La « *data viz* » : donnez vie à vos données !

Arnaud Sallaberry
Maître de conférences à l'Université Paul Valéry, Montpellier 3
Pascal Poncelet
Professeur à l'Université de Montpellier

De la carte thématique teintée, hachurée ou coloriée d'un livre de géographie (on parle de carte *choroplèthe*) à l'infographie d'un journal d'information, du plan de métro au diagramme statistique, la visualisation de données, appelée aussi la *data visualisation* ou *dataviz* en anglais, a envahi notre quotidien. Son objectif ? Améliorer la lecture de données numériques, textuelles ou topologiques de façon à en tirer rapidement de l'information. Comme nous le dit en effet l'expression, *« a picture is worth ten thousand words »* (une image vaut mieux que dix mille mots)…

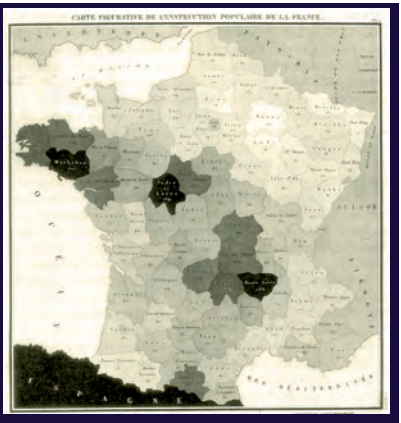
L'émergence des cartes, déjà avant notre ère

Bien que l'on puisse faire remonter la naissance de la représentation graphique aux peintures préhistoriques ou aux premiers systèmes d'écritures basés sur des pictogrammes (comme les hiéroglyphes), c'est au cinquième siècle avant notre ère que l'on voit apparaître les premières visualisations de données, en l'occurrence de données géographiques. Certes, des cartes locales avaient déjà été produites, mais c'est durant ce siècle que des cartes dont l'objectif n'est plus simplement de se déplacer mais de transmettre des informations globales sont réalisées. La carte d'Anaximandre, décrite par Marcel Conche dans son livre *Anaximandre : Fragments et témoignages* (Presses universitaires de France, 1991), est à cet égard un exemple frappant. Un autre spécimen bien connu de cette époque est la carte de Babylone, conservée au British Museum.

Carte babylonienne du monde.
© *British Museum 92687*

La représentation de données géographiques va ensuite se développer jusqu'à nos jours, comme l'illustre l'impressionnante carte topographique réalisée par la famille Cassini au XVIII^{e} siècle. La première carte choroplèthe est quant à elle produite en 1826 par le mathématicien, ingénieur et homme politique français Charles Dupin. Elle montre la distribution et l'intensité de l'illettrisme en France. Pour la première fois, des données numériques géo-localisées sont représentées sur une visualisation.

La carte des Cassini.
© *Connaissance des temps, 1763, BNF*



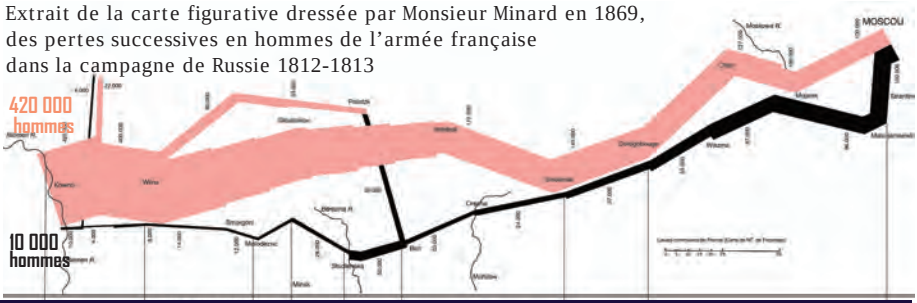
L'illettrisme en France.
© *Charles Dupin, 1826*

Les arbres, diagrammes et autres graphiques

Le Moyen Âge voit l'apparition des premières visualisations de données non géographiques. C'est le cas par exemple des *diagrammes en arcs*, dans lesquels des relations entre éléments sont représentées à l'aide d'arcs qui les relient. C'est le cas aussi des *arbres* (voir des exemples à l'adresse http://treevis.net), qui connaîtront un énorme succès pour la représentation des généalogies et des hiérarchies (pensez à l'organigramme d'une société). Les *lignes de temps* (voir sur https://vcg.informatik.uni-rostock.de/~ct/timeviz/timeviz.html) font également leur apparition pour la représentation de données chronologiques. Elles sont maintenant déclinées en *frises historiques*, en *diagrammes de Gantt*, en *séries chronologiques*…

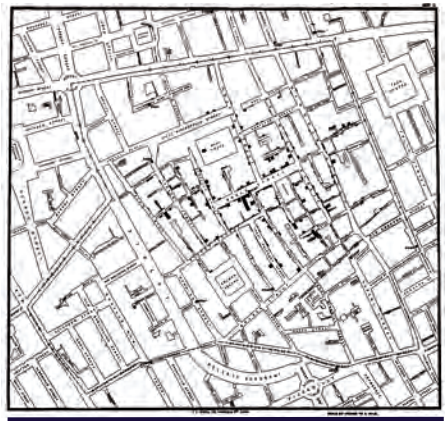
Un pas décisif dans la visualisation de données statistiques est réalisé par William Playfair à la fin du XVIII^{e} siècle. Il introduit trois grands types de diagrammes : la série chronologique, l'histogramme et le *diagramme circulaire*, aussi appelé *camembert*.

