POUR LA SCIENCE
Édition française de Scientific American
HORS-SÉRIE POUR LA SCIENCE
HORS-SÉRIE

M 01930 - 98H - F: 7,50 € - RD

Février-Mars 2018
N° 98

STATISTIQUES
COMMENT DÉJOUER LES PIÈGES

FAKE NEWS
COMMENT LIMITER LEUR PROPAGATION

ALGORITHMES
LES DÉFIS DE L'APPRENTISSAGE PROFOND

ORDINATEURS
LA COURSE AUX SUPER-CALCULATEURS

DANIEL COHEN
LE REGARD AIGU D'UN ÉCONOMISTE

BIG DATA

VERS UNE RÉVOLUTION DE L'INTELLIGENCE ?

BEL: 8,9 € - CAN: 12,5 CAD - DOM/S: 8,9 € - ESP: 8,5 € - GR: 8,5 € - LUX: 8,5 € - MAR: 100 MAD - TOM: 2290 XPF - TOM/S: 1260 XPF - PORT. CONT.: 8,5 € - CH: 16,2 CHF

DAP

DATAANALYTICSPOST.COM

ILLUSTRATION : VINCENT DEVILLARD

JE M'ABONNE !

DEUX FOIS PAR MOIS,
RECEVEZ GRATUITEMENT PAR EMAIL
UN ÉCLAIRAGE SCIENTIFIQUE, TECHNIQUE,
OPÉRATIONNEL... SUR LES DATA SCIENCES
AINSI QUE DES OFFRES D'EMPLOIS

POUR LA SCIENCE HORS-SÉRIE
Édition française de Scientific American

www.pourlascience.fr
170 bis boulevard du Montparnasse – 75014 Paris
Tél. 01 55 42 84 00

GROUPE POUR LA SCIENCE
Directrice des rédactions: Cécile Lestienne

HORS-SÉRIE POUR LA SCIENCE
Rédacteur en chef adjoint: Loïc Mangin
Maquettiste: Céline Lapert

POUR LA SCIENCE
Rédacteur en chef: Maurice Mashaal
Rédactrice en chef adjointe: Marie-Neige Cordonnier
Rédacteurs: François Savatier, Sean Bailly

Développement numérique: Philippe Ribeau-Gésippe
Directrice artistique: Céline Lapert
Maquette: Pauline Bilbault, Raphaël Queruel, Ingrid Leroy
Réviseuse: Anne-Rozenn Jouble
Marketing & diffusion: Laurence Hay et Arthur Peys
Direction financière et direction du personnel: Marc Laumet
Fabrication: Marianne Sigogne et Olivier Lacam
Directrice de la publication et gérante: Sylvie Marcé
Anciens directeurs de la rédaction: Françoise Pétry et Philippe Boulanger
Conseiller scientifique: Hervé This
Ont également participé à ce numéro: Denis Bosq, Maud Bruguière et Aline Gestner

PRESSE ET COMMUNICATION
Susan Mackie
susan.mackie@pourlascience.fr • Tél. 01 55 42 85 05

PUBLICITÉ France
Directeur de la Publicité: Stéphanie Jullien (stephanie.jullien@pourlascience.fr)

ABONNEMENTS
Abonnement en ligne: http://boutique.pourlascience.fr
Courriel: pourlascience@abopress.fr
Tél.: 03 67 07 98 17
Adresse postale: Service des abonnements
Pour la Science – 19 rue de l'Industrie – BP 90053
67402 Illkirch Cedex

Tarifs d'abonnement 1 an (16 numéros)
France métropolitaine: 79 euros – Europe: 95 euros
Reste du monde: 114 euros

DIFFUSION
Contact kiosques: À Juste Titres ; Benjamin Boutonnet
Tél. 04 88 15 12 41
Information/modification de service/réassort: www.direct-editeurs.fr

SCIENTIFIC AMERICAN
Editor in chief: Mariette DiChristina
President: Dean Sanderson
Executive Vice President: Michael Florek

Origine du papier : Italie
Taux de fibres recyclées : 0%
« Eutrophisation » ou « Impact sur l'eau » : Ptot 0.008kg/tonne
Ce produit est issu de forêts gérées durablement et de sources contrôlées.

PEFC 10-32-2813 / Certifié PEFC / pefc-france.org

ÉDITORIAL

LOÏC MANGIN
Rédacteur en chef adjoint

Soulever le capot !

En 2015, 23 % des Français ne jugeaient pas crédible le taux de natalité officiel de l'hexagone ! Et 55 % pensaient que l'indice des prix établi par l'Insee ne reflétait pas bien la réalité. En novembre 2017, Chamath Palihapitiya, ancien vice-président de Facebook, s'alarmait de ce que le célèbre réseau social « déchirait le tissu social qu'avait fondé notre société ». Que ce soit à propos des statistiques ou du *big data* et des algorithmes, ces trois exemples témoignent d'une crise majeure dans notre rapport aux chiffres, aux données, à leur exploitation…

On peut avancer plusieurs raisons à cela. À en croire Martine Durand, directrice des statistiques de l'OCDE, la magie des chiffres ne fonctionne plus, car « derrière les moyennes se cachent des transformations et des inégalités de plus en plus importantes. […] On ne reconnaît plus les difficultés de sa vie quotidienne dans les indicateurs. […] Ceci est un danger pour la démocratie ». Un autre élément d'explication réside dans la disruption. Selon le philosophe Bernard Stiegler, cette accélération de l'innovation, nourrie d'algorithmes et de *big data*, « prend de vitesse les organisations sociales » et entraîne l'apparition brutale de nombreuses instabilités, notamment « une perte du sentiment d'exister, qui provoque de la frustration ».

Pour remettre de l'humain dans le numérique, lutter contre le désenchantement, Dominique Cardon, directeur du Médialab de Sciences Po, propose d'améliorer la compréhension par tous des mécanismes de traitement des données. En d'autres termes, on doit « soulever le capot » !

Ce numéro est là pour ça. Les articles de ce *Hors-Série* donnent les clés pour se réapproprier les données, ne plus se faire piéger par les statistiques, comprendre les algorithmes… En un mot, il s'agit de faire en sorte que les données deviennent intelligentes sans pour autant nous rendre idiots. Bonne lecture !

SOMMAIRE

POUR LA SCIENCE HORS-SERIE

N° 98
Février-Mars 2018

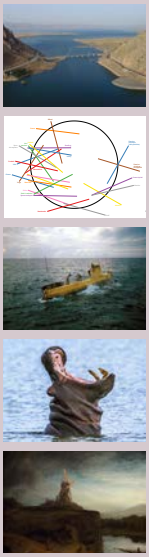
BIG DATA

VERS UNE RÉVOLUTION DE L'INTELLIGENCE ?

STASTISTIQUES : MODE D'EMPLOI

LES DÉFIS DES BIG DATA

RESTER MAÎTRE DU JEU



RENDEZ-VOUS

par Loïc Mangin

Déchiffrer les nombres

Quand il s'agit de manipuler des données, que ce soit en statistiques ou dans le *big data*, des termes techniques surgissent rapidement. Pour s'y retrouver, un petit glossaire s'impose.

DISTRIBUTION GAUSSIENNE

Elle correspond à la loi de probabilité (qui donne la fréquence d'apparition d'un résultat aléatoire) la plus adaptée aux phénomènes naturels. La courbe correspondante est en forme de cloche : un résultat proche de la moyenne des valeurs possibles (le sommet de la cloche) est plus probable qu'un autre éloigné (les bords de la cloche).

FAUX POSITIF ET NÉGATIF

Un faux positif est un résultat, par exemple à un test médical, positif alors qu'il devrait être négatif. Un faux négatif est l'inverse.

INTERVALLE DE CONFIANCE

Cette marge d'erreur rend compte de la précision des mesures d'un paramètre statistique. Par exemple, un intervalle à 95 % contient la valeur réelle du paramètre avec une probabilité de 95 %.

PROBABILITÉ CONDITIONNELLE

Quand deux événements sont liés, la probabilité du second dépend de celle du premier.

RISQUE ABSOLU

En médecine, probabilité qu'une personne développe une maladie durant une période de temps donnée.

RISQUE RELATIF

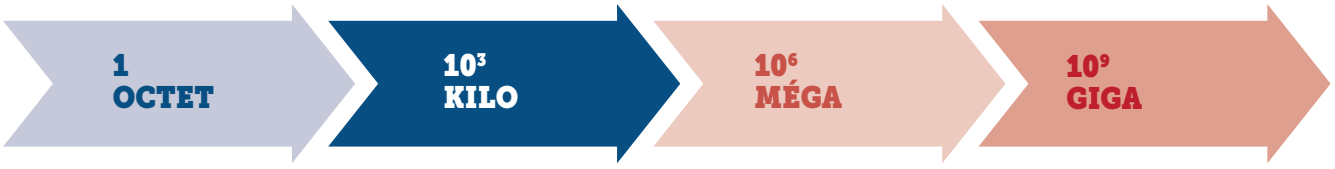
En médecine, rapport entre le risque dans un groupe étudié et celui dans un groupe témoin.

VARIABLE ALÉATOIRE

Le résultat d'une expérience aléatoire (par exemple, un jet de dé) auquel on affecte une probabilité.

VARIANCE & ÉCART-TYPE

Ces deux indices, très répandus, caractérisent la dispersion d'un ensemble de données autour de la moyenne. Plus ils sont faibles, plus les valeurs sont regroupées autour de la moyenne. La variance est égale au carré de l'écart-type, que l'on préfère souvent.



ALGORITHME

Une séquence logique d'actions, traduisible sous la forme d'un programme informatique, que l'on exécute sur des données initiales pour obtenir un résultat. La multiplication est un exemple simple d'algorithme.

BASE DE DONNÉES

Ensemble de données structurées stockées sur un support et gérées par un système dédié (un système de gestion des données, ou SGBD).

BIAIS

Cette erreur de méthode, dans l'origine de données utilisées ou dans leur traitement, conduit à des erreurs dans le résultat final.
Par exemple, le traitement automatique de données discriminatoires par un algorithme transmettra ce biais aux conclusions qui en seront tirées.

BIT & OCTET

Le bit est l'unité fondamentale d'information, elle prend 0 ou 1 comme valeur. L'octet correspond au regroupement de 8 bits et peut donc coder des valeurs numériques, jusqu'à $2^8 = 256$, ou 256 caractères différents. Ses multiples (la frise en bas de cette page) sont les unités de mesure des données.

CLOUD COMPUTING

Service de calcul dispersé sur des machines reliées par un réseau. Il nécessite aujourd'hui une programmation particulière.

CALCUL HAUTE PERFORMANCE

Recours à des supercalculateurs fondés sur le parallélisme des tâches à accomplir (des calculs) pour gagner en vitesse et en puissance. On peut alors mettre en œuvre des algorithmes complexes nécessitant de longs temps de calcul. Les plus performants aujourd'hui effectuent plusieurs millions de milliards d'opérations à la seconde. Le HPC désigne également la science développée autour des équipements nécessaires (matériels, logiciels…).

CORRÉLATION

Elle dénote un lien entre deux variables qui évoluent de la même façon, mais en aucun cas un lien de causalité. Une corrélation n'est pas une connaissance.

EXPLICABILITÉ

Elle est nécessaire pour décrypter et comprendre le fonctionnement des algorithmes d'apprentissage et faire en sorte qu'ils ne soient pas des « boîtes noires ». Pour les applications critiques (médicales, judiciaire, militaires…), elle est devenue un enjeu majeur.

TRAÇABILITÉ

Elle implique le suivi des actions d'un système qui apprend à partir des données. Elle est essentielle pour déterminer les responsabilités et fonder, le cas échéant, un recours juridique.

VÉLOCITÉ

Fréquence à laquelle les données sont générées, traitées et mises en réseau. Cette fréquence étant de plus en plus élevée, il est souvent nécessaire d'employer les ressources du calcul haute performance.

DANIEL COHEN

« Avec le *big data*, la matérialité des corps et des objets pourrait disparaître. Comme dans le film *Matrix*. »

En tant qu'économiste, que vous inspire l'ère du *big data* qui s'ouvre ?

Daniel Cohen : Plusieurs aspects de cette révolution du *big data* en cours s'inscrivent en fait dans les tendances antérieures. Elle prolonge à sa manière la société de consommation traditionnelle, en prônant l'avènement d'une société du sur-mesure, au plus près des désirs des consommateurs, mais au terme de laquelle il s'agit toujours d'acheter du dentifrice ou des voitures.

Les économistes ont parlé à cet égard du passage du fordisme au toyotisme. Dans le premier, toutes les voitures étaient noires. Ensuite, grâce à une révolution industrielle, les constructeurs, à commencer par Toyota, ont pu fabriquer des voitures de la couleur souhaitée par l'acheteur.

Ce changement, qui date des années 1960, explique la victoire de Toyota sur Ford, son grand rival américain. Avec le *big data*, on pousse à l'extrême la personnalisation des biens vendus. On passe d'une consommation de masse uniforme à une consommation ciblée. C'est mieux, mais ça reste des voitures…

Aujourd'hui, dans une société saturée de biens industriels, les coûts de fabrication des biens à proprement parler ont notablement baissé. On peut aller flatter le consommateur dans son souci de singularité… c'est ce que promet le *big data*. En s'approchant des goûts de chacun, en promettant du personnalisé, il relance une société de consommation qui s'essouffle depuis la fin des années 1970. C'est en cela que le *big data* prolonge et d'une certaine façon fait aboutir les tendances antérieures.

On a compris depuis longtemps que les consommateurs ont besoin de se distinguer. Dès les années 1920, General Motors avait déjà concurrencé Ford en multipliant les marques. Et aujourd'hui, les technologies et l'évolution de la société permettent de diversifier l'offre sur une échelle toujours plus grande. Nous sommes toujours

BIO EXPRESS

16 JUIN 1956
Naissance à Tunis.

1976
Diplômé de l'École normale supérieure (ENS) et agrégé de mathématiques.

2006
Fondation de l'École d'économie de Paris.

2008
La Prospérité du vice, chez Flammarion.

2015
Le monde est clos et le désir infini, chez Albin Michel.

AUJOURD'HUI
Professeur et directeur du département d'économie à l'ENS.

dans une consommation de biens matériels, mais ils sont vendus à travers des stratégies de marketing de plus en plus sophistiquées, qui anticipent même nos désirs. Amazon révolutionne la logistique, nous suggère des produits qui pourraient nous intéresser, tout en continuant d'apporter à domicile des biens que nous pourrions tout aussi bien trouver en magasin.

Néanmoins, des bouleversements plus drastiques ne s'annoncent-ils pas ?

Daniel Cohen : En effet, je crois qu'il faut voir au-delà de cette première vision du *big data*, même si c'est un peu comme ça qu'il se donne à voir. Des ruptures plus considérables se mettent en place. La société du *big data* permet tout d'abord de pousser plus loin la logique d'optimisation des coûts, qui habite depuis toujours le fonctionnement de la société industrielle. Le covoiturage, le partage d'appartements, les sites de rencontres… révolutionne la gestion des interactions sociales qui étaient auparavant laissées au petit bonheur la chance, au chaos des accidents de la vie.

Nous sommes ici dans ce qu'on peut appeler une société postindustrielle, au sens où elle assure le service après-vente de la précédente. Les interactions créées par la société industrielle, et en particulier ce que les économistes nomment les externalités négatives, sont peu à peu gérées autrement. Restons sur l'exemple de la voiture. On doit à la société industrielle les problèmes de trafic, les encombrements… Le covoiturage, en organisant les choses scientifiquement, fluidifie la gestion de ces interactions et améliore l'interface sociale que la société industrielle avait fabriquée de façon un peu chaotique.

Toutefois, à nouveau, comme dans le Toyotisme, le service final reste identique : même avec le covoiturage, il s'agit d'aller d'un point A à un point B avec une automobile. Ce n'est pas encore un changement systémique à proprement parler, bien que le mode d'optimisation mis en œuvre commence à faire apparaître une forme de socialité différente.

Des statistiques établies par Blablacar montrent bien ce qui est en train de surgir. Les Français ont en général un taux d'allergie à la vie sociale considérable par rapport à d'autres pays, notamment les États-Unis. Or dès que l'on est dans la sphère du covoiturage, la confiance accordée à des inconnus atteint des niveaux record, de l'ordre de ceux que, toujours en France, on ne connaît que pour sa famille et ses proches.

On assiste à la transition de la gestion optimisée d'une société de consommation dont les fondamentaux sont toujours les mêmes (ce que l'on partage reste des appartements, des voitures…) à un modèle qui ferait émerger de nouvelles formes de socialité, spécifiques du numérique.

« Le *big data* nous demande de devenir nous-mêmes des informations, traitées par des algorithmes »

Ces modèles fondés sur le *big data* n'ont-ils pas été favorisés par les crises ?

Daniel Cohen : La crise a joué son rôle, en poussant les consommateurs à chercher des stratégies de réduction de leurs dépenses. Je ne pense pas, toutefois, que l'on puisse véritablement parler de société post-matérialiste, même si la critique de la société de consommation, amorcée dans les années 1960 et 1970, reste prégnante.

Après la crise des années 1970 et les années Reagan, on a mis au placard les attentes nées dans les années 1960, de *peace and love*, et on a remis le travail au cœur des valeurs morales. Cependant, quelque chose a été préservé de la mise à distance d'une société de consommation dont on avait perçu les limites. Travailler toute sa vie pour acheter une voiture ne faisait plus envie. On voit aujourd'hui les réminiscences de cette critique dans cette facilité avec laquelle les jeunes générations se fondent dans un moule où la propriété n'est plus aussi essentielle qu'avant.

En résumé, à mesure que la société se numérise et que le *big data* l'envahit, on a d'abord optimisé la demande du consommateur vis-à-vis des biens eux-mêmes, puis géré les interactions sociales que leur consommation génère. La première étape relève encore de l'ancien monde, la seconde dessine déjà la frontière d'un nouveau monde.

À quoi ressemble ce nouveau monde ?

Daniel Cohen : Celui-là m'inquiète davantage… Cette fois, le *big data* et les algorithmes ne sont plus des instruments d'optimisation d'une vie qui ressemble par ailleurs à celle d'avant. Ils nous demandent d'entrer dans ce monde, de devenir nous-mêmes des informations pour être traités par des algorithmes.

On passerait d'un monde où un capteur nous dit : « Votre pouls s'accélère, allez voir votre médecin » à un autre où le capteur lui-même prescrit la solution optimale. L'individu, totalement numérisé, serait géré par un algorithme. Dans ce monde, la matérialité des corps et des objets disparaîtrait, ne laissant plus qu'informations et algorithmes. On n'existe plus en chair, un peu comme dans le film *Matrix*.

Je prends ça au sérieux. Le *big data* se prépare à nous faire entrer dans un monde entièrement numérique, où nous devenons des informations gérées par des informations. On peut certes refuser d'y entrer, mais au risque d'un ostracisme général (moyens de paiement, de guérison, d'éducation…). Or une fois à l'intérieur, on devient un flux d'informations qui nous rend tout à fait différents des objets du monde antérieur. C'est un thème que je compte développer dans mon prochain livre.

Nous sommes loin de la société postindustrielle que l'on imaginait.

Daniel Cohen : Dans les années 1950 et 1960, on pensait que le monde postindustriel marquerait le retour à une simplicité humaine des relations sociales. Après les sociétés agraires qui cultivaient la terre et la société industrielle qui « cultivait » la matière, on pensait que dans une société postindustrielle, l'humain allait se cultiver lui-même. L'éducation, la santé, les loisirs seraient au cœur du projet social. C'était l'espoir de l'économiste Jean Fourastié par exemple. L'histoire était vue comme un cycle : aux chasseurs-cueilleurs avait succédé la malédiction, comme disait Jean-Jacques Rousseau, de l'agriculture et de la métallurgie, mais on reviendrait à un monde humain. Un autre scénario est en train de se dessiner.

Peut-on donc dire que l'idéal des Lumières a été abandonné ?

Daniel Cohen : L'idée des Lumières était qu'en allant au bout de la civilisation, >

> on aurait la possibilité de retrouver la simplicité des origines. Alors oui, elles se sont trompées, je pense qu'on peut le dire maintenant. Une illustration en est que les emplois tournés vers l'humain (instituteur, aide aux personnes âgées...) se sont prolétarisés, ils ne scintillent plus comme l'avenir radieux de l'humanité.

La société s'est polarisée. Les emplois qui dans le monde d'hier consistaient à gérer les informations (dans les administrations, les banques, les assurances) sont appelés à disparaître et à être remplacés par des algorithmes. Les survivants sont, tout en haut, ceux qui font tourner le système et, tout en bas, ceux dont le métier est encore inaccessible aux logiciels. Les seconds sont sous la pression de la prolétarisation des gens du milieu, qui restent pour l'essentiel leur clientèle. L'espérance d'une société tirée par le monde des services, où chacun est le coiffeur ou le docteur d'un autre, est en passe d'être complètement balayée tandis que le monde *big data* est en train de construire de façon complètement endogène un univers cohérent, où la personne s'efface derrière l'information qui définit ses caractéristiques (son bulletin de santé ou ses goûts pour telle ou telle série). Dès lors, on comprend pourquoi le grand débat sur le rôle du travail humain à l'heure des algorithmes est totalement relancé.

Cela ne pose-t-il pas la question du revenu universel ?

Daniel Cohen : J'ai soutenu en effet cette idée, que Benoît Hamon a reprise lors de la dernière campagne présidentielle, mais sous une forme différente. Le revenu universel a souvent été interprété comme la contrepartie nécessaire à la disparition du travail. Je n'y crois pas. Peut-être suis-je naïf, mais je pense que les êtres humains ont envie de travailler, notamment pour interagir socialement, pour participer à un projet collectif. L'homme est un animal social, il a besoin d'être avec les autres, et le travail est une façon de répondre à ce désir.

Je perçois une résistance, qu'il faut aider, au passage vers une société du pur algorithme où nous serions désœuvrés. L'humanité n'y est pas prête. Dans ce contexte, à mes yeux, le revenu universel donne aux individus un pouvoir de négociation face à une société où ils sont constamment menacés, mis en chantage par des alternatives. Le revenu universel devient important pour revaloriser le travail, pas pour le remplacer.

Avec le *big data*, l'automatisation des métiers est donc inévitable ?

Daniel Cohen : Au début du XIXe siècle s'est déjà tenu un débat sur le rôle des machines. À cette époque, l'économiste suisse Sismonde de Sismondi s'inquiétait : « Que se passera-t-il le jour où le roi aura une manivelle lui permettant d'actionner tous les emplois du royaume ? » Les luddites britanniques, les canuts lyonnais ont bien vu que les machines détruisaient leurs emplois et se sont révoltés. Cependant, on s'est vite rendu compte que ces craintes étaient injustifiées, qu'au bout du compte la machine rendait les ouvriers plus productifs et augmentait leurs salaires.

La destruction créatrice selon Schumpeter était au rendez-vous, mais le sera-t-elle encore ?

Daniel Cohen : On peut et on doit s'interroger. L'idée que les destructions ici provoquent des créations là-bas ne doit pas nous dispenser de demander : où, et à quel prix. Le fait que la transition d'un type d'emploi à un autre ait fonctionné hier ne suffit pas à comprendre pourquoi et comment ce fut le cas. Si la mécanisation n'a pas, finalement, paupérisé les ouvriers au XIXe et au XXe siècle, c'est parce que les machines et les humains ont été complémentaires : les premières ont eu besoin des seconds pour fonctionner. Ces liens ont créé un effet de levier qui a tiré l'ensemble vers le haut. Mais que se passerait-il si, au lieu d'être complémentaires, les machines se substituaient aux humains ? Si le travail humain était mis en concurrence avec des robots, des machines, des algorithmes ? C'est la grande question du monde contemporain.

D'une façon un peu futuriste, on peut imaginer deux types de sociétés. Dans l'une, des élites technofinancières inventent et commercialisent des produits, des logiciels, des algorithmes... qui font tourner cette société totalement numérisée. On peut reconnaître la Silicon Valley. Le travail des humains tel que nous le connaissons aujourd'hui sera superflu, car tout sera numérisé, y compris la santé, l'éducation... Nous serons devenus des corps numériques dispensés des interactions qui ont cours aujourd'hui.

Dans cette société profondément inégalitaire, le peu de travail qui restera sera confié à une domesticité entourant l'élite. C'est le retour à un système féodal où le luxe sera d'échapper au monde que l'on réserve aux autres. Plus on s'éloignera du sommet, plus le travail sera dévalorisé.

Dans l'autre monde possible, et je pense qu'il adviendra (je ne peux pas croire le contraire !), nous retrouverons des complémentarités. L'architecte qui peut concevoir des maisons totalement différentes, et les faire visiter virtuellement, le professeur qui réinventera ses méthodes d'enseignements, etc. L'usage des machines ne nous dispensera pas de faire jaillir de l'intelligence humaine.

De quelle façon ?

Daniel Cohen : C'est le travail de la génération à venir, des jeunes actuels. Je crois à la loi de Jean-Baptiste Say, économiste du tournant du XIXe siècle, selon laquelle l'offre crée sa propre demande, parce qu'elle s'est toujours vérifiée dans l'histoire. Si les jeunes veulent travailler, et tirer le meilleur profit des technologies à leur disposition, ils y parviendront, mais à condition évidemment que la société, son système de formation notamment, leur en donne les moyens.

Pour cette raison, je suis très attentif à la réforme de l'université. J'aimerais

« Le revenu universel donne un pouvoir de négociation dans une société où le numérique domine »

tant qu'elle soit vraiment un lieu d'intelligence collective. Aujourd'hui, en France, au sortir du baccalauréat, vers 18 ans, les jeunes doivent choisir une voie (avocat, médecin…). C'est ridicule et revient à les envoyer au « peloton d'exécution », car on est certain que les métiers auxquels on les destine sont déjà en voie de disparition.

On devrait plutôt leur donner les moyens d'être beaucoup plus agiles afin qu'ils puissent inventer de nouvelles complémentarités avec les machines et éviter d'être entièrement numérisés. En un mot, j'aspire à une révolution démocratique par laquelle la créativité humaine garderait le dernier mot.

Êtes-vous optimiste ?

Daniel Cohen : On doit se rappeler que le système dans lequel nous vivons vise à gagner de l'argent en vendant des produits à des consommateurs. Or ces consommateurs sont des humains. Certains comparent parfois le travail des humains à celui des chevaux. Au début du XIXe siècle, on a pensé que ces animaux allaient disparaître, remplacés par le chemin de fer. Pourtant, il n'y a jamais eu autant de chevaux qu'au XIXe siècle. Pourquoi ? Parce que cette période fut celle d'une transition. Le chemin de fer était efficace pour vous emmener d'un point A à B, mais pour rejoindre A, pour se déplacer autour de B, le cheval restait indispensable. À la fin du XIXe siècle les villes comme Paris étaient le paradis des chevaux. Il a fallu attendre l'automobile, la bicyclette, les routes pour que d'un coup les chevaux disparaissent, dans les années 1950.

Alors de même, avec les « routes numériques », lorsque tout sera numérisé, l'humain deviendra-t-il obsolète ? Cette métaphore ne fonctionne que jusqu'à un certain point, parce que le consommateur final n'était pas un cheval, mais bien un humain. Le cheval n'a pas eu son mot à dire.

Je pense que l'humain fera de la résistance, car il a besoin de voir des gens, d'aller au concert, de voir les œuvres d'art en vrai… Le frisson, le plaisir, l'émotion empêcheront l'abandon de toute matérialité.

« Je vois dans le numérique la promesse d'une réconciliation avec l'exigence écologique »

Pour reprendre ce que disent les philosophes, notamment mon ami Francis Wolff, l'humain n'est ni dieu ni bête, il est entre les deux : il a un corps et un esprit. L'idée de le faire migrer vers l'esprit uniquement, car au fond c'est un peu ça la numérisation, butera sur le fait qu'il a un corps.

Mon optimisme n'est pas béat. Des catastrophes peuvent advenir, notamment le développement d'une schizophrénie. D'ailleurs, on la voit déjà à l'œuvre. Dans le monde de Facebook où l'on n'est en quelque sorte qu'un pur esprit, mais tout en extériorité, on se montre le plus beau possible, on se met en scène *via* des « cartes postales ». À l'inverse, dès que l'anonymat règne, une violence incroyable se déchaîne, quand le ça prend le pas sur le surmoi. La pulsion est une pathologie possible dans le monde algorithmique.

Pour récapituler, le *big data* et le numérique nous aident à éviter les embouteillages, puis devancent nos désirs grâce à des publicités et des recommandations qui correspondent à nos souhaits. On est là dans une optimisation, intelligente, de la société industrielle, de ses coûts de fonctionnement et de ses externalités négatives, mais dont la base sociale ne change pas.

Toutefois, ce ne sont là que les prémices d'un monde cybernétique, dans lequel on est clairement en train d'entrer. Nous deviendrons, si l'on n'y prend pas garde, un paquet d'informations qui pourra être traité par des algorithmes.

Que devient l'enjeu écologique dans cette révolution ?

Daniel Cohen : Pendant les premières phases que nous avons décrites, la gestion optimisée des interactions sociales, l'écologie est cruciale dans l'essor des *smart cities*, du covoiturage, de l'économie circulaire. À partir du moment où la société industrielle ne fait plus rêver, on entre dans une période de plus grande sobriété par rapport au monde des objets. Néanmoins, la situation est ambiguë voire paradoxale. D'abord, on sait que le monde numérique est un très gros consommateur d'énergie. Ensuite, dans une société où le coût des produits industriels et le temps de travail humain nécessaire à leur fabrication ont beaucoup diminué, la production industrielle n'a pas cessé de croître. En d'autres termes, il y a de plus en plus d'objets, mais ils sont de moins en moins chers. La question énergétique continue de se poser avec une acuité croissante. Je vois quand même dans le numérique la promesse d'une réconciliation possible avec l'exigence écologique.

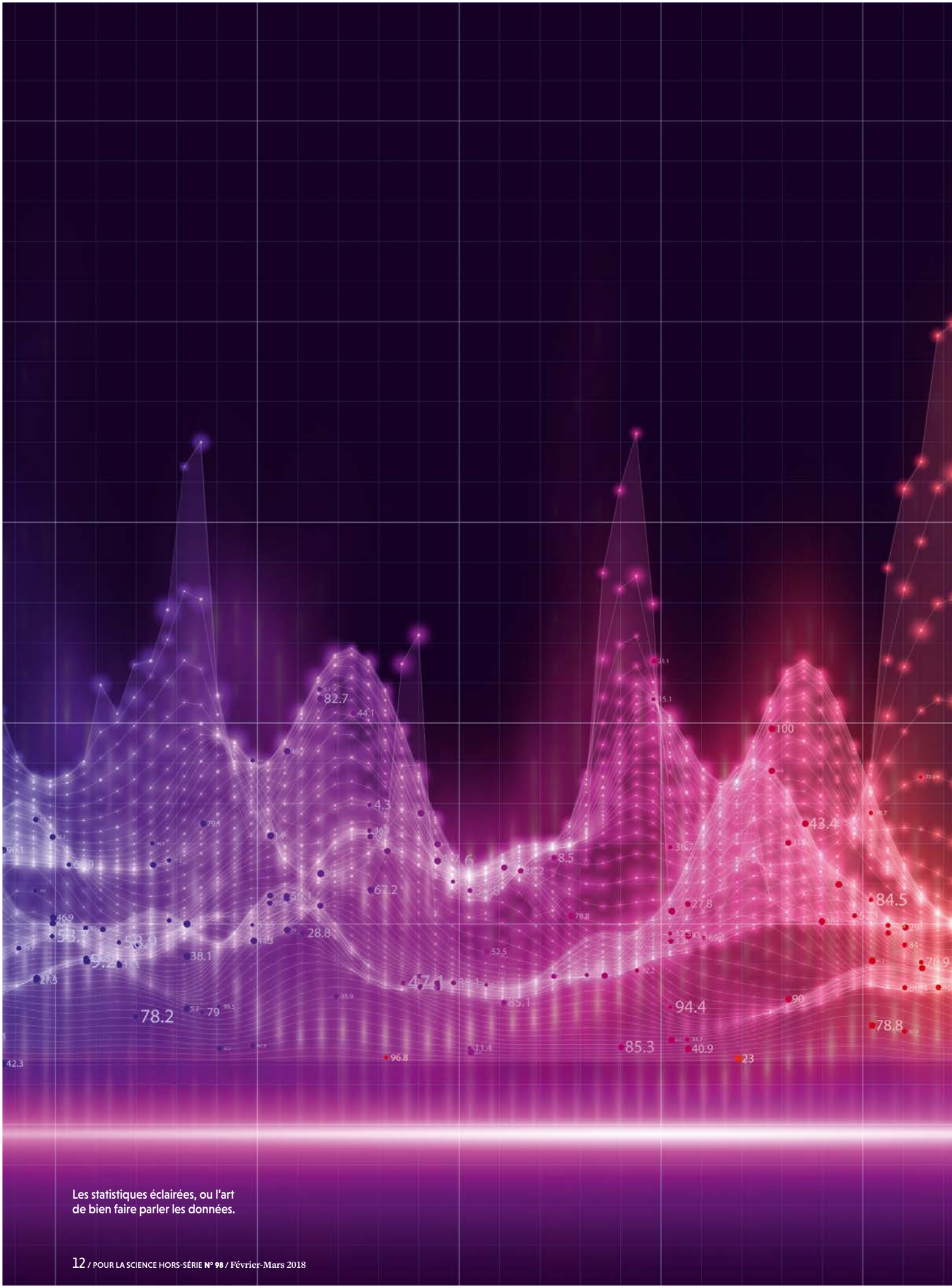
En 2008, dans *La Prospérité du vice*, vous appeliez à ce que le cybermonde soit à l'origine d'une nouvelle solidarité. Où en est-on ?

Daniel Cohen : J'y voyais la promesse que l'humanité devenait une, prenait conscience d'elle-même et était d'une certaine façon capable d'internaliser la question écologique, de comprendre le monde et les interactions. Cette promesse fut minée par un autre aspect de la société de l'information, qui est de produire de l'entre-soi, ce que j'appelle l'endogamie sociale. Internet attise un biais de renforcement des croyances, qui fait perdre ce qu'il pouvait y avoir de mixité sociale dans l'ancien monde.

