

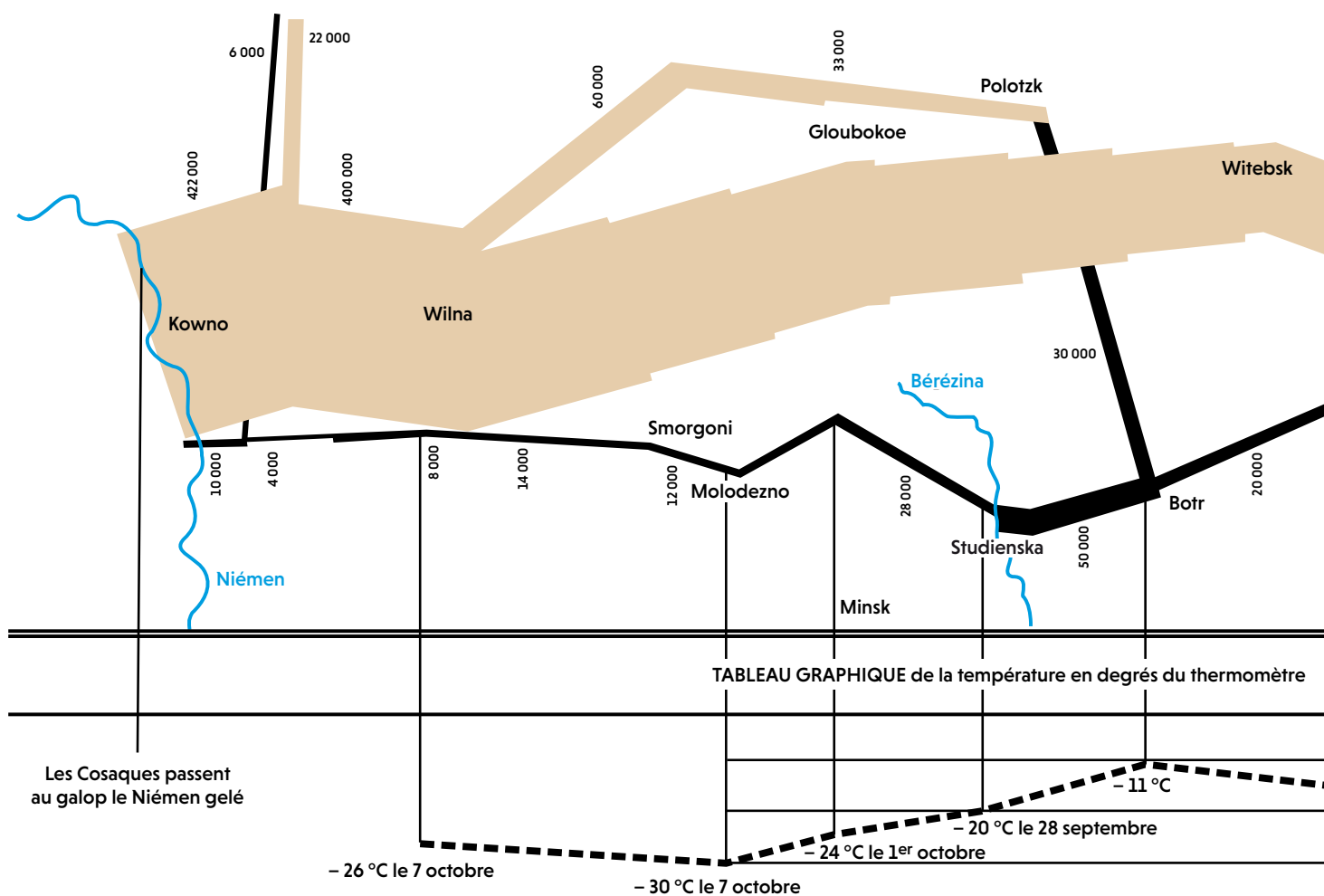
JE VOIS, donc je comprends

Les algorithmes sont aveugles et leur fonctionnement est opaque. La visualisation des données complexes est donc un complément indispensable au traitement des *big data*. C'est grâce à elle que les chercheurs détectent des caractéristiques inattendues, formulent des hypothèses, confirment des résultats... La preuve en images !

