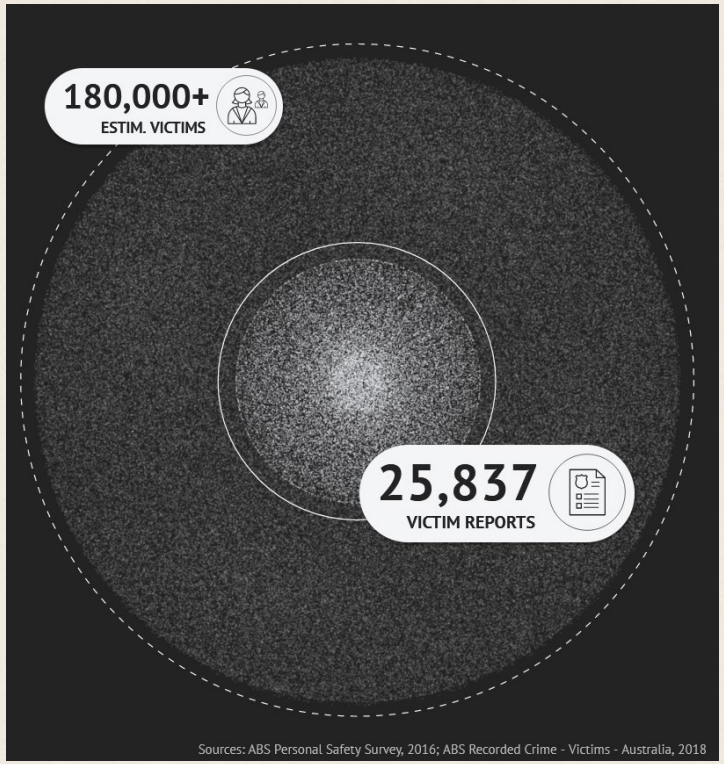


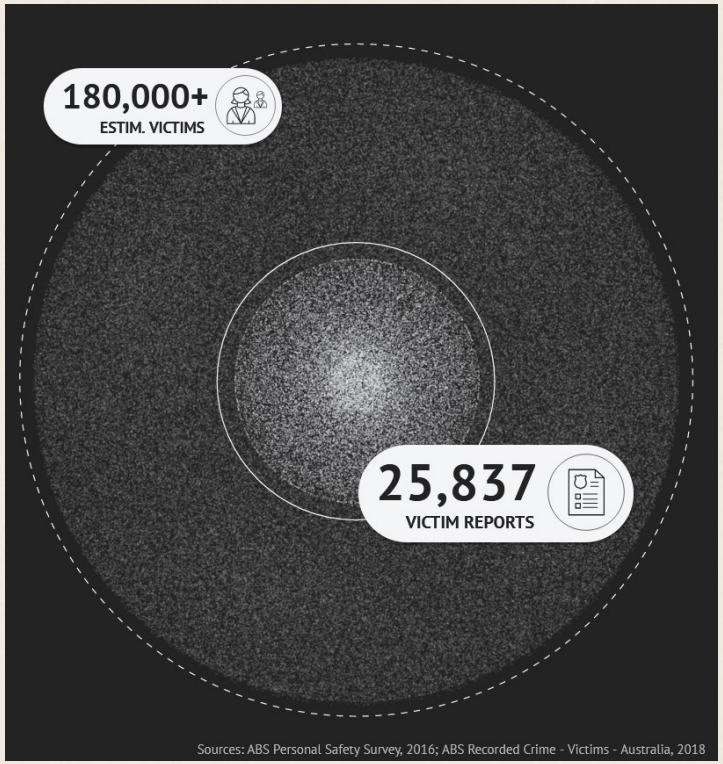
[Carte originale réalisée par John Snow en 1854]