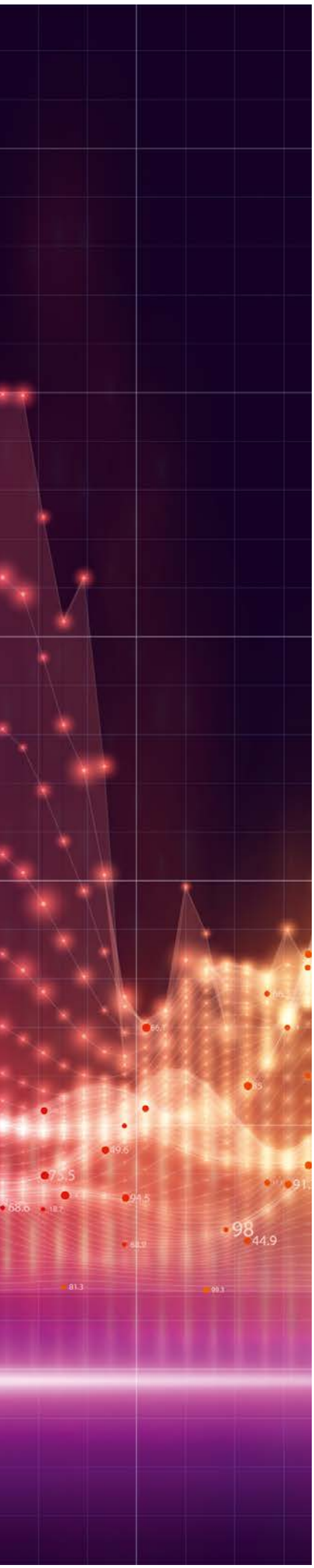
Le *big data* influe sur l'évolution du monde, mais peut-il aussi aider les économistes à mieux le comprendre ?

Daniel Cohen : Sans aucun doute. Grâce aux innombrables données disponibles, les économistes font de plus en plus de tests empiriques pour traquer les causalités au-delà des simples corrélations. Cette tâche est à l'agenda de Hal Varian, qui se voit comme un ingénieur social capable d'extraire des informations causales sur le fonctionnement de la société. Qui est-il ? Le chef économiste de… Google. ■

PROPOS RECUEILLIS PAR LOÏC MANGIN



Les statistiques éclairées, ou l'art de bien faire parler les données.



© Shutterstock.com/GarryKillian

STATISTIQUES : MODE D'EMPLOI

Avant le *big data*, les statistiques sont les premiers outils utilisés pour faire parler les données et leur donner du sens. De fait, un résultat présenté sous la forme de statistiques devient aux yeux de beaucoup une vérité gravée dans le marbre. Pourtant, la « première des sciences inexactes » est riche en chausse-trappes qui imposent une vigilance permanente : paradoxes, mauvaises interprétations, biais dans les données, échantillons insuffisants… Elle est également à la peine face aux événements rares, tel un krach boursier ou une catastrophe naturelle. Les statisticiens eux-mêmes, échaudés par des déconvenues, remettent en cause certains de ses piliers.

L'ESSENTIEL

- On croit tous bien maîtriser les statistiques, les pourcentages, les résultats de sondages...
- Pourtant, ils recèlent des pièges et des chausse-trappes qui sont parfois difficiles à repérer et à déjouer.
- Ces nombres sont aussi porteurs de paradoxes qui défient le bon sens et autorisent parfois des interprétations contradictoires... en apparence.
- Averti de ces dangers, on doit examiner avec une grande attention les données chiffrées pour éviter les arnaques.

L'AUTEUR

JEAN-PAUL DELAHAYE
est professeur émérite à l'université de Lille et chercheur au Centre de recherche en informatique, signal et automatique de Lille (Cristal).

© Shutterstock.com/Phongphan

Déjouer LES PIÈGES des statistiques

Les statistiques sont parées d'une aura scientifique qui les élève au rang de vérité absolue. Pourtant, les pièges qu'elles tendent sont nombreux. La subtilité de certains d'entre eux nous ravit.

Lucide, Mark Twain écrivait: «Les faits sont têtus, il est plus facile de s'arranger avec les statistiques.» Il ajoutait: «Il y a trois sortes de mensonges: les mensonges, les sacrés mensonges et les statistiques.» Le domaine des calculs et des raisonnements statistiques est plein de pièges, d'évidences trompeuses, et même d'arnaques: soyons sur nos gardes, car l'intuition est souvent mauvaise conseillère.

Dans *Statistiques. Méfiez-vous!*, le mathématicien et psychologue Nicolas Gauvrit, du laboratoire Cognitions humaine et artificielle, recense toutes les erreurs, astuces et idées fausses dont il faut être averti. Nous lui emprunterons plusieurs exemples. Commençons par les pourcentages que nous croyons maîtriser depuis l'école primaire. Vérifions avec quelques questions (ni machine ni crayon ne sont nécessaires).

Question 1. Si le prix de l'essence augmente de 25%, il reviendra à son prix initial après une baisse de: (*a*) 25%, (*b*) 20% ou (*c*) 16,67%?

Question 2. Deux produits A et B ont le même prix. Le prix de A augmente de 12%, puis baisse de 23%. Celui de B baisse de 23%, puis augmente de 12%. Au final, (*a*) A est plus cher que B, (*b*) B est plus cher que A ou (*c*) A et B valent à nouveau le même prix.

Question 3. On augmente votre salaire de 2% par an. En dix ans, il a augmenté de: (*a*) 21,90%, (*b*) 20%, ou (*c*) 18,62%?

Question 4. Un membre du gouvernement assure que «l'augmentation de la dette qui était de 15% l'année dernière a été ramenée à 14% cette année». L'opposition prétend pourtant que «le déficit qui était de 15 milliards d'euros

l'année dernière a encore augmenté cette année de plus d'un milliard d'euros ». L'un des deux ment-il (a) ou est-ce possible (b) ?

Question 5. Dans une université, à l'examen de la licence de biologie, les filles ont mieux réussi que les garçons et à l'examen de la licence de physique, les filles ont, là encore, mieux réussi que les garçons. Pourtant, en regroupant les résultats des deux licences, on découvre que les garçons ont mieux réussi que les filles. Y a-t-il eu une malversation sexiste ?

Pour répondre correctement aux trois premières questions, une perception multiplicative des pourcentages s'impose.

Quand le prix P de l'essence augmente de 25%, plutôt que d'additionner P à $25P/100$, multipliez P par 1,25 $(1 + 0{,}25)$ pour calculer le nouveau prix. Pour qu'il revienne à son prix initial, il faut le multiplier par l'inverse de 1,25, c'est-à-dire 0,80 soit $(1 - 0{,}20)$. Il faut donc baisser le prix de 20% pour le ramener à sa valeur précédente. La bonne réponse de la question 1 est donc (b).

Pour la question 2, la bonne réponse est (c), car le prix de A a été multiplié par 1,12 puis 0,77, ce qui est équivalent à une multiplication par 0,77 puis par 1,12, l'opération étant commutative. Plus généralement, si vous avez des augmentations et des baisses à calculer, l'ordre de calcul n'a pas d'importance pour le résultat final. En revanche, il ne faut pas simplifier en disant que +12% et +23% fait +35% !

Pour la question 3, dix augmentations consécutives de 2% correspondent à une multiplication par $(1{,}02)^{10}$. Sans faire de calcul, on sait que l'augmentation totale est supérieure à 20%, car $(1 + x)^n > 1 + nx$. La bonne réponse est (a).

Les deux dernières questions du test sont plus subtiles. Les deux affirmations de la question 4 peuvent être vraies simultanément. Les 15 milliards d'euros du déficit de l'année dernière correspondent à 15% de la dette initiale (d'il y a deux ans). Celle-ci était donc de 100 milliards d'euros. L'année dernière, la dette est ainsi passée de 100 milliards à 115 milliards. Si, comme l'indique >

> la première affirmation, l'augmentation de la dette, c'est-à-dire le déficit, a été de 14%, cette année, l'augmentation a donc atteint 14% de 115 milliards, soit 16,1 milliards. C'est bien conforme à la deuxième affirmation selon laquelle le déficit a augmenté de plus d'un milliard. Les deux affirmations sont parfaitement compatibles: l'augmentation de la dette peut diminuer en pourcentage chaque année en même temps qu'elle s'accroît en valeur absolue.

LES SIMPSON À L'ÉCOLE

La question 5 correspond à une situation remarquable qui froisse l'intuition et peut fausser l'analyse de certains chiffres réels. En effet, il est tout à fait possible que les filles réussissent mieux que les garçons en licence de biologie et mieux en licence de physique, et que, globalement, les garçons réussissent mieux que les filles!

Prenons l'exemple (fictif) d'une situation où 100 candidates filles et 100 candidats garçons se présentent aux examens des licences de biologie ou de physique.

	PHYSIQUE		BIOLOGIE		CUMUL	
	G	F	G	F	G	F
RÉUSSITE	80	10	4	50	84	60
ÉCHEC	10	0	6	40	16	40
TOTAL	90	10	10	90	100	100

Les filles réussissent mieux en physique puisque 100% ont obtenu leur diplôme contre seulement 88% des garçons. De même, en biologie, 55,5% des filles et seulement 40% des garçons réussissent. Pourtant, globalement, 84% des garçons ont un diplôme, contre 60% chez les filles. Comment l'expliquer?

Les filles sont plus nombreuses en licence de biologie et les garçons plus nombreux en licence de physique. Or le taux de réussite est meilleur en physique qu'en biologie. Les filles tentent donc en moyenne un examen plus difficile que les garçons. Ceux-ci ne gagnent, au total, que parce qu'ils optent pour la facilité! Quatre-vingt-dix filles sur 100 tentent la licence de biologie qui a un taux de réussite de 55% alors que 90 garçons sur 100 tentent l'examen de physique qui a un taux de réussite de 90%. Cette situation est nommée paradoxe de Simpson ou effet de Yule-Simpson. Elle a été décrite par Edward Simpson en 1951... et par George Yule en 1903.

Cet effet peut avoir d'ennuyeuses conséquences, comme le montre une étude comparée sur l'efficacité de deux traitements différents contre les calculs rénaux, menée en 1986 par C. Charig, D. Webb, S. Payne et O. Wickham.

Globalement le traitement A avait conduit à 273 succès (78%) sur 350 cas alors que le traitement B sur 350 cas en donnait 289, soit 83%. Le traitement B semblait donc meilleur que A. Pourtant en y regardant de plus près, on constatait que dans le cas des petits calculs rénaux, le traitement A était meilleur que le traitement B: A obtenait 80 succès sur 87, soit 93%, contre 234 succès sur 270, soit 87%, pour le traitement B. Et il en allait de même pour les gros calculs rénaux où A obtenait 73% de réussites (192 succès sur 263) alors que B n'en obtenait que 69% (55 succès sur 80)!

Un autre cas réel de paradoxe de Simpson a été rapporté en 1975 par P. Bickel, E. Hammel et J. O'Connell à propos de l'admission des étudiants dans les diverses facultés de l'université de Berkeley, aux États-Unis. L'analyse globale des données indiquait un biais en faveur des garçons qui, comme dans l'exemple fictif précédent, réussissaient mieux que les filles. Pourtant l'étude détaillée des admissions, faculté par faculté, ne montrait rien de tel: les facultés favorisant les garçons n'étant pas plus nombreuses que celles favorisant les filles.

L'explication du phénomène était semblable à celle de notre exemple: les filles postulaient dans des facultés plus difficiles en moyenne que celles que tentaient les garçons. Doit-on conclure qu'on peut faire dire une chose et son contraire aux statistiques? Risque-t-on toujours de rencontrer de telles situations? La réponse est non: dans l'agrégation des données qui engendre le paradoxe de Simpson, on ne mélange pas des données correspondant à des effectifs égaux pour les sous-cas. Si l'on agrège le résultat des examens de 100 filles passant la biologie, 100 filles passant la physique, 100 garçons passant la biologie, 100 garçons passant la physique, alors le paradoxe de Simpson disparaît. Si l'on veut obtenir des conclusions sensées, l'agrégation des résultats doit respecter certaines règles d'homogénéité.

Notons que le paradoxe de Simpson se généralise en prenant plus de deux catégories d'étudiants. Des données, selon qu'on les regarde d'une façon ou d'une autre, peuvent conduire à des classements exactement inversés. Poursuivons par quelques exemples proposés par Nicolas Gauvrit qui montrent que des paradoxes analogues à celui de Simpson sont plus fréquents qu'on ne l'imagine.

Bienvenue au sein de l'entreprise Marchive dont la situation salariale se résume ainsi:

		OUVRIERS	CADRES
2016	SALAIRE	200 €	2000 €
	EFFECTIF	1000	100
2017	SALAIRE	180 €	1800 €
	EFFECTIF	600	500

Un conflit oppose les syndicats et le patron. Les premiers disent: «Les salaires des ouvriers et ceux des cadres ont baissé cette année de 10%.» Le patron répond: «Nos calculs indiquent que le salaire moyen dans l'entreprise a augmenté. Il est passé de 363,64 euros par semaine >

CITATIONS

«La statistique est la première des sciences inexactes.»

Edmond et Jules de Goncourt

«Fêter les anniversaires est bon pour la santé. Les statistiques montrent que les personnes qui en fêtent le plus deviennent les plus vieilles.»

Den Hartog

«Les statistiques, c'est comme le bikini. Ce qu'elles révèlent est suggestif. Ce qu'elles dissimulent est essentiel.»

Aaron Levenstein

«Dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales.»

Alfred Sauvy

«La mort d'un homme est une tragédie. La mort d'un million d'hommes est une statistique.»

Joseph Staline

«Tout comme certaines sciences occultes, les statistiques possèdent leur propre jargon, volontairement mis au point pour dérouter les non-initiés.»

G. O. Ashley

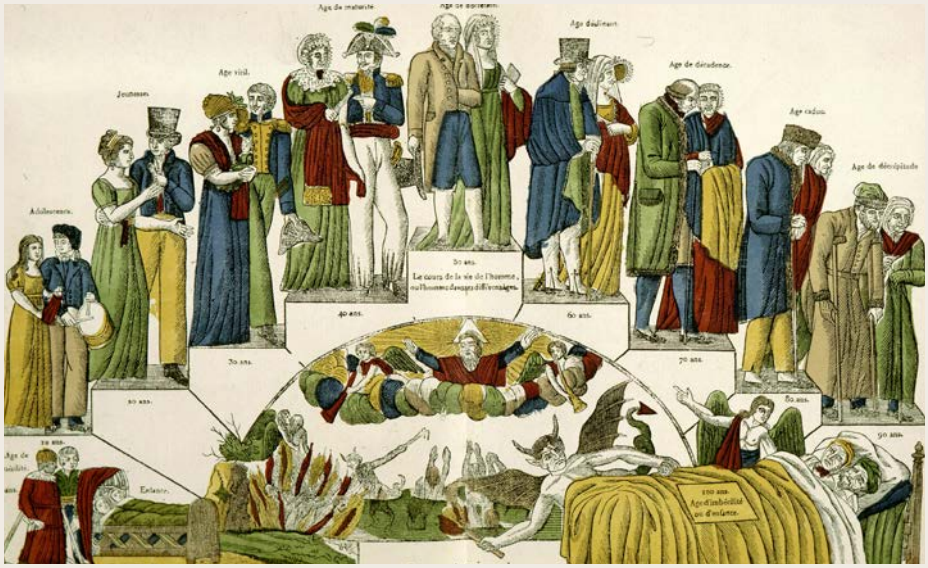
«Les statistiques ont une particularité majeure: elles ne sont jamais les mêmes selon qu'elles sont avancées par un homme de droite ou par un homme de gauche.»

Jacques Mailhot

LES PARADOXES DE L'ESPÉRANCE DE VIE

Le sort des nouveau-nés se répercute sur l'espérance de vie de tous. Pour calculer celle d'un ensemble d'individus, la méthode la plus naturelle serait d'attendre qu'ils soient tous morts, puis de calculer l'âge moyen de leur mort. Dans la pratique, son calcul utilise une autre méthode, rationnelle, mais susceptible d'engendrer incompréhensions et paradoxes.

Pour calculer l'espérance de vie en 2016, on imagine une population fictive d'individus qui naîtraient tous en 2016 et qui, chaque année de leur future vie, auraient une probabilité de mourir égale à celle constatée en 2016 pour cette tranche d'âge. On imagine par exemple 100 000 individus qui naissent en 2016, si 1,3% des enfants ayant entre 0 et 1 an sont morts en 2016, on considère que, dans notre population fictive, il en sera de même. Ensuite, pour les individus de la population fictive qui franchissent leur premier anniversaire (98,7%), on considère qu'ils mourront au cours de leur seconde année dans la même proportion que les enfants qui, en 2016, avaient entre 1 et 2 ans. Et ainsi de suite. L'âge moyen du décès de tous les individus de cette population fictive est par définition l'espérance de vie à la naissance pour l'année 2016. Pour simplifier les calculs, on considère que ceux qui meurent dans leur première année meurent à 0,5 an, que ceux qui meurent dans leur seconde année meurent à 1,5 an, etc.



L'espérance de vie à l'âge de 30 ans se calcule de la même façon en se donnant au départ une population fictive d'individus de 30 ans qui ensuite mourront année après année en se conformant aux chiffres de mortalité constatés en 2016, etc. Ces espérances de vie pour 2016 dépendent donc des conditions de mortalité de l'année 2016 et de nulle autre. Elles ne donnent pas comme on le croit naïvement la durée de vie moyenne des gens vivants en 2016, car les taux de mortalité par âge évolueront dans l'avenir. Plusieurs paradoxes résultent de ce mode de calcul.

Le massacre des innocents

Si un massacre des innocents éradique tous les bébés de moins de 1 an le jour de Noël 2016 sans tuer personne d'autre, alors l'espérance de vie à la naissance en 2016 sera de 0,5 an, car les individus fictifs envisagés par le calcul seront tous morts dès leur première année et seront donc comptabilisés comme vivant 0,5 an.

La bombe atomique

En 2016, l'espérance de vie en France était d'environ 79 ans pour les hommes et 85 pour les femmes. Cela n'aura rien de faux même si en 2017 une bombe atomique tue tous les Français et que la moyenne de la durée de vie des Français ayant vécu en 2016 est donc en réalité d'environ 41 ans!

Le médicament de un an

Le paradoxe du médicament de un an est encore plus frappant. Imaginons qu'un nouveau médicament-miracle empêche totalement de mourir dans l'année qui suit son absorption, sauf ceux de la classe d'âge la plus élevée (laquelle mourra dans l'année), mais qu'il n'ait aucun effet au-delà. Une seule prise du médicament est efficace; imaginons aussi que tous les Français aient pris le médicament le 1er janvier 2016 et donc qu'aucun ne soit mort en 2016 sauf les individus de 114 ans (l'âge du plus vieux des Français, Honorine Rondello). Alors l'espérance de vie en 2016 sera exactement de 114 ans, bien qu'en réalité, la vie de chaque Français aura été prolongée au plus de une année! En effet, avec nos hypothèses, le taux de mortalité par classe d'âge est de 0% quelle que soit la classe d'âge, sauf pour la dernière, et donc tous les individus de la population fictive qui sert au calcul de l'espérance de vie pour l'année 2016 atteindront l'âge maximal constaté en France, âge auquel ils décéderont tous. L'espérance de vie à la naissance ou à n'importe quel âge est donc de 114 ans exactement pour les données de 2016... et redevient normale dès 2017. La méthode de calcul de l'espérance de vie n'a rien d'absurde, mais il est bon de ne pas lui attribuer plus de sens qu'elle n'en a.

> à 916,34 euros, ce qui correspond à une augmentation de 152%. Pourtant, à nouveau personne ne ment. Comment est-ce possible?

Le tableau répond: le salaire hebdomadaire des ouvriers, passant de 200 euros à 180 euros, a baissé de 10%. Celui des cadres, passant de 2000 à 1800 euros, a lui aussi diminué de 10%. Les syndicats ont donc raison d'affirmer que les salaires ont baissé de 10%.

De son côté, le patron ne triche pas! Le salaire versé aux 1100 employés de l'entreprise en 2006 était chaque semaine de 400000 euros ($1000 \times 200 + 100 \times 2000$), soit 363,64 euros par employé. En 2007, le salaire versé aux 1100 employés (l'effectif global est inchangé) s'est élevé à 1008000 euros ($180 \times 600 + 1800 \times 500$), soit 916,34 euros par employé. Le salaire moyen a donc bien augmenté de 152%.

L'arnaque, car il y en a une, est que les effectifs par catégorie ont changé entre 2016 et 2017. Il en résulte que la baisse du salaire dans chaque catégorie est compensée par l'augmentation du nombre des cadres qui sont mieux rétribués. Une baisse de 10% du salaire de chaque catégorie de personnel est parfaitement compatible avec une augmentation du salaire moyen des employés.

Cette situation n'est pas tant fictive que cela, car chaque année l'État français insiste sur les chiffres de la masse salariale des fonctionnaires et sur le salaire moyen d'un fonctionnaire qui évoluent plus favorablement que le point d'indice utilisé pour payer les salaires des fonctionnaires. Ce point d'indice détermine, à grade fixé, le salaire d'un fonctionnaire et bien sûr c'est à lui que les syndicats préfèrent se fier. L'augmentation de l'âge moyen des fonctionnaires – due en particulier à un fort recrutement dans l'enseignement dans les années 1970 et 1980 – entraîne mécaniquement une augmentation du grade moyen (comme dans l'entreprise fictive de l'exemple indiqué plus haut) et conduit donc à une divergence entre les chiffres portant sur le salaire moyen d'un fonctionnaire et ceux portant sur le point d'indice, que ministres et syndicats utilisent de la façon qui les arrange le mieux.

Toujours à propos des fonctionnaires, voyons un autre paradoxe apparent. On peut affirmer sans se contredire que: (*a*) le salaire moyen dans la fonction publique est supérieur à celui dans le privé; (*b*) la majorité des fonctionnaires gagneraient plus s'ils partaient dans le privé. L'explication réside à nouveau dans la répartition des emplois du secteur public et du secteur privé: dans le premier, le nombre d'emplois qualifiés est plus important, ce qui a pour effet d'augmenter le salaire moyen du secteur public, sans pour autant contredire qu'à diplôme égal on gagne moins dans le public que dans le privé.

La statistique utilise des indices aux définitions précises qui résument en un seul nombre une masse parfois considérable de chiffres. Ces indices sont inévitables: sans eux, on ne pourrait synthétiser des tableaux de données. Cependant les indices des statisticiens tendent des pièges et peuvent engendrer des résultats totalement absurdes. Soit que celui qui les examine les comprenne imparfaitement, soit qu'il prenne mal en compte des situations particulières.

L'espérance de vie est un indice piège (*voir l'encadré page précédente*). Le seuil de pauvreté tel que le définit l'Insee est tout aussi troublant. Par définition, le seuil de pauvreté dans un pays est la moitié du revenu médian, c'est-à-dire la moitié du revenu X tel qu'il y a autant de gens gagnant plus de X que de gens gagnant moins dans le pays. L'indice de pauvreté est le pourcentage de gens vivant sous le seuil de pauvreté.

Si maintenant quelqu'un vous annonce que dans tel pays 60% des gens vivent sous le seuil de pauvreté, c'est une ânerie. Par définition, ce n'est pas possible: la moitié (à quelques unités près) des gens ont un revenu inférieur au revenu médian, et nécessairement moins de la moitié ont un revenu inférieur à la moitié du revenu médian.

La conséquence dramatique est que même dans un pays où tout le monde mourrait de famine, le nombre de pauvres resterait inférieur à 50%, alors qu'à l'opposé, dans un pays composé uniquement de milliardaires, il pourrait tout à fait y avoir 45% de gens vivant sous le seuil de pauvreté. La pauvreté et la richesse sont des concepts relatifs, certes, mais tout de même!

ÊTRE PAUVRE EN COCAGNE

Nicolas Gauvrit imagine la petite histoire suivante qui montre par l'absurde combien l'utilisation de l'indice de pauvreté est dangereuse. Au pays de Cocagne, une pomme coûte un millième de sol. Un logement en coûte 5. On mange bien pour 0,2 sol. On vit dans le confort pour 100 sols par mois, et comme un nabab pour 200 sols par mois. Il y a dans ce pays deux types de travailleurs: les ouvriers qui ont un revenu de 1000 sols par mois, et les penseurs qui ont un revenu mensuel de 3000 sols. Il y a autant d'ouvriers que de penseurs, exactement. Tous vivent en harmonie et dans une opulence que les pays voisins jalousent. Seule exception à la règle des deux salaires: le président du pays touche, pour sa part, une rémunération de 2800 sols. Le revenu médian est alors de 2800 sols, comme vous le confirmerait tout statisticien. La moitié du revenu médian est donc de 1400 sols, et la moitié de la population (sans compter le président) touche une rétribution inférieure. L'indice de pauvreté est donc de 50% et c'est la valeur la plus élevée possible de cet indice.

Mais un beau jour de mai 2068, le président accepte de baisser ses indemnités, car, dit-il «il n'y a pas de raison pour que je réclame plus qu'un ouvrier! Je suis un travailleur comme les autres». Tenant compte de ses responsabilités, il s'accorde toutefois une petite rallonge et se déclare satisfait de son nouveau revenu de 1400 sols mensuels.

CITATIONS

«Le loto, c'est un impôt sur les gens qui ne comprennent pas les statistiques.»

Anonyme

«Il est statistiquement prouvé que sur dix personnes atteintes de bronchite, une seule va chez son médecin et les neuf autres dans une salle de spectacle.»

Anonyme

«En moyenne chaque personne possède un testicule.»

Anonyme

«À la question: "Faites-vous encore confiance aux sondages?", 64% des Français répondent OUI et 59% répondent NON.»

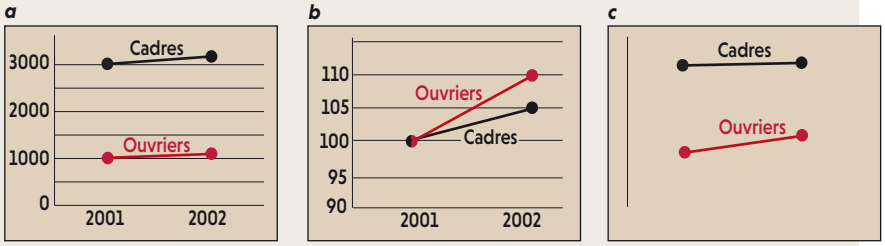
Philippe Geluck

L'ART DU DESSIN

Il y a mille façons de projeter graphiquement des données, chacune permettant de mettre en avant un fait, d'en masquer un autre, ou de donner à croire autre chose que ce que les données indiquent. L'exemple suivant a d'abord été proposé dans la revue *The Economist*. Entre 2001 et 2002, les salaires des cadres et des ouvriers ont évolué ainsi :

	SALAIRES MENSUELS MOYENS	
	2001	2002
OUVRIERS	1000 €	1100 €
CADRES	3000 €	3150 €

En moyenne, les cadres et les ouvriers gagnent plus en 2002 qu'en 2001. En valeurs brutes, l'augmentation des cadres (150 euros) est supérieure à celle des ouvriers (100 euros). Cependant, la situation est inversée en valeur relative : 5 % pour les cadres contre 10 % pour les ouvriers. Ces points de vue différents ont leurs équivalents graphiques, respectivement en *a* et en *b*.
Le commentaire naturel de la représentation *a* est : « Les ouvriers et les cadres ont été augmentés. Les cadres recevaient plus, et reçoivent encore plus que les ouvriers. L'écart entre les salaires empire. »
La représentation *b* s'attache aux augmentations relatives. Elle montre que les cadres ont été moins augmentés en pourcentage que les ouvriers, mais elle ne renseigne pas sur les éventuelles modifications de rémunération en valeur absolue. Cette courbe donne l'impression que les cadres sont moins gâtés que les ouvriers…
Peut-on représenter les évolutions relatives et les différences entre les professions ? Oui, en ayant recours à une échelle logarithmique en ordonnée *(c)*. Avec cette graduation, une même différence de hauteur correspond à un même coefficient multiplicateur.
La lecture de ce graphique suggère cette fois : « Les cadres recevaient et reçoivent encore plus que les ouvriers, mais les inégalités se réduisent. »
Ces trois représentations étant toutes parfaitement honnêtes, on peut commenter l'évolution des salaires selon sa préférence.



Le revenu médian en Cocagne passe alors illico à 1400 sols, et le demi-revenu médian à 700 sols, et par un miracle statistique à couper le souffle, l'indice de pauvreté passe aussitôt, en mai 2068, à 0, le minimum possible. »

Créé suite à une polémique sur l'indice de pauvreté, le BIP 40 ou baromètre des inégalités et de la pauvreté est un indicateur synthétique des inégalités et de la pauvreté. Cet indice a été proposé par le Réseau d'alerte sur les inégalités en 2002 et il élimine certains des problèmes mentionnés ci-dessus. Toutefois, il est assez complexe, si bien qu'il est délicat d'en comprendre le sens et qu'on ne peut garantir qu'il évitera toutes les absurdités des indices plus simples.

UNE AFFAIRE DE FAMILLE

Parmi tous les pièges que détaille le livre de Nicolas Gauvrit, l'un d'eux est assez subtil et mérite une attention particulière, nous le nommerons le paradoxe du nombre moyen d'enfants. Une enquête exhaustive menée dans une ville lointaine indique que les familles ayant des enfants de moins de 18 ans se répartissent de la manière suivante : 10 % de familles à 1 enfant, 50 % à 2 enfants, 30 % à 3 enfants, 10 % à 4 enfants. Le nombre moyen d'enfants par famille (parmi celles qui ont des enfants) est donc de $(10 + 100 + 90 + 40)/100 = 2{,}4$.

Pour contrôler cette statistique, les autorités administratives procèdent à un sondage. On interroge 1000 enfants de moins de 18 ans soigneusement pris au hasard et on leur demande combien il y a d'enfants dans leur famille, eux compris. En faisant la moyenne des réponses, on obtient… 2,68 ! Cela semble absurde. On recommence donc le sondage en interrogeant cette fois 10000 enfants, on trouve maintenant 2,67. Un troisième sondage sur 100000 enfants donne 2,668 à nouveau. Pourquoi cet écart si important avec les 2,4 de la statistique qui prenait en compte toutes les familles ayant des enfants ?

La réponse tient dans le fait qu'en interrogeant des enfants au hasard, vous interrogerez 4 fois plus d'enfants des familles à 4 enfants que vous n'en interrogerez dans les familles à 1 enfant, ce qui fausse la moyenne. S'il y a 1000 familles, il y aura 100 enfants uniques, 1000 enfants appartenant à une famille de 2 enfants, 900 enfants appartenant à une famille de 3 enfants, 400 enfants appartenant à une famille de 4 enfants. Au total, les réponses données par ces 2400 enfants conduiront au résultat de 2,666… enfants par famille.

Les sondages opérés n'évaluent pas le nombre moyen d'enfants d'une famille prise au hasard, mais le nombre moyen d'enfants qu'on trouve dans la famille d'un enfant pris au hasard. « Prendre une famille au hasard » et « Prendre un enfant au hasard » n'est pas la même chose.

Nous le savons depuis Condorcet pour les votes, il est bien difficile de synthétiser un ensemble de nombres en un seul. Nous espérons que les pièges de la statistique, des représentations graphiques, des indices synthétiques, des sondages présentés ici – et ceux que vous trouverez dans le livre de Nicolas Gauvrit – vous aideront à mieux comprendre les vérités cachées derrière l'inquiétant et fluctuant monde des nombres. ■

BIBLIOGRAPHIE

N. GAUVRIT, Statistiques. Méfiez-vous, *Éditions Ellipses*, 2014.

I. EKELAND, Statistiques incroyables, *Pour la Science* n° 334, p. 6, août 2005.

P. BICKEL *ET AL*, Sex Bias in Graduate Admissions: Data from Berkeley, *Science*, vol. 187, n° 4175, pp. 398-404, 1975.

G. YULE, Notes on the Theory of Association of Attributes in Statistics, *Biometrika*, vol. 2, n° 2, pp. 121-134, 1903.

L'ESSENTIEL

- La loi des grands nombres et le théorème de la limite centrale décrivent le comportement de la moyenne d'un grand nombre de petites contributions indépendantes.
- Ils échouent cependant lorsque des événements rares ou extrêmes dominent.
- D'autres outils statistiques – lois de Lévy ou des valeurs extrêmes – sont alors plus indiqués.
- Mais quand le risque n'est pas quantifiable, ces outils se révèlent à leur tour impuissants.