Extrait de la carte figurative dressée par Monsieur Minard en 1869, des pertes successives en hommes de l'armée française dans la campagne de Russie 1812-1813



L'exemple le plus emblématique de visualisation est sans doute la carte de la campagne de Russie de Napoléon réalisée par Charles Joseph Minard au XIX^{e} siècle. Il faut la voir en grand format ! Pas moins de six types de données y sont représentées en une carte concise et facilement mémorisable : le fond représente les données géographiques, une bande représente le chemin pris par la grande armée, son épaisseur représente l'importance de l'armée en nombre d'hommes, sa couleur représente la direction, une ligne en pointillés représente les températures lors de la retraite. *© Charles Joseph Minard, 1869*

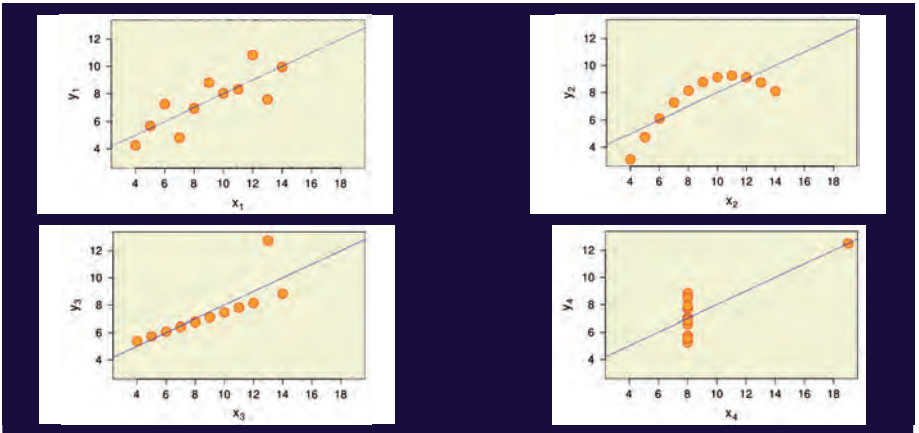
Un autre exemple marquant de la visualisation a été réalisé par le médecin britannique John Snow en 1855. Il s'agit d'une carte de Londres sur laquelle ont été placés des points noirs symbolisant les personnes mortes lors d'un épisode de choléra. En observant la densité de points, il est facile de trouver le puits infecté qui est à l'origine de l'épidémie. La visualisation montre ici sa capacité à non seulement permettre d'accéder à des données («tant de morts à telle adresse») mais aussi à en extraire de nouvelles informations («le puits infecté se trouve à tel endroit»).



Morts lors d'une épidémie de choléra à Londres en 1854.
© John Snow, 1854

Cette potentialité qu'ont les graphiques à nous montrer de l'information initialement cachée rappelle les objectifs des approches statistiques. La complémentarité des deux domaines a très bien été illustrée par le statisticien britannique Francis Anscombe en 1973 avec son célèbre *quartet*. Celui-ci est composé de quatre séries de valeurs x et y ayant des propriétés statistiques «très proches» (moyennes, variances, coefficients de corrélation…).

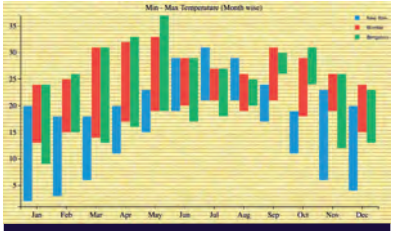
Or, si l'on positionne les éléments dans le plan, on observe que les quatre séries sont très différentes ! Cet exemple célèbre démontre que la visualisation est un outil complémentaire à la statistique dans la mesure où elle peut aider à découvrir des informations que l'on ne sait pas forcément calculer.