Les cas de choléra sont surlignés en noir

Transmettre l'information

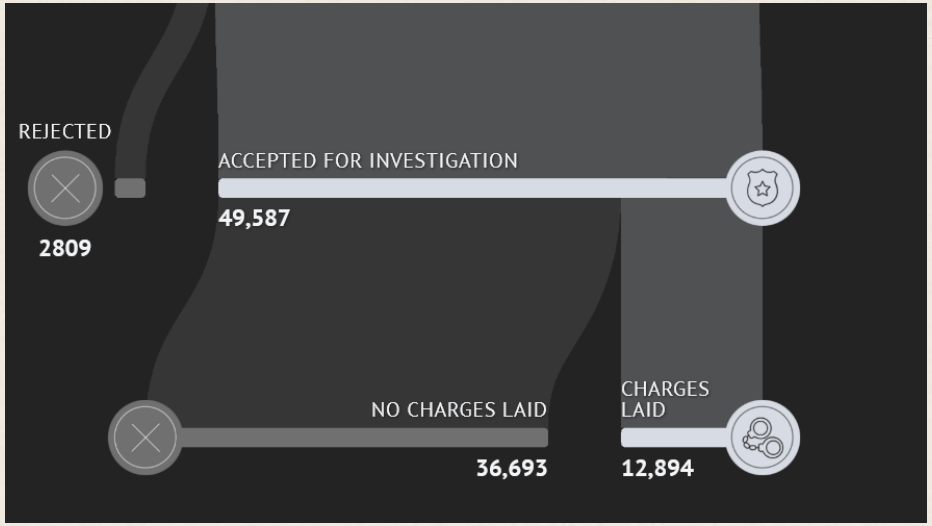
la visualisation de données peut permettre une médiation de l'information



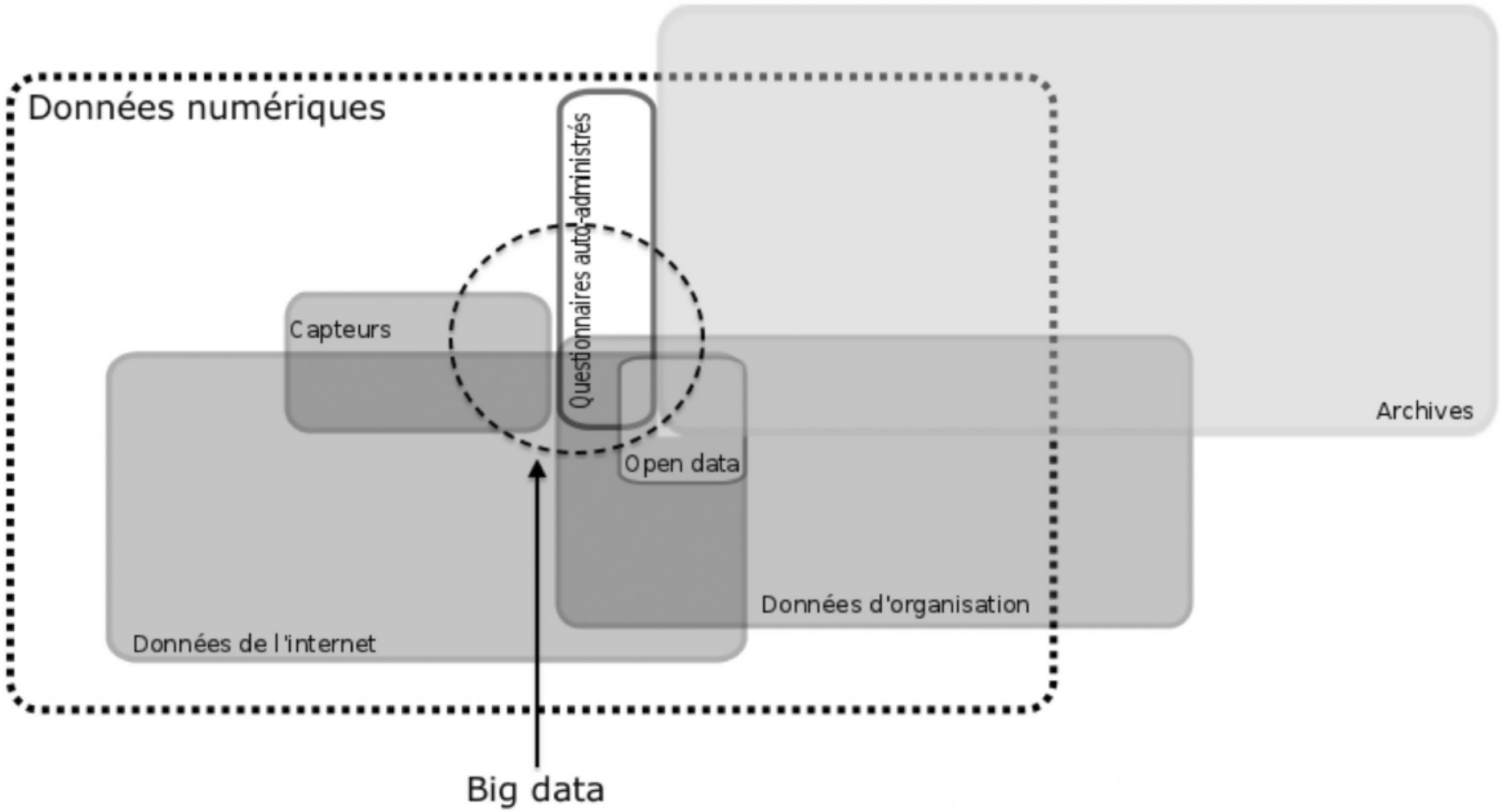


[The Age, « Are we failing victims of sexual violence ? »]

Infographie en scrollytelling



L'explosion des données



Au delà des big data, les principales sources numériques en sciences sociales

Étienne Ollion et Julien Boelaert, « Au-delà des big data », *Sociologie* (2015)

L'EXPLOSION DES DONNÉES

- Comment faire sens des données ?
- Comment utiliser ces données dans les processus de décision ?
- Comment ne pas être surchargé ?