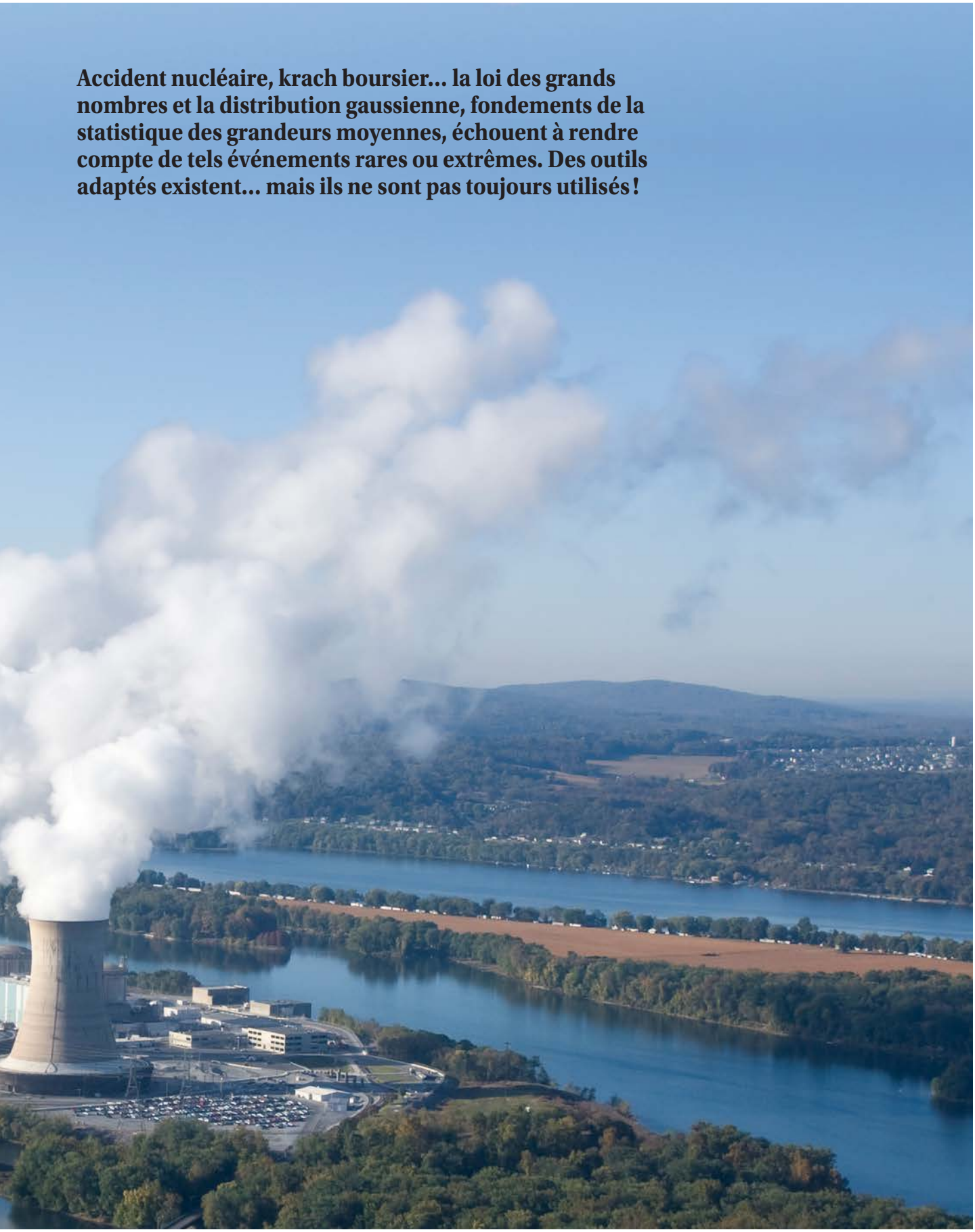
L'AUTEUR

RAMA CONT
est directeur de recherche CNRS au Laboratoire de probabilités et modèles aléatoires, à Paris. Ses travaux portent sur les processus aléatoires et la modélisation mathématique des risques financiers.

Prévoir l'IMPROBABLE

L'accident de la centrale nucléaire de Three Mile Island, en 1979, résulte d'un enchaînement improbable de petites erreurs – dont une humaine –, chacune étant difficilement quantifiable. Était-il prévisible ?

Accident nucléaire, krach boursier… la loi des grands nombres et la distribution gaussienne, fondements de la statistique des grandeurs moyennes, échouent à rendre compte de tels événements rares ou extrêmes. Des outils adaptés existent… mais ils ne sont pas toujours utilisés !



Le 15 septembre 2008, la banque américaine Lehman Brothers se déclare en faillite. Cet événement, consécutif de la crise des *subprimes* l'année précédente, ébranle le système financier mondial et affole les bourses de la planète. Les gouvernements et les banques centrales enchaînent les plans de relance, mais rien n'y fait, le monde plonge dans la crise financière la plus grave depuis celle de 1929. Dans quelle mesure aurait-on pu prévoir ce cataclysme ?

De tels événements rares et catastrophiques (aux krachs boursiers, ajoutons les tremblements de terre, les inondations, les accidents nucléaires…) fascinent par leur caractère imprévisible et l'importance des enjeux sociaux et scientifiques qu'ils représentent. Ils sont aussi des défis à la modélisation scientifique. Exceptionnels presque par définition, ou parfois mis de côté comme « aberrations statistiques », les événements rares sont néanmoins accessibles à la théorie statistique… pour peu qu'elle sollicite des outils mieux adaptés que ceux du quotidien.

Dans beaucoup d'applications, les statistiques se résument au calcul de moyennes et d'écarts-types, et la distribution des observations se répartit sagement selon la fameuse « courbe en cloche » ou distribution gaussienne.

LES STATS POUR LES NULS

Le statisticien représente souvent un échantillon d'observations (températures, précipitations, indices boursiers…) comme une suite de tirages indépendants d'une « variable aléatoire » notée X et représentant la grandeur en question. On s'intéresse alors à la moyenne, ou espérance mathématique, de cette variable aléatoire, à son écart-type (la dispersion des valeurs autour de la moyenne) et, plus généralement, à sa distribution de probabilité.

La loi des grands nombres, établie pour la première fois dans sa version la plus simple au XVIIIe siècle par Jacques Bernoulli, assure que, lorsque la taille de l'échantillon est assez grande, la moyenne empirique de l'échantillon tend vers la moyenne théorique de la variable aléatoire qu'il représente. De même, l'écart-type empirique converge vers l'écart-type de la variable aléatoire. En garantissant qu'un échantillon fournit une bonne estimation d'une variable aléatoire réelle, la loi des grands nombres s'impose comme un des piliers de la statistique.

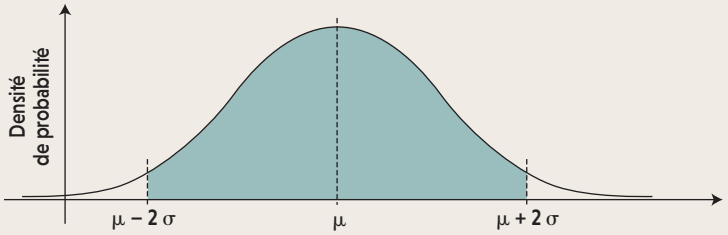
La loi des grands nombres est une loi asymptotique : la moyenne d'un échantillon ne converge vers l'espérance mathématique que si la taille N de l'échantillon est grande. La différence entre les deux valeurs – l'erreur statistique – est régie par le « théorème de la limite centrale » (*voir l'encadré page ci-contre*). Ce théorème, second pilier des statistiques, est connu sous des formes diverses depuis le XVIIIe siècle, mais sa forme définitive, due au mathématicien français Paul Lévy, date du début du XXe siècle. Il affirme que, si la variable aléatoire X a un écart-type fini, alors pour un échantillon de N observations indépendantes de X, la différence entre la moyenne de l'échantillon et la moyenne de X est bien représentée par une distribution gaussienne dont l'écart-type est proportionnel à $1/\sqrt{N}$.

La distribution gaussienne apparaît donc comme une description statistique de toute quantité que l'on peut représenter comme la somme d'un grand nombre de petites contributions indépendantes.

Ces deux principes ont une grande portée. La loi des grands nombres est le fondement des sondages, car elle montre qu'il suffit de sonder un échantillon assez grand pour approcher les caractéristiques statistiques de la population réelle. Le théorème de la limite centrale permet, pour sa part, de contrôler l'erreur ainsi commise : l'écart entre la moyenne empirique et la moyenne théorique suit une loi gaussienne.

La moyenne et l'écart-type n'ont plus forcément de sens pour les distributions inégalitaires

LOI DES GRANDS NOMBRES ET THÉORÈME DE LIMITE CENTRALE



Le théorème de la limite centrale permet de quantifier la probabilité que la moyenne empirique approche la moyenne réelle avec une précision donnée.

La loi des grands nombres énonce que lorsqu'on échantillonne une variable aléatoire, sous certaines conditions, lorsque la taille de l'échantillon est assez grande, la moyenne empirique de cet échantillon tend vers la moyenne théorique de la variable aléatoire échantillonnée.
C'est sur cette loi que reposent notamment les sondages. Considérons un échantillon d'observations $x_1, x_2, \ldots x_n$ d'une variable aléatoire X, d'espérance μ et d'écart-type σ finis. La loi forte des grands nombres affirme que quand n tend vers l'infini, la moyenne empirique $M_n = (x_1 + x_2 + \ldots + x_n)/n$ converge presque sûrement vers μ, c'est-à-dire que la probabilité que la moyenne empirique converge vers la moyenne théorique est égale à 1.

$$\mathbb{P}\left(\lim_{n\to+\infty} M_n = \mu\right) = 1$$

En d'autres termes, la moyenne d'un grand nombre d'observations aléatoires indépendantes n'est plus aléatoire ! La loi des grands nombres assure ainsi que la moyenne empirique est un estimateur convergent (ou « consistant ») de l'espérance mathématique. Cela reste vrai tant que la contribution de chaque terme x_i à la moyenne reste négligeable : aucune observation n'est assez grande pour dominer la somme. Cette situation est qualifiée de hasard « sage ».
Le théorème de la limite centrale est l'autre pilier de l'échantillonnage. Considérons des variables aléatoires X_i, indépendantes et de même distribution (d'espérance μ et d'écart-type σ). Alors la somme centrée $(X_1 + X_2 + \ldots + X_n - n\mu)/\sqrt{n}$ tend vers une variable aléatoire normale (ou gaussienne) d'espérance nulle et d'écart-type σ quand n tend vers l'infini. La densité de probabilité d'une variable gaussienne de moyenne m et écart-type σ est définie par la fonction

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

et sa représentation graphique est la célèbre courbe en cloche.
Le théorème de la limite centrale reste valable même si les distributions individuelles sont différentes, pour autant que l'écart-type de chacun des termes soit négligeable vis-à-vis de l'écart-type de la somme.
Ce théorème implique que la différence entre la moyenne empirique M_n de l'échantillon et l'espérance μ de la variable aléatoire X est représentée par une distribution gaussienne dont l'écart-type est proportionnel à $1/\sqrt{n}$. Cela permet de contrôler l'erreur faite en approximant la moyenne théorique μ par la moyenne empirique M_n. La probabilité que M_n se trouve dans l'intervalle $[\mu - t\sigma, \mu + t\sigma]$ est représentée, pour n assez grand, par l'aire sous la courbe en cloche comprise entre les abscisses $\mu - t\sigma$ et $\mu + t\sigma$. Il y a par exemple 95 % de chances que l'écart entre la moyenne empirique M_n et l'espérance μ soit inférieur à 2σ.

Par exemple, si un sondage électoral porte sur 10 000 personnes et que 46 % d'entre elles se déclarent favorables au candidat A, le théorème de la limite centrale indique qu'il y a 95 % de chance que lors du vote, il recueille entre 45 et 47 % des voix. Concrètement, il n'est pas nécessaire de sonder un très grand échantillon de la population pour avoir une estimation fiable.

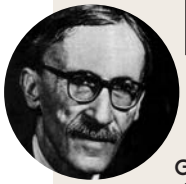
Le théorème de la limite centrale est invoqué pour justifier la représentation par une distribution gaussienne dans des domaines allant des sciences sociales à la physique, en passant par l'économie, la biologie ou la finance. La courbe en cloche est mise à toutes les sauces. On l'utilise même pour « normaliser » la distribution des notes dans les concours…

Cependant, l'omniprésence de la distribution gaussienne en statistique (d'où son qualificatif de « loi normale ») est peut-être due à la trop grande confiance des statisticiens plutôt qu'à la réalité : tout phénomène ne peut pas être ramené à une somme de petites contributions indépendantes. De fait, en sciences économiques et sociales, la loi normale paraît plutôt être l'exception que la règle.

Dès le début du XX^{e} siècle, l'économiste Vilfredo Pareto s'est par exemple intéressé à la richesse nationale italienne, et a constaté que les 20 % les plus riches de la population détenaient 80 % de la richesse totale du pays. Cette situation d'inégalité n'est manifestement pas bien représentée par une distribution gaussienne des richesses, où la richesse individuelle serait concentrée autour de la richesse moyenne et où les individus très riches ou très pauvres seraient très rares.

Pareto a proposé une distribution plus réaliste qui porte aujourd'hui son nom : la part d'individus ayant une richesse supérieure à un niveau u est proportionnelle à $1/u^{\alpha}$. Plus l'exposant α de cette loi de puissance est petit, plus >

LES LOIS STABLES DE LÉVY

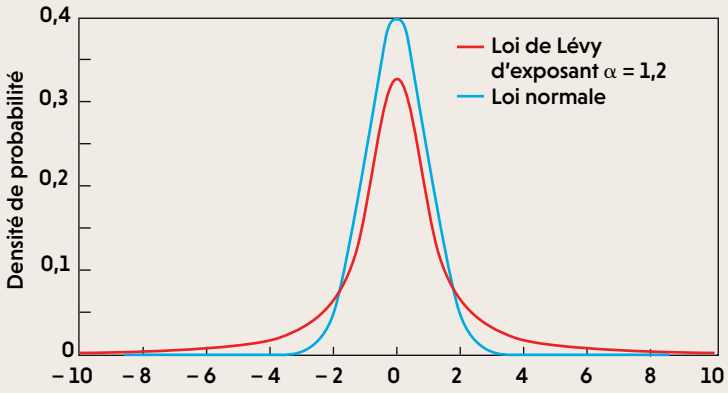


Paul Lévy était un pionnier de la théorie moderne des processus aléatoires. Il découvrit dans les années 1920, en même temps que Boris Gnedenko en Union soviétique, les distributions « stables » qui portent aujourd'hui son nom et étendit le théorème de la limite centrale au cas du « hasard sauvage ».

Soit une suite $X_1, X_2, \ldots, X_n$ de variables aléatoires indépendantes centrées, dont la queue de distribution se comporte comme une loi de Pareto $1/x^{\alpha}$ avec $0 < \alpha < 2$ (l'écart-type est alors infini) ; alors la somme $X_1 + X_2 + \ldots + X_n$ n'obéit plus au théorème de la limite centrale.

La distribution de $(X_1 + X_2 + \ldots + X_n)/n^{1/\alpha}$ se comporte, pour n assez grand, non comme une loi gaussienne, mais comme une « loi stable de Lévy » d'exposant. À la différence du théorème de la limite centrale, le facteur de normalisation est ici $n^{1/\alpha}$ au lieu de n.

Les distributions stables n'ont pas de formules analytiques explicites, sauf dans le cas $\alpha = 1$ (loi de Cauchy) et $\alpha = 1/2$. Leur queue de distribution se comporte comme $1/|x|^{\alpha}$ pour $|x|$ assez grand : la somme a la même queue de distribution que chacun de ses termes.



> cette proportion est grande et plus la distribution des richesses est inégalitaire. Face à de telles distributions, les notions de moyenne et d'écart-type se révèlent inadaptées et offrent une description trompeuse.

LES MATHS DES INÉGALITÉS

En effet, la richesse moyenne et l'écart-type peuvent par exemple rester identiques alors que la distribution des richesses se modifie (ce qui est impossible dans le cadre d'une distribution gaussienne).

Pire, si l'exposant de la loi de Pareto est inférieur à deux, alors l'écart-type de la distribution est infini. S'il est inférieur à un, c'est aussi le cas pour la moyenne. Ces indicateurs sont donc dénués de sens pour les lois de Pareto.

Ce fait a conduit les économistes à définir d'autres indicateurs pour mieux caractériser les distributions inégalitaires. Le coefficient de Gini, proposé par le statisticien italien Corrado Gini en 1912, est ainsi un nombre compris entre 0 et 1 qui mesure le degré d'inégalité de la distribution des richesses. Plus il est proche de 1, plus la répartition est inégalitaire. Dans le cas d'une loi de Pareto d'exposant α, le coefficient de Gini vaut $1/(2\alpha - 1)$. Ces indicateurs se focalisent sur les plus grands événements dans un échantillon, événements d'occurrence faible, mais d'importance prépondérante. Ces « événements rares » forment des queues de distributions.

Dès lors que la moyenne et l'écart-type n'ont plus forcément de sens pour les distributions inégalitaires, qu'advient-il de la loi des grands nombres et du théorème de la limite centrale ? Dans les années 1920, Lévy fut l'un des premiers à remarquer que les hypothèses qui les sous-tendent peuvent cesser d'être valables dans le cas où des événements rares influent de façon prépondérante sur la moyenne de l'échantillon.

Lévy étendit le théorème de la limite centrale à ces situations : il montra que pour un échantillon de variables centrées dont la queue de distribution se comporte comme une loi de Pareto avec un exposant compris entre 0 et 2, la moyenne empirique se comporte non plus comme une variable gaussienne, mais selon une distribution que Lévy appela loi stable, dont la queue se comporte comme une loi de Pareto de même exposant. En d'autres termes, la loi limite est une loi de Lévy, et non une loi normale.

Pour ces distributions de Lévy, variance et écart-type n'ont plus de sens : ils sont infinis. Benoît Mandelbrot, qui fut élève de Lévy à l'École polytechnique, montra en 1963 que, loin d'être une curiosité mathématique, les situations décrites par les lois de Lévy étaient omniprésentes en économie, en finance et en assurance : les rendements des matières premières, la richesse des individus ou la taille des entreprises suivent des lois de Pareto.

Mandelbrot qualifia ces situations, où le comportement de l'échantillon n'est plus correctement caractérisé par la moyenne et l'écart-type, de hasard sauvage, par opposition au hasard sage de la loi gaussienne.

DOMPTER LE HASARD SAUVAGE

Ce comportement sauvage est manifeste dans les rendements d'indices boursiers (*voir l'encadré page 27*). Alors que l'amplitude des fluctuations d'un échantillon gaussien reste de l'ordre de son écart-type, certains indices boursiers présentent des variations d'amplitude erratiques atteignant plusieurs dizaines de fois l'écart-type ! De nombreuses autres situations se conformant aux lois de Lévy ont été rapidement mises en évidence en physique statistique.

L'épithète « sauvage » suggère l'incontrôlabilité du hasard, mais une bonne compréhension de l'aléa aide parfois à l'apprivoiser. C'est le cas dans certains systèmes physiques, ainsi que dans des situations, où,

indépendamment du fait que l'aléa sous-jacent soit sage ou sauvage, la moyenne ou l'écart-type d'un échantillon ne sont tout simplement pas les indicateurs pertinents, même si on peut les estimer avec précision.

Il en est ainsi des études de fiabilité, où, pour déterminer la probabilité de défaillance d'un système, on cherche à estimer celle de son « maillon le plus faible ». Par exemple, pour choisir la dimension et les caractéristiques d'un barrage, il faut avoir une idée précise de la pression maximale que l'ouvrage pourra être amené à supporter, et non de la pression moyenne qu'il subira en conditions normales. Pour cela, il faut caractériser directement les valeurs extrêmes dans l'échantillon d'observations. Les notions d'écart-type, de moyenne et le théorème de la limite centrale ne sont alors pas d'un grand secours...

Parallèlement aux travaux de Lévy sur le théorème de la limite centrale, un groupe de statisticiens, dont sir Ronald Fisher et Leonard Tippett en Grande-Bretagne, Boris Gnedenko en Union soviétique, Maurice Fréchet en France et Ludwig von Mises en Autriche, ainsi qu'Emil Julius Gumbel, mathématicien allemand réfugié en France, développaient une nouvelle branche des statistiques dont le but est d'étudier non plus les valeurs moyennes des échantillons, mais leurs valeurs extrêmes, c'est-à-dire maximales ou minimales.

De même que le théorème de la limite centrale décrit le comportement de la moyenne d'un grand échantillon, cette théorie des valeurs extrêmes décrit le comportement du maximum d'un échantillon. Ces statisticiens ont montré que, selon la nature de la variable aléatoire étudiée, la valeur maximale de l'échantillon ne peut obéir qu'à l'une parmi trois distributions différentes : les distributions de Fréchet, de Gumbel ou de Weibull (*voir l'encadré ci-contre*).

Comme la loi normale et la loi de Lévy, qui sont des candidats naturels pour modéliser les sommes de variables indépendantes, les lois de valeurs extrêmes sont des candidats privilégiés pour la modélisation des valeurs maximales d'un échantillon. Gumbel présenta en 1941 une application de ces concepts à l'analyse statistique des crues fluviales. Cette théorie allait être mise en application dans un contexte dramatique.

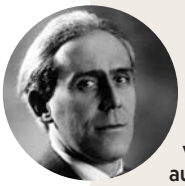
APPRIVOISER LA RARETÉ... ET LA MER

Les Pays-Bas ont toujours vécu sous la menace de la montée des eaux. Les nombreuses digues qui parsemaient le pays ont longtemps suffi à le protéger. Mais dans la nuit du 31 janvier au 1er février 1953, une montée des eaux due à une tempête eut raison de leur résistance et submergea le Sud-Ouest du pays, tuant plus de 1800 personnes et des milliers de têtes de bétail.

Sous le choc, le gouvernement néerlandais lança peu après un vaste projet de construction de barrages. La commission chargée de revoir les normes de sécurité exigea de fixer la hauteur des digues de façon à ce que la probabilité que la marée les dépasse soit inférieure à 1/10000. Une équipe de scientifiques, armée de la récente théorie des valeurs extrêmes, se mit alors à étudier marées et crues passées afin d'estimer la distribution de probabilité de la hauteur maximale des inondations. Ils montrèrent qu'elle suit une loi de Gumbel, et que la hauteur requise dépassait 5 mètres, une nette augmentation par rapport aux digues de l'époque. Sur la base de ces estimations, d'énormes barrages furent construits (*voir la photo page 26*) ; ces édifices protègent encore le pays aujourd'hui.

Dans une ambiance de foi grandissante dans les méthodes quantitatives, la théorie des valeurs extrêmes fut par la suite appliquée aux >

LES LOIS DE VALEURS EXTRÊMES

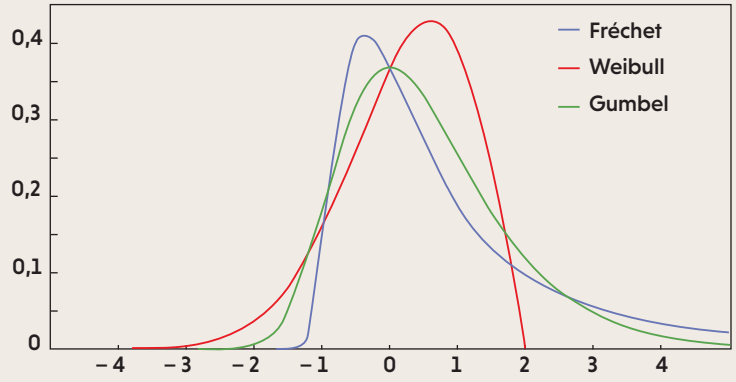


Considérons un échantillon $x_1, x_2, \ldots, x_n$ de tirages indépendants d'une distribution de probabilité G. La théorie des valeurs extrêmes s'intéresse au comportement de la valeur maximale $U_n = \max(x_1, x_2, \ldots, x_n)$ d'un échantillon de grande taille (n grand). Un des résultats fondamentaux de la théorie, le théorème de Fisher-Tippett-Gnedenko, montre que s'il existe des constantes de normalisation a_n et b_n telles que la distribution de $(U_n - a_n)/b_n$ ait une limite quand n tend vers l'infini, alors la distribution limite de la suite normalisée de maxima des variables aléatoires est nécessairement de la forme :

$$F(x;\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

définie pour $1+\xi(x-\mu)/\sigma > 0$, et $\sigma > 0$.

On distingue alors trois cas.
Le premier, quand $\xi < 0$, nommé distribution de Weibull, décrit le cas où les variables x_i sont bornées.
Le second, lorsque $\xi = 0$, correspond à la distribution de Gumbel. Elle est au maximum d'un échantillon ce que la loi normale est à la moyenne : elle représente le cas où le hasard décrit par la distribution G est sage.
Enfin, si les variables x_i suivent une loi de Pareto d'exposant α, alors $\xi = 1/\alpha > 0$. La distribution, dite de Fréchet, décrit le maximum d'un échantillon qui obéit au hasard sauvage. Plus ξ est grand, plus la queue de distribution est importante, c'est-à-dire plus les événements extrêmes sont fréquents.
Ces trois lois de probabilité sont adaptées aux distributions de valeurs extrêmes.



> assurances, à la gestion des risques financiers et celle des risques environnementaux. Les assureurs l'utilisent par exemple pour estimer les plus grosses pertes qu'ils peuvent subir en vendant des assurances contre les catastrophes naturelles.

Le scénario catastrophe par excellence étant l'explosion d'une centrale nucléaire, toutes les idées sur la mesure des risques extrêmes ont trouvé tôt ou tard un terrain d'application dans le domaine de la sûreté nucléaire. Dans les années 1970, sont apparues les études probabilistes de sûreté (EPS), une méthodologie d'estimation de la probabilité des scénarios d'accident dans les installations nucléaires. La méthode consiste à identifier les séquences d'événements pouvant mener à un résultat catastrophique, telle la fusion du cœur de réacteur ; et assigner des probabilités à ces séquences à partir de principes physiques ou de l'expérience acquise. L'objectif, comme pour les digues hollandaises, est de définir les normes à satisfaire pour garantir une probabilité d'occurrence de l'événement catastrophique inférieure à un seuil donné.

Mais en 1979, l'accident de la centrale nucléaire de Three Mile Island, en Pennsylvanie, aux États-Unis, rappela brutalement que, dans un système complexe formé de multiples composantes interdépendantes comme une centrale nucléaire, une petite erreur – humaine en l'occurrence – peut provoquer une catastrophe (*voir la figure pages 20-21*).

Or l'occurrence de telles erreurs, qui ne sont pas liées à un facteur physique obéissant à une loi bien identifiée, est difficile, voire impossible, à quantifier. Dès lors, quelle foi accorder aux « certitudes quantitatives » produites par les études de sûreté et sur les certifications techniques qui en résultent ?

Frank Knight, économiste américain et théoricien du risque, distinguait déjà en 1921 dans *Risk, Uncertainty and Profit* le risque probabilisable, correspondant à des événements dont on peut évaluer la probabilité avec une certitude raisonnable, et l'incertitude, qui recouvre des événements dont on ignore jusqu'aux probabilités d'occurrence. Lorsqu'on s'éloigne des systèmes physiques pour aller vers des domaines où les facteurs humains sont plus importants, la part de l'incertitude augmente.

Données d'observation incomplètes, hypothèses abusivement simplificatrices (telle l'indépendance des facteurs de risques, fréquemment postulée dans les modèles statistiques), ou encore omission de certains facteurs de risque, les sources d'incertitudes dans les modèles de risque sont nombreuses. Elles sont autant de limitations qui invitent à rester modestes dans nos attentes envers les modèles statistiques lorsque ceux-ci ne s'appuient pas sur des principes physiques bien ancrés.

LA CHASSE AU CYGNE NOIR

Que faire alors ? Tout au plus peut-on compléter les résultats des modèles par l'analyse au cas par cas de scénarios extrêmes. C'est l'objet de la méthode de *stress test*, qui apporte un complément utile aux analyses statistiques de risque. Le CEA et EDF ont retenu cette option en complément des études EPS pour définir la politique de sûreté nucléaire.

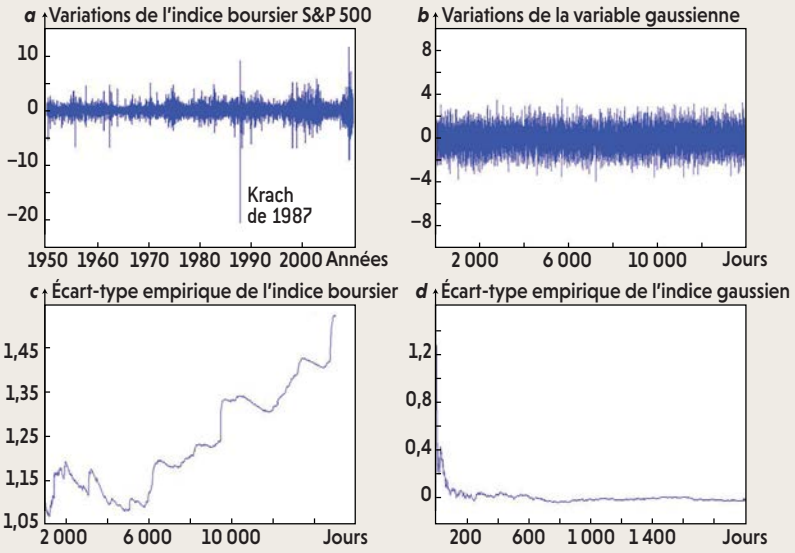
Les statistiques montrent là leurs limites. Faut-il pour autant abandonner l'ambition de quantifier les événements rares en dehors des sciences physiques, où les principes physiques permettent de calculer leurs probabilités ?

C'est ce que certains semblent suggérer, tels Nassim Taleb ou Nicolas Bouleau, de l'École nationale des ponts et chaussées, à la

Le barrage d'Oosterschelde, aux Pays-Bas, a été conçu pour que la probabilité qu'une marée le dépasse soit inférieure à 1/10 000. La théorie des valeurs extrêmes indiqua que la hauteur des digues devait être de 5 mètres au moins.

HASARD SAUVAGE CONTRE HASARD SAGE

Benoît Mandelbrot *(ci-contre)* fut le premier à mettre en évidence dans les années 1960 les manifestations du hasard sauvage en économie et en finance. On peut illustrer la différence entre hasard sauvage et hasard sage en comparant les rendements journaliers d'un indice boursier réel (l'indice S&P 500 de la Bourse de New York) de janvier 1950 à septembre 2009, et un échantillon de 14 000 observations simulées d'une variable aléatoire gaussienne de même moyenne (0) et même écart-type (1 %) que les données réelles. Sur le graphique de l'indice boursier *(a)*, on voit que la variabilité est grande et que l'indice connaît des bouffées de volatilité. En octobre 1987, l'indice a perdu 20 % de sa valeur en une journée, soit 20 fois son écart-type ! *A contrario*, les variations de la série gaussienne *(b)* restent concentrées autour de la moyenne : la plus grande perte est de 4 %, soit seulement 4 fois l'écart-type journalier. Les écarts-types empiriques des deux séries évoluent aussi de façons très différentes en fonction du nombre d'observations. Alors que l'écart-type de l'échantillon gaussien *(d)* converge doucement vers sa valeur théorique, comme le prévoit la loi des grands nombres, l'écart-type empirique des rendements de l'indice boursier *(c)* continue à osciller à mesure que de nouvelles observations sont ajoutées à l'échantillon, sans donner aucun signe de convergence. Ces observations suggèrent que les aléas financiers relèvent du hasard sauvage, domaine où les événements extrêmes jouent un rôle déterminant. Mandelbrot proposa ainsi de modéliser les mouvements boursiers à l'aide des distributions stables de Lévy.



L'indice boursier S&P 500 obéit à un hasard sauvage, alors qu'une variable aléatoire gaussienne suit un hasard sage.

lumière notamment de l'apparent échec des méthodes quantitatives de gestion de risques financiers lors de la crise financière dont nous avons parlé. Selon Nassim Taleb, il est en effet impossible de quantifier le risque d'événements rares catastrophiques dont nous n'avons jamais connu d'exemple par le passé. Ces « cygnes noirs », tels qu'il les qualifie, invalideraient les approches statistiques de modélisation du risque.

ÉVITER LA CRISE ?

Cependant, la récente crise financière est-elle un cygne noir au même titre que l'accident de Three Mile Island, où, malgré les précautions, l'erreur humaine, non quantifiable, mit en échec le système de sûreté ? Ou ressemble-t-elle plutôt à la tempête qui ravagea les Pays-Bas en 1953, événement rare, mais dont on pouvait mesurer la probabilité ?

Je penche en faveur de la seconde option : les institutions financières sont loin d'avoir mis en pratique des méthodes adéquates pour mesurer les risques de leurs portefeuilles. Plus de quarante ans après les travaux de Mandelbrot, nombre d'établissements bancaires, par facilité, utilisent encore la loi normale dans leurs calculs de risques : difficile alors de blâmer les modèles statistiques… qui ne sont pas utilisés !

Sans aller jusqu'à affirmer, comme le faisait récemment un journaliste du *Nouvel Observateur*, que si la banque d'affaires Lehman Brothers avait calculé ses risques avec les distributions de Lévy, elle aurait évité la faillite, il est certain que l'utilisation de modèles plus réalistes ne pourra qu'améliorer la gestion des risques financiers.

Mais Nicolas Bouleau soulève un autre point, d'ordre sociologique. Les énormes enjeux sociaux de certains événements catastrophiques – météorologiques, géologiques ou économiques – ont pour conséquence d'engendrer une pression importante des décideurs publics sur les « experts » pour, sinon prévoir ces événements rares, au moins quantifier leur risque d'occurrence. La tentation est alors grande de préférer des indicateurs statistiques inadaptés plutôt que de reconnaître que les connaissances scientifiques actuelles n'apportent par de réponse à certaines questions.

C'est sans doute un aspect de la défaillance des systèmes de gestion des risques financiers lors de la crise récente : indépendamment de leur précision, ils ont entretenu l'illusion d'une sécurité aux yeux de leurs utilisateurs, qui ont de ce fait baissé leur garde. Ne rejetons pas les méthodes statistiques de modélisation des risques, utilisons-les avec discernement ! ■

BIBLIOGRAPHIE

P. DEHEUVELS, *La Probabilité, le hasard et la certitude*, « Que Sais-Je ? », Presses Universitaires de France, 2008.

B. MANDELBROT ET R. HUDSON, *Une approche fractale des marchés. Risquer, perdre et gagner*, Odile Jacob, 2005.

F. BARDOU, J.-PH. BOUCHAUD, A. ASPECT ET C. COHEN-TANNOUDJI, *Lévy Statistics and Laser Cooling: How Rare Events Bring Atoms to Rest*, Cambridge University Press, 2002.

N. BOULEAU, Splendeurs et misères des lois de valeurs extrêmes, *Risques*, vol. 3, pp. 85-92, 1991.

L. DE HAAN, Fighting the arch-enemy with mathematics, *Statistica Neerlandica*, vol. 44(2), pp. 45-68, 1990.

L'ESSENTIEL

- Certains résultats fracassants en matière de statistiques sont parfois fondés sur de petits effets difficiles à interpréter.
- La pertinence statistique de tels résultats requiert une estimation précise des incertitudes et une méthode d'analyse adaptée.
- Les analyses fréquentistes et bayésiennes sont parmi les méthodes les plus fiables.
- Autre précaution nécessaire: une estimation des petits effets n'a de sens que si la taille des échantillons est grande.