L'AUTEUR

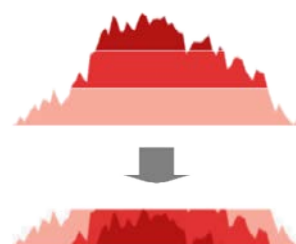


MICHEL BEAUDOUIN-LAFON est professeur d'informatique à l'université Paris-Sud et membre de l'Institut universitaire de France.



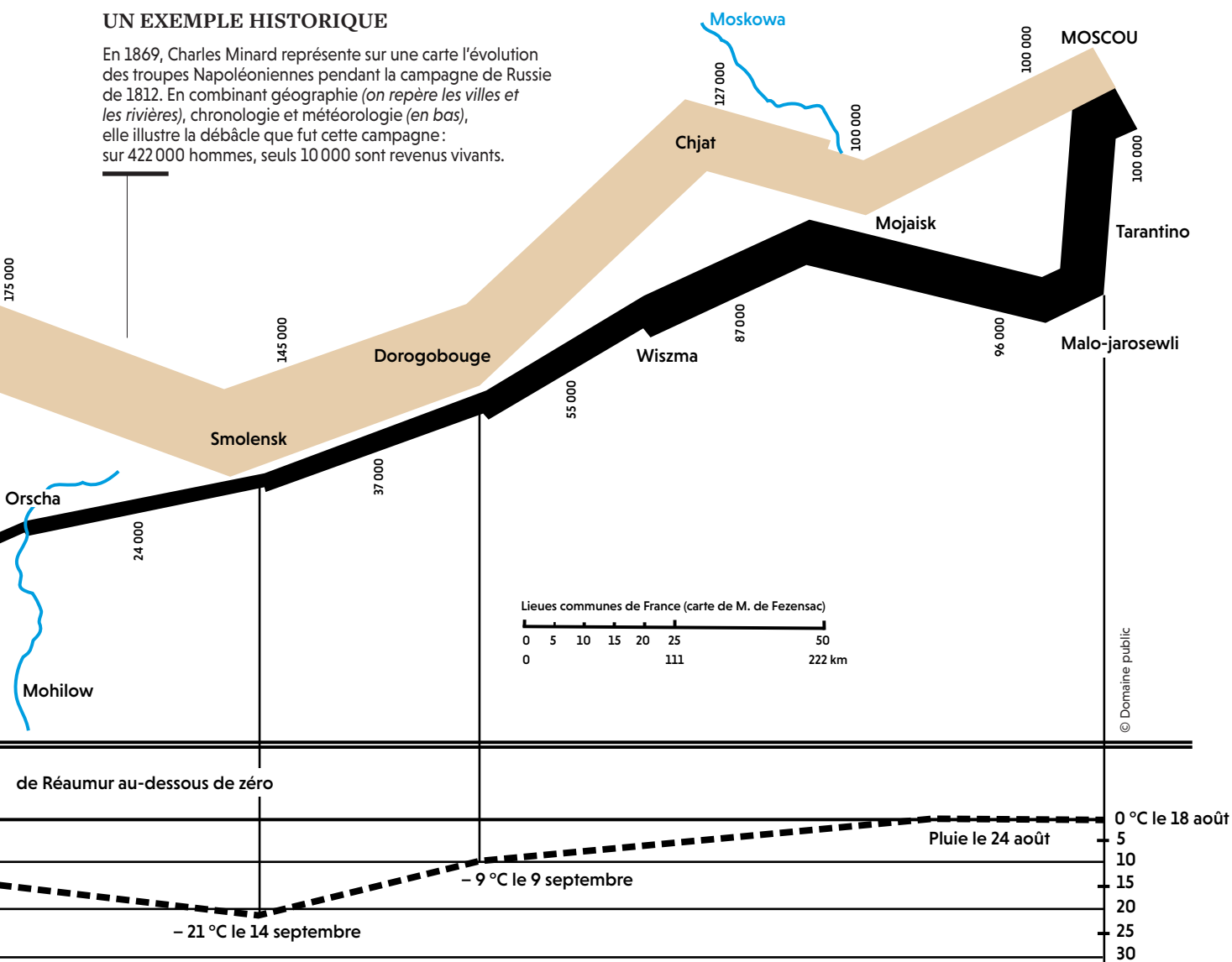
SÉRIES TEMPORELLES

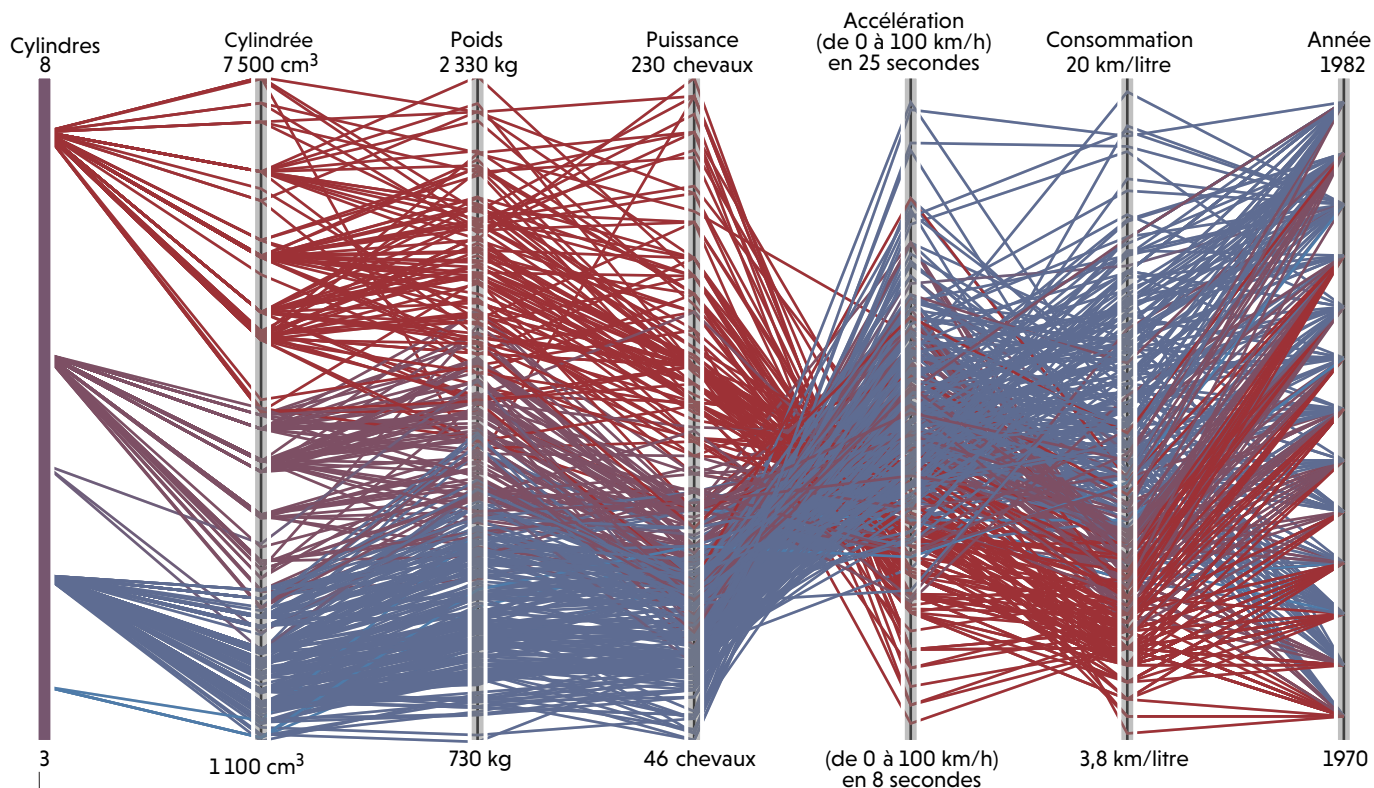
Ce type de représentation montre l'évolution d'indicateurs, ici boursiers, en fonction du temps. Les graphes d'horizon (ci-contre) mettent en évidence de grandes variations de façon compacte (voir le principe ci-dessous). Les variations positives sont en rouge, les négatives en bleu.



UN EXEMPLE HISTORIQUE

En 1869, Charles Minard représente sur une carte l'évolution des troupes Napoléoniennes pendant la campagne de Russie de 1812. En combinant géographie (on repère les villes et les rivières), chronologie et météorologie (en bas), elle illustre la débâcle que fut cette campagne: sur 422 000 hommes, seuls 10 000 sont revenus vivants.