Défi: transformer les données en connaissance (découverte, compréhension) pour qu'elles deviennent utiles.

Les règles de la visualisation de données

- Gagner du temps
- Objectif clair :
  - Quelles sont nos données ?
  - Que veut-on montrer ?
  - Qui sont les futurs utilisateurs ?
  - Quels sont leurs besoins ?

- Encodage approprié des données :
  - Comment montrer, comment dessiner ?
  - Comment manipuler ? Quelles interactions ?

Les types de visualisations

Visualisation d'une variable :

- graphique linéaire
- diagramme à barres
- histogramme
- graphique circulaire
- diagramme en radar/en toile d'araignée

Visualisation de deux variables :

- graphique en nuage de points
- heatmap/carte de fréquentation
- carte

Visualisation de plus de deux variables :

- matrice de nuage de points
- graphique en mosaïque/Marimekko chart
- Coordonnées parallèles

Visual vocabulary

Designing with data

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

FT graphic: Alan Smith, Chris Campbell, Ian Bott, Liz Faunce, Graham Parrish, Billy Ehrenberg-Shannon, Paul McCallum, Martin Stabe
Inspired by the Graphic Continuum by Jon Schwabish and Severino Ribecca

ft.com/vocabulary

Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar
A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar
Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine
Splits a single value into two contrasting components (eg male/female).

Surplus/deficit filled line
The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment; income and life expectancy

Scatterplot
The standard way to show the relationship between two continuous variables, each of which has its own axis.

Column + line timeline
A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot
Usually used to show how the relationship between 2 variables has changed over time.

Bubble
Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap
A good way of showing the patterns between 2 categories of data, less effective at showing fine differences in amounts.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

Ordered bar
Standard bar charts display the ranks of values much more easily when sorted into order.

Ordered column
See above.

Ordered proportional symbol
Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot
Dots placed in order on a strip are a space-efficient method of laying out values across multiple categories.

Slope
Perfect for showing how ranks have changed over time or vary between categories.

Lollipop
Lollipops draw more attention to the data value than standard bar/column and can also show rank and value effectively.

Bump
Effective for showing changing rankings across multiple dates. For large datasets, consider grouping lines using colour.

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (age/sex) distribution, revealing inequality

Histogram
The standard way to show a statistical distribution – keep the gaps between columns small to highlight the 'shape' of the data.

Dot plot
A simple way of showing the change or range (min/max) of data across multiple categories.

Dot strip plot
Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Barcode plot
Like dot strip plots, good for displaying all the data in a table, they work best when highlighting individual values.

Boxplot
Summarise multiple distributions by showing the median (centre) and range of the data

Violin plot
Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple average).

Population pyramid
A standard way for showing the age and sex breakdown of a population distribution; effectively, back to back histograms.

Cumulative curve
A good way of showing how unequal a distribution is; y axis is always cumulative frequency, x axis is always a measure.

Frequency polygons
For displaying multiple distributions of data. Like a regular line chart, best limited to a maximum of 3 or 4 datasets.

Beeswarm
Use to emphasise individual points in a distribution. Points can be sized to an additional variable. Best with medium-sized datasets

Change over Time

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

Example FT uses
Share price movements, economic time series, sectoral changes in a market

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

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

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

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

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

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

Fan chart (projections)
Use to show the uncertainty in future projections - usually this grows the further forward to projection.

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

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

Priestley timeline
Great when date and duration are key arrangement of the elements of the story in the data.

Circle timeline
Good for showing discrete values of varying size across multiple categories (eg earthquakes by continent).

Vertical timeline
Presents time on the Y axis. Good for displaying detailed time series that work especially well when scrolling on mobile.

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

Streamgraph
A type of area chart; use when seeing changes in proportions over time is more important than individual values

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Example FT uses
Commodity production, market capitalisation, volumes in general

Column
The standard way to compare the size of things. Must always start at 0 on the axis.

Bar
See above. Good when the data are not time series and labels have long category names.