LES AUTEURS

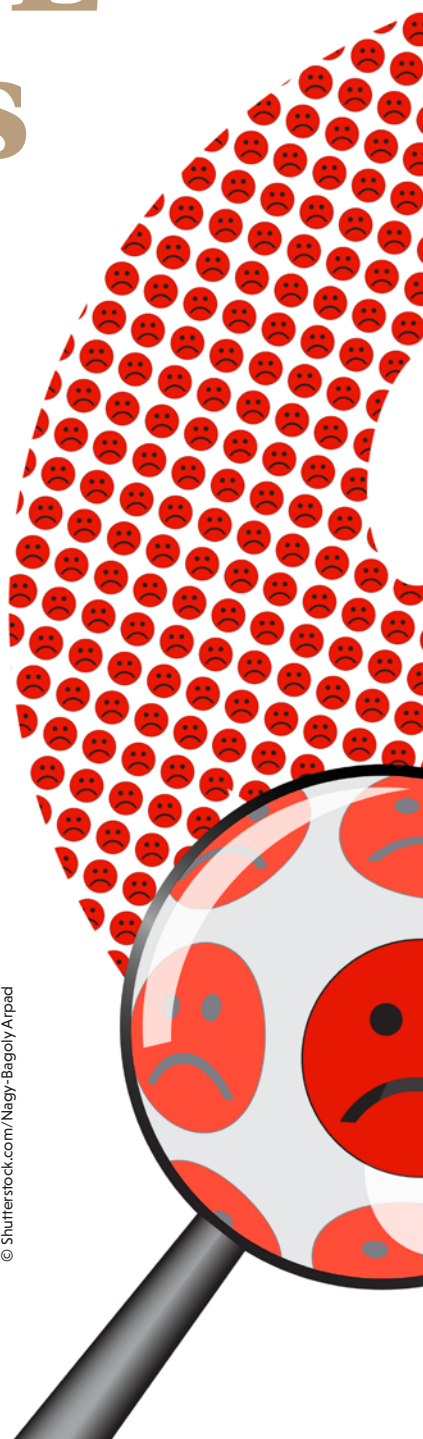
ANDREW GELMAN
professeur de statistiques et de sciences politiques, dirige le Centre de statistiques appliquées de l'université Columbia, à New York.

DAVID WEAKLIEM
est professeur de sociologie à l'université du Connecticut.

Nous remercions la revue *American Scientist* de nous avoir autorisés à publier cet article.

LE CASSE-TÊTE des petits effets

On accorde parfois à de petites différences un sens qu'elles n'ont pas. Distinguer les petits effets statistiquement pertinents de ceux relevant du hasard est un défi !



© Shutterstock.com/Nagy-Bagoly Arpad

« **Les ingénieurs ont plus de garçons, les infirmières plus de filles », «Les hommes violents ont plus de garçons», «Les individus séduisants ont plus de filles»...** On doit ces phrases définitives à Satoshi Kanazawa, psychologue évolutionniste à l'École d'économie de Londres. L'auteur est certes controversé, mais il a néanmoins publié ces «conclusions» pour le moins insolites dans des revues scientifiques sérieuses. L'un de ses derniers articles, parus en 2017, est intitulé «Les individus sont d'autant moins séduisants que leur père était âgé au moment de la conception». Qu'en pensez-vous?

Les spécialistes ont rendu leur avis: les analyses statistiques sur lesquelles ces articles étaient fondés ont été invalidées, par exemple à cause de biais d'échantillonnage. Ainsi, ces «découvertes» ne sont pas significatives et peuvent n'être que le fruit du hasard: elles n'auraient jamais été publiées si leur pertinence statistique avait été correctement évaluée.

Faut-il pour autant rejeter en bloc ces travaux? Non, et d'ailleurs, certaines hypothèses de Satoshi Kanazawa s'appuient sur des théories reconnues par la communauté scientifique. Cependant, le bulldozer de la médiatisation a écrasé toute nuance et précaution pourtant indispensables à l'interprétation des petites différences statistiques relevées par le psychologue. Elles relèvent de ce que l'on nomme les petits effets. Dans quelle mesure ces derniers ont-ils un sens statistique? Comment interpréter des résultats non significatifs (et que l'on cherche malgré tout à interpréter)? Quelles peuvent être les conséquences d'une interprétation erronée de ces petits effets?

Avant de répondre, voyons ce que signifient «statistiquement significatif», et «écart-type», deux notions centrales dans l'étude des petits effets. Nous utiliserons l'exemple de l'étude montrant que «Les individus séduisants ont plus de filles». L'écart-type mesure la dispersion de la variable aléatoire étudiée (la différence des proportions de filles entre deux catégories de parents, les beaux et les autres) autour de sa valeur moyenne. Plus l'écart-type est grand, plus la dispersion des valeurs trouvées est importante. Comment l'estimer? L'écart-type est la racine carrée de la variance, laquelle se calcule à partir de la taille des groupes. Pour une proportion, elle est inversement proportionnelle à la taille de l'échantillon testé. On a donc tout intérêt à choisir de grands échantillons pour réduire l'incertitude sur les mesures. Par exemple, pour un échantillon de 100 couples, l'écart-type de la proportion de filles est égal à 5%. Pour un échantillon de 3000 couples, l'écart-type n'est plus que de 0,9%.

Dans le cas qui nous intéresse, nous cherchons à vérifier l'hypothèse suivante: la proportion de filles est plus élevée dans le groupe de parents séduisants, que dans le groupe de parents jugés moins beaux. Quand peut-on dire que le résultat trouvé est statistiquement significatif? Pour l'affirmer, il faut qu'il soit suffisamment éloigné d'un résultat plausible si les parents jugés beaux ont autant de chances d'avoir une fille que les autres parents (environ une chance sur deux). La différence des proportions entre deux groupes n'est pas statistiquement significative quand elle est probable, du seul fait des fluctuations d'échantillonnage, même s'il n'y a pas de réelle différence entre les deux groupes.

Prenons l'exemple plus simple d'une séquence de 20 lancers de pièce. Imaginons que nous obtenions 8 faces pour 12 piles. La proportion observée de faces serait de 40%, pour un écart-type de 11%. L'estimation obtenue n'étant pas assez éloignée de 50% (moitié faces et moitié piles), le résultat obtenu peut être attribué au hasard et n'est donc pas significatif. Dans les études dont il est question ici, on approximera la loi de la variable aléatoire par une loi gaussienne, et on se contentera de dire, sans détailler le calcul, qu'une différence est significative si elle est supérieure à son écart-type multiplié par 1,96.

FILLE OU GARÇON?

Voyons de plus près l'analyse de Kanazawa. La beauté des sujets fut évaluée sur une échelle de 1 à 5 et le sexe de leurs enfants répertorié. Pour les quelque 3 000 parents étudiés, Kanazawa rapporta une différence des proportions de 8%, paraissant significative: la proportion de filles était égale à 52% pour les parents les plus attirants (notés 1), contre 44% pour la moyenne des quatre autres catégories (de 2 à 5). En fait, comparer la première catégorie aux quatre autres n'est qu'une des nombreuses approches possibles. On aurait pu, par exemple, comparer les deux groupes les plus beaux aux deux groupes les moins beaux: cette fois, la significativité disparaît.

Voilà un bel exemple de résultat sociologique suggestif, mais qui n'est pas statistiquement significatif: il pourrait tout à fait être le fruit du hasard. Pourtant, il semble étayer le modèle proposé. Face à ce type de problème statistique, nous devons tenir compte de l'amplitude des effets attendus. >

BEAUTÉ ET DESCENDANCE

Le magazine *People* publie chaque année la liste des cinquante célébrités mondiales les plus belles. On a répertorié entre 1995 et 2000 le sexe de leurs enfants, ainsi que l'écart-type des données obtenues. Difficile d'en déduire que plus on est beau, plus on a de filles.

ANNÉE DE PUBLICATION	NOMBRE DE FILLES	NOMBRE DE GARÇONS	PROPORTION DE FILLES	ÉCART-TYPE
1995	32	24	57,1 %	6,7 %
1996	45	35	56,2 %	5,6 %
1997	24	35	40,7 %	6,5 %
1998	21	25	45,7 %	7,4 %
1999	23	30	43,4 %	6,9 %
2000	29	25	53,7 %	6,8 %
1995-2000	157	172	47,7 %	2,8 %

> On s'attend à ce que les effets étudiés ici soient faibles, ce qu'attestent les multiples études sur les variations du rapport filles-garçons à la naissance. Ce rapport varie de 1% (la probabilité d'avoir une fille passant par exemple de 48,5 à 49,5%), selon divers facteurs: le groupe ethnique, l'âge des parents, le rang de naissance, le poids de la mère, le statut du couple et la saison de la naissance. Les conditions socioéconomiques, notamment la pauvreté et la sous-alimentation, ont une influence plus marquée, atteignant 3%, car les fœtus mâles sont plus fragiles.

Compte tenu de ces données scientifiques, on s'attendrait à ce que l'effet de la beauté des parents sur le rapport filles-garçons à la naissance soit inférieur à 1%, comparable aux variations observées couramment. Vérifions si c'est bien le cas en nous fondant sur deux approches statistiques: l'analyse dite fréquentiste et l'analyse bayésienne.

L'ANALYSE FRÉQUENTISTE

Dans la première approche, on se fixe des hypothèses, puis on traite statistiquement les données pour savoir si elles sont plus compatibles avec l'une des hypothèses: celle-ci sera alors éventuellement déclarée vraie. Dans la seconde, on tient compte d'une information *a priori*, (ici, les effets de la beauté des parents sur le rapport des sexes à la naissance ne peuvent être grands) sous la forme d'une distribution des valeurs plausibles de l'effet. L'information que l'on tire de l'expérience est la loi conditionnée par l'observation, donc modifiée par elle, qui est appelée loi *a posteriori*: c'est une nouvelle distribution des effets plausibles.

En reprenant l'étude de Kanazawa, nous avons d'abord suivi une méthode d'analyse fréquentiste pour estimer la probabilité de naissance de filles en fonction de la beauté des parents. Nous avons estimé une différence de probabilité de 4,7% entre les deux groupes, pour un écart-type de 4,3%, ce qui reste cohérent avec le résultat de Kanazawa (*voir la figure page 32*). Avec ces valeurs, on peut calculer l'intervalle de confiance qui contient la vraie valeur avec une probabilité de 95%, et qui est, en pourcentages: [-3,9; 13,3]. Comment interpréter statistiquement l'intervalle de confiance?

Comment reconnaître les données fiables? En collectant toujours plus de données!

Il contient la valeur zéro qui correspond à l'absence d'effet de la beauté des parents sur le rapport des sexes à la naissance. Notre estimation à 4,7% n'est donc pas significative et il faut poursuivre les analyses statistiques avant de conclure (si c'est possible!). Et en dehors des bornes de l'intervalle de confiance à 95%, que se passe-t-il? Que représentent les 5% de probabilité restants?

Pour que l'effet trouvé soit statistiquement significatif, il faudrait qu'il soit plus grand que 1,96 fois l'écart-type, c'est-à-dire au-dessus de 8,4% (l'intervalle de confiance ne contiendrait alors pas 0). Exprimé autrement: la probabilité de rejeter à tort la valeur nulle est de 5%.

Mais est-ce bien raisonnable d'envisager des effets significatifs aussi grands que 8,4%? Non, bien sûr. C'est ce que l'on nomme une erreur de magnitude. L'étude est construite de telle façon que tout résultat statistiquement significatif surestime le véritable effet (qui ne peut dépasser 1%). S'y ajoutent des erreurs de signe quand l'estimation trouvée est de signe opposé au véritable effet (ici: «les parents beaux ont plus de garçons que de filles»). Ainsi, on distingue deux types d'effet: positif, si les parents «beaux» ont plus de filles que les autres, et négatif, s'ils en ont moins.

Pour illustrer les probabilités associées à ces erreurs, envisageons quatre scénarios fondés sur des écarts-types égaux à 4,3% (*voir l'encadré ci-contre*). Ils montrent qu'une étude à partir de cette taille d'échantillon (3000 couples environ) n'est pas pertinente pour estimer des petits effets de l'ordre du 1%. Cela est dû notamment à la valeur de l'écart-type qui est particulièrement élevée. C'est pourquoi les études de la répartition des sexes des nouveau-nés utilisent des échantillons beaucoup plus grands, fondés sur de vastes bases de données démographiques, comptant plus d'un million d'individus.

L'ANALYSE BAYÉSIENNE

On résume tout ce que l'on sait sur l'effet à analyser, en utilisant des sources extérieures déjà connues, par une distribution *a priori*. Cette distribution sera modifiée (conditionnée) par le résultat de l'expérience pour donner la distribution *a posteriori*. Si l'on ne savait rien d'avance, on pourrait prendre une distribution *a priori* non informative et le résultat correspondrait à celui de l'approche fréquentiste. Ici, la distribution obtenue *a posteriori* serait alors approximativement une gaussienne, de moyenne 4,7% et d'écart-type 4,3%, ce qui correspondrait à une probabilité de 86% environ que l'effet réel soit positif. En général, plus la distribution *a priori* est concentrée autour de zéro (hypothèse selon laquelle l'effet réel sur la différence des sexes est faible), plus la probabilité *a posteriori* est proche de 50%.

Choisissons, par exemple, une distribution *a priori* gaussienne (en forme de cloche), centrée sur zéro avec une forme telle que la différence réelle de probabilité d'avoir une fille (selon que les parents sont beaux ou non) soit proche de zéro, avec des probabilités 50%, 90% et 94% d'être respectivement dans les intervalles: $[-0{,}3; 0{,}3]$, $[-1; 1]$, et $[-3; 3]$. Pourquoi centrer la distribution sur zéro? Parce que nous n'avons pas d'*a priori* sur le signe de la différence réelle de probabilité de naissance de filles en fonction de la beauté des parents.

À l'étape suivante, nous calculons, à partir de cette distribution *a priori* et des données, la distribution *a posteriori* de l'effet. Pour résumer, la distribution *a posteriori* donne une probabilité que l'effet soit positif (les parents beaux ont plus de filles) de seulement 58%, dont 45% que cette différence positive soit inférieure à 1%. Cette analyse dépend – mais peu – de la distribution *a priori*; par exemple, si l'on décide d'élargir la courbe de distribution autour de zéro, la probabilité que l'effet réel soit positif n'augmente que de 7% (pour atteindre 65%). Le fait de changer de famille de courbes de distribution a peu d'effets sur les résultats: les effets réels restent faibles, ce que confirment les données.

L'idéal scientifique dans l'inférence sur une quantité (ici, le lien entre la beauté des parents et le rapport des sexes à la naissance) est la >

QUATRE SCÉNARIOS POUR COMPRENDRE LES ERREURS

On peut étudier les erreurs liées aux résultats de Kanazawa en déterminant les probabilités associées *via* quatre scénarios. Dans le premier, que nous nommerons «Différence exactement nulle», supposons qu'il n'y a aucune corrélation entre la beauté des parents et le sexe de leur enfant (il n'y a pas de différence entre les deux groupes de parents). Il existe toujours une petite probabilité – 5% – que le résultat trouvé soit statistiquement significatif. Mais ce résultat sera quand même faux, on a cinq% de chances de se tromper sur la valeur et ce quel que soit le signe de l'effet, positif ou négatif.

Dans le scénario dit de 0,3% exactement, les parents beaux ont 0,3 % de probabilité supplémentaire d'avoir une fille – valeur plausible. On a toujours une probabilité de 5% de trouver un résultat statistiquement significatif à l'issue de l'étude. Comment se répartissent ces 5% de part et d'autre de la valeur réelle? Dans ces conditions, on a une probabilité de 3% d'observer un effet statistiquement significatif positif et 2% d'observer un effet statistiquement significatif négatif. Mais dans chaque cas, l'effet estimé – au moins 8,4% pour qu'il soit significatif – sera beaucoup trop éloigné de la valeur réelle de 0,3%. Ainsi, l'erreur de magnitude sera importante. Par ailleurs, la probabilité d'aller dans le mauvais sens – erreur de signe – sera égale à 2/5. Imaginons que l'on se retrouve entre les bornes de l'intervalle de confiance à 95%, donc proches de la valeur réelle. Dans ce cas, l'erreur de magnitude serait faible, mais la probabilité de faire une erreur de signe n'est pas négligeable: 47,5 %, proche des 50% observés en l'absence d'effet. Ainsi, le signe observé n'apporte pratiquement aucune information sur le sens d'une éventuelle différence.

Troisième scénario: si les parents séduisants ont une probabilité d'avoir une fille supérieure de 1% (ce qui semble être l'effet maximal possible), alors on a réciproquement des probabilités de 4 et 1% d'avoir des effets positif et négatif statistiquement significatifs. Globalement, il y a une probabilité de 40% de faire une erreur sur le signe de l'estimation; là encore, l'estimation donne peu d'informations sur le signe ou l'ordre de grandeur de l'effet réel.

Enfin, envisageons le quatrième scénario: même si la différence réelle était de 3% – valeur improbable –, il n'y aurait encore que 10% de probabilité d'obtenir un résultat statistiquement significatif. Dans ce cas, nous aurions une probabilité de 24% de faire une erreur de signe. Ce type d'étude est donc peu informatif.

> description de l'incertitude qui est résumée ici par une distribution de probabilités. Des chercheurs peuvent collecter des données ou analyser des données qui ont déjà été publiées de façon créative (comme l'a fait Kanazawa) et publier leurs résultats. Des métaanalyses peuvent être faites pour revoir tous ces résultats conjointement; elles lisseront les variations qui sont inhérentes à ces études sur de petits échantillons dans lesquelles la probabilité d'un effet positif peut passer de 50% à 58%, puis peut-être redescendre à 38% et ainsi de suite.

Comment reconnaître les données fiables? En collectant toujours plus de données. Chaque année, le magazine américain *People* publie une liste des 50 plus belles célébrités mondiales. Nous avons répertorié le sexe de leurs enfants pour les numéros parus entre 1995 et 2000 (*voir l'encadré page 30*). En 1995, par exemple, on a dénombré 32 naissances de filles pour 24 garçons, soit 57,1% de filles, ce qui correspond à 8,6% de plus que dans la population générale (48,5%). Un résultat en accord avec l'hypothèse de Kanazawa. Mais l'écart-type étant de 6,7%, l'estimation de 8,6% n'est pas statistiquement significative. Pour le confirmer, nous avons comparé ce résultat avec ceux des années suivantes. Entre 1995 et 2000, les plus belles personnes selon People ont eu 157 filles sur un total de 329 enfants, soit 47,7% de filles (pour un écart-type de 2,8%), ce qui est seulement 0,8% inférieur au chiffre obtenu pour la population générale. On ne peut rien en conclure...

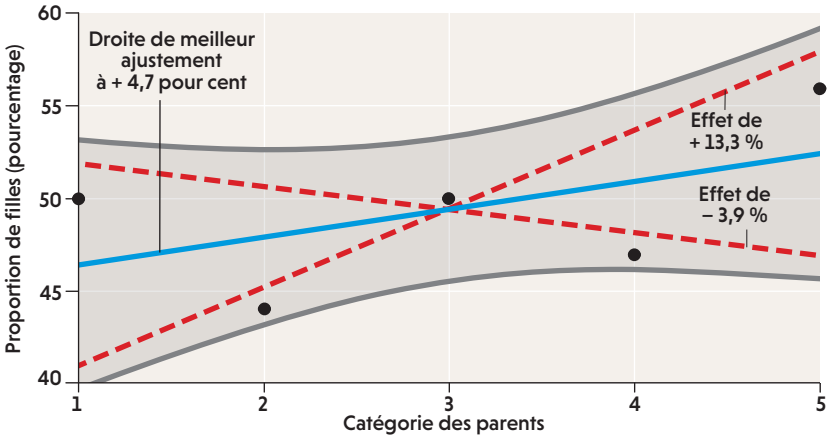
Pourquoi perdre notre temps à étudier des erreurs statistiques que personne n'a repérées? Pour deux raisons. D'abord, les résultats qui semblent avoir un sens sans être statistiquement significatifs sont les plus problématiques. Ensuite, certains médias et diverses publications scientifiques, par leur intérêt pour certains sujets de sociologie et leur sélection des résultats, biaisent la recherche en sciences sociales.

De fait, les résultats de Kanazawa ont tout de suite suscité l'intérêt des médias, jusqu'à des blogs du *New York Times*. Cette publication dans une revue à comité de lecture semblait être une caution suffisante balayant les doutes éventuels.

UN BRUIT ASSOURDISSANT

Qui plus est, l'effet n'a cessé d'augmenter. Ainsi, l'estimation – statistiquement non significative – de 4,7% que nous avons faite est passée à 8% dans l'analyse de Kanazawa (comparaison du groupe le plus beau à la moyenne des quatre groupes les moins séduisants), pour atteindre la valeur de 26% après une étude complémentaire introduisant d'autres corrections, avant de grimper à 36% pour des raisons encore floues!

Cette inflation nous surprit, ce chiffre étant 10 à 100 fois supérieur à tous les



Une analyse fréquentiste des données détermine la courbe de meilleur ajustement *(la droite bleue)* aux données de S. Kanazawa *(points en noir)*, pour un écart-type de 4,3%. Un intervalle de confiance à 95% montre que des effets aussi faibles que – 3,9% et aussi grands que 13,3% sont compatibles avec les données, car ils encadrent l'estimation à 4,7%.

rapports filles-garçons publiés dans la littérature. Nous en avons conclu que, dans cette étude, le bruit (les données parasites) était supérieur au signal pertinent. La puissance statistique désigne la capacité de détecter une différence lorsqu'elle existe. Les études avec des échantillons plus importants ont toujours plus de puissance. Ainsi, si l'on veut affirmer quelque chose à propos d'effets de l'ordre de 1%, on a tout intérêt à partir de données pertinentes et à réaliser des tests qui exploitent bien les données. Cet exemple illustre bien le fait que les études qui manquent de puissance statistique ont peu de chances de parvenir à une pertinence statistique et, plus important encore, elles surestiment la taille des effets. Autrement dit, avec ces études, le bruit devient plus fort que le signal, c'est-à-dire l'effet observé.

Comment échapper à ce type de problèmes en sociologie? Aujourd'hui, la plupart des sujets de sociologie ont été passés au crible, et les chercheurs en sont donc réduits à étudier les petits effets. L'étude du rapport des sexes à la naissance est un sujet de société proche de nos préoccupations. Présenté sous forme d'une «vérité politiquement incorrecte», le résultat de Kanazawa, parce qu'il concerne les naissances, touche à des questions sensibles telles que l'avortement, le congé parental, le rôle de l'homme et de la femme dans la société.

On a vu que les études dont la pertinence statistique est insuffisante produisent des résultats aléatoires, parfois statistiquement significatifs, mais le plus souvent intuitifs. C'est un des points faibles de la psychologie évolutionniste: elle interprète des résultats aléatoires sans reconnaître la fragilité des explications qu'elle donne. Par exemple: les personnes jugées séduisantes auraient plus de chances d'être en bonne santé, riches et issues de groupes ethniques dominants, et plus

DES PETITS EFFETS PASSÉS AU CRIBLE

Il est difficile de mettre en évidence un petit effet, car les données n'apportent toujours qu'une information limitée. Plus l'effet que l'on cherche à estimer est petit, plus la quantité de données nécessaire pour le mettre en évidence est grande pour qu'il ressorte des fluctuations dues au hasard. N'y a-t-il pas d'autre issue qu'une simple augmentation de la quantité de données, coûteuse et parfois impossible ? Autrement dit : « Comment exploiter au mieux un ensemble de données ? » Donnons quelques exemples. Dans un essai thérapeutique concernant une nouvelle préparation que l'on souhaite comparer à une ancienne, les patients sont souvent très hétérogènes. On constitue des paires de patients aussi proches que possible pour toutes les caractéristiques susceptibles d'influer sur le résultat (sexe, âge, catégorie socioprofessionnelle, origine ethnique...). Pour éviter tout biais, on choisit au hasard dans chaque paire le patient qui reçoit le traitement à tester et celui qui reçoit l'ancien. La comparaison des deux traitements se fonde ainsi sur un ensemble de comparaisons élémentaires beaucoup plus efficaces que si l'on avait choisi les patients sans tenir compte de leur statut. Dans les sondages d'opinion, on sait que l'âge, la catégorie socioprofessionnelle influencent le résultat. On répartit la population à sonder en différents groupes plus homogènes que l'on échantillonnera séparément (échantillonnage stratifié). Si l'effectif des groupes est déjà connu, on montre que l'on peut ainsi améliorer la précision par rapport à un échantillonnage aléatoire simple de même taille. On dispose en plus d'une information beaucoup plus riche sur la répartition des opinions. En outre, les statisticiens ont développé la théorie des plans d'expérience pour améliorer la collecte des données. Il s'agit, dans les exemples cités, de concevoir, par exemple, l'allocation des mesures ou le choix des unités expérimentales, de façon à ce que pour un coût donné la précision de l'analyse sur les quantités d'intérêt soit la meilleure possible. Cette optimisation repose sur le modèle d'analyse et en exploite les propriétés. Une analyse statistique n'est jamais menée sans une connaissance minimale du problème, et la prise en compte, dans le modèle statistique, de toute information pertinente déjà disponible, permet aussi un gain de précision. Les statisticiens ont développé beaucoup d'outils pour permettre une inférence efficace, mais sans jamais lever le caractère intrinsèquement probabiliste de la démarche. Il reste toujours une incertitude : statistique n'est pas divination. Un exemple historique est celui des élections présidentielles américaines de 1948. Les sondeurs d'opinion, rendus trop confiants par leur précédent succès, annoncent la victoire de Dewey, mais Truman l'emporte et le jour de son investiture, les sénateurs – déçus – de l'Indiana observeront une minute de silence « à la mémoire du Dr Gallup », le fondateur de l'institut de sondage qui porte son nom !

François Rodolphe
Laboratoire de mathématique, informatique et génome (INRA), *Jouy-en-Josas*

généralement de présenter des caractéristiques valorisées par la société. Elles auraient du pouvoir, ce qui d'après certaines théories sociologiques, serait plus bénéfique pour les hommes que pour les femmes. Il serait donc « naturel » que des parents attrayants aient plus de garçons. Nous ne prétendons pas que c'est vrai ; nous disons simplement que cela pourrait l'être, mais qu'on pourrait tout aussi bien imaginer une argumentation aboutissant au fait qu'ils ont plus de filles... Ce qui n'est pas sans poser quelques difficultés !

TROUVER DES DIFFÉRENCES LÀ OÙ IL N'Y EN A PAS

Comparez ces deux affirmations : « Les parents beaux ont plus de filles » et « Il n'est pas prouvé que des parents beaux aient plus ou moins de filles ». Nul doute que la première, sensationnelle, ferait davantage les gros titres ! Les éditeurs des revues sérieuses où l'affirmation de Kanazawa a été publiée ont-ils eux-mêmes été influencés ? Sans doute, et à cela deux raisons possibles. D'une part, les erreurs statistiques sont parfois difficilement détectables, même par des spécialistes. D'autre part, la signification statistique n'est pas directement liée à la taille des échantillons quand les effets testés sont petits. Avec un échantillon suffisamment grand, on peut presque toujours trouver de petits effets statistiquement significatifs. Mais quand les effets étudiés sont infimes, les études faites sur des cohortes trop petites conduisent à des interprétations abusives.

L'étude du rapport des sexes à la naissance n'est pas neuve. Par exemple, dans son ouvrage *Probabilité, statistiques et vérité*, publié en 1957, Richard von Mises étudia ce rapport pour les naissances de 1907 et 1908 à Vienne, et trouva moins de variations qu'on en attendrait du simple hasard. Il l'attribua à des répartitions des sexes différentes selon les groupes ethniques. Pourtant, l'incertitude obtenue sur les mesures n'était ni plus ni moins que celle d'un hasard pur. Que faire face à cette volonté de trouver des différences là où il n'y en a pas ? Pour éviter ces biais, il faut montrer que les résultats observés représentent des effets réels indépendants de la sélection des échantillons, et trouver un argument biologique confirmant que des effets de l'ordre de 1 % sont importants.

Lorsque nous devons estimer des petits effets, les statisticiens doivent garder un regard critique sur les estimations obtenues. Mais les méthodes d'analyses ne sont pas exemptes de failles méthodologiques : les calculs fréquentistes ne tiennent pas compte des tailles des effets ; les analyses bayésiennes ne sont pas souvent meilleures en utilisant surtout des distributions a *priori* gaussiennes, et ignorent souvent les problèmes de puissance statistique. D'où l'importance d'estimer correctement les incertitudes, notamment en calculant les statistiques sur le signe des effets et sur leur amplitude. Nul doute que l'échange de méthodes et d'idées ouvrira la voie à une meilleure estimation des petits effets. ■

BIBLIOGRAPHIE

A. GELMAN *ET AL.*
Letter to the editor regarding some papers of Dr. Satoshi Kanazawa,
Journal of Theroretical Biology, vol. 245, pp. 597-599, 2007.

S. KANAZAWA
Beautiful parents have more daughters : a further implication of the generalized Trivers-Willard hypothesis,
Journal of Theoretical Biology, vol. 244, pp. 133-140, 2007.

R. VON MISES
Probability, Statistics and Truth, *Dover*, 2006.

L'ESSENTIEL

- La valeur-*p* désigne la probabilité qu'un résultat statistique ne soit pas le fait du hasard.
- Plébiscitée par les chercheurs, elle souffre néanmoins de nombreux travers. Ainsi, elle ne dit rien de la taille de l'effet que l'on étudie.
- Elle se laisse aussi facilement manipuler pour asseoir les résultats recherchés : c'est le «*p*-hacking».
- La recherche gagnerait en qualité et en reproductibilité si l'on tenait compte de la probabilité que l'effet étudié existe réellement.

L'AUTEUR

REGINA NUZZO
est journaliste indépendante et professeure de statistiques à l'université Gallaudet de Washington.

La malédiction de la VALEUR-P

Depuis des décennies, on mesure la qualité des résultats statistiques grâce à la valeur-*p*. Mais cet étalon-or n'est pas aussi fiable qu'on veut bien le croire. Il est temps de le remettre en question.

Un mauvais tireur incapable d'atteindre une cible *(à gauche)* peut néanmoins se déclarer doué en peignant la cible à l'endroit où se trouve par hasard la majorité des impacts de tir *(à droite)*. Cette métaphore illustre le «*p*-hacking», un comportement qui entache de nombreuses publications scientifiques.

En court instant, Matt Motyl s'est tenu dans l'antichambre de la gloire scientifique. En 2010, il avait découvert lors d'une expérience que les extrémistes voyaient le monde en noir et blanc – littéralement. Ses résultats étaient «clairs comme de l'eau de roche», se souvient le doctorant en psychologie de l'université de Virginie, à Charlottesville. Ses travaux sur 2 000 participants avaient montré que parmi eux, les extrémistes de gauche ou de droite avaient plus de mal à différencier des nuances de gris que les personnes politiquement plus modérées. L'étudiant était confiant et croyait en la solidité de ses résultats. La publication dans une revue renommée semblait à portée de main. Mais rien ne s'est passé comme prévu.

De fait, la robustesse des résultats statistiques avait été calculée par l'indice rituel, l'arme de tous les statisticiens, l'incontournable valeur-*p* (*voir l'encadré ci-contre*). Ici, il était de 0,01, ce qui indique un résultat «très significatif». Tout allait pour le mieux. Hélas, par précaution, Motyl et son encadrant, Brian Nosek, répétèrent l'expérience. Avec de nouvelles données, la valeur-*p* bondit à 0,59, bien au-delà du seuil de 0,05 sous lequel un résultat est considéré comme significatif. Ainsi disparut l'effet… et le rêve de gloire de Motyl. Le test de la valeur-*p* condamnait l'étude aux oubliettes. Et si nous avions affaire à une erreur judiciaire ? Et si la valeur-*p* n'était pas aussi fiable qu'on veut bien le penser ?

LA VALEUR-*P*, CE MOUSTIQUE

De fait, l'étude sur les extrémistes ne souffrait pas de problème de collecte de données ou d'erreur de calcul. La faute revenait bien plus à la nature trompeuse de la valeur-*p* elle-même. Cette dernière n'est en effet pas aussi fiable et objective que le pensent la plupart des scientifiques. Stephen Ziliak, économiste de l'université Roosevelt de Chicago et critique régulier de la façon dont les statistiques sont utilisées, va encore plus loin. Selon lui, «les valeurs-*p* ne font pas leur travail, car elles ne le peuvent pas».

C'est préoccupant, particulièrement au regard des inquiétudes sur la reproductibilité des résultats qui traversent la communauté scientifique, et que le cas de Motyl illustre. John Ioannidis, épidémiologiste à l'université Stanford, avait affirmé en 2005 que la majorité des résultats publiés étaient faux et avait avancé des explications à ce phénomène. Depuis, la reproduction d'études a échoué dans de nombreux cas célèbres, ce qui a contraint les scientifiques à reconsidérer leurs méthodes. En d'autres termes, la façon dont les résultats sont évalués est-elle fiable ? Existe-t-il des alternatives, c'est-à-dire des méthodes avec lesquelles toutes les fausses alertes sont identifiées, sans occulter un vrai résultat ?

La critique de la valeur-*p* n'a rien de nouveau. Depuis sa formalisation par le mathématicien britannique Karl Pearson au début du XX^{e} siècle, on l'a dénigrée en la comparant à un «moustique» (ennuyant et impossible de s'en débarrasser), aux «habits neufs de l'empereur» (il y a un problème, mais personne n'en parle)… Charles Lambdin, d'Intel Corporation, a même proposé de rebaptiser la méthode «Statistical Hypothesis Inference Testing», probablement pour l'acronyme que forment les initiales…

La valeur 0,05 devint le seuil séparant le significatif de ce qui ne l'est pas. Mais ce n'était pas son rôle

Ironie de l'histoire, lorsque Ronald Fisher (1890-1962), l'un des pères de la statistique moderne, introduisit la valeur-*p* dans les années 1920, il n'avait nullement en tête un test qui déterminerait tout de manière définitive. Il y voyait plus un moyen de juger si un résultat était significatif, ce terme étant à prendre dans un sens non scientifique : le résultat signifie quelque chose, suffisamment pour que l'on y regarde de plus près. Son idée était de mener une expérience et de vérifier ensuite si les données sont cohérentes avec ce que le hasard seul aurait produit.