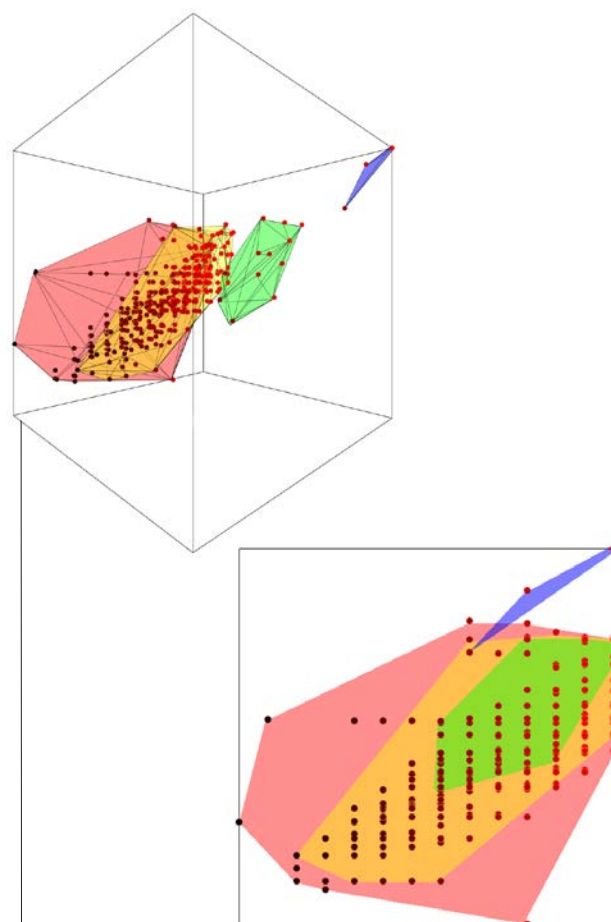
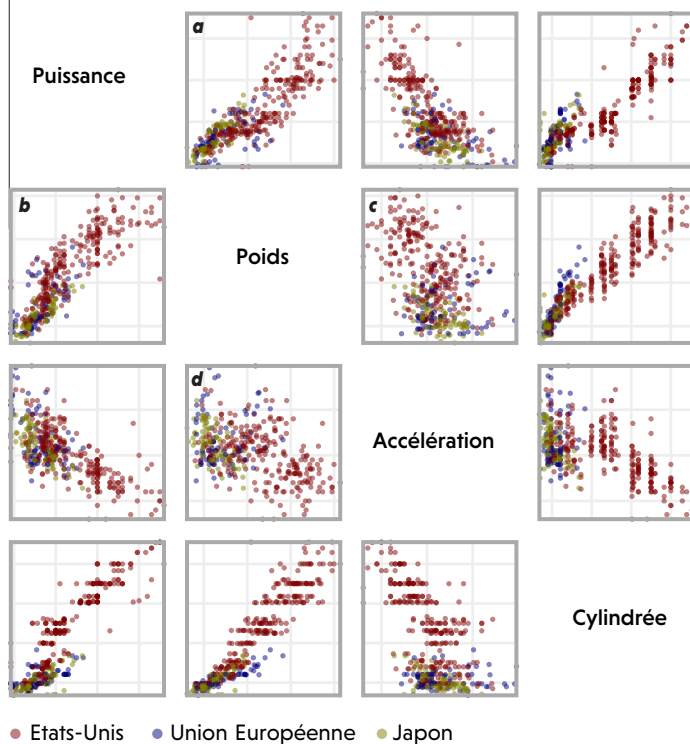