Le quartet d'Anscombe. © *Schutz, 2010*

Avec le XXe siècle, la sémiologie graphique

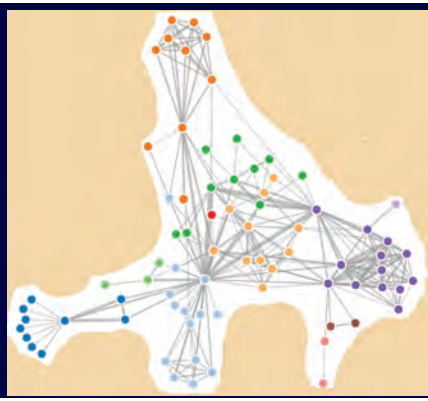
Malgré les nombreuses visualisations proposées jusqu'alors, on ne peut pas encore parler au début du XXe siècle de « langage graphique », c'est-à-dire de système cohérent comportant des signes définis et des règles syntaxiques pour les combiner. C'est en 1967 que Jacques Bertin révolutionne le monde de la visualisation dans son célèbre ouvrage *Sémiologie graphique* (Mouton et Gauthier-Villars) en posant les fondements d'un tel système. Selon lui, une visualisation se compose d'objets graphiques caractérisés par leur *type d'implantation* : ponctuelle (comme une ville sur une carte), linéaire (par exemple une route entre deux villes) ou zonale (un pays).



Températures annuelles de trois villes (la couleur des barres représente les villes, leur longueur la différence entre les températures minimale et maximale). © *Mike Bostock, 2011, https://d3js.org*

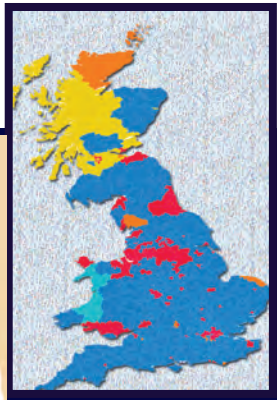
Ces éléments sont disposés sur un plan selon un *type d'imposition* défini par la nature des données à représenter : les diagrammes (camemberts, diagrammes en barres, courbes…) permettent de représenter

des données tabulaires, les graphes permettent de représenter des réseaux, les cartes permettent de représenter des données géographiques.



Réseau
des personnages
du roman
les Misérables
(Victor Hugo, 1862).

© Mike Bosock, 2011,
https://d3js.org



Élections de 2017
au Royaume-Uni.
© Mike Bostock, 2017,
https://d3js.org

Enfin, les objets graphiques prennent différents aspects (forme, couleur, taille, orientation…), appelés *variables visuelles*, qui reflètent des valeurs contenues dans les données initiales.

En définissant ainsi les fondations d'une véritable science, Jacques Bertin a posé les bases formelles qui ont alimenté les recherches en cartographie et en visualisation tout au long du XX^{e} siècle.

Au centre du processus : l'utilisateur

L'avènement de l'informatique, et son utilisation en visualisation à partir des années 1990, marquent une étape cruciale du domaine. Les ordinateurs présentent deux atouts majeurs qui permettent cette révolution : une quantité de données prodigieuse à la portée de tous, en particulier grâce à l'émergence d'Internet, et une capacité de calcul et de rendu permettant de traiter ces gros volumes de données et de générer automatiquement des représentations graphiques interactives. Des types de graphiques innovants et de nouvelles variables visuelles font leur apparition. Ils sont complétés par des techniques d'interaction ouvrant la voie à de véritables outils d'exploration de larges volumes de données.

Parallèlement, l'utilisateur est replacé au centre du processus de conception. Une visualisation correcte des données n'est plus considérée comme intéressante si elle ne lui permet pas d'en tirer des informations.

Dans ce contexte, les sciences cognitives jouent un rôle primordial en explicitant les règles régissant notre système perceptif. Des expérimentations se multiplient afin de définir l'efficacité des différentes variables visuelles. Par exemple, on est capable de dire si la longueur d'une barre d'un diagramme est deux fois plus importante que la longueur d'une autre barre, alors que ce type de comparaison est très approximatif quand il s'agit d'aires. Pour représenter des données numériques, mieux vaut utiliser des longueurs que des aires !

Comparatif des caractéristiques de différentes voitures.
© *tulip (http://tulip.labri.fr/TulipDrupal/)*

Le concepteur est amené à se poser trois questions. *Quoi ?* (Quels sont les types de données à visualiser : graphes, textes, données géographiques ?) *Pourquoi ?* (Quelles sont les tâches que l'utilisateur aura à accomplir avec la visualisation ? Pour un même jeu de données mais des tâches différentes, une seule visualisation n'est pas forcément adaptée.) *Comment ?* (Il existe une multitude d'encodages visuels et de techniques d'interaction.) Selon les réponses, le concepteur doit sélectionner les approches les plus appropriées. Le livre *Visualization Analysis and Design* de Tamara Munzner (A.K. Peters et CRC Press, 2014) recense les principaux types de données, de tâches, de représentations et de systèmes d'interaction utilisés à ce jour, et constitue donc une excellente introduction aux outils mis à la disposition des concepteurs de visualisations…

ou des curieux intéressés par le domaine !

A.S. & P.P.