Pour cela, un chercheur formule d'abord une hypothèse, dite nulle, qui exprime l'inverse

VALEUR-P

La valeur-*p* indique dans quelle mesure il est probable que le résultat présenté dans une étude soit vrai et ne résulte pas du hasard. Ainsi, une valeur-*p* inférieure à 0,05 signifie que nous aurions raison 95 fois sur 100 de croire qu'un effet observé n'est pas une coïncidence.

de ce qu'il veut prouver, par exemple que les individus qui ont reçu un médicament ne sont pas en meilleure santé que ceux ayant reçu un placebo. Le chercheur suppose ensuite que l'hypothèse nulle est vraie et calcule sous cette condition préalable la probabilité d'obtenir des résultats au moins aussi marqués que ceux qu'il a réellement observés. Cette probabilité est la valeur-*p*. Plus elle est petite, d'après Fisher, plus grande est la probabilité que l'hypothèse nulle soit fausse, et donc qu'elle doive être rejetée, ce que le chercheur souhaitait au début, car l'effet étudié est ainsi confirmé.

On peut calculer la valeur-*p* avec autant de chiffres après la virgule que l'on souhaite, mais cette précision apparente est trompeuse, car les hypothèses étudiées sont loin d'être aussi fines. Pour Fisher, la valeur-*p* ne représentait qu'un maillon dans une chaîne qui relie les observations et les connaissances de base pour parvenir à une conclusion scientifique.

Ces précautions ont été balayées par un mouvement dont le but était de rendre la prise de décision fondée sur les preuves aussi rigoureuse et objective que possible. Les pionniers de ce revirement étaient, à la fin des années 1920, le mathématicien polonais Jerzy Neyman et le statisticien britannique Egon Pearson, les rivaux les plus acharnés de Fisher. Leur système inclut des concepts comme « faux positif », « faux négatif » (*voir les Repères, page 6*) et « puissance » d'un test statistique, que l'on trouve aujourd'hui dans tous les cours de statistiques pour débutants. Neyman et Pearson mirent cependant volontairement de côté la valeur-*p*.

Pendant que les représentants des deux camps s'écharpaient, d'autres chercheurs perdirent patience et rédigèrent eux-mêmes des manuels de statistiques. Malheureusement, beaucoup de ces auteurs n'étaient pas suffisamment versés en la matière pour apprécier les subtilités philosophiques des deux visions. Ils intégrèrent donc la valeur-*p* de Fisher, facile à calculer, dans le système régulé, rigoureux et sécurisant de Neyman et Pearson. Depuis, du haut de son piédestal, la valeur-*p* trône toujours. La valeur de 0,05 devint le seuil séparant le significatif de ce qui ne l'est pas. Mais ce n'était pas son rôle.

GUEULE DE BOIS ET TUMEUR

En conséquence, il règne une grande confusion aujourd'hui sur ce que la valeur-*p* exprime réellement. L'étude de Motyl en est une bonne illustration. La plupart des scientifiques interpréteraient sa première valeur-*p* de 0,01 comme une probabilité de fausse alerte de 1 %. C'est pourtant faux. La valeur-*p* ne livre qu'un résumé sommaire des données en considérant comme vraie une hypothèse nulle spécifique. >

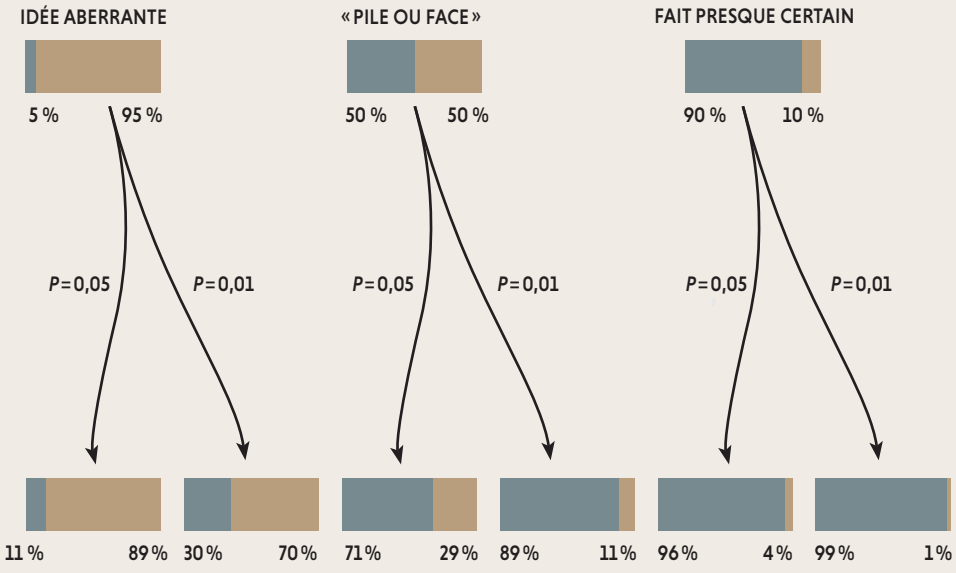
PLAUSIBLE OU NON ?

La valeur-*p* mesure la probabilité qu'un résultat observé soit dû au hasard. Mais elle néglige la question essentielle : quelle est la probabilité que l'hypothèse testée soit correcte ? La réponse dépend de la robustesse du résultat observé et, plus encore, de la plausibilité de l'hypothèse testée.

Avant l'expérience, la plausibilité de l'hypothèse testée (la probabilité qu'elle soit vraie) peut être estimée à partir d'études antérieures, de mécanismes supposés… Ici, on envisage trois cas.

Une valeur-*p* égale à 0,05 traduit conventionnellement un résultat « statistiquement significatif ». À 0,01, c'est « très significatif ».

Après l'expérience, une faible valeur-*p* peut rendre l'hypothèse plus plausible, mais pas nécessairement de façon… significative.



> Une information supplémentaire décisive manque : la probabilité que l'effet en question existe réellement (*voir l'encadré page précédente*). En faire abstraction est semblable à se réveiller le matin avec une migraine et en incriminer une tumeur rare au cerveau ; c'est possible, mais assez peu probable, car bien plus de preuves sont nécessaires pour exclure des explications plus courantes. Moins l'hypothèse est plausible et plus un résultat excitant, une fausse alerte en fait, surgira fréquemment, et cela complètement indépendamment de la valeur-*p*.

Pour le praticien, une telle affirmation est difficile à appréhender. Comment doit-il évaluer la probabilité que l'effet existe ? S'agit-il de la fréquence relative avec laquelle il est constaté au sein d'un ensemble d'études (on parle d'interprétation fréquentielle) ? Ou bien un tel nombre traduit-il les connaissances partielles du chercheur, qui seraient à améliorer par l'expérience seulement (interprétation bayésienne) ? Quoi qu'il en soit, avant une expérience, un chercheur estime grossièrement la probabilité que l'hypothèse étudiée soit vraie.

En établissant cette probabilité dès le départ, et en calculant avec des hypothèses additionnelles judicieuses comment les valeurs-*p* se comportent dans ces conditions, les résultats sont alors moins spectaculaires. Avec une probabilité préalable de 50 % attribuée à l'hypothèse de Motyl, la valeur-*p* de 0,01 signifie alors : dans un cas sur neuf, son expérience lui fait miroiter l'illusion qu'il existe un effet qui n'existe absolument pas.

Par ailleurs, la probabilité que ses collègues puissent reproduire son expérience n'est pas de 99 %, comme on pourrait le penser, mais se situe plutôt autour de 73 %, voire de seulement 50 %, si l'on souhaite à nouveau une valeur-*p* de 0,01. En d'autres termes, le fait que sa seconde expérience soit non concluante est à peu près aussi surprenant que s'il avait perdu au jeu de pile ou face.

L'EFFET DE LA TAILLE DE L'EFFET

De nombreux critiques estiment aussi que la valeur-*p* nuit au raisonnement, en particulier parce qu'elle détourne l'attention de la taille réelle de l'effet. Prenons un exemple. En 2013, une étude de John Cacioppo, de l'université de Chicago, portant sur plus de 19 000 participants montra que les mariages nés d'une rencontre en ligne étaient plus robustes ($p < 0{,}002$) que les autres. En outre, les individus encore mariés avaient tendance à être plus satisfaits de leur mariage que ceux qui s'étaient rencontrés hors ligne ($p < 0{,}001$). Ces valeurs impressionnent, mais l'effet mesuré était infime : une rencontre sur Internet faisait baisser le taux de séparation de 7,67 à 5,96 % et augmentait la satisfaction du couple de 5,48 à 5,64 sur une échelle de 7.

Une autre erreur, sans doute la plus grave, se cache derrière ce que le psychologue Uri Simonsohn, de l'université de Pennsylvanie, nomme le «*p*-hacking». Ce biais consiste à manipuler les données jusqu'à l'obtention du résultat souhaité (*voir la figure page 35*), même sans mauvaises intentions. Ainsi, un changement minime de méthode lors de l'analyse des données peut faire monter le taux de faux positifs d'une étude à 60 %.

Il est difficile d'évaluer à quel point ce problème est répandu, mais selon Simonsohn, le *p*-hacking se multiplierait, parce qu'il serait courant de rechercher de très faibles effets dans des données «bruitées». En analysant des études en psychologie, il a découvert que les valeurs-*p* publiées étaient nombreuses à se situer aux environs de 0,05. Est-ce étonnant si dans les faits les chercheurs partent à la pêche aux valeurs-*p* significatives jusqu'à ce que l'une d'elles, qui les intéresserait, tombe dans leur filet ?

Avec toutes ces critiques, les choses ont-elles changé ? Peu. John Campbell, aujourd'hui chercheur en psychologie à l'université du Minnesota à Minneapolis, le déplorait déjà en 1982, lorsqu'il était éditeur du *Journal of Applied Psychology* : «Il est pratiquement impossible d'arracher les auteurs à leurs valeurs-*p*. Et plus il y a de zéros derrière la virgule, plus ils s'y accrochent.»

RECETTE POUR LA VALEUR-P

D'abord, on identifie les résultats attendus d'une expérience, par exemple à partir d'études précédentes. Imaginons qu'en France, les voitures rouges soient deux fois plus souvent verbalisées que les bleues. Vous souhaitez vérifier que votre région reflète bien l'échelon national. Votre échantillon consiste en 150 amendes : le résultat attendu est donc 100 amendes pour les rouges et 50 pour les bleues. Les observations sont de respectivement 90 et 60. Les différences sont-elles le fruit du hasard ? Le calcul de la valeur-*p* s'impose.

On détermine d'abord le degré de liberté qui rend compte de la variabilité dans la recherche. Il correspond à *n*-1, *n* étant le nombre de variables (ici, 2, rouge et bleu). Le degré de liberté est donc 1.

On calcule ensuite le χ^2 (prononcez *Khi-2*) qui mesure la différence entre la théorie et les observations :

$\chi^2 = \Sigma((o-e)^2/e)$,

o correspondant aux données observées et *e* à celles attendues ou théoriques. On obtient ici :

$\chi^2 = ((90-100)^2/100) + ((60-50)^2/50)$

$\chi^2 = ((-10)^2/100) + (10)^2/50)$

$\chi^2 = (100/100) + (100/50) =$

$\chi^2 = 1 + 2 = 3$

Enfin, dans des registres statistiques de référence (accessibles sur Internet) reliant le degré de liberté et le χ^2, on trouve la valeur-*p* associée. Elle est ici comprise entre 0,05 et 0,1. Supérieure à 0,05, elle ne permet pas de conclure. On sait juste que les résultats observés ont entre 5 et 10 % de chances d'être dus au hasard. Ce n'est pas significatif.

© D'après WikiHow

Chaque tentative de réforme devra combattre des habitudes fermement établies : la façon dont les statistiques sont enseignées dans les universités, celle dont les résultats des études sont exploités et interprétés et comment ils sont ensuite relayés dans les revues spécialisées. Mais au moins, de nombreux scientifiques ont admis l'existence d'un problème. Grâce à des chercheurs comme John Ioannidis, les préoccupations des statisticiens ne sont plus vues uniquement comme de la pure théorie.

Il est pratiquement impossible d'arracher les auteurs à leur sacro-sainte valeur-*p*

Pour améliorer la situation, les statisticiens ont quelques outils à leur disposition. Par exemple, les chercheurs pourraient toujours publier également les tailles d'effet et les intervalles de confiance (*voir les Repères, page 6*) obtenus. Ces valeurs expriment ce que ne peut exprimer la valeur-*p* seule : la portée et l'importance relative de l'effet.

Beaucoup plaident aussi pour remplacer la valeur-*p* par des méthodes fondées sur la vision de Thomas Bayes, mathématicien britannique du XVIII^e^ siècle. Cette approche décrit une façon de penser les probabilités comme la plausibilité d'un résultat, plutôt que la fréquence potentielle de ce résultat. On introduit certes par là une subjectivité dans les statistiques, que les pionniers du début du XX^e^ siècle voulaient éviter à tout prix. Pourtant, la statistique bayésienne simplifie l'intégration des connaissances contextuelles sur le monde et permet de calculer comment les probabilités sont modifiées par des preuves nouvellement acquises.

D'autres soutiennent une approche plutôt pluraliste : les scientifiques devraient utiliser plusieurs méthodes sur un même ensemble de données. Lorsque les résultats différeront, les chercheurs devront être plus créatifs pour découvrir à quoi cela est dû. La compréhension de la réalité sous-jacente y gagnerait.

Aux yeux de Simonsohn, la meilleure protection pour un chercheur est de tout montrer. Les auteurs devraient garantir leur travail « sans *p*-hacking » en explicitant le choix de la taille de leurs échantillons, les données éventuellement mises de côté ainsi que les manipulations mises en œuvre. Aujourd'hui, aucune de ces informations n'est disponible ni vérifiable.

DES ÉTUDES EN DEUX ACTES

À l'instar de la mention « Les auteurs n'ont pas d'intérêts financiers dépendants du contenu de cet article » encore courante aujourd'hui, ces informations feraient la différence entre faute scientifique involontaire ou délibérée. Lorsqu'une telle déclaration sera monnaie courante, le *p*-hacking sera éliminé. Ou au moins son absence sera-t-elle remarquée par le lecteur et lui permettra un jugement plus éclairé.

L'analyse en deux étapes ou « *preregistered replication* » (réplication préenregistrée) est une idée qui va dans ce sens, et elle gagne du terrain. Dans cette approche, les études exploratoires et confirmatoires sont abordées différemment et clairement identifiées. Au lieu, par exemple, de conduire quatre petites expériences et de réunir les résultats dans un article, les chercheurs balaieraient d'abord le domaine pour récolter des observations intéressantes avec deux petites études exploratoires, sans trop se préoccuper des fausses alertes. C'est seulement après, sur la base de ces données, qu'ils concevraient une étude qui confirmera, peut-être, leurs résultats, et préenregistreraient publiquement leurs intentions dans une banque de données publique, telle l'Open Science Framework. Ils publieraient ensuite leurs résultats, accompagnés de ceux de l'étude exploratoire dans un article d'une revue habituelle. Une telle approche laisse beaucoup de liberté, explique le politologue et statisticien Andrew Gelman, de l'université Columbia à New York. Mais elle est aussi suffisamment rigoureuse pour diminuer le nombre de fausses découvertes.

Plus généralement, l'heure est venue pour les scientifiques de prendre conscience des limites des statistiques conventionnelles. Avant tout, une estimation scientifique sérieuse de la plausibilité des résultats devrait être effectuée dès leur analyse : quels résultats ont été apportés par des recherches similaires ? Existe-t-il un mécanisme qui pourrait les expliquer ? Les résultats se recoupent-ils avec l'expérience clinique ? Voilà les questions décisives ! En y répondant dans de prochains travaux, Motyl peut encore espérer accéder au succès qu'il espère ! ■

BIBLIOGRAPHIE

D. BENJAMIN *ET AL.*,
***Redefine statistical significance,* 2017**
https://psyarxiv.com/mky9j

B. NOSEK *ET AL.*,
Restructuring incentives and practices to promote truth over publishability,
***Perspect. Psychol. Sci.*,**
vol. 7, pp. 615-631, 2012.

C. LAMBDIN
Significance tests as sorcery : Science is empirical – significance tests are not,
***Theory & Psychology*,**
vol. 2, pp. 67-90, 2012.

Le *big data*, un flot de données venant de toutes parts.

© Shutterstock.com/Dmitriy Rybin

LES DÉFIS DES BIG DATA

Comment faire face à ce déluge de données qui constitue les *big data* ? Comment les stocker, les exploiter, les analyser, les valoriser... ? Les algorithmes et les calculateurs à qui ces tâches incombent doivent être à la hauteur. La recherche sur ces deux aspects essentiels du traitement des *big data* est particulièrement active. Le développement des intelligences artificielles, « éduquées » par des quantités énormes de données, est une illustration de ces efforts. Ces systèmes sont parfois si performants que certains y voient une menace sur la façon de faire de la science.

L'ESSENTIEL

- L'intelligence artificielle telle qu'on l'imaginait dans les années 1950 a été plus difficile à développer que prévu.
- Depuis quelques années, le domaine a connu un grand renouveau avec les techniques de l'apprentissage profond, inspirées des réseaux de neurones du cerveau.
- Un réseau de neurones artificiels à apprentissage profond acquiert sans cesse de l'expertise à partir de nouvelles données.
- Un tel réseau a déjà battu un joueur de go professionnel, d'autres reconnaissent des images ou la parole...

L'AUTEUR

YOSHUA BENGIO
est professeur d'informatique à l'université de Montréal. Il est l'un des pionniers du développement des méthodes d'apprentissage profond.

La révolution de l'APPRENTISSAGE PROFOND

Pour créer une intelligence artificielle, pourquoi ne pas s'inspirer d'une intelligence naturelle ? C'est le pari des réseaux multicouches de neurones artificiels, des « machines » qui apprennent à partir de grandes quantités de données. Leurs succès sont saisissants !

En 2016, le programme AlphaGo, mis au point par Google DeepMind, faisait sensation en battant au jeu de go, par le score de 4 à 1, le Sud-Coréen Lee Sedol, réputé le meilleur joueur du monde. Un an après, en octobre 2017, AlphaGo s'est fait humilier avec un score de 100 à 0 par… AlphaGo Zéro, le nouveau logiciel de Google DeepMind !

La différence entre les deux logiciels ? Là où AlphaGo a été nourri par des millions de parties humaines, AlphaGo Zéro s'est contenté des règles du jeu et des positions des pierres sur le plateau. Il a appris en jouant des millions de parties contre lui-même, d'abord au hasard, puis en affinant sa stratégie. En quarante jours, il est devenu le meilleur !

Les deux logiciels ont toutefois un point commun : ils fonctionnent grâce à l'apprentissage profond, aboutissement de décennies de recherche sur l'intelligence artificielle. Aujourd'hui, les machines apprennent grâce à une architecture inspirée de celle du cerveau et d'un nombre considérable de données (AlphaGo Zéro les a générées lui-même). Un retour historique s'impose.

Dans les années 1950, le programme élaboré par Arthur Samuel pour jouer aux dames commence à damer le pion de joueurs de bon niveau. Stupéfaction dans le monde de l'informatique naissante. Le programme en question >

© Shutterstock.com/kc look

> est le premier à inclure un autoapprentissage. L'enthousiasme voire l'euphorie gagne les laboratoires. La conférence de Dartmouth de 1956 consacre la naissance de l'« intelligence artificielle ». L'un des pionniers, Marvin Minsky – en 1951, il avait construit la première machine à réseau neuronal, le Snarc – n'hésite pas à déclarer en 1970, dans *Life* : « Dans trois à huit ans nous aurons une machine avec l'intelligence générale d'un être humain ordinaire. » Cet optimisme était un tantinet précipité…

LA CHASSE AU SNARC

Les logiciels de l'époque, par exemple pour aider les médecins dans leur diagnostic, n'ont pas tenu leurs promesses. Les algorithmes étaient trop simples et manquaient des données nécessaires pour parfaire leur apprentissage. La puissance de traitement informatique était également insuffisante pour mener à bien les calculs complexes requis pour imiter, même approximativement, des subtilités de la pensée humaine.

Au milieu des années 2000, le rêve de construire des machines aussi intelligentes que des humains avait presque été abandonné. À cette époque, même le terme d'« intelligence artificielle » semblait avoir déserté le domaine de la science sérieuse. Les chercheurs décrivent la période allant des années 1970 au milieu des années 2000 par l'expression « AI winters », une sorte de long hiver glaciaire de l'intelligence artificielle où tous les espoirs avaient été brisés.

Les choses ont commencé à changer en 2005. Les perspectives dans le domaine de l'intelligence artificielle ont radicalement changé avec l'apprentissage automatique et l'émergence de l'« apprentissage profond », qui revisite le connexionnisme des années 1960 et s'inspire des neurosciences. Les réalisations actuelles de l'apprentissage profond sont à la hauteur des promesses d'antan, et les grandes entreprises des technologies de l'information, y consacrent désormais des milliards de dollars.

Qu'entend-on par apprentissage profond ? Il s'agit d'un traitement effectué par un grand nombre de neurones artificiels (imitant de façon très simplifiée les neurones biologiques) qui, par leurs interactions, permettent au système d'apprendre progressivement à partir d'images, de textes ou d'autres données. L'apprentissage repose sur des principes mathématiques généraux. Le résultat de l'apprentissage est une représentation (par exemple, « Cette image contient des éléments différents »), une décision (par exemple « Cette image représente Jeanne Dupont ») ou une transformation (par exemple la traduction d'un texte dans une autre langue).

La technique de l'apprentissage profond a transformé la recherche en intelligence artificielle, ranimant des ambitions oubliées pour la vision par ordinateur, la reconnaissance automatique de la parole et la robotique. Les premières applications ont vu le jour en 2012 pour la compréhension de la parole (Siri, qui équipe les iPhone en une illustration). Et peu après sont arrivés des logiciels qui identifient le contenu d'une image, une fonctionnalité qu'intègre maintenant le moteur de recherche de Google.

Tous ceux qui sont frustrés par les menus malcommodes de leur téléphone peuvent apprécier les avantages d'utiliser Siri ou un autre « assistant personnel » sur leur appareil. Et pour ceux qui se rappellent combien la reconnaissance d'objets était mauvaise il y a quelques années seulement, les avancées dans le domaine de la vision artificielle ont été phénoménales : aujourd'hui, moyennant certaines conditions, les ordinateurs savent reconnaître sur des images un chat, une pierre ou des visages presque aussi bien que les humains. Les logiciels d'intelligence artificielle font désormais partie du quotidien de millions d'individus, qui par exemple, n'écrivent plus de messages, mais les dictent à leur téléphone !

Ces progrès ont ouvert la voie à de nouvelles réalisations, à des applications commercialisées, et l'intérêt ne cesse de croître. La concurrence fait rage entre les entreprises pour attirer les jeunes talents… et les moins jeunes. De nombreux professeurs d'université spécialistes du domaine (la majorité, d'après certaines estimations) sont passés du milieu académique à l'industrie, séduits par des équipements de pointe et par de généreux salaires.

Les efforts déployés pour répondre aux défis de l'apprentissage profond ont débouché sur des réussites époustouflantes, les performances d'AlphaGo Zéro en sont l'illustration. Les applications vont s'étendre à d'autres domaines de l'expertise humaine, et pas seulement aux jeux. Par exemple, un algorithme d'apprentissage profond serait capable de diagnostiquer des insuffisances cardiaques sur des images d'IRM aussi bien qu'un cardiologue.

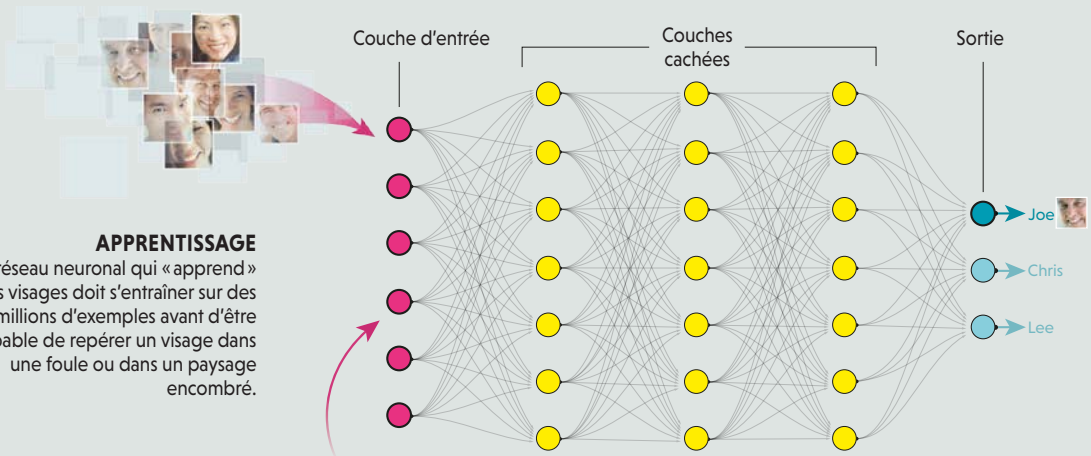
Ce succès récent de l'apprentissage profond a de quoi surprendre quand on se penche sur l'histoire de l'intelligence artificielle. Pourquoi celle-ci a-t-elle buté sur tant d'obstacles dans les décennies précédentes ? Parce qu'apprendre des choses nouvelles est, pour une machine, difficile. L'essentiel de la connaissance que nous avons du monde autour de nous n'est pas formalisé dans un langage écrit sous forme d'un ensemble de tâches explicites, ce qui est indispensable pour écrire un programme informatique. C'est pourquoi nous n'avons pas pu directement programmer un ordinateur pour effectuer la plupart des tâches qui nous sont

APPRENTISSAGE PROFOND ET ÉTUDE DE L'ADN

Dans le domaine de la génomique, l'accumulation de données nécessite des algorithmes puissants pour en extraire des informations. En 2015, Brendan Frey, de l'université de Toronto, et ses collègues ont développé DeepBind, un logiciel qui recourt à l'apprentissage profond pour analyser la façon dont des protéines se lient à l'ADN et à l'ARN. Les chercheurs détectent ainsi des mutations qui perturbent les processus cellulaires entraînant des maladies.

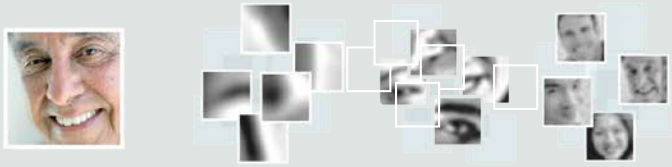
DES RÉSEAUX DE NEURONES QUI DÉVELOPPENT LEUR EXPERTISE

Les connexions entre neurones dans le cortex cérébral ont inspiré la création d'algorithmes d'apprentissage qui imitent ces liens complexes. On peut apprendre à un tel réseau de neurones artificiels à reconnaître un visage en l'entraînant avec un très grand nombre d'images. Le réseau détermine les traits qui lui permettent de distinguer un visage d'une main, par exemple, et reconnaît la présence de visages dans une image. Il utilise ensuite cette connaissance pour identifier des visages qu'il a déjà vus, même si l'image de la personne est un peu différente de celle sur laquelle il s'est entraîné. Pour reconnaître un visage dans une image, le réseau commence par analyser les pixels d'une image qui lui est présentée au niveau de la couche d'entrée. Puis il discerne les formes géométriques caractéristiques du visage au niveau de la couche suivante. En remontant la hiérarchie, des yeux, une bouche et d'autres traits du visage apparaissent. Enfin, une forme composite émerge et le réseau tente de « deviner » au niveau de la couche de sortie s'il s'agit du visage de Joe, Chris ou Lee.



APPRENTISSAGE
Un réseau neuronal qui « apprend » des visages doit s'entraîner sur des millions d'exemples avant d'être capable de repérer un visage dans une foule ou dans un paysage encombré.

RECONNAISSANCE
Une photographie de visage injectée dans le réseau est transformée séquentiellement par les différentes couches de neurones, permettant en sortie d'identifier le visage.



Chaque couche identifie progressivement des traits plus complexes

Jen Christiansen (graphique) - Punchstock (visages)

évidentes ou faciles, qu'il s'agisse de comprendre la parole, d'interpréter des images ou de conduire une automobile. Les tentatives en ce sens (organiser des ensembles de faits en bases de données élaborées afin de doter les ordinateurs d'un fac-similé d'intelligence) ont rencontré peu de succès.

UNE BONNE DÉCISION

C'est là qu'entre en scène l'apprentissage automatique – le cadre général de l'apprentissage profond. Il se fonde sur des principes généraux, permettant aux systèmes d'utiliser les données disponibles pour apprendre à bien décider, à acquérir de bonnes connaissances et, *in fine*, à rechercher de nouvelles données pour apprendre mieux. Mais qu'est-ce qu'une « bonne » décision?

Pour les animaux, au regard de l'évolution, une bonne décision optimise les chances de survie et de reproduction. Dans les sociétés humaines, une bonne décision est plus subtile à définir et peut inclure des interactions sociales qui confèrent un statut élevé ou une sensation de bien-être. Pour une voiture autonome la qualité de la prise de décision sera d'autant meilleure que le véhicule reproduira avec fidélité les comportements de bons conducteurs humains.

Les connaissances nécessaires pour prendre une bonne décision dans un contexte particulier ne sont pas nécessairement évidentes et faciles à traduire en langage informatique. Une souris, par exemple, connaît bien son environnement et a un sens inné des endroits où elle doit flairer, sait instinctivement comment bouger ses pattes, trouver de la nourriture, éviter >

> les prédateurs… Aucun informaticien ne serait capable de spécifier un programme étape par étape pour produire ces comportements. Et pourtant, ces connaissances sont codées dans le cerveau du rongeur.

Avant de créer des ordinateurs capables d'apprendre, les scientifiques doivent répondre à des questions fondamentales, concernant en particulier le partage entre l'inné et l'acquis. Depuis les années 1950, les chercheurs ont étudié et tenté d'affiner les principes généraux qui permettent aux animaux et aux humains (ainsi que, en l'occurrence, aux machines) d'acquérir des connaissances par expérience. L'apprentissage automatique vise à établir des procédures, des algorithmes d'apprentissage, qui confèrent à une machine la capacité d'apprendre à partir d'exemples qu'on lui présente.

La science de l'apprentissage automatique fait face à un résultat négatif formel surnommé *No free lunch* («Rien n'est gratuit»): si tous les problèmes sont équiprobables, il n'existe pas d'algorithme universel, c'est-à-dire meilleur que tous les autres sur l'ensemble des problèmes. Il faut donc élaborer des algorithmes distincts pour s'attaquer à différentes catégories de problèmes (par exemple reconnaître un coucher de soleil ou traduire un texte en ourdou). L'apprentissage automatique repose ainsi sur deux piliers, théorique et expérimental: un problème réel ne satisfait pas toujours les hypothèses théoriques de l'algorithme, et la validation de l'algorithme sur des données réelles est toujours nécessaire.

Or il semble que notre cerveau incorpore des algorithmes généraux qui nous permettent d'apprendre une multitude de tâches auxquelles l'évolution n'avait pas préparé nos ancêtres: jouer aux échecs, construire des ponts ou faire de la recherche en intelligence artificielle.

Ces compétences supplémentaires suggèrent que l'intelligence humaine pourrait encore être source d'inspiration pour créer des machines dotées d'une forme d'intelligence générale. C'est exactement pourquoi les développeurs ont adopté le modèle du cerveau comme guide pour concevoir des systèmes intelligents sous la forme de réseaux de neurones.

DES CERVEAUX VIRTUELS

La principale unité de calcul du cerveau est le neurone. Cette cellule transmet un signal sous la forme d'une impulsion électrique qui se propage jusqu'à la synapse, la zone de communication avec un autre neurone. Des molécules nommées neurotransmetteurs sont libérées, puis réabsorbées par le neurone cible, qui prend alors le relais. La propension d'un neurone à transmettre un signal vers un autre neurone est la force synaptique. À mesure qu'un neurone «apprend», sa force synaptique augmente, et il a plus de chances d'envoyer des signaux à ses voisins quand il est stimulé par une impulsion électrique.

Les neurosciences ont influencé l'émergence des réseaux de neurones artificiels, où ces éléments sont connectés de façon matérielle ou logicielle. Les premiers programmeurs de cette sous-discipline de l'intelligence artificielle, connue sous le nom de connexionnisme, postulaient que les réseaux neuronaux seraient capables d'apprendre des tâches complexes en modifiant progressivement les connexions entre neurones, de telle façon que les schémas d'activité neuronale coderaient le contenu des données livrées sous forme d'images, de sons… À chaque exemple soumis au réseau, le processus d'apprentissage se poursuivrait en modifiant les forces synaptiques entre neurones connectés. Ces valeurs convergeraient petit à petit vers la configuration du réseau qui représente le mieux, par exemple, les images de couchers de soleil.

Les réseaux de neurones actuels sont des versions améliorées des travaux pionniers issus du connexionnisme. Cependant, les algorithmes d'apprentissage correspondants requièrent une participation active de l'homme. La plupart d'entre eux utilisent un apprentissage supervisé, où chaque exemple d'apprentissage est accompagné d'une étiquette indiquant le sujet de l'apprentissage, tel que «coucher de soleil».

Dans ce cas, l'objectif de l'algorithme d'apprentissage supervisé est, à partir d'une donnée d'entrée constituée par une photographie, de produire comme sortie le nom de l'objet central de l'image. L'opération mathématique qui transforme une entrée en une sortie est une fonction. Les valeurs numériques qui définissent cette fonction, telles que les forces synaptiques, constituent une solution de la tâche d'apprentissage.

Dans trois à huit ans nous aurons une machine avec l'intelligence générale d'un être humain ordinaire… pouvait-on lire en 1970

Il serait facile pour un système d'apprendre par cœur les réponses correctes sur les exemples connus (il suffit d'une mémoire suffisante). Mais cela n'a pas grand intérêt. Même si on accumule des millions d'images de coucher de soleil, le nombre d'images possibles d'une telle scène est infini. L'objectif de l'apprentissage est d'être capable de donner de bonnes réponses sur de nouveaux exemples, inconnus du système, en généralisant les exemples connus. Le bon niveau de généralisation dépend du contexte. Ainsi, on peut vouloir reconnaître la notion d'arbre en général, mais on peut aussi vouloir distinguer les chênes des hêtres, ou vouloir reconnaître le hêtre d'un jardin donné...