© J. M. Heer/ U. Washington

DONNÉES MULTIDIMENSIONNELLES

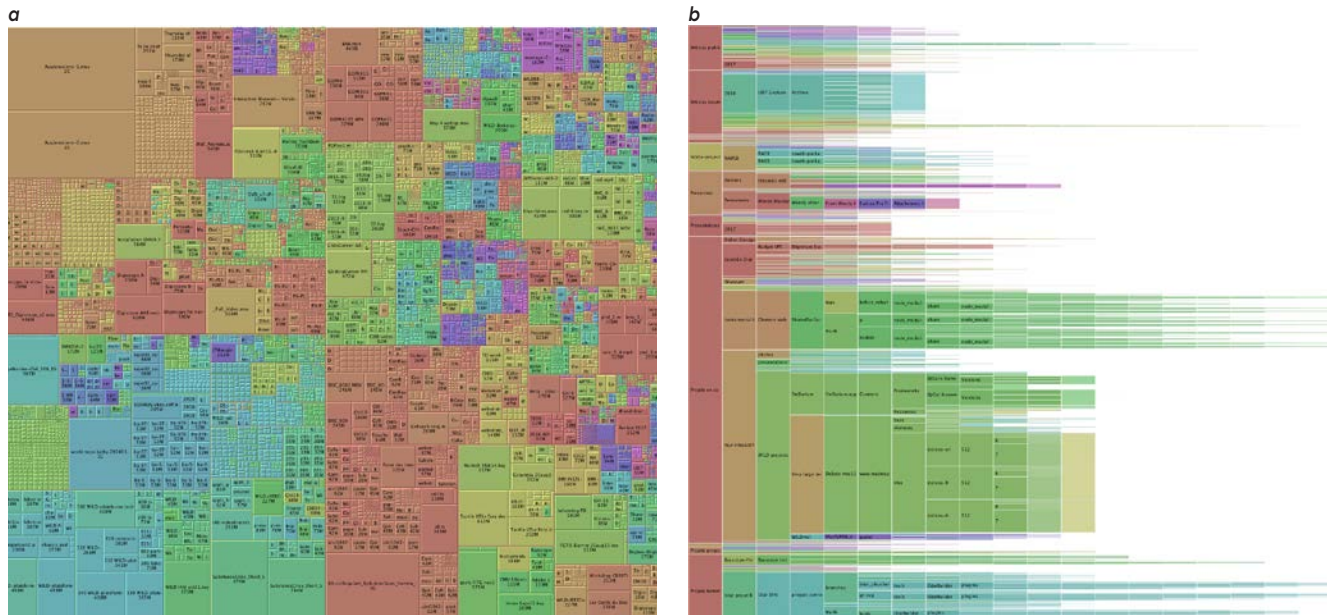
Quand les données concernent plusieurs caractéristiques d'un ensemble d'éléments (ici, des voitures), on peut les afficher avec des coordonnées parallèles (*ci-dessus*) : chaque voiture est une ligne brisée qui relie ses caractéristiques selon chacune des dimensions (*les barres verticales*). Ce graphe révèle des motifs, telle la corrélation inverse entre puissance et accélération. Pour une analyse plus fine, on préfère les matrices de corrélation (*ci-dessous*), qui montrent pour chaque paire de dimensions le nuage de points correspondant. On repère une corrélation forte entre le poids et la puissance (a et b), mais pas entre le poids et l'accélération (c et d).



«ROLLING THE DICE»