Paired column
As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar
See above.

Marimekko
A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol
Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictogram)
Excellent solution in some instances – use only with whole numbers (do not slice off an arm to represent a decimal).

Lollipop
Lollipop charts draw more attention to the data value than standard bar/column – does not have to start at zero (but preferable).

Radar
A space-efficient way of showing value of multiple variables- but make sure they are organised in a way that makes sense to reader.

Parallel coordinates
An alternative to radar charts – again, the arrangement of the variables is important. Usually benefits from highlighting values.

Bullet
Good for showing a measurement against the context of a target or performance range.

Grouped symbol
An alternative to bar/column charts when being able to count data or highlight individual elements is useful.

Part-to-whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

Stacked column/bar
A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Marimekko
A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie
A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut
Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg total).

Treemap
Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments.

Voronoi
A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Arc
A hemicycle, often used for visualising parliamentary composition by number of seats.

Gridplot
Good for showing % information, they work best when used on whole numbers and work well in small multiple layout form.

Venn
Generally only used for schematic representation.

Waterfall
Can be useful for showing part-to-whole relationships where some of the components are negative.

Spatial

Aside from locator maps only used when precise locations or geographical patterns in data are more important to the reader than anything else.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

Basic choropleth (rate/ratio)
The standard approach for plotting data on a map – should always be rates rather than totals and use a sensible base geography.

Proportional symbol (count/magnitude)
Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map
For showing unambiguous movement across a map.

Contour map
For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values

Equalised cartogram
Converting each unit on a map to a regular and equally-sized shape – good for representing voting regions with equal value.

Scaled cartogram (value)
Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density
Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map
Grid-based data values mapped with an intensity colour scale. As choropleth map – but not snapped to an admin/political unit.

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrants, lawsuits, information; relationship graphs.

Sankey
Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

Waterfall
Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord
A complex but powerful diagram which can illustrate 2-way flows (and net winner) in a matrix.

Network
Used for showing the strength and inter-connectedness of relationships of varying types.

Les données - définition

Mot, nombre, signal, chaîne de caractères, séquence de bits, morceau de matière ou tout autre élément brut enregistré dans un système d'information où il pourra être corrélé à d'autres objets et interprété pour constituer une information.

Informations factuelles utilisées comme base de raisonnement, de discussion, ou de calcul.

Information ou savoir représenté ou codé sous une forme permettant un usage ou un traitement facilité.

→ Il faut donc commencer par transformer ses données pour les rendre exploitables

Structurer ses données

Données non structurées : tout ce qui n'est pas organisé en base de données, c'est-à-dire la bureautique, la messagerie, les images, les vidéos, etc.

Ex : texte, image, vidéo, sons, nuage de points 3D, imagerie médicale…

Données structurées : informations (mots, signes, chiffres…) contrôlées par des référentiels et présentées dans des cases (les champs d'une base de données) qui permettent leur interprétation et leur traitement par des machines.

Ex : fichier CSV (comma seperated values)

Trouver la bonne visualisation

L’abstraction de données :

Décrire les données pour définir les méthodes appropriées pour la visualisation.

Réduction d'un ensemble de données spécifique en une représentation simplifiée.

Comment ? En connaissant les types et les caractéristiques des attributs de votre jeu de données.

Trouver la bonne visualisation

Les différents types de jeux de données :

Jeu de données : ensemble d'éléments caractérisés par des attributs.

Tables

Attributs + éléments
Le + commun
Ensemble de colonnes et de lignes
Une ligne = un élément
Une colonne = un attribute / variable

Réseaux et Arbres

Nœuds (éléments) + liens + attributs
Les Nœuds sont connectés par les liens
Les attributs peuvent être associés aux nœuds **et** aux liens