Un tel algorithme doit aussi s'appuyer sur certaines hypothèses relatives aux données et à ce que pourrait être une solution possible à un problème donné. Ainsi, le logiciel doit intégrer comme principe que si des données d'entrée d'une fonction particulière sont semblables, leur sortie ne devrait pas être radicalement différente: modifier quelques pixels sur une image de chat ne devrait pas transformer l'animal en chien.

De telles hypothèses faites sur des images sont utilisées par les réseaux de neurones dits convolutifs. Ces programmes sont à l'origine du renouveau de l'intelligence artificielle. Les réseaux neuronaux convolutifs utilisés dans l'apprentissage profond comprennent de nombreuses couches de neurones organisées de sorte que le logiciel sera robuste vis-à-vis des changements dans l'objet qu'il tente d'analyser. Il reconnaîtra l'élément même s'il a un peu bougé. Ainsi, un réseau bien entraîné identifiera un visage quel que soit l'angle de vue.

UNE STRUCTURE MULTICOUCHE

La configuration d'un réseau convolutif s'inspire de la structure en couches multiples du cortex visuel, c'est-à-dire la partie de notre cerveau qui reçoit les signaux des yeux. Ces nombreuses couches de neurones virtuels justifient le qualificatif de «profond» donné à ces réseaux et confèrent à ces derniers une meilleure capacité à appréhender le monde environnant.

L'apprentissage profond est devenu envisageable il y a une dizaine d'années grâce à certaines innovations, alors que l'intérêt pour l'intelligence artificielle était au plus bas. Un organisme canadien financé par le gouvernement et par des fonds privés, l'Icra (Institut canadien de recherches avancées), a contribué à ranimer la flamme en soutenant un programme dirigé par Geoffrey Hinton, de l'université de Toronto, et dont faisaient partie Yann LeCun, de l'université de New York et du centre de recherche de Facebook à Paris, Andrew Ng, de l'université de Stanford, Bruno Olshausen, de l'université de Californie à Berkeley, moi-même et plusieurs autres. En raison du scepticisme ambiant, il était difficile à cette époque de publier des articles sur le sujet et même de persuader des étudiants de faire leur thèse dans ce domaine. Mais nous étions convaincus qu'il fallait persévérer dans cette voie.

La réticence vis-à-vis de l'intelligence artificielle découlait initialement de l'idée que la création de réseaux neuronaux était sans espoir à cause de la difficulté à optimiser leur performance pour apprendre efficacement.

L'optimisation est une branche des mathématiques qui vise à trouver la meilleure combinaison de paramètres, ici les poids synaptiques des neurones virtuels, pour atteindre un certain objectif. Lorsque la relation qui lie les paramètres et l'objectif est assez simple (on dit que l'objectif est une «fonction convexe» des paramètres) alors on peut améliorer progressivement les paramètres et l'on parle d'optimisation convexe. La procédure d'entraînement est répétée jusqu'à ce que les paramètres s'approchent aussi près que possible des valeurs qui produisent le meilleur résultat. On cherche en fait un minimum global, à savoir le jeu de paramètres qui permet d'atteindre la valeur la plus basse (et la meilleure) de l'écart à l'optimum.

En général, la situation n'est pas aussi simple, la fonction reliant les paramètres à l'objectif qu'ils atteignent n'étant pas convexe. Or l'optimisation non convexe représente un défi bien plus difficile. De nombreux chercheurs >

APPRENTISSAGE PROFOND ET ART PSYCHÉDÉLIQUE

L'apprentissage profond est performant pour la reconnaissance d'objets dans une image, mais il est en réalité assez difficile de comprendre ce qui se passe au cours de l'analyse à chaque niveau du réseau de neurones. La société Google a développé le logiciel DeepDream qui utilise l'apprentissage profond. Les chercheurs peuvent analyser le résultat à différentes étapes du processus en demandant au logiciel de faire ressortir ce qu'il «voit» dans l'image. Ainsi, telles des paréidolies – l'illusion de voir, par exemple, un chien dans un nuage qui en a vaguement la forme –, le logiciel, qui s'est, par exemple, entraîné à reconnaître des images d'animaux, verra des oiseaux et des rongeurs dans une photographie d'un champ de maïs. Chacun peut créer ses propres images psychédéliques sur le site Deepdreamgenerator.com.

© Stoonn/Shutterstock.com

© Deepdreamgenerator.com

pensaient qu'il n'était pas possible de le relever. L'algorithme d'apprentissage pourrait en effet se retrouver coincé dans un minimum local, d'où il ne pourrait sortir en ajustant les paramètres par petites touches.

Or avec mes collègues, nous avons montré que si le réseau de neurones est assez grand, le problème des minima locaux est fortement réduit. Dans un tel réseau, la plupart des minima locaux correspondent à un niveau d'apprentissage quasi équivalent à celui du minimum global.

Bien que les problèmes théoriques de l'optimisation soient, en principe, résolubles, la construction de grands réseaux comportant plus de deux ou trois couches avait souvent échoué. À partir de 2005, les recherches financées par l'Icra ont commencé à porter leurs fruits et les obstacles ont été surmontés progressivement. En 2006, nous sommes parvenus à entraîner des réseaux neuronaux plus profonds en utilisant une technique qui opérait couche par couche.

Par la suite, en 2011, nous avons trouvé un moyen d'entraîner des réseaux encore plus profonds (avec davantage de couches de neurones virtuels) en modifiant les opérations effectuées par chacun des neurones, ce qui les rapprochait davantage des neurones biologiques dans leur façon de traiter l'information. Nous avons aussi découvert qu'en injectant du bruit aléatoire dans les signaux transmis entre les neurones au cours de l'apprentissage – comme ce qui a cours dans le cerveau –, les réseaux apprenaient mieux à identifier correctement une image ou un son.

Par ailleurs, deux facteurs essentiels ont contribué au succès des techniques d'apprentissage profond. Le premier est l'augmentation d'un facteur 10 de la puissance de calcul des ordinateurs grâce aux processeurs dédiés au traitement d'image, ce qui a permis d'entraîner des réseaux plus grands en un temps raisonnable. Le second facteur est que l'on avait désormais accès à d'énormes bases de données étiquetées, avec lesquelles les algorithmes d'apprentissage ont pu s'exercer à reconnaître un « chat », par exemple.

Une autre raison aux récents succès de l'apprentissage profond est que le réseau est capable d'apprendre à effectuer un ensemble de calculs qui construisent ou analysent pas à pas une image, un son ou un autre type de donnée. Avec une profondeur suffisante, ces dispositifs excellent dans beaucoup de tâches de reconnaissance visuelle ou auditive. Des travaux théoriques et expérimentaux récents ont d'ailleurs montré qu'il est impossible d'effectuer efficacement certaines de ces opérations mathématiques sans un réseau assez profond.

Mais que se passe-t-il au cœur du réseau de neurones profond ? Chaque couche transforme son entrée et produit une sortie qui est envoyée à la couche suivante (*voir l'encadré page 45*). Les premières couches se concentrent sur les détails aux échelles les plus petites, puis les couches suivantes agrandissent les échelles considérées. Plus les couches sont profondes, plus elles vont représenter des concepts abstraits.

RECONNAÎTRE DES VIDÉOS

Jusqu'à récemment, les réseaux de neurones artificiels se distinguaient en grande partie par leur capacité à effectuer des tâches telles que la reconnaissance de motifs dans des images statiques. Mais d'autres types de réseaux neuronaux affichent des résultats intéressants sur des événements dynamiques. Les « réseaux neuronaux récurrents », par exemple, ont démontré leur capacité à effectuer une séquence de calculs pour traiter la parole ou une vidéo. Les données séquentielles sont constituées d'unités (phonèmes ou mots dans le cas de la parole) qui se suivent. La façon dont les réseaux neuronaux récurrents traitent leurs entrées ressemble un peu au fonctionnement du cerveau. Les signaux qui se propagent parmi les neurones changent constamment à mesure que les entrées sensorielles sont traitées.

Les réseaux récurrents parviennent à prédire quel sera le mot suivant dans une phrase et peuvent ainsi produire une nouvelle séquence de mots, l'un après l'autre. Ils peuvent aussi s'atteler à des tâches plus complexes. Après avoir « lu » tous les mots d'une phrase, le réseau est à même de deviner le sens de la phrase entière. Un réseau récurrent distinct peut alors utiliser le traitement sémantique du premier réseau pour traduire la phrase dans une autre langue.

La recherche sur les réseaux neuronaux récurrents a connu sa propre période de stagnation à la fin des années 1990 et au début des années 2000. Mes propres travaux théoriques suggéraient qu'ils rencontreraient des difficultés à apprendre à récupérer de l'information du passé lointain, à savoir les premiers éléments de la séquence en cours de traitement. Mais certains de ces problèmes ont été résolus grâce à des techniques consistant à stocker l'information pour qu'elle persiste durablement. Ces réseaux neuronaux utilisent la mémoire temporaire (mémoire cache) d'un ordinateur pour traiter des informations multiples et dispersées, comme des idées contenues dans différentes phrases réparties au fil d'un document.

Le retour en force des réseaux neuronaux à apprentissage profond après le long sommeil de l'intelligence artificielle n'est pas uniquement un triomphe technique. C'est aussi une leçon de sociologie des sciences. En particulier, il souligne la nécessité de soutenir des idées qui font fi du *statu quo* technologique et d'encourager la tenue d'un portefeuille de recherches diversifié, laissant de la place à des champs provisoirement passés de mode. ■

BIBLIOGRAPHIE

D. SILVER *ET AL.*, ***Mastering the game of Go without human knowledge,*** **Nature, vol. 550, pp. 354–359, 2017.**

Y. LECUN *ET AL.*, ***Deep learning*****, Nature, vol. 521, pp. 436-444, 2015.**

Y. BENGIO *ET AL.*, ***Representation learning: A review and new perspectives*****, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35(8), pp. 1798-1828, 2013.**

ImageNet classification with deep convolutional neural networks, 26th Annual Conference on Neural Information Processing Systems (NIPS 2012), 2012.

SUR LE WEB

Cours de Yann LeCun au Collège de France, 2015-2016: https://www.college-de-france.fr/site/yann-lecun/index.htm


© Air France


© Pixabay


© SNCF

« Quand les “ data scientists ” nous simplifient la vie »

En ce début de XXIe siècle, la science des données s’immisce tous les jours un peu plus dans notre quotidien. Reconnaissance vocale, moteurs de recherche, publicités ciblées, recommandations d’achats… Elle est bien sûr déjà omniprésente dans nos activités numériques. Mais cette *data science* gagne aussi progressivement tous les secteurs d’acti-vités économiques. Décoller en A380, emprunter sereinement son train quotidien ou bien encore se remettre d’une catastrophe naturelle: voici trois cas très concrets où des algorithmes aussi discrets qu’efficaces sont déjà à la manœuvre. Autant d’applications qui nécessitent les compétences de *data scientists* chevronnés… Un secteur d’activité florissant qui pourrait créer trois millions d’emplois en Europe! Et grâce au nombre et à la qualité de ses formations, la France est bien placée pour ne pas laisser filer sa part du gâteau… Rien que sur l’Hexagone, le recrutement de 130 000 *data scientists* sera nécessaire d’ici à 2020. En ferez-vous partie?

SOMMAIRE

Cahier spécial réalisé en partenariat avec

AIRFRANCE MAIF assureur militant SNCF

Décoller à l'heure

Les Airbus A380 d'Air France qui atterrissent à Roissy sont aujourd'hui analysés par des algorithmes de maintenance prédictive. Objectif : empêcher toute panne inopinée source d'importants retards… Reportage.

Lundi 11 décembre 2017, 10 heures du matin. Dans un froid glacial, un Airbus A380 d'Air France en provenance de Washington D.C. atterrit sur le tarmac de l'aéroport parisien Roissy-Charles-de-Gaulle. Les passagers commencent à s'ébrouer puis quittent leurs sièges, tandis que tout un ballet se met en place autour du géant des airs pour sortir les bagages des soutes, récupérer les déchets, faire des contrôles sur l'appareil, etc. Mais ce que les passagers ne savent pas, c'est qu'un autre ballet démarre en même temps, numérique celui-là…

En effet, à peine a-t-il posé les roues sur la piste que notre A380 commence à transmettre par wifi toutes les données enregistrées en continu sur son dernier aller-retour *via* les centaines de milliers de capteurs dont ses équipements sont truffés. Consommation électrique, vitesse, angle, température… L'ensemble de ces *data* (1 gigaoctet en moyenne) est alors traité par un logiciel de maintenance prédictive baptisé Prognos. Ce dernier transmet ensuite ses résultats en temps réel aux ingénieurs du département maintenance avion d'Air France-KLM.

1 Go de données par vol

C'est là que nous rencontrons Paul-Louis Vincenti, *data analyst* au département recherche opérationnelle chez Air France Industries-KLM Engineering & Maintenance. «*Grâce à une infrastructure big data distribuée dédiée* [1]*, Prognos détecte en moins de trente minutes les éventuelles signatures de prémices de pannes cachées dans cette masse de données*», explique-t-il. Le cas échéant, le service maintenance agit de manière préventive… De quoi éviter toute panne inopinée qui empêcherait l'avion de décoller, générant un retard important, voire une annulation de dernière minute.

Dans la pratique, les algorithmes [2] de Prognos mettent en œuvre des méthodes de *machine learning* par apprentissage supervisé * qui s'entraînent sur l'historique de précédents vols. «*Nous sommes en effet très attachés à l'explicabilité des résultats,* précise P. L. Vicenti. *Voilà pourquoi nous évitons volontairement le* deep learning*, et les réseaux de neurones* **: *nous voulons toujours pouvoir justifier d'un point de vue physique les décisions prises*».

En 2017, grâce à Prognos, les équipes d'Air France ont remplacé dans les A380 huit pompes d'alimentation moteur, et quatre capteurs de rotation logés dans le nez des trains d'atterrissage. «*Jusqu'à présent,* confirme P. L. Vincenti*, toutes ces déposes concernaient bien des équipements déclarés défectueux après inspection*». Fort de ces résultats, l'outil de maintenance prédictive a aussi été mis en place sur les dix-huit Boeing 747 de KLM.

* Apprentissage automatique à partir d'exemples annotés
** Méthodes mimant la profondeur des couches d'un cerveau

1. Hadoop Distributed File System
2. Développés en langages Python et Spark, ils s'appuient sur la librairie Scikit learn

© Air France

« Les algorithmes de maintenance prédictive des A380 doivent toujours rester explicables »

PAUL-LOUIS VICENTI
***Data analyst* chez Air France-KLM**

AMÉLIORER LA RELATION AVEC LES PASSAGERS

Chez Air France-KLM, une seule et même plateforme *big data* centralise aussi toutes les données que la compagnie collecte sur ses clients *via* divers canaux: *call centers*, site web, réseaux sociaux, dans l'aéroport, lors de salons… ou bien encore à bord des avions. Toutes ces données sont passées à la moulinette d'algorithmes qui en tirent des recommandations adaptées au profil de chaque client: options pour le menu à bord du prochain vol, propositions de destinations et de tarifs personnalisés, temps de transport jusqu'à l'aéroport de départ, guidage du passager en correspondance dans l'aéroport… Sur ordinateur ou tablette, tous les personnels au sol ou en cabine auront bientôt accès au dossier de chaque client, et à ses éventuels échanges en cours avec d'autres services de la compagnie. Le tout se fait bien sûr dans le respect de la vie privée définie par la Cnil.

Des sinistrés pris en charge plus vite

Grâce à la *data science*, on commence à prédire l'étendue des dégâts… avant même qu'une catastrophe naturelle s'abatte sur une région donnée. De quoi mieux anticiper la prise en charge des futurs sinistrés. Explications…

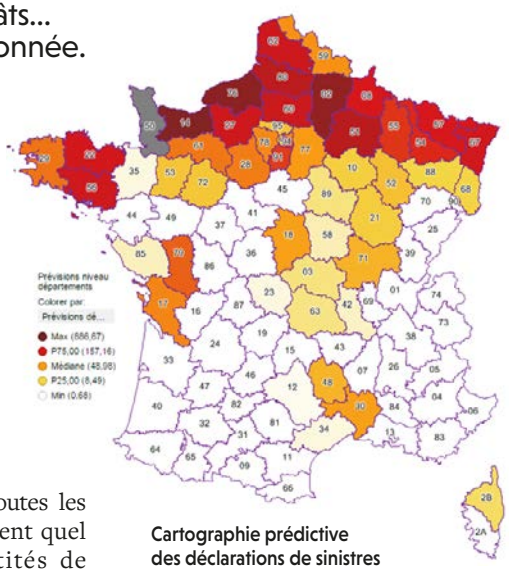
Inondations, tempêtes, cyclones, ouragans… Entre 1988 et 2013, les catastrophes naturelles ont coûté 1,9 milliard d'euros en indemnisations, avec 431 000 sinistrés en moyenne chaque année. Quand des catastrophes de ce type s'abattent sur un territoire, les assureurs activent immédiatement leurs plans de gestion de crise. Principal objectif: traiter au plus vite l'arrivée en masse de dossiers de demande d'indemnisation d'assurés impactés par la catastrophe et qui risquent d'affluer en agences. Mais aujourd'hui les assureurs veulent aller plus loin…

«Data visualisation» des futurs sinistrés

ORLANE MONNET

«*À la MAIF, nous développons par exemple un outil capable de prédire le nombre de déclarations de sinistres sur telle ou telle zone, avant même qu'elle ne soit touchée par une tempête hivernale ou une inondation cévenole*», lance Orlane Monnet du département Études statistiques de cet assureur. Calibré sur la sinistralité des événe-ments climatiques passés les plus significatifs (tempêtes Klaus, Xynthia, Joachim…), ce modèle de prévision utilise deux grands types de données. Tout d'abord, des données sur les communes et les assurés qui risquent d'être touchés: maisons ou appartements, domiciles ou résidences secondaires, propriétaires ou locataires, densité de population, existence ou non d'un plan de prévention des risques, etc. Ensuite, le modèle utilise aussi bien sûr des données météorologiques: force du vent prévue, durée estimée, quantité de précipi-tations à venir, etc.

«*Écrit en langage R*, notre outil se lance et s'utilise* via *une seule et unique interface de* data visualisation** *nommée Tibco Spotfire,* indique Orlane Monnet. *Elle permet de visualiser sur une carte le nombre de déclarations de sinistres prévisibles – plus les zones sont touchées plus elles apparaissent foncées.*» Par simple clic sur des curseurs, elle permet aussi de mesurer la sensibilité du modèle de façon interactive; exemple, si on majore par sécurité toutes les rafales de 10 km/h, on voit directement quel pourcentage augmente les quantités de déclarations de sinistres sur la carte. La MAIF a commencé à utiliser son outil sur des inondations fin 2016 et sur de petites tempêtes début 2017.



Cartographie prédictive des déclarations de sinistres par département

Anticiper les réparations

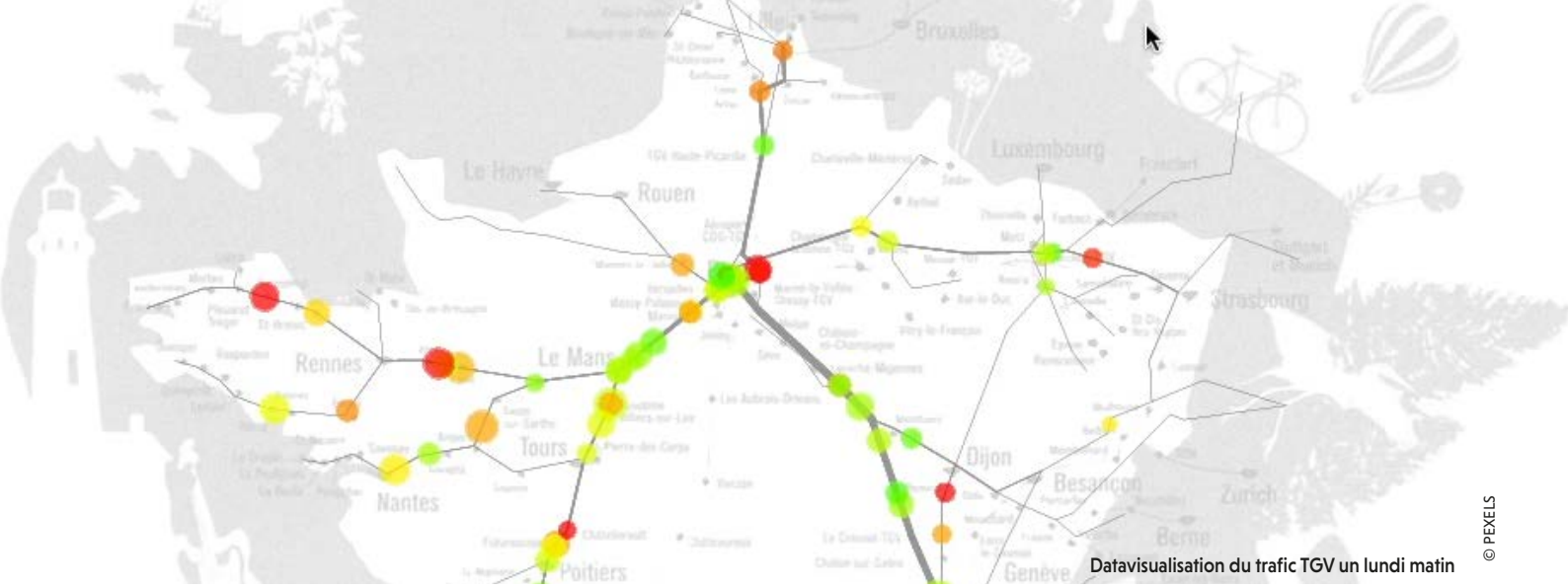
À l'avenir, ce type d'outil pourrait encore être amélioré pour évaluer plus précisément les coûts attendus, et même commencer à mobiliser les entreprises réparatrices (loueurs de motopompes et de groupes électrogènes, réparateurs de carrosseries de voitures…). Mais pour cela, il faudrait collecter des données encore plus précises sur les assurés: équipements, habite à tel étage, arbres sur son terrain, type de toiture, parking aérien ou souterrain, terrain en forte ou faible pente… Enfin, ce type d'outil pourrait aussi servir à alerter par mail ou SMS les assurés concernés et leur prodiguer des conseils de prévention personnalisés.

« L'usage d'algorithmes de machine learning nous aidera à encore mieux assurer les biens de nos sociétaires »

© Darri MAIF

STÉPHANE RENOUX
responsable du programme Data & intelligence artificielle à la MAIF

* Langage informatique dédié aux statistiques et à la science des données
** En mettant en images les données, la « dataviz » est un moyen d'identifier des tendances et corrélations



Datavisualisation du trafic TGV un lundi matin

© PEXELS

Améliorer l'information temps réel

SNCF met au point des algorithmes de *big data* capables de prédire en temps réel l'affluence à bord d'un train, et l'heure prévisionnelle de ses futurs passages en gare en cas de situation perturbée. Explications avec Maguelonne Chandesris qui dirige l'équipe Innovation & Recherche « Statistique, Économétrie et Datamining » de SNCF.

MAGUELONNE CHANDESRIS

Des futurs du passé... au futur du présent

Quel est l'objectif principal de ce projet ?

M. C. Baptisé RELUANCE (qui contient « retards et affluence en avance »), ce projet vise deux grands objectifs : prédire en temps réel le niveau de confort à bord (par exemple si on trouvera une place assise) et à quelle heure est prévue l'arrivée d'un train dans telle ou telle gare en cas de situation perturbée... Si le second point intéresse l'ensemble des 5 millions de voyageurs qui empruntent nos trains chaque jour, le premier concerne ceux qui prennent des trains régionaux, sans réservation. Pour y parvenir, nous avons développé des algorithmes de *machine learning* déjà testés sur plus de 3000 trains en circulation. Chaque jour, ils s'alimentent de centaines de milliers de données, à partir desquelles ils fournissent des millions de prédictions, au fur et à mesure de la remontée en temps réel des informations !

Sur quels types de données se basent-ils ?

M. C. Dans la pratique, nos algorithmes utilisent deux grands types de données. Tout d'abord celles qui concernent la localisation précise des trains en temps réel, collectée par des capteurs placés au sol et à bord des trains. Mais aussi des données sur le nombre de passagers qui montent et descendent des trains à chaque gare, obtenues par des capteurs installés à bord, au niveau des portes. Toutes ces données sont acheminées en temps réel, *via* des interfaces de programmation (API), vers nos algorithmes de calcul, hébergés sur nos serveurs situés près de la gare de Lyon.

Comment fonctionnent ces algorithmes ?

M. C. Leur principe général est facilement compréhensible. Très concrètement, ils vont chercher dans l'historique de ce type de données les situations les plus semblables à la situation actuelle. Une fois celles-ci identifiées, ils s'appuient sur ce qui est advenu après ces situations semblables passées (les « futurs du passé »)... pour prédire ce qui va se passer maintenant (le « futur du présent »). Point important : ces algorithmes dits « d'apprentissage non-paramétrique » réalisent leur phase d'apprentissage en continu, sans aucun reparamétrage à effectuer, facilitant la maintenance industrielle. Au final, ils améliorent sensiblement les prédictions de l'heure d'arrivée des trains à destination, et une marge d'erreur sur l'affluence inférieure à vingt voyageurs... pour des trains pouvant transporter plus de neuf cents personnes.

Quelles sont les prochaines étapes ?

M. C. Forts de ces résultats, nous avons déjà démarré les travaux d'industrialisation. Il nous faut notamment implémenter ces algorithmes de manière robuste dans « l'immeuble numérique » du groupe SNCF, afin que ces prédictions puissent alimenter tous nos systèmes d'analyse et de diffusion : affichages en gare, applications clients, centres opérationnels de supervision...
Ces informations en temps réel seront en effet aussi très utiles aux agents qui régulent la circulation des trains et les flux de passagers en gare. Au final, tous nos trains devraient bénéficier de ces prédictions algorithmiques à l'horizon 2020.

© SNCF

« Et si vous deveniez “ data scientist ” ? »

Comme l'illustrent les trois précédents articles de ce cahier, les entreprises ont de plus en plus besoin de *data scientists*. Mais comment se former à cette nouvelle profession ? Le point sur les différentes voies pour y parvenir…

« *Métier le plus sexy du XXIe siècle* ». C'est en ces termes que la prestigieuse Harvard Business Review qualifie le job de *data scientist* ! Un métier en plein boom consistant à extraire de nouvelles connaissances à partir de données (*data*), grâce à des techniques issues des mathématiques appliquées, de la statistique et de l'informatique. En quelques années à peine, les formations se sont multipliées. Alors, comment s'y retrouver ?

Les formations de qualité remplissent généralement trois critères : elles sont souvent d'un niveau master, adossées à des laboratoires ou des chaires… et à un écosystème industriel. L'objectif étant de pouvoir ainsi adapter les enseignements au rythme des évolutions de la recherche et de l'industrie. On trouve de tels masters dans les universités et les grandes écoles, le plus réputé restant le MVA (Mathématiques, Vision, Apprentissage) de l'ENS Paris-Saclay… qui a fêté ses vingt ans en 2017 ! Certains sont plutôt à dominante « maths », d'autres plus à dominante informatique.

Trois casquettes

Car dans la pratique, le métier de *data scientist* en regroupe plusieurs. Il y a bien sûr le concepteur d'algorithmes, nécessitant une solide formation générale en maths, en informatique scientifique et en modélisation. Mais il y a aussi l'« intégrateur » ou « développeur informatique » qui lui met en œuvre des algorithmes et des méthodes d'apprentissage, et doit gérer problèmes opérationnels et informatiques. Sans oublier le « spécialiste métier » ou « utilisateur » : véritable expert dans la manipulation de ces outils, et capable de comprendre le contexte d'utilisation (médical, marketing, transports…).

CEPE, École Polytechnique, TelecomParisTech, université Paris-Descartes… de nombreux organismes proposent aussi des formations continues qui se sont fortement développées ces dernières années, notamment dans le domaine du *big data*. Enfin, même s'ils ne remplacent pas les formations classiques, les cours en ligne ouverts à tous (Mooc) peuvent aussi se révéler très utiles.

« *Les* data scientists *devront de plus en plus se spécialiser.* »
NICOLAS VAYATIS
Responsable du master MVA de l'ENS Paris-Saclay

« *L'enjeu est à la fois de faire évoluer nos statisticiens en interne vers ces nouveaux métiers et de recruter en externe.* »
OLIVIER BAES
Responsable du Datalab de la MAIF

« *Nous avons créé avec l'École des ponts et chaussées la première Chaire en France regroupant optimisation et apprentissage.* »
ISABEL GOMEZ GARCIA DE SORIA
Directrice de la Recherche Opérationnelle chez Air France

« *Nous initions des concours en sciences des données depuis 2014.* »
MAGUELONNE CHANDESRIS
***Data scientist* chez SNCF Innovation & Recherche**

En chiffres

- La France aura besoin de **130 000** *data scientists* d'ici à 2020
- La *data science* pourrait créer **3 millions d'emplois** en Europe
- Un *data scientist* gagne en moyenne de **45 000 à 55 000 euros** par an

DATA CHALLENGES : DE VRAIS DÉFIS !

Aujourd'hui, de nombreuses formations intègrent des « data challenges », compétitions dans lesquelles des équipes d'étudiants data scientists s'affrontent pour résoudre des problèmes très concrets : prévoir la consommation électrique d'une zone, la fréquentation d'une gare, l'intérêt de consommateurs pour tel ou tel produit, améliorer des algorithmes de recommandations d'achat en ligne, etc. Souvent proposés par des entreprises, y compris les grandes (Air France, SNCF…), ces *data challenges* permettent aussi à certains étudiants de se faire repérer pour décrocher un stage… voire un job !

© Shutterstock

L'ESSENTIEL

- Les données, de diverses natures et toujours plus abondantes, sont interrogées *via* des requêtes ou analysées pour y dénicher des corrélations, des indicateurs...
- Les applications nécessitent en amont des algorithmes qui optimisent la collecte, la gestion et le prétraitement des données.
- Ces traitements sont facilités par divers outils, tels que l'algèbre relationnelle, le calcul distribué, le modèle MapReduce...
- Au-delà du calcul proprement dit, la gestion des données passe aussi par leur indexation et leur protection.

LES AUTEURS

MOKRANE BOUZEGHOUB
est professeur à l'université de Versailles et directeur adjoint scientifique de l'INS2I, au CNRS.

ANNE DOUCET
est professeure à l'université Pierre-et-Marie-Curie, déléguée scientifique à l'international de l'INS2I, au CNRS.

Traiter les BIG DATA avec raffinement

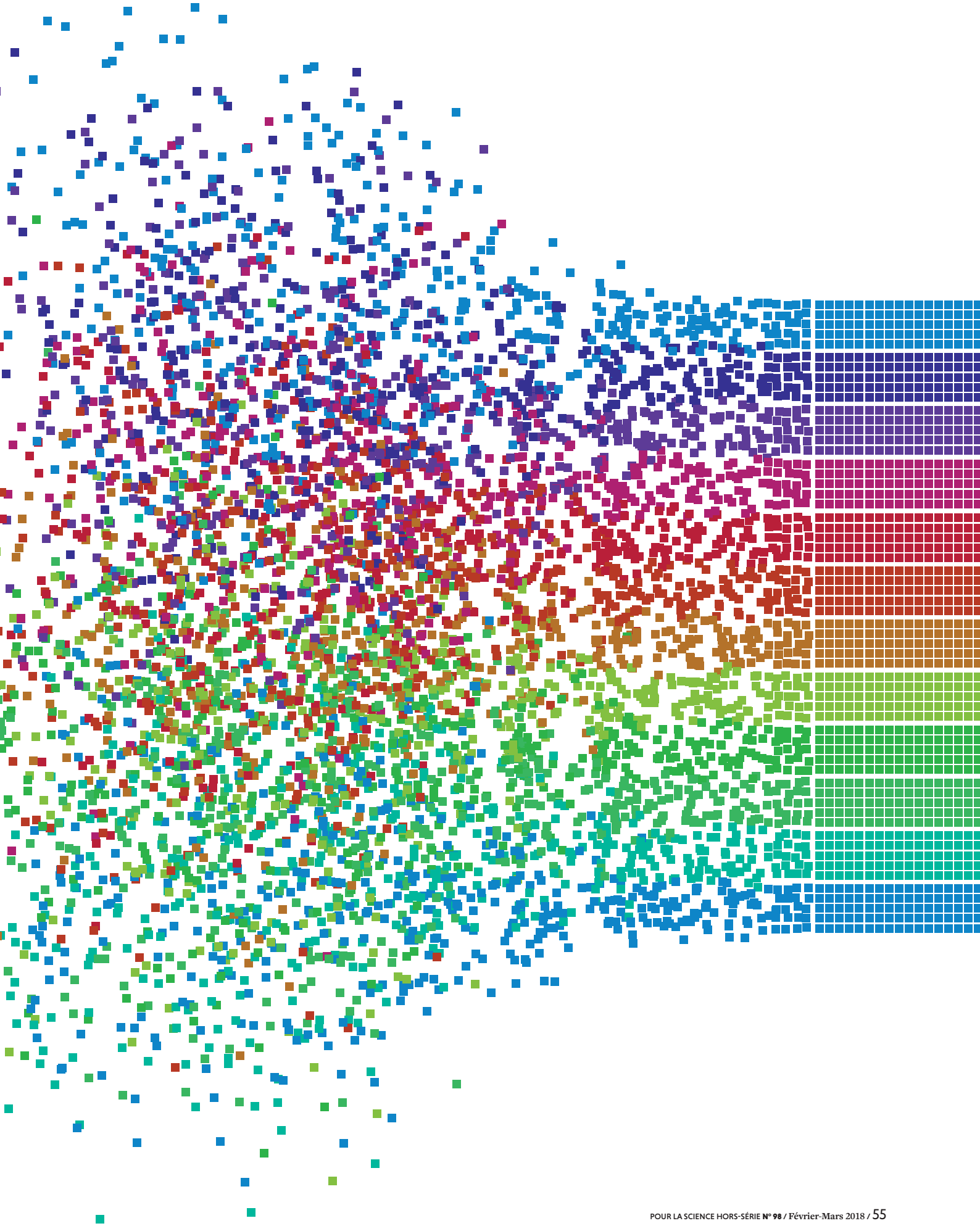
La gestion des *big data*, de la collecte jusqu'à l'analyse, requiert de nombreux algorithmes. Tous sont confrontés à des problèmes de volume, d'hétérogénéité et de qualité des données... Comment relever le défi ?