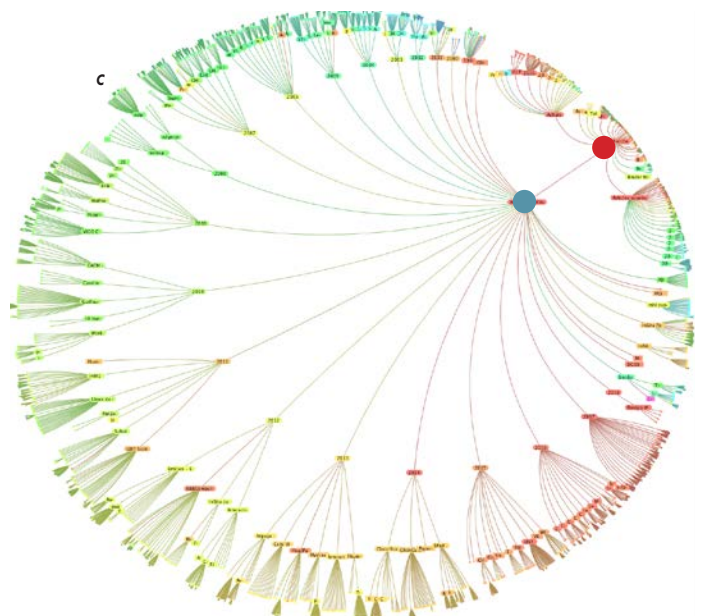
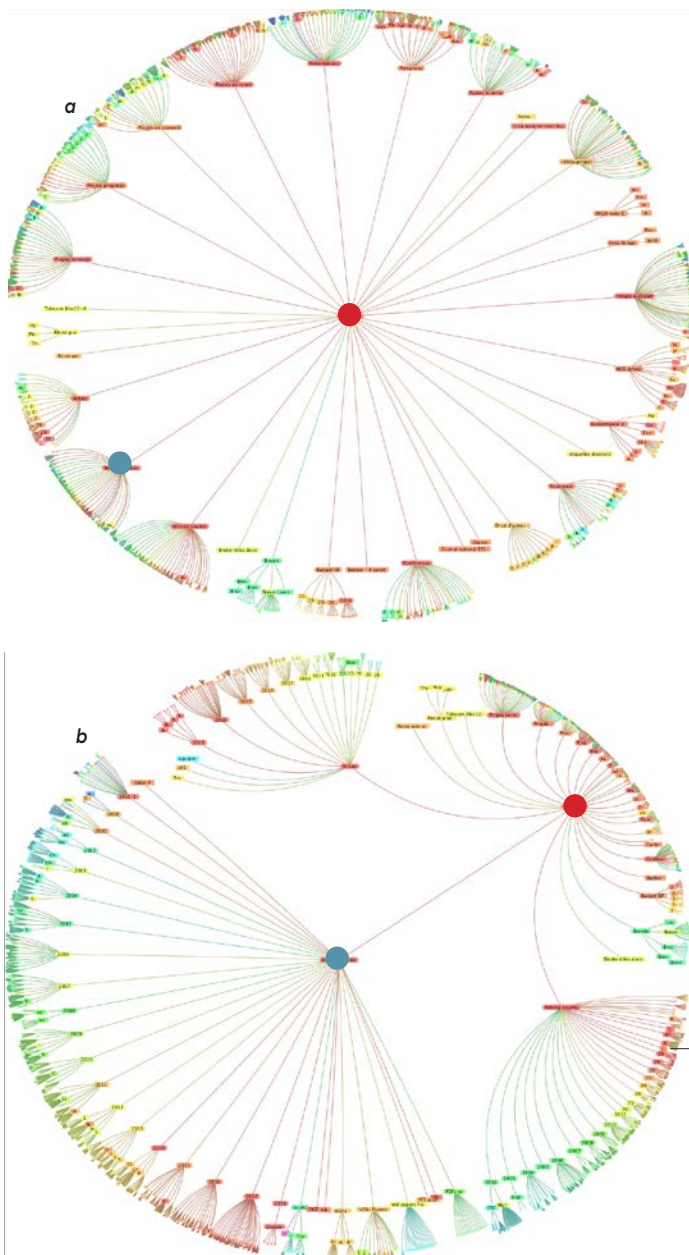
Avec cette technique d'animation 3D (littéralement « Faire rouler les dés »), on peut projeter dans l'espace les matrices de corrélations (*ci-contre*) et passer de l'une à l'autre afin de mieux comprendre leurs relations. On peut voir cette animation ici : <http://bit.ly/RolDice>

© N. Elmqvist, P. Dragicevic et J.-D. Fekete



UNE FORÊT D'ARBRES

La représentation traditionnelle des arbres (des graphes sans cycle où chaque nœud a un ou plusieurs descendants) sous forme... d'arbre n'est pas très lisible lorsqu'ils sont gros. On préfère des représentations qui occupent mieux l'espace, comme les arbres planaires, ou *treemap* (a), où la taille des rectangles est proportionnelle au volume de son contenu en données. On peut par exemple visualiser l'espace qu'occupe chacun des éléments du disque dur d'un ordinateur et l'arborescence selon laquelle ils sont organisés. La représentation en stalactites (b) est plus proche de la version classique.



L'HYPERBOLE

En utilisant le disque de Poincaré, un modèle de la géométrie hyperbolique, on peut représenter un arbre de taille arbitraire (les bords du disque sont mathématiquement situés à l'infini). La navigation dans un tel arbre hyperbolique (de a à c, on pousse le nœud rouge vers le Nord-Est) offre une vue détaillée d'une partie quelconque de l'arbre, sans perdre la vue d'ensemble.