Trouver la bonne visualisation

Les types d'attributs

- catégoriels ou nominaux

  Ex : attributs décrivant le genre → F, M, A

- ordinaux ou ordonnés

  Ex : attributs décrivant le statut économique → élevé, moyen, haut

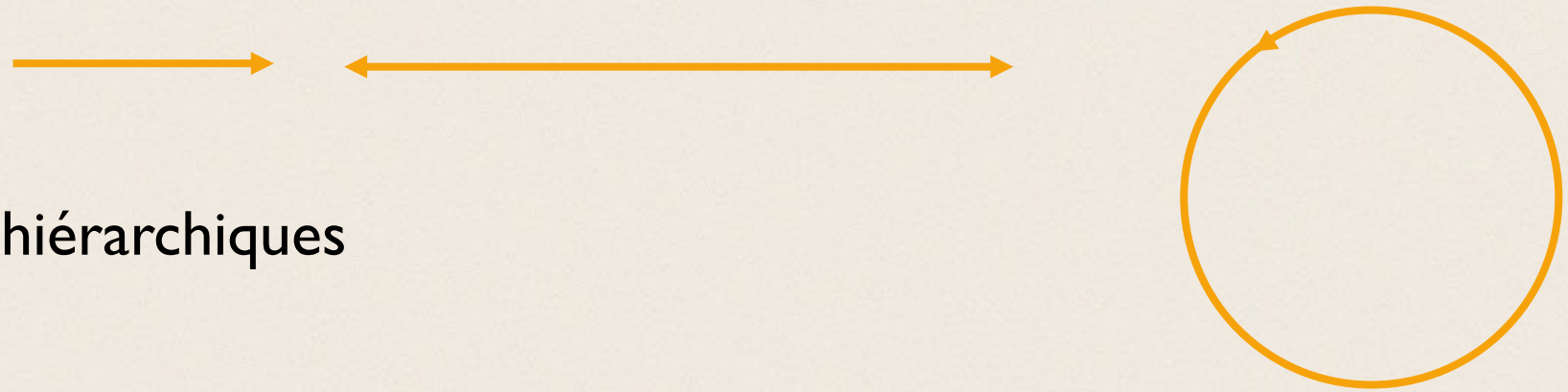
- quantitatifs

  Ex : attributs décrivant un prix, une taille, une quantité, un poids...

Trouver la bonne visualisation

La sémantique des attributs

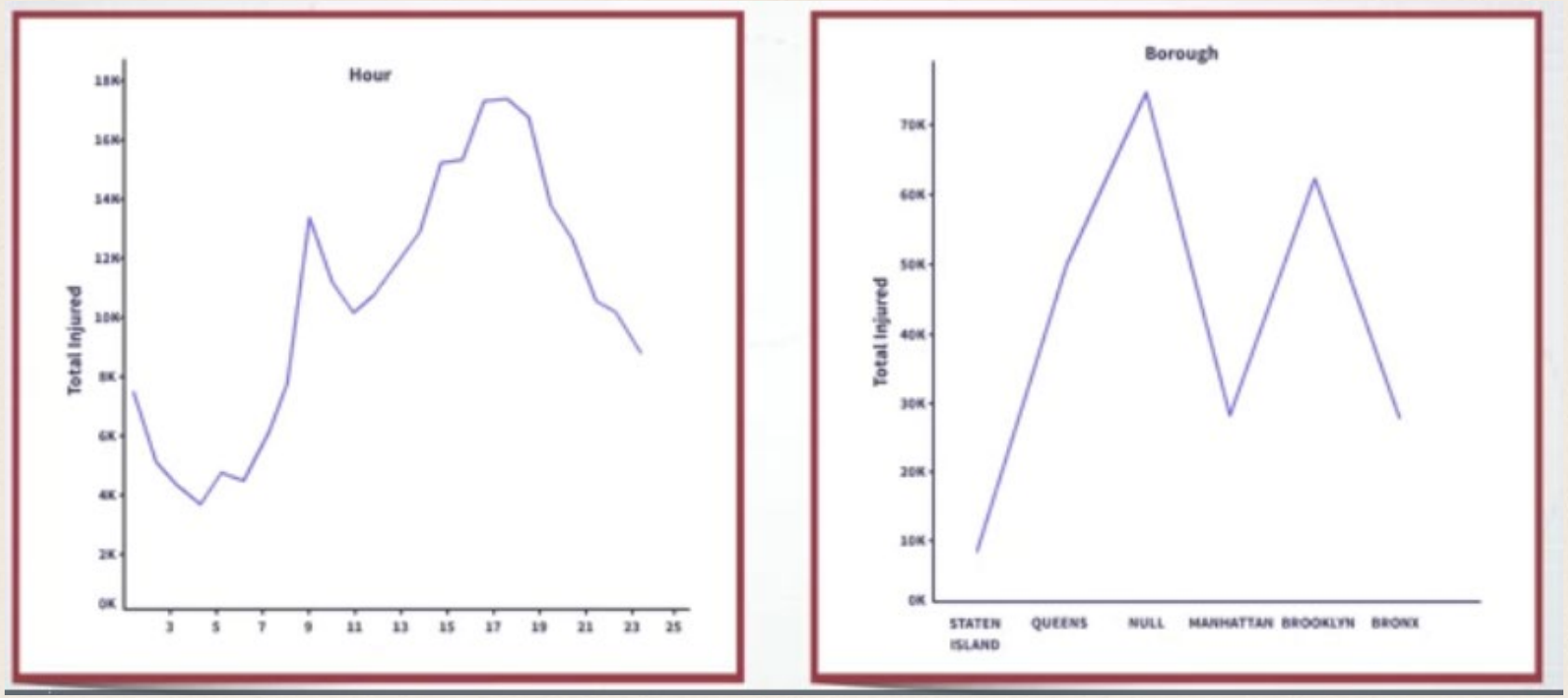
- spatiaux et temporels (régions, coordonnées…)
- séquentiels, divergents (températures…) ou cycliques (mois…)