Au sortir d'un puits, le pétrole est un mélange complexe de nombreux hydrocarbures, d'autres composés et d'impuretés. Pour être utile, ce cocktail doit être raffiné de façon à séparer les produits exploitables: bitume, fioul, kérosène, gazole, essence... De même, les données réunies sous le terme de *big data* subissent un long processus de raffinage, de sélection et de transport avant de pouvoir livrer une quelconque information, c'est-à-dire d'être interrogées ou analysées. En quoi consistent ces traitements?

Lorsqu'on évoque les *big data*, on pense aussitôt à l'analyse de données plus qu'à leur exploration par des requêtes, comme si le volume croissant des données, en atteignant un certain niveau, rendait caduque leur interrogation directe. En réalité, ce sont deux approches complémentaires, toutes deux contraintes par les limites technologiques dues au traitement massif des données. Qui plus est, l'analyse au sens large nécessite souvent une exploration et un prétraitement individualisé des données. La chaîne de valorisation des *big data* est longue et semée d'embûches, la partie analyse ne représentant qu'un petit maillon de l'ensemble en bout de chaîne.

Les algorithmes sous-jacents à la collecte des données, leur indexation, leur nettoyage, leur stockage, leur annotation... peu visibles de l'extérieur, constituent pourtant l'ossature du pipeline >

Les données disparates doivent rentrer dans le rang pour être exploitables. C'est le rôle des algorithmes.



> qui alimente les applications de décision ou d'apprentissage. Qu'il s'agisse d'algorithmes simples ou complexes, déterministes ou prédictifs, opérant sur des données exactes ou bruitées, leur compréhension est indispensable pour entrevoir en quoi consiste l'exploitation des *big data*.

Occultés par les succès récents de l'apprentissage machine et autres programmes de jeux, ces algorithmes enfouis sont donnés pour acquis, avec un coût nul pour leur mise en œuvre et leur exécution. Nous verrons que ce n'est pas le cas, un travail permanent de recherche et d'ingénierie est nécessaire pour les adapter aux nouveaux contextes applicatifs, les étendre, les optimiser et en inventer de nouveaux.

Il y a autant de processus de gestion et de valorisation des données que de raisons de valoriser ces données. On y retrouve cependant de grandes activités génériques (*voir l'encadré page ci-contre*), sur chacune desquelles existent des verrous importants.

Pour réaliser ces grandes activités, il est important de spécifier les tâches constitutives en fonction des applications visées. La consultation intensive des données n'a pas les mêmes besoins qu'un traitement des flux de données en temps réel. Une application d'aide à la décision n'a pas les mêmes besoins qu'une application de prédiction du panier d'un consommateur ou d'aide au diagnostic d'une tumeur. La nature de l'application et des tâches algorithmiques qui en découlent doivent être mises en perspective avec le modèle de calcul le plus adapté.

DEUX MODÈLES DE CALCUL

La gestion des données n'est pas constituée d'algorithmes *ad hoc* réalisés au coup par coup, dans un langage donné, en fonction des applications. C'est une science qui a ses fondements théoriques, ses propres objets d'étude, ses propres abstractions et ses propres méthodes. Les technologies produites ont pour vocation de réduire l'effort nécessaire à la gestion des données, de contrôler l'accès à ces données et de faciliter l'évolution des systèmes sans réécriture de leur code. Pour simplifier, il y a deux modèles de calcul sur les grands volumes de données: celui dédié à la recherche d'information et celui adapté à l'analyse de données. Même si les deux approches partagent de nombreux concepts et techniques et peuvent se combiner, elles diffèrent d'un point de vue sémantique.

Dans le premier cas, le sens est porté par la requête de l'utilisateur; si la requête est pertinente, les données délivrées seront bien conformes aux critères de recherche spécifiés. Par exemple, «rechercher les références de produits vendus en quantité supérieure à 200 dans chacun des magasins de la région parisienne durant la période de Noël», est une requête complexe, mais qui définit clairement le calcul à faire et les critères de sélection de données. Ainsi spécifié, le système de gestion de données se chargera de générer le code nécessaire à l'exécution de cette requête. D'autres requêtes sont plus approximatives, leurs critères étant spécifiés de façon lâche. C'est le cas des recherches sur le Web. Ainsi, «rechercher les magasins parisiens pratiquant les prix les moins chers pour le chocolat et le champagne durant la période de Noël» est une requête floue. Elle impose un traitement approché du critère «moins cher» obligeant à présenter les résultats dans un certain ordre.

Dans le second cas, la question du sens est plus complexe, car elle interroge à la fois la méthode utilisée pour l'analyse des données et l'interprétation par l'utilisateur des résultats produits. Pour une requête d'analyse de l'évolution des cours de bourse des entreprises du CAC40 ou de prédiction des ventes de smartphones par région et par période de l'année, les résultats ne sont pas interprétables sans connaissance du métier et des modèles de prédiction sous-jacents.

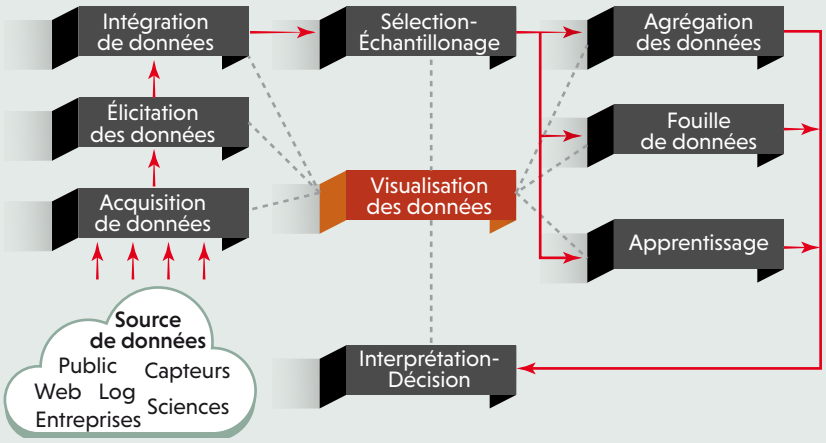
Ce problème peut devenir plus aigu dans des algorithmes encore moins transparents. Par exemple, dans un système de reconnaissance fondé sur l'apprentissage, la confusion entre un chat et un caniche fait sourire, mais celle d'un honnête citoyen avec un criminel peut être dramatique. Les algorithmes sous-jacents à ces systèmes posent des défis à la société.

Il y a autant de processus de valorisation des données que de raisons de valoriser ces données

D'un point de vue conceptuel, un modèle de calcul est défini par les structures de données qu'il reconnaît et les opérations qu'il offre sur ces structures de données. Les structures de données peuvent être des tables (ou matrices), des arbres, des graphes ou des séquences finies ou infinies. Les opérations sont associées à chaque structure: les opérations de tri, projection, produit et union sur

LES MILLE ET UNE VIES DES DONNÉES

Le traitement des *big data* implique des activités que l'on retrouve quelle que soit l'application. Passons-les en revue. Le stockage pose la question du choix des données brutes à enregistrer, de leur acquisition et des données à résumer, ainsi que les méthodes d'indexation pour accélérer leur recherche ultérieure. La validation des données, c'est-à-dire leur nettoyage et leur annotation (on parle d'élicitation), nécessite souvent une interaction avec un expert, ce qui est hors de portée d'un humain si le volume de données est considérable. L'intégration de données est une phase très complexe car elle concerne des sources de données distribuées à grande échelle où l'hétérogénéité est très élevée. Le modèle d'intégration dépendra des objectifs assignés au système : la recherche d'information, la prise de décision, l'extraction de connaissances… C'est là que se situent les principaux verrous liés aux performances des systèmes et à la pertinence des données produites. La sélection et l'échantillonnage constituent des vues sur les données intégrées, c'est-à-dire une façon de réduire les données aux sous-ensembles nécessaires à un type de requête ou un type d'analyse. Un mauvais choix de ce sous-ensemble hypothèque la qualité de la tâche ultérieure. Quant à l'analyse elle-même, elle peut être réalisée par différentes méthodes : l'agrégation pour générer des indicateurs statistiques (somme, moyenne, tendance…), la fouille de données pour identifier des règles d'association ou des motifs fréquents, l'apprentissage automatique (profond ou par renforcement) pour définir un modèle de prédiction permettant de décider ultérieurement si un nouvel objet peut être classé de façon pertinente. Enfin, la visualisation des données est utile à toutes les étapes. Elle offre souvent une vue synthétique des grandes données et aide à leur compréhension.



Le cycle de vie des données au terme duquel elles peuvent livrer des informations.

les tables ; la recherche d'éléments ou le parcours de chemin sur un arbre ; l'appariement ou l'extraction de sous-graphes ; une copie instantanée d'une séquence. Ces structures et ces opérations ont été étudiées depuis des décennies pour formaliser leurs sémantiques, expliciter leurs propriétés et trouver les meilleurs programmes qui leur correspondent dans différents environnements.

Plusieurs facteurs interviennent dans leur mise en œuvre : la hiérarchie et la taille des mémoires (mémoire principale, mémoire cache, disque…), la localisation des données (un site centralisé, de multiples serveurs…), les ressources de calcul mobilisables (une ou plusieurs machines…) et les caractéristiques techniques de ces ressources (puissance de calcul, partage de mémoire…).

DES TABLEAUX MALLÉABLES

L'algèbre relationnelle a représenté un saut technologique et qualitatif dans la gestion des données. En quoi consiste-t-elle ? Rappelons d'abord qu'une table relationnelle est décrite par un ensemble de variables (ou attributs) désignant les colonnes et un ensemble de valeurs à ces variables organisées en lignes, chacune représentant un objet du réel. En d'autres termes, il s'agit de tableaux à deux dimensions, on parle aussi de tableau à double entrée. Or avec cinq opérateurs de base (*voir la figure page suivante*), on sait aujourd'hui transformer ces tableaux et en dériver d'autres.

C'est d'une simplicité confondante, et pourtant, la combinaison de ces opérations permet d'exprimer une large palette de requêtes, sans programmer une seule ligne de code, puisque le code correspondant à chacune des opérations est défini une fois pour toutes et pour toutes les applications. Cette façon de représenter le réel par un ensemble de tableaux indépendants et d'exprimer une requête sur ces tableaux au moyen d'une expression algébrique a notablement modifié notre rapport à la programmation.

Parmi ces opérateurs, certains sont plus sensibles que d'autres au volume des données. Ainsi la jointure de deux tables comportant des milliards de lignes pose de réels problèmes de calcul pour comparer ces milliards de lignes. De nombreux travaux de recherche proposent des algorithmes de jointure tenant compte ou non du tri des données, de la taille de la mémoire centrale, de l'indexation des données, de la distribution des données sur différents disques, du parallélisme des machines et du débit sur les réseaux.

Ce coût dépend aussi du choix initial de la représentation d'une table selon ses lignes ou ses colonnes. Par exemple, pour les requêtes >

PROJECTION

MAGASIN	PRODUIT	PRIX	DATE	SITE
Auchan	Iphone	700	12/9/2017	Paris Est
Auchan	Iphone	760	12/9/2017	Paris Est
Leclerc	IPad	480	27/9/2017	Essonne
Fnac	IPad	870	30/9/2017	La Défense
Carrefour	PC Dell	600	24/9/2017	Paris Ouest
Leclerc	Galaxy	400	28/9/2017	Essonne
Boulanger	Galaxy	310	4/10/2017	Plaisir
Carrefour	Galaxy	470	4/11/2017	St. Quentin
Leclerc	MacBookAir	970	7/10/2017	Yvelines
Auchan	MacBookAir	1100	7/10/2017	Yvelines
Darty	PC Dell	860	19/10/2017	Velizy
Darty	PC Dell	640	23/10/2017	Parly
Darty	iPad	480	27/10/2017	Parly

RESTRICTION

MAGASIN	PRODUIT	PRIX	DATE	SITE
Auchan	Iphone	700	12/9/2017	Paris Est
Auchan	Iphone	760	23/9/2017	Paris Est
Carrefour	PC Dell	600	24/9/2017	Paris Ouest
Leclerc	IPad	400	27/9/2017	Essonne
Leclerc	Galaxy	400	28/9/2017	Essonne
Fnac	IPad	870	30/9/2017	La Défense
Fnac	iMac	1220	1/10/2017	La Défense
Fnac	iMac	2100	3/10/2017	Montparnasse
Boulanger	Galaxy	310	4/10/2017	Plaisir
...	...	...	...	...
...	...	...	...	...

PRODUIT	MARQUE	BREVET
Iphone	Apple	AP2056XY
PC Dell	Dell	DL3467GH
Galaxy	Samsung	SM734DG
IPad	Apple	AP205VY

JOINTURE

MAGASIN	PRODUIT	PRIX	DATE	SITE	MARQUE	BREVET
Auchan	Iphone	700	12/9/2017	Paris Est	Apple	AP2056XY
Auchan	Iphone	760	23/9/2017	Paris Est	Apple	AP2056XY
Carrefour	PC Dell	600	24/9/2017	Paris Ouest	Dell	DL3467GH
Leclerc	IPad	400	27/9/2017	Essonne	Apple	AP205VY
Leclerc	Galaxy	400	28/9/2017	Essonne	Samsung	SM734DG
Fnac	IPad	870	30/9/2017	La Défense	Apple	AP205VY
Boulanger	Galaxy	310	4/10/2017	Plaisir	Samsung	SM734DG
...	...	...	...	...		
...	...	...	...	...		

L'algèbre relationnelle offre cinq opérateurs de base grâce auxquels on peut exprimer un grand nombre de requêtes sur des données. À partir d'une table relationnelle *(le tableau bleu)*, la projection réduit le nombre de colonnes, la restriction celui des lignes, l'union fusionne deux tables de même structure, l'intersection calcule les lignes communes à deux tables de même structure, la jointure apparie deux tables sur un ou plusieurs critères.

> d'exploration de données, les tables seront rangées par lignes alors que pour les requêtes d'analyse agrégeant les valeurs, la représentation par colonne est plus adaptée.

Dans le même élan de définition de ce modèle élégant, des extensions ont été apportées pour traiter des objets plus complexes que les tables relationnelles, notamment les tables imbriquées et les tables à plus de deux dimensions. L'algèbre relationnelle a également été étendue à d'autres domaines spécialisés grâce à l'introduction, par exemple, d'opérations géométriques ou d'autres sur des données temporelles. Le champ d'investigation du modèle relationnel est loin d'être couvert !

DES ARBRES ET DES GRAPHES

Avec le Web, les bases de données ont évolué pour intégrer des données XML dont la structure est formalisée par des arbres. Une algèbre est aussi associée à ces arbres pour effectuer des recherches d'éléments et pour transformer leurs structures. La sélection et la projection permettent d'extraire des sous-arbres vérifiant certaines conditions portant soit sur les données (les valeurs d'un élément), soit sur les métadonnées (les éléments descriptifs de ces valeurs). Il est possible également de restructurer un arbre, ou de l'aplatir, pour obtenir une organisation différente des données et les visualiser sous des angles variés.

Enfin, l'appariement permet de comparer deux structures, une opération très utile pour de nombreuses applications, comme la bioinformatique, l'analyse d'images, les bases de documents... On associe à ce type d'opération une mesure de similarité qui peut être le nombre de transformations à effectuer pour obtenir un arbre à partir d'un autre. Plus la distance est petite, plus les arbres sont similaires. Cette notion est cruciale dans le contexte des *big data*. En effet, le calcul de ces distances est un problème difficile, dont la complexité augmente fortement avec le volume des données et avec la diversité structurelle des graphes à comparer. Les données dont nous avons parlé sont durables, mais qu'en est-il des flux de données, qui disparaissent juste après leur observation ?

LES FLUX DE DONNÉES

La logique de gestion est ici bien différente. Cette fois, ce sont les requêtes qui sont persistantes. Elles sont définies une fois pour toutes et agissent comme des filtres, déterminant les données observées qu'on souhaite conserver. Le premier opérateur indispensable sur un flux définit la taille de la fenêtre d'observation et un pas de glissement pour suivre le flux. Ces paramètres peuvent être définis par des durées, par le nombre d'événements attendus ou la combinaison des deux. Les données observées peuvent être stockées temporairement dans des tables relationnelles et filtrées par des requêtes définies en amont.

Pour faciliter l'interprétation des flux, il est souvent nécessaire de les corréler avec des données de référence persistantes dans des catalogues, ce qui impose des opérations de jointures potentiellement coûteuses si ces catalogues sont gigantesques. Par exemple, l'observation d'une température de −110 °C dans un bâtiment est sans doute liée à la défaillance d'un capteur. En revanche, une forte température trahit sans

doute l'explosion d'une chaudière située dans ce bâtiment. Cette interprétation n'est possible qu'avec un catalogue des immeubles avec leurs caractéristiques. Avec des millions de capteurs, le passage à l'échelle du traitement des flux envoyés est un défi. Le déploiement des compteurs Linky est un exemple.

LES LIMITES DES ALGÈBRES

Malgré tous ses avantages, les modèles de calcul «à la relationnelle» ne couvrent pas les besoins de l'ensemble des applications, et notamment ceux de deux grands domaines. Dans le premier, on trouve des applications dont les données sont conceptuellement relationnelles, mais dont les performances et certaines contraintes (telle la représentation en ligne des tables) ne sont pas satisfaisantes. C'est le cas des applications nécessitant une représentation des tables par colonnes plus adaptée aux calculs d'agrégats, ou dont la taille des tables, gigantesques, conduit à des performances médiocres de la part des systèmes de gestion de base de données.

Ce domaine est illustré par les applications décisionnelles faisant de gros calculs sur les colonnes et par des applications en astrophysique faisant des calculs sur des tables immenses. Dans les deux cas, même si conceptuellement on manipule des tables, on développe des algorithmes *ad hoc* opérant sur des données massivement distribuées et exploitant des architectures parallèles et des *clusters* de machines.

MapReduce permet une parallélisation des tâches en distribuant les données plutôt que le code

Le second domaine d'applications est illustré, au-delà des volumes de données importants, soit par des données non relationnelles (documents, images…), donc sans connaissance préalable d'une structure, soit par des logiques de traitement qui ne s'expriment pas aisément avec l'algèbre relationnelle. C'est le cas des algorithmes de classification ou d'apprentissage. Là encore, on a recours à une programmation *ad hoc*.

Notons que dans les deux domaines d'applications, le relationnel n'est pas exclu, mais seulement limité à quelques tâches. Par exemple, dans un projet de *machine learning*, la phase de prétraitement peut s'appuyer sur une technologie relationnelle, alors que la phase d'apprentissage proprement dite sera faite par un algorithme approprié.

Cette approche, notée NoSQL, n'est pas nouvelle, des interfaces entre les langages de requêtes et les langages de programmation ont toujours existé pour développer des applications complexes, sous-traitant aux systèmes de gestion de base de données ce qu'ils peuvent faire de mieux et laissant le programmeur réaliser les autres tâches. On la retrouve également dans les systèmes à base de *workflows* où l'on coordonne un ensemble d'activités pouvant être réalisées par des technologies différentes.

Aujourd'hui, ce mode d'ingénierie est toujours pratiqué, avec un regard plus spécifique sur ces tâches. Il s'agit en général d'opérations complexes dont l'optimisation nécessite la mobilisation de ressources de calcul particulières. Ce type de développement est généralement guidé par des patrons (ou *patterns*) de programmation qui intègrent l'invocation de services de données existant et le code de l'utilisateur.

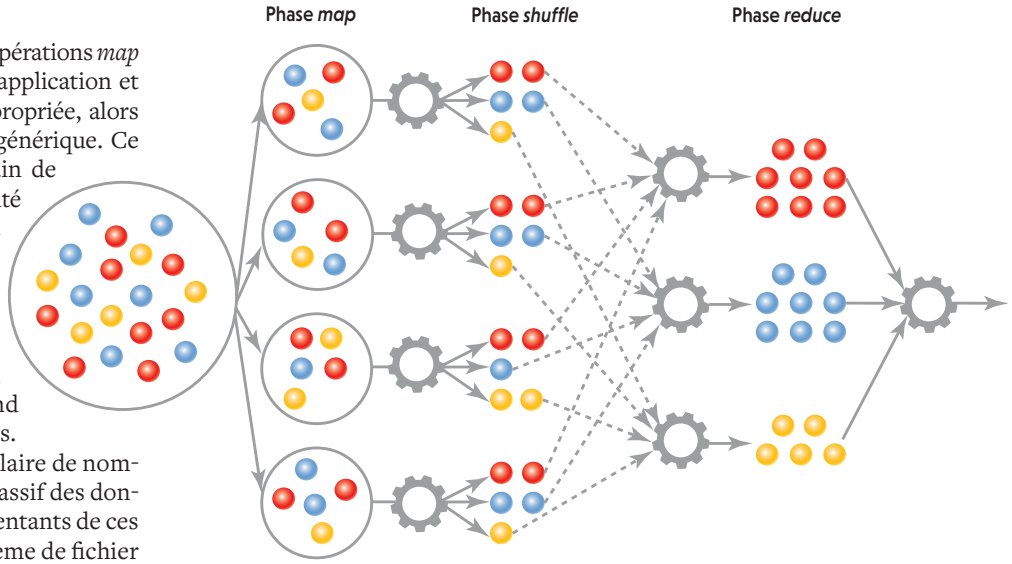
Un de ces *patterns*, connu depuis longtemps en programmation fonctionnelle, est MapReduce (*voir la figure page suivante*). Il permet une parallélisation des tâches à grande échelle en distribuant les données au lieu de distribuer le code. Pour illustrer ce *pattern*, imaginons une opération visant à classifier un sac de billes en fonction de leurs couleurs et à évaluer ensuite la taille de chaque classe et le nombre total de billes. Cette opération peut être réalisée par une seule personne, mais elle sera longue, surtout si le sac de billes est très grand. On recrute alors dix enfants entre qui on partage les billes. En augmentant le parallélisme, on accélère le processus, mais on a à la fin dix classements, chacun produisant des paquets de billes selon leurs couleurs.

Après un travail intermédiaire de regroupement des paquets par couleur, on recrute un nouveau groupe d'enfants et on leur demande de compter le nombre de billes de chaque couleur. À la fin, une seule personne fait la somme des résultats intermédiaires.

Dans ce principe général du modèle de calcul MapReduce, la phase *map* répartit les données (les billes) sur plusieurs serveurs (les enfants), chacun réalisant la même tâche (reconnaissance de la couleur et classement des billes). La phase *reduce* prend les paquets de données intermédiaires et fait la tâche suivante (le comptage). Entre les deux, une «boîte noire» (le *shuffle*) regroupe les paquets par couleur, avant de les acheminer aux serveurs >

> suivants qui feront le *reduce*. Les opérations *map* et *reduce* sont dépendantes de l'application et donc programmées de façon appropriée, alors que le *shuffle* est une opération générique. Ce modèle de calcul permet un gain de temps significatif grâce à sa capacité de mobilisation de ressources à chacune de ses phases d'exécution, mais le prix à payer est au niveau du développement qui nécessite une compétence pointue et au niveau de la mise en œuvre qui dépend d'un grand nombre de paramètres techniques.

Ce modèle est la pierre angulaire de nombreux systèmes de traitement massif des données. Hadoop est l'un des représentants de ces systèmes, combiné avec un système de fichier massivement distribué, de nombreuses extensions ont été réalisées soit au cœur du système pour améliorer ses performances soit au-dessus du système (on parle de surcouches) pour offrir des interfaces de développement plus riches et plus simples.



MapReduce accélère le processus de traitement des données grâce au parallélisme, c'est-à-dire à une distribution des tâches. Ici, elle consiste en une répartition des données pour les classer. D'abord, le paquet initial est divisé en quatre parties *(à gauche)*, chacune étant ensuite classée *(au centre)*. Les résultats partiels sont enfin réunis en un seul *(à droite)*.

AU-DELÀ DU CALCUL

Si les opérations de calcul sur les données sont importantes et cristallisent à juste titre l'attention des spécialistes et des programmeurs, la gestion des données ne se limite pas à ces opérations. L'indexation des données et les techniques de hachage, c'est-à-dire la répartition des données sur divers disques, sont cruciales pour la gestion des données, quel que soit le modèle de calcul choisi. Indexer des tables comportant des milliards de lignes ou des documents de plusieurs centaines de millions de pages est une tâche coûteuse. Les techniques mises en œuvre rivalisent d'ingéniosité à la fois dans la façon de construire l'index pour accélérer les accès aux données et dans les performances de cette construction.

Outre l'indexation et le hachage, la réplication des données est aussi un bon moyen d'augmenter la disponibilité de l'information en multipliant les exemplaires et les points d'accès. La réplication est très importante dans un système distribué afin de réduire les transferts de données, souvent coûteux, sur le réseau. Cependant, les avantages de l'indexation et de la réplication sont vite anéantis quand les données sont fréquemment mises à jour, ce qui n'est heureusement pas le cas dans de nombreuses applications des *big data*.

La protection des données reste un sujet sensible lorsqu'il s'agit de données personnelles, de santé ou concernant des processus industriels. De nombreux systèmes intégrés de gestion de données sont dotés de mécanismes d'anonymisation (*voir L'art de préserver l'anonymat, par T. Allard, page 98*), de cryptage, ou de restriction d'accès à certaines données. Des hiérarchies de droits sont souvent associées à des hiérarchies d'utilisateurs contrôlant ainsi l'accès aux données sensibles. Les algorithmes associés à ces garde-fous sont souvent complexes.

Dans le registre des risques, ajoutons que la complexité des algorithmes mis en œuvre dans l'exploitation des *big data* peut entraîner des effets de bord qui, si l'on n'y prend pas garde, peuvent poser des problèmes d'éthique, violer la vie privée, réduire la liberté et avoir un impact fortement intrusif dans la vie des personnes. Ce sont des questions essentielles mises en débat récemment (*voir l'entretien avec N. Boujemaa, page 104*).

La sensibilité des données n'est pas seulement liée à leur usage, elle peut être liée à la technologie. Les erreurs de matériel ou logicielles ne sont pas rares. Ces incidents ne doivent pas mettre le système de données dans un état incohérent, avec des risques de perte d'informations. Des mécanismes de fiabilisation sont développés, notamment au niveau des systèmes de gestion de base de données pour protéger ces systèmes et permettre un redémarrage à chaud ou à froid, sans perte d'informations et sans incohérence sur les données globales. Ces algorithmes de protection des données à tous les niveaux occupent une place stratégique. Ils sont complexes et leur mise en œuvre n'est pas sans incidence sur les coûts de calcul.

En fin de compte, les algorithmes destinés à la gestion des *big data* constituent une vaste faune, toujours plus diversifiée et riche à mesure des développements. C'est que, pour filer la métaphore, leur monde, celui des *big data*, ne cesse de croître. D'ailleurs, la comparaison avec le pétrole trouve ici ses limites. Les ressources en cette énergie fossile sont inéluctablement appelées à se tarir, alors que celles en données sont en augmentation permanente. Et une donnée peut être utilisée autant de fois que nécessaire sans se consumer. ■

BIBLIOGRAPHIE

M. BOUZEGHOUB ET R. MOSSERI (DIR.), ***Les Big Data à découvert,*** **CNRS Éditions, 2017.**

S. ABITEBOUL, «Sciences des données: de la logique du premier ordre à la Toile», Leçon inaugurale de la Chaire d'Informatique et sciences numériques au Collège de France, 2012.

ABONNEZ-VOUS À POUR LA SCIENCE

NOUVELLE FORMULE

FORMULE PASSION -27%

79€

FORMULE DÉCOUVERTE -25%

59€

FORMULE INTÉGRALE -41%

99€

BULLETIN D'ABONNEMENT

À renvoyer accompagné de votre règlement à : Pour la Science - Service abonnements - 19 rue de l'industrie - BP 90 053 - 67 402 Illkirch cedex

❑ OUI, JE M'ABONNE À *POUR LA SCIENCE* POUR 1 AN, JE CHOISIS MA FORMULE CI-DESSOUS :

❑ FORMULE **INTÉGRALE**
12 n° de *Pour la Science* + 4 Hors-Séries
+ Accès illimité aux archives en ligne depuis 1996
I1A99E

99€ Au lieu de ~~168,45€~~*

❑ FORMULE **PASSION**
12 n° de *Pour la Science*
+ 4 Hors-Séries
P1A79E

79€ Au lieu de ~~108,45€~~*

❑ FORMULE **DÉCOUVERTE**
12 n° de *Pour la Science*
D1A59E

59€ Au lieu de ~~78,45€~~*

J'INDIQUE MES COORDONNÉES

❑ M. ❑ Mme

Nom : _____ Prénom : _____

Adresse : _____

_____

Code postal |_|_|_|_|_| Ville : _____

Téléphone |_|_|_|_|_|_|_|_|_|_|

• Mon e-mail : (à remplir en majuscule)

_____@_____

J'accepte de recevoir les informations de *Pour la Science* ❑ OUI ❑ NON

et de ses partenaires ❑ OUI ❑ NON

• Je choisis mon mode de règlement :

❑ Par chèque à l'ordre de *Pour la Science*

❑ Par carte bancaire

N° |_|_|_|_| |_|_|_|_| |_|_|_|_| |_|_|_|_|

Date d'expiration : |_|_| |_|_| Clé (Les 3 chiffres au dos de votre CB) : |_|_|_|

Signature obligatoire

PAG18098STDC

* Par rapport au prix de vente en kiosque, l'accès numérique aux archives. Délai de livraison : dans le mois suivant l'enregistrement de votre règlement. Offre réservée aux nouveaux abonnés, valable jusqu'au 30/06/2018 en France métropolitaine uniquement. Pour un abonnement à l'étranger, merci de consulter notre site www.pourlascience.fr. Conformément à la loi "Informatique et libertés" du 6 janvier 1978, vous disposez d'un droit d'accès et de rectification aux données vous concernant en adressant un courrier à *Pour la Science*. Photos non contractuelles.

Groupe Pour la Science - Siège social : 170 bis, boulevard du Montparnasse, CS20012, 75680 Paris cedex 14 - Sarl au capital de 32 000€ - RCS Paris B 311 797 393 - Siret : 311 797 393 000 23 - APE 58.14 Z

CALCULER plus vite,

© Shutterstock.com/chombosan

plus haut, plus fort

Les besoins en calcul et en analyse des données explosent dans notre société du *big data*, par exemple pour le déploiement des voitures autonomes. Matériel, logiciels et algorithmes seront-ils à la hauteur des attentes ?

L'ESSENTIEL

- **L'essor du *big data* repose sur deux piliers, le calcul à haute performance (HPC) et l'analyse de données à haute performance (HPDA).**
- **Ces deux aspects se rapprochent dans leurs contraintes techniques, théoriques et matérielles, notamment au travers de l'intelligence artificielle.**
- **Les progrès récents concernent le développement de nouveaux processeurs, d'algorithmes adaptés, de langages de programmation inédits…**
- **La physique quantique et les neurosciences sont aussi des sources d'inspiration pour l'innovation dans ces domaines.**

L'AUTEUR

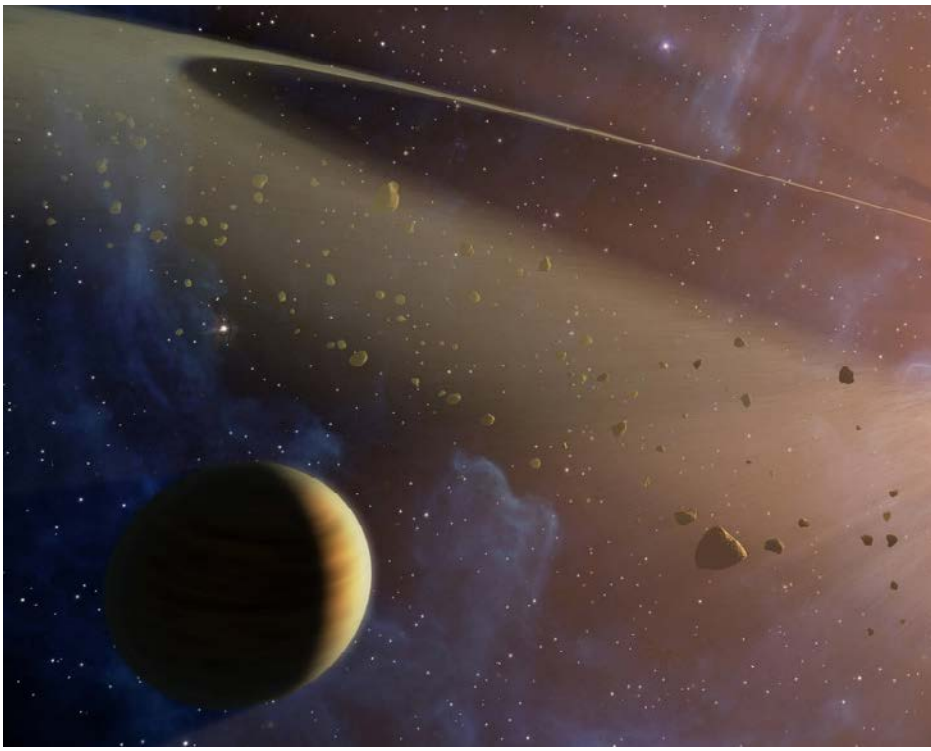
JEAN-LAURENT PHILIPPE est spécialiste du HPC, chez Intel.

Le 14 avril 2003, après bien des péripéties, notamment une course entre un consortium public et une entreprise privée, le séquençage complet de l'ADN du génome humain était annoncé. Un exploit faramineux qui avait coûté la modique somme de 100 millions de dollars et nécessité plusieurs années de travail intensif. Aujourd'hui, cet acte est pratiqué en routine, en une journée et pour seulement une centaine de dollars. L'idée de médecine personnalisée est désormais à notre portée.