L'ARBRE DE LA VIE

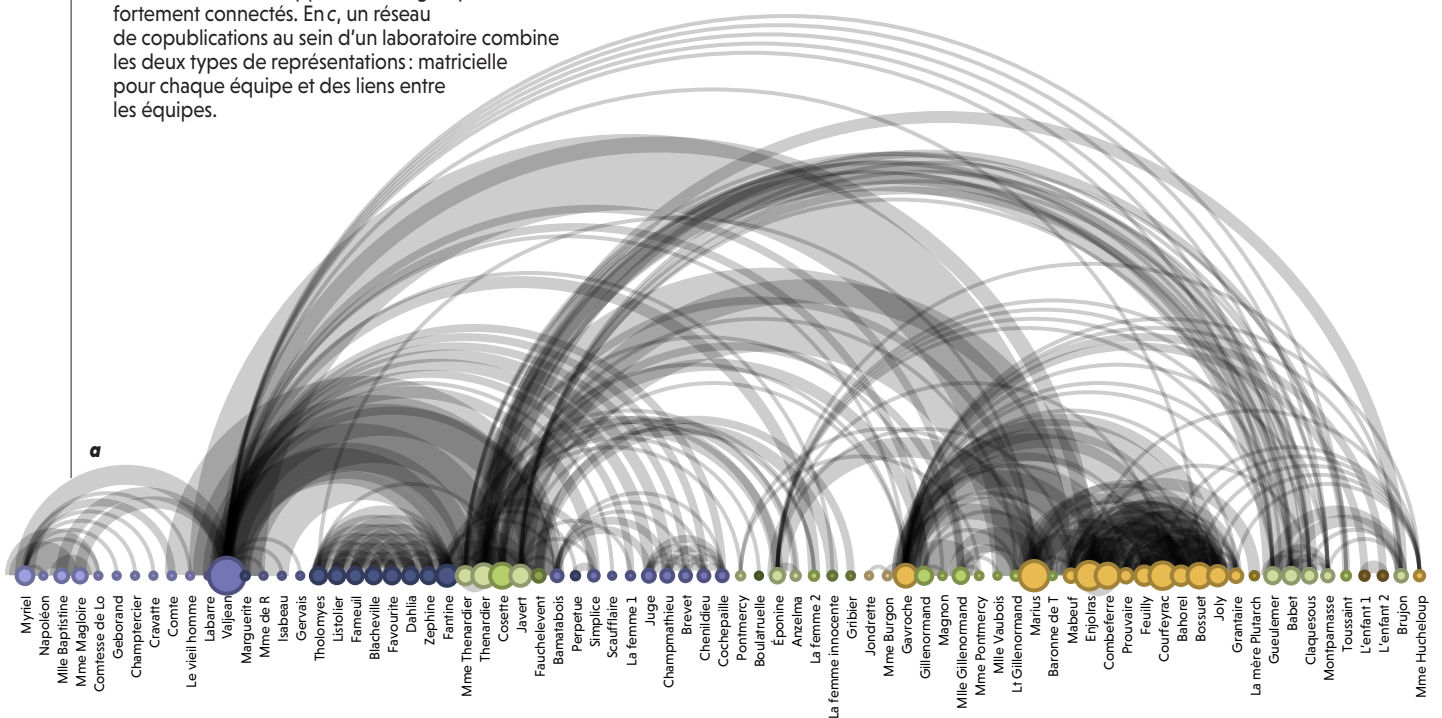
Sur le mur d'écrans Wilder du projet Digiscope (<http://digiscope.fr>), de six mètres de largeur et de 70 millions de pixels, s'affiche l'arbre planaire rassemblant près de 200 000 espèces d'organismes vivants. La technique des images hybrides permet de lire de loin les étiquettes des grandes catégories, mais aussi les détails lorsque l'on s'approche (dans le cartouche).



© Michel Beaudouin-Lafon

NON AUX SPAGHETTIS!

Les données en réseau sont difficiles à représenter car on obtient en général un « plat de spaghettis » illisible. Dans la représentation *a*, les arcs relient les personnages des *Misérables* qui apparaissent en même temps dans le livre. Ils ont été ordonnés et rassemblés par couleur de façon à rendre plus lisibles les regroupements. Plus un personnage intervient, plus le cercle correspondant est grand. En *b*, ce même réseau est représenté sous la forme de matrice où l'agencement des lignes et des colonnes fait apparaître des groupes fortement connectés. En *c*, un réseau de copublications au sein d'un laboratoire combine les deux types de représentations : matricielle pour chaque équipe et des liens entre les équipes.



Images 1 et 2 : © J. M. Heer/ U. Washington
Image 3 : © 2007, N. Henry, P. Dragicevic, J.-D. Fekete et Inria

Je vois, donc je comprends