- hiérarchiques

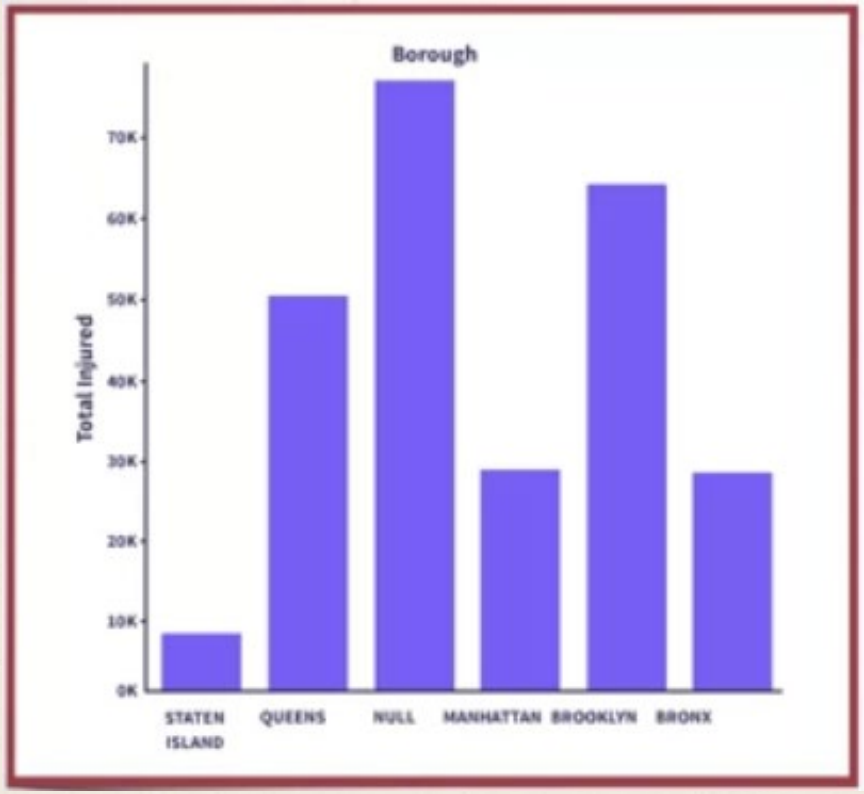
Trouver la bonne visualisation

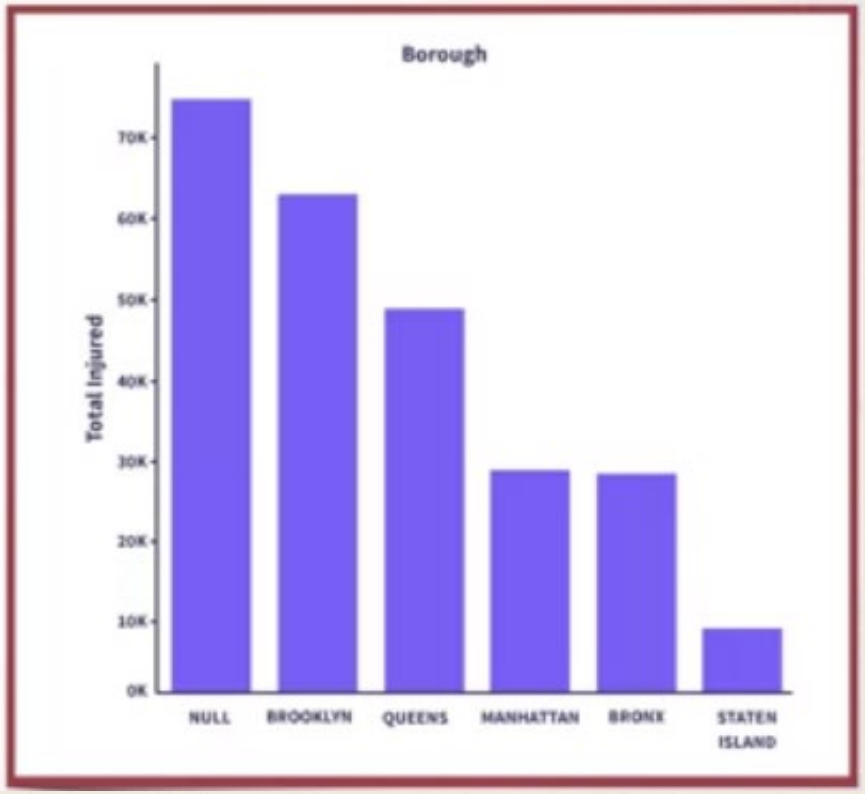
Graphique linéaire : uniquement des attributs **ordinaux** ou **quantitatifs**



Trouver la bonne visualisation

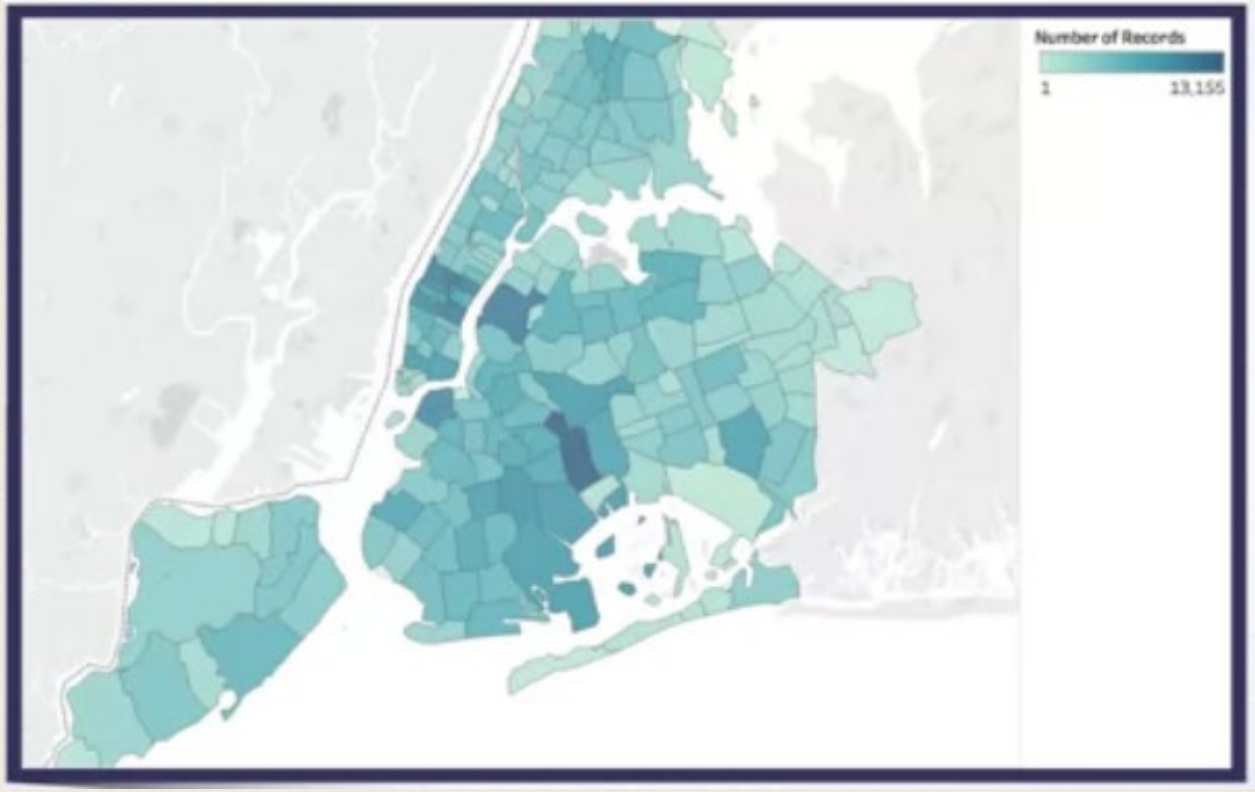
Diagrammes en barres : approprié pour les attributs catégoriels qui peuvent être réorganisés de manière arbitraire





Trouver la bonne visualisation

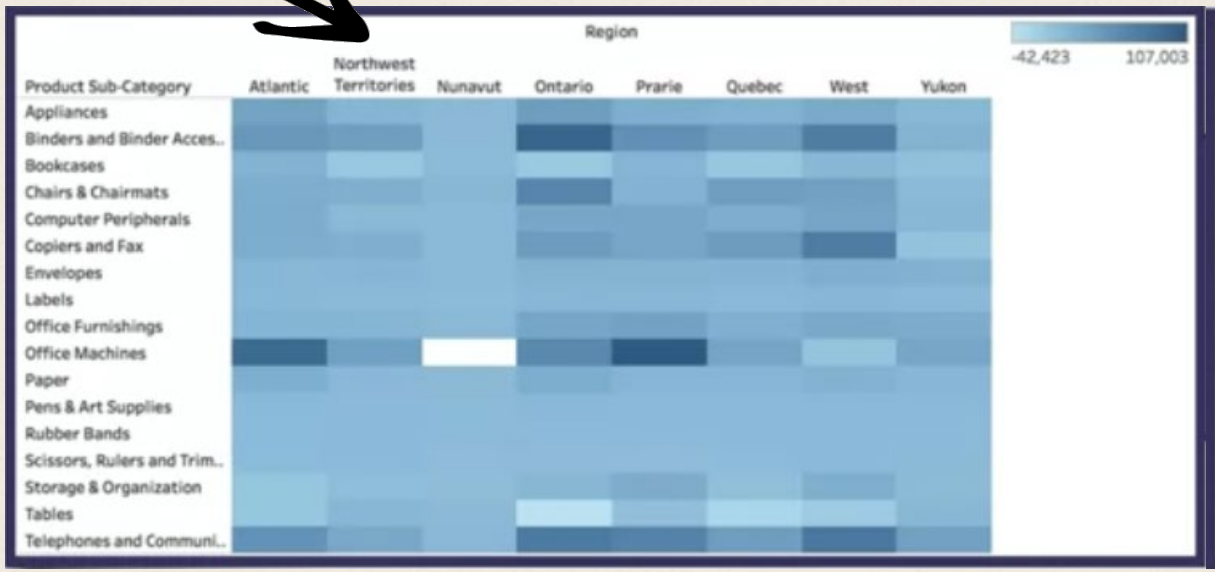
Cartes : appropriées pour les attributs spatiaux
mais ne sont pas toujours la bonne solution !



Trouver la bonne visualisation

Les couleurs divergentes sont plus appropriées qu'un gradient d'une seule couleur pour les attributs divergents quantitatifs

Ici : la seconde heat map utilise le rouge pour le négatif et le bleu pour le positif et le blanc pour zéro : plus lisible



Pour récapituler

- Que sont les données ? Eléments et attributs / Données structurées et non structurées
- Quel rôle ont les données dans la data visualisation ?
- Pourquoi visualiser ? Gain de temps, remplacer la cognition par la perception
- Abstraction de données : une méthode pour identifier les caractéristiques des attributs grâce aux types d'attributs et leur sémantique

Pour aller plus loin...

Liste des références utiles

- Le cours de Lev Manovich : https://docs.google.com/document/d/1BN-V07YMvKf-WZ4LoDJMrOwseWhB7arhc-z3cKRnJmY/edit

- Les slides de Nicolas Bonneel, Aurélien Tabard et Romain Vuillemot de Lyon 1 : https://lyondataviz.github.io/teaching/lyon1-m2/2016/pdfs/1_introduction.pdf

- Les slides du CERN : https://www.slideshare.net/eamonnmag/principles-of-data-visualization-71834041

Liste des références utiles

- Étienne Ollion et Julien Boelaert, « Au-delà des big data », *Sociologie* [En ligne], N°3, vol. 6 | 2015, mis en ligne le 20 janvier 2016, consulté le 01 juillet 2021. URL : http://journals.openedition.org/sociologie/2613

- Lemercier Claire, Zalc Claire, *Méthodes quantitatives pour l'historien*. La Découverte, « Repères », 2008, 128 pages. ISBN : 9782707153401. DOI : 10.3917/dec.lemer.2008.01. URL : https://www.cairn.info/methodes-quantitatives-pour-l-historien--9782707153401.html

- Sylvain Genevois, « Comment différencier infographie et data visualisation ? », *Cartographie(s) numérique(s),* 11/02/2019. URL : http://cartonumerique.blogspot.com/2019/02/infographie-datavisualisation.html

Les outils :

- Palladio, une application de nettoyage et d'exploration de données complexes, temporelles et spatiales développée par Stanford University
- CARTO, une application de cartographie, période d'essai de 30 jours
- Gephi, un logiciel de visualisation et d'exploration interactive pour toutes sortes de réseaux et de systèmes complexes, de graphiques dynamiques et hiérarchiques
- Omeka, un logiciel de gestion de bibliothèque numérique et de diffusion de données numériques (photos, vidéos, textes numérisés…)
- Draw.io, outil pour créer des diagrammes et visualisations
- D'autres outils et infos ici : https://medialab.sciencespo.fr/outils/