On doit ces progrès fulgurants à la formidable augmentation des puissances de calcul, aux progrès dans la conception d'algorithmes et aux avancées dans les capacités de traitement des gros volumes de données. En 2018, un séquenceur de génome produira quotidiennement 3 à 4 téraoctets (To) de données par jour. Cela tient en deux sigles, HPC et HPDA, les deux piliers du *big data*.

Le premier, le calcul à haute performance (*High Performance Computing*), correspond à la science des supercalculateurs, c'est-à-dire des ordinateurs. Ainsi, les machines actuelles exécutent de nombreuses opérations en parallèle, de façon à diminuer le temps de calcul. Ceux installés en France dépassent la puissance de 1 pétaflops, soit 1 million de milliards d'opérations par seconde, ou 10^{15} flops (*floating point operations per second*, soit opérations avec des décimales par seconde). Par comparaison, un Mac Pro récent de chez Apple atteint 7×10^{12} flops. Le second, le HPDA (*High Performance Data Analytics*, ou analyse de données à haute performance), concerne la science du traitement et de l'analyse des gros volumes de données, voire très gros volumes : les 150 millions de capteurs du LHC fournissent des zettaoctets (10^{21}) de données par an dans lesquels on doit extraire des informations pertinentes !

Ces chiffres donnent le vertige, mais ils ne doivent pas cacher que les besoins en ces deux domaines sont immenses. Par exemple, les capacités en HPC actuelles sont encore insuffisantes pour soulager les riverains des aéroports gênés par le bruit des avions au décollage. En effet, pour diminuer l'empreinte acoustique d'un avion dans sa totalité, les concepteurs doivent modéliser la



© Nasa

contribution des réacteurs, mais aussi celle des mouvements d'air turbulents autour de l'appareil. Or pour une simulation adaptée, il faudrait compter environ une semaine de calcul sur une machine de quelque 500 000 processeurs dédiés. Pure théorie, car le plus gros calculateur installé en France, la machine Pangea chez Total Exploration Production, n'a que 18 400 processeurs de ce type. Sur cette machine, la simulation d'un avion requerrait des mois voire des années de calcul. Impensable !

La future voiture autonome illustre quant à elle les besoins en traitement de gros volumes de données (HPDA). Toutes les 90 minutes, un tel véhicule générera environ 4 To de données par jour, venues des radars, sonars, GPS, lidars, caméras embarquées... En déplacement dans un environnement complexe, la voiture devra traiter les données en temps réel, et prendre des décisions immédiates afin d'éviter des accidents. Il lui faudra donc aussi une capacité de calcul importante à bord.

VERS LA CONVERGENCE

Ces exemples montrent que les attentes en capacités de calcul sont importantes, surtout dans une société où le *big data* sera un carburant central (*voir l'entretien avec D. Cohen, page 8*). Les efforts de recherche pour les améliorer sont à la hauteur des attentes, et l'on pourrait même assister à une convergence entre HPC et HPDA. Pour quelles raisons ?

Les gros centres de calcul, autrefois dédiés au HPC et équipés pour effectuer des opérations très nombreuses sur des données souvent peu nombreuses, ont cependant des capacités de stockage très importantes. Or le HPDA est justement le traitement de gros volumes de données. De plus, outre l'infrastructure de stockage des données, les centres de calcul ont également celle nécessaire à leur traitement, même si parfois les capacités requises pour les gros volumes de données du HPDA sont légèrement différentes des capacités de traitement pour le HPC.

Plus précisément, un centre de calcul HPC dispose d'un espace de stockage rapide des données, de nombreux nœuds de calcul, d'un système d'interconnexion rapide entre ces nœuds ainsi qu'avec le système de stockage. Les nœuds sont par ailleurs équipés pour faire du calcul décimal très efficacement (pour la résolution d'équations mathématiques bien souvent).

Puisque les infrastructures pour le HPC sont en place, il est tentant de leur faire exécuter des programmes de traitement de type HPDA, puisque tous les ingrédients existent et permettent d'effectuer les traitements HPDA. Il suffirait de les adapter en augmentant l'espace de stockage des données et en améliorant l'efficacité des interconnexions entre les nœuds de calcul et avec le stockage. En fin de compte, la convergence HPC-HPDA permettrait de mieux utiliser les ressources matérielles existantes.

On observe également un rapprochement des communautés HPC et intelligence artificielle, notée IA, car elles exécutent les mêmes types d'applications intensives en données ou en calcul que le HPDA, que ce soit sur de très gros supercalculateurs, sur des petits groupes institutionnels de machines (des *clusters*), ou dans le nuage (on parle alors de *cloud computing*).

À quoi ressemble l'atmosphère des exoplanètes ? RobERt vous répond. Il a appris à les sonder grâce au *deep learning*.

L'ENVOL DE L'IA

La révolution de l'IA, par laquelle les ordinateurs arrivent à créer des modèles prédictifs à partir de données (*voir La révolution de l'apprentissage profond, par Y. Bengio, page 42*), conduit à l'adoption rapide de cette technologie HPC qui a également rendu possibles de nombreux développements scientifiques, algorithmiques, logiciels et matériels.

Le fait que la communauté IA utilise l'infrastructure HPC est un élément important dans la convergence HPC-HPDA, car le matériel, très cher, est déjà en place et suffisant pour du traitement HPDA. Les *data scientists* n'ont donc pas à investir dans du matériel, tandis que les utilisateurs du HPC peuvent maximiser l'utilisation de leurs ressources en les mettant à la disposition d'autres applications.

On peut en conséquence s'attendre à un essor de l'IA : les machines nous aideront de plus en plus à prendre des décisions complexes. Pour ce faire, des jeux de données de plus grande taille et plus complexes doivent être utilisés lors de la phase d'apprentissage de l'IA. L'IA sera présente dans les voitures autonomes, >

> mais aussi en science par exemple pour identifier des cellules cancéreuses sur des images biomédicales, trouver des événements clés en physique des hautes énergies, repérer des événements biomédicaux rares…

Afin que les scientifiques puissent extraire de la richesse des volumes énormes de données mesurées, modélisées ou simulées, l'évolutivité est la clé. Cette évolutivité consiste à répartir les calculs sur un grand nombre de processeurs (ou nœuds de calcul), et donc d'effectuer les calculs plus rapidement ou bien de traiter des volumes de données plus importants. Les programmes doivent alors être capables de tirer parti d'autant de processeurs que possible.

MONTE-CARLO, PATRIE DU HASARD

Par exemple, dans la communauté HPC, les scientifiques ont tendance à se concentrer sur la modélisation et la simulation avec des techniques telles que Monte-Carlo afin de produire des représentations précises de ce qui se passe dans la nature. Rappelons que les méthodes Monte-Carlo sont une famille de procédés algorithmiques visant à calculer une valeur numérique approchée en utilisant des techniques probabilistes. L'idée essentielle est d'utiliser le hasard pour résoudre des problèmes qui seraient déterministes en principe. Ils sont souvent utilisés dans des problèmes physiques ou mathématiques et sont plus utiles lorsqu'il est difficile voire impossible d'utiliser d'autres approches.

Les méthodes de Monte-Carlo sont intéressantes, car elles peuvent être parallélisées, et sont donc théoriquement rapides (chaque élément fourni par le processus probabiliste étant indépendant, les calculs peuvent se faire sur autant de processeurs que disponibles). Cependant, elles ne convergent que si la taille de l'échantillonnage est suffisamment grande. Chaque processeur doit donc traiter séquentiellement plusieurs éléments, augmentant d'autant le temps de calcul total.

En pratique, les physiciens des hautes énergies et les astrophysiciens examinent des images à identifier grâce à des techniques d'apprentissage profond (le *deep learning*), c'est-à-dire en les comparant à celles d'une base de données importante. Pour ce faire, les images de référence ont été étiquetées avec des informations les décrivant, puis livrées à l'ordinateur pour qu'il apprenne à « voir » ce qu'elles représentent. Ainsi éduquée, la machine peut reconnaître ce que montrent des images inconnues (on parle d'inférence).

Toujours en astrophysique, le programme RobERt (Robotic Exoplanet Recognition) traque la composition de l'atmosphère des exoplanètes. Mis au point à l'University College de Londres, il fonctionne grâce à une phase d'apprentissage à partir de larges bases de données non structurées recueillies par les instruments de mesure (*voir la figure page précédente*).

Dans d'autres domaines, l'apprentissage consiste à examiner des séries, faire du traitement du signal, des analyses harmoniques, de la modélisation et de la prédiction avec des systèmes non linéaires…

De même, l'IA tente de générer des représentations précises de ce qui se passe dans la nature à partir de données non structurées recueillies par des capteurs, des caméras, des microscopes électroniques, des séquenceurs… La nature inconnue de ces données non structurées conduit les *data scientists* à passer une grande partie de leur temps à extraire et nettoyer les informations utiles pour entraîner le système IA.

Les programmes écrits en langage productif, tel Julia, parviennent rarement à tirer parti du supercalculateur

La convergence HPC-IA concerne aussi bien les algorithmes qui doivent s'adapter à la taille des données et à celle de l'ordinateur cible, que la façon dont les calculs sont répartis sur plusieurs processeurs, le matériel, le logiciel, le prétraitement des données. Les langages de programmation à haute productivité sont également un moteur de rapprochement. De tels langages, comme Python et Julia, permettent de développer des programmes rapidement, car ils sont plus intuitifs dans leur écriture. L'objectif est que les utilisateurs puissent implémenter leurs modèles de calcul, indépendamment du volume de données disponible et de la plateforme de calcul cible.

Soulignons que les seules mesures de performance qui comptent pour la phase d'apprentissage du *deep learning* sont le temps nécessaire pour obtenir le modèle, et la précision de ce dernier. On sait depuis des décennies que cette étape d'apprentissage se comporte de façon quasi linéaire en fonction du nombre de nœuds de calcul, c'est-à-dire que le temps de

calcul diminue proportionnellement au nombre de nœuds de calcul.

L'évolutivité des applications d'apprentissage pour le *deep learning* sur des architectures distribuées est bien plus limitée. L'une des raisons est l'indispensable communication des résultats intermédiaires entre les différents processeurs de l'architecture distribuée, ce qui est très coûteux en temps. Ainsi, on a utilisé des mesures de performance liées au matériel, à savoir les sous-systèmes mémoire de stockage des données au plus proche du processeur, les caches qui accélèrent cet accès aux données, et les calculs flottants (portant sur des nombres décimaux, par opposition aux calculs sur des entiers) pour identifier les solutions matérielles plus susceptibles de supporter des performances élevées, et d'atteindre rapidement l'objectif, en l'occurrence fournir le modèle.

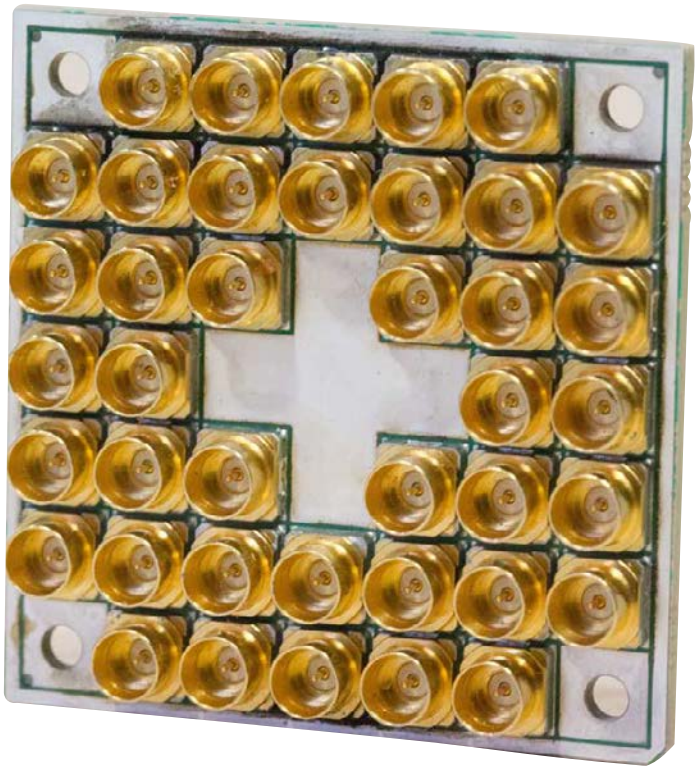
TOUJOURS PLUS DE FLOPS

Le fossé qui nous sépare de la convergence HPC-IA idéale ou au moins attendue par les communautés doit être rapidement comblé. La collaboration avec le monde de la recherche qui a permis d'atteindre une performance de 15 pétaflops avec 9 600 nœuds équipés de processeurs Intel® Xeon® Phi™ sur le supercalculateur Cori du National Energy Research Scientific Computing Center (le NERSC), un organisme dédié au HPC, à Berkeley, aux États-Unis, est un exemple de progrès matériel.

En effet, auparavant, on ne savait pas faire travailler ensemble et efficacement autant de nœuds de calcul : au-delà d'un nombre seuil de nœuds, le temps de calcul total ne diminuait plus. Les supercalculateurs actuels ont quelques dizaines de milliers de processeurs, et ceux qui permettront d'atteindre la prochaine barrière de l'exaflops devront sans doute faire travailler ensemble plus de 100 000 processeurs, soit un ordre de grandeur de plus.

Les principaux efforts concernant les logiciels pour l'IA consistent à aider les *data scientists* à utiliser des outils logiciels familiers fonctionnant aussi bien sur des petits que des très gros systèmes, et sur le matériel actuel aussi bien que sur le matériel du futur.

Un autre défi logiciel important dans la convergence du HPC et de l'IA est l'unification des modèles de programmation. Les programmeurs HPC peuvent être des gourous de la programmation parallèle, c'est-à-dire de la répartition des calculs effectués en même temps par plusieurs processeurs, mais l'IA est principalement programmée avec des outils de type MATLAB, à savoir des outils de représentation dans des langages de haut niveau et un environnement interactif qui permet d'exécuter du calcul intensif plus rapidement qu'avec les langages de programmation traditionnels tels que C, C ++ et Fortran.



© INTEL

Cette puce quantique contient 17 qubits supraconducteurs. Elle a été mise à disposition de l'université technologique de Delft, aux Pays-Bas.

La communauté IA essaie de résoudre le difficile problème d'obtenir des performances élevées sur des architectures évolutives en taille, sans avoir besoin de former les *data scientists* à la programmation parallèle de bas niveau, c'est-à-dire proche du matériel et exploitant explicitement ses caractéristiques : en d'autres termes, les *data scientists* doivent s'abstraire des machines.

Dans cette optique, en collaboration avec des partenaires universitaires, Intel a multiplié par 10 les performances d'Apache Spark (un moteur rapide et général pour le traitement de données à grande échelle), grâce notamment au développement de bibliothèques mathématiques optimisées. De plus, Intel, en collaboration avec Julia Computing et le MIT, a réussi à accélérer de façon significative les programmes écrits en Julia aussi bien au niveau d'un nœud de calcul (monoprocesseur ou biprocesseur) qu'au niveau d'un *cluster* (multiprocesseur), permettant ainsi l'évolutivité tant recherchée par les utilisateurs.

Des outils open source (ParallelAccelerator ou HPAT) transforment des programmes écrits en langages productifs (Julia et Python) en codes qui seront exécutables sur des supercalculateurs, et qui fourniront de très hautes performances. De la sorte, on pallie un inconvénient majeur : le plus souvent, les programmes écrits en langage productif ne parviennent pas à tirer parti du supercalculateur car ils ne prennent pas en compte ses caractéristiques.

D'autres développements matériels sont en cours. Intel a annoncé en octobre 2017 la mise au point de la famille de processeurs Intel® >

> Nervana™ Neural Network Processors (NNP), dont l'architecture est inspirée des réseaux de neurones du cerveau. Une collaboration avec Facebook aidera à développer des méthodes et globalement l'écosystème IA lié à ces dispositifs. De nouveaux jeux d'instruction pour des processeurs à venir ont été annoncés à l'été 2017, ils accéléreront certains calculs du *deep learning*. Côté logiciel, le développement d'outils de productivité donnera accès plus facilement aux puissances de calcul disponibles.

QUANTIQUE ET NEUROMORPHIQUE

L'IA et le HPDA mis en œuvre sur l'infrastructure HPC conduiront à des avancées importantes dans de nombreux domaines grâce à l'extraction d'informations dans de gros volumes de données, par exemple pour le déploiement de la conduite automobile autonome. Plus en amont, d'autres segments de l'informatique émergent, pour l'instant encore cantonnés au stade de la recherche, et qui, associés à l'analyse de données, sont théoriquement prometteurs.

À ce titre, deux secteurs se distinguent : l'ordinateur quantique et les puces neuromorphiques pourraient accélérer une partie des calculs, valider des décisions prises, supprimer des choix possibles… En effet, si les supercalculateurs savent calculer rapidement certains algorithmes, l'ordinateur quantique serait idéal pour les calculs combinatoires. De son côté, la puce neuromorphique peut apprendre seule. En repensant les algorithmes pour tirer avantage de chacune de ces techniques, on réduirait notablement les temps de calcul.

Un ordinateur quantique exploite le phénomène physique de superposition d'états. Alors qu'un bit classique est soit dans l'état 0 soit dans l'état 1, l'équivalent quantique, un qubit, est dans une combinaison des deux états. Exploiter cette propriété accélérerait de façon exponentielle les calculs en parallèle. De la sorte, les ordinateurs quantiques pourraient s'attaquer à des problèmes inaccessibles aux ordinateurs classiques : simuler la nature pour faire progresser la recherche en chimie, en science des matériaux et en modélisation moléculaire ; créer un supraconducteur à température ambiante ; découvrir de nouveaux médicaments.

En novembre 2017, à l'occasion du colloque QuTech sur l'architecture des ordinateurs quantiques qui s'est tenu à Delft, aux Pays-Bas, Intel a annoncé la mise à disposition d'une puce de test supraconductrice de 17 qubits à son partenaire académique de recherche sur le sujet, l'université de technologie de Delft. Cet événement illustre les progrès dans la recherche et le développement d'un système informatique quantique fonctionnel. De fait, une puce de 49 qubits est prévue d'ici la fin de l'année 2018.

Quant aux puces neuromorphiques, Intel a mis au point une puce autodidacte unique en son genre. Baptisée Loihi, elle imite le mécanisme du cerveau en apprenant à fonctionner en analysant la réaction de son environnement. L'informatique neuromorphique s'inspire de l'architecture du cerveau et de ses calculs associés, notamment du fait que les réseaux neuronaux relaient les informations, modulent le poids des interconnexions, et stockent ces changements localement aux interconnexions. Cette puce extrêmement économe en énergie, qui utilise les données pour simultanément apprendre et décider, devient plus intelligente avec le temps et n'a pas besoin d'être entraînée de façon traditionnelle.

Les puces neuromorphiques utilisent les données pour apprendre et décider en même temps

Les avantages des puces d'autoapprentissage sont vertigineux. Par exemple, avec les informations de rythme cardiaque d'un individu dans diverses conditions (après le jogging, après un repas ou avant d'aller au lit), un système neuromorphique en déduirait ce qu'est un rythme cardiaque « normal ». Le système serait alors apte à surveiller en continu des données cardiaques et surtout à identifier un signal « anormal» pouvant trahir une pathologie.

Ce type de logique pourrait également être appliqué à d'autres domaines, comme la cybersécurité où une anomalie ou une différence dans les flux de données pourrait identifier une violation ou un piratage puisque le système a appris le «normal» dans divers contextes.

L'IA n'en est sans doute qu'à ses balbutiements. D'autres architectures et méthodes continueront d'émerger et amélioreront encore les performances. Ainsi, les modèles d'apprentissage automatique ont récemment fait d'énormes progrès, mais ces systèmes ne sont le plus souvent pas capables de faire des généralisations. C'est un des défis à relever. ■

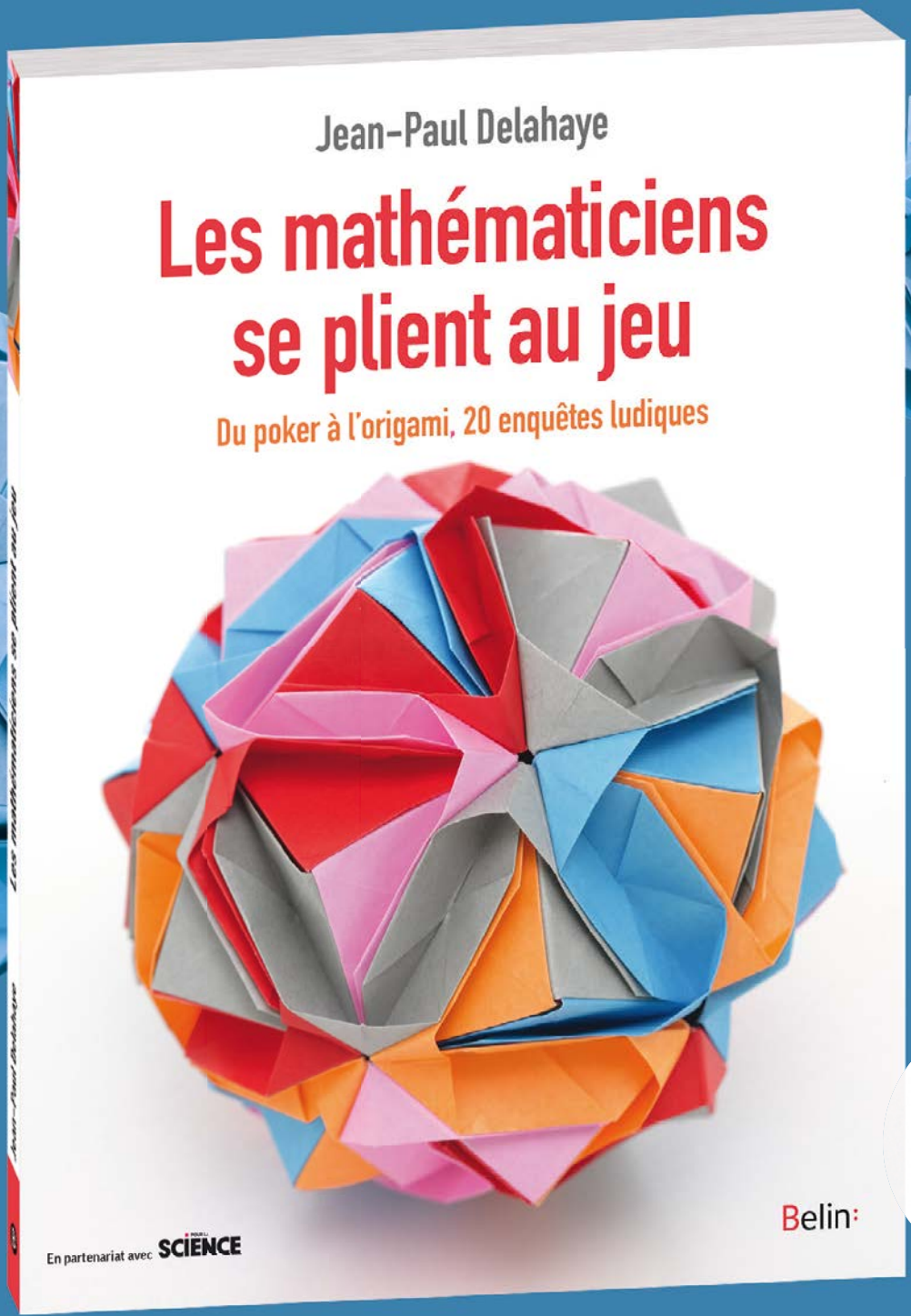
BIBLIOGRAPHIE

M. ANDERSON *ET AL.*, **Bridging the Gap Between HPC and Big Data Frameworks,** ***Proceedings of the VLDB Endowment*, vol. 10, pp. 901-912, 2017.**

J. PARK *ET AL.*, **Faster CNNS with direct sparse convolutions and guided pruning, *ICLR 2017 Conference paper*, 2017.**

T. KURTH *ET AL.*, ***Deep Learning at 15PF : Supervised and Semi-Supervised Classification for Scioentific Data*** **Cornell University Library, 2017.**

Le nouveau livre de Delahaye
sous le signe du jeu



Disponible en librairie

Belin ÉDITEUR

Suivez-nous et abonnez-vous à notre newsletter sur www.belin-editeur.com

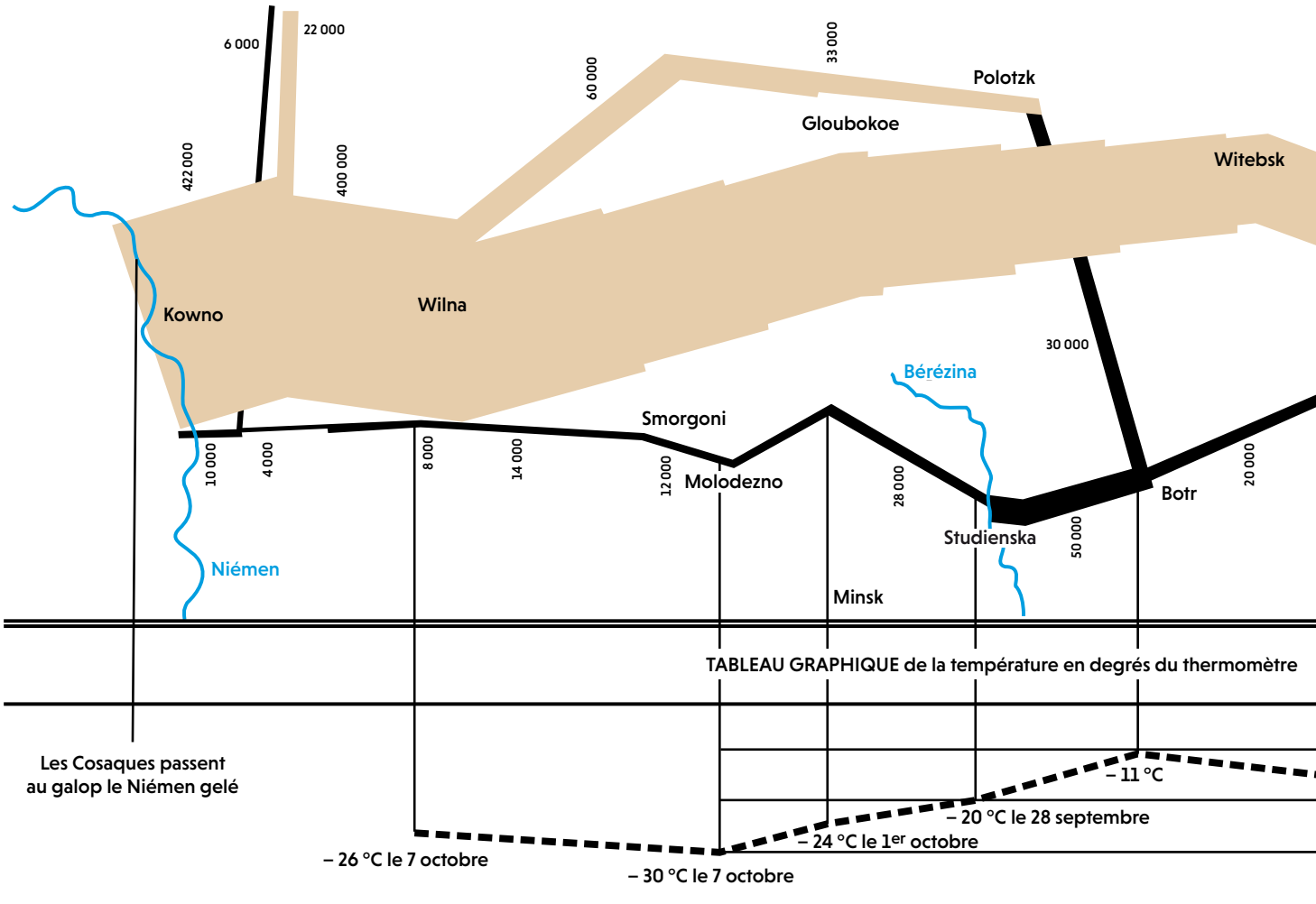
/EditionsBelin • @editions_belin

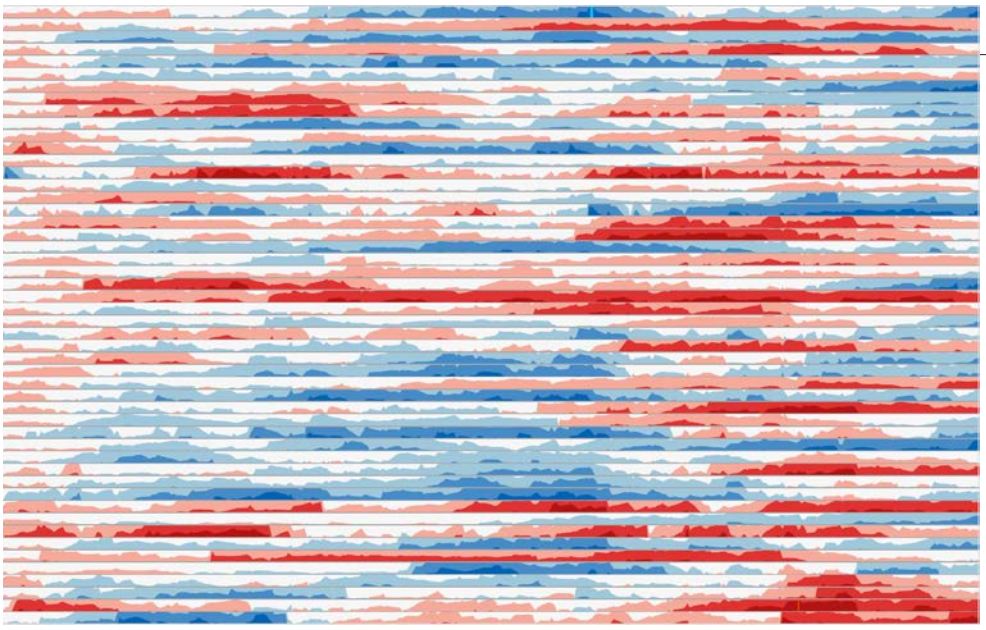
JE VOIS, donc je comprends

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.

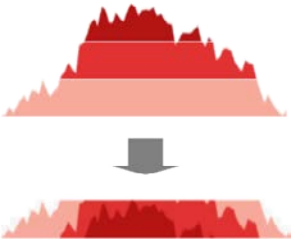




© S. Few, Perceptual Edge

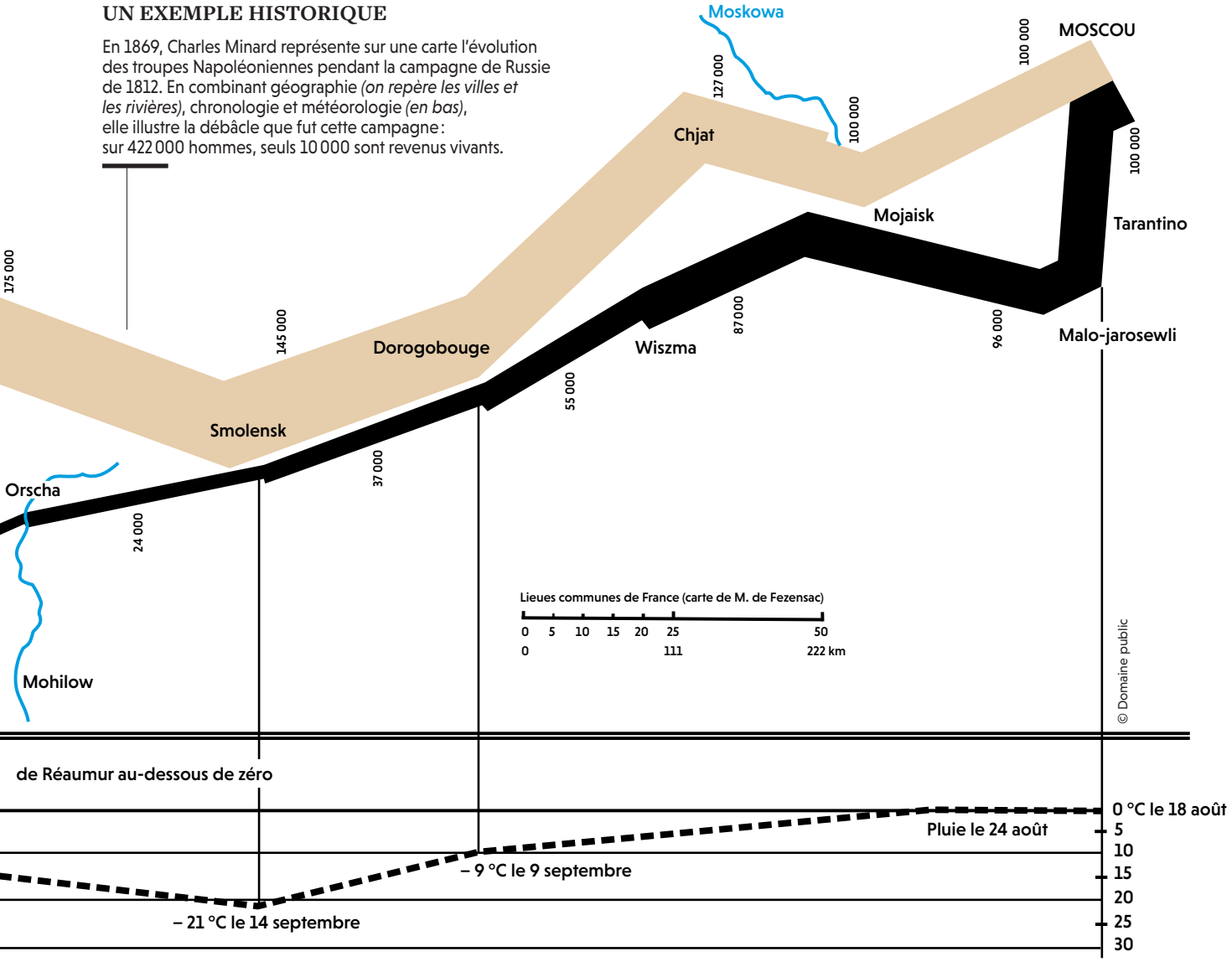
SÉRIES TEMPORELLES

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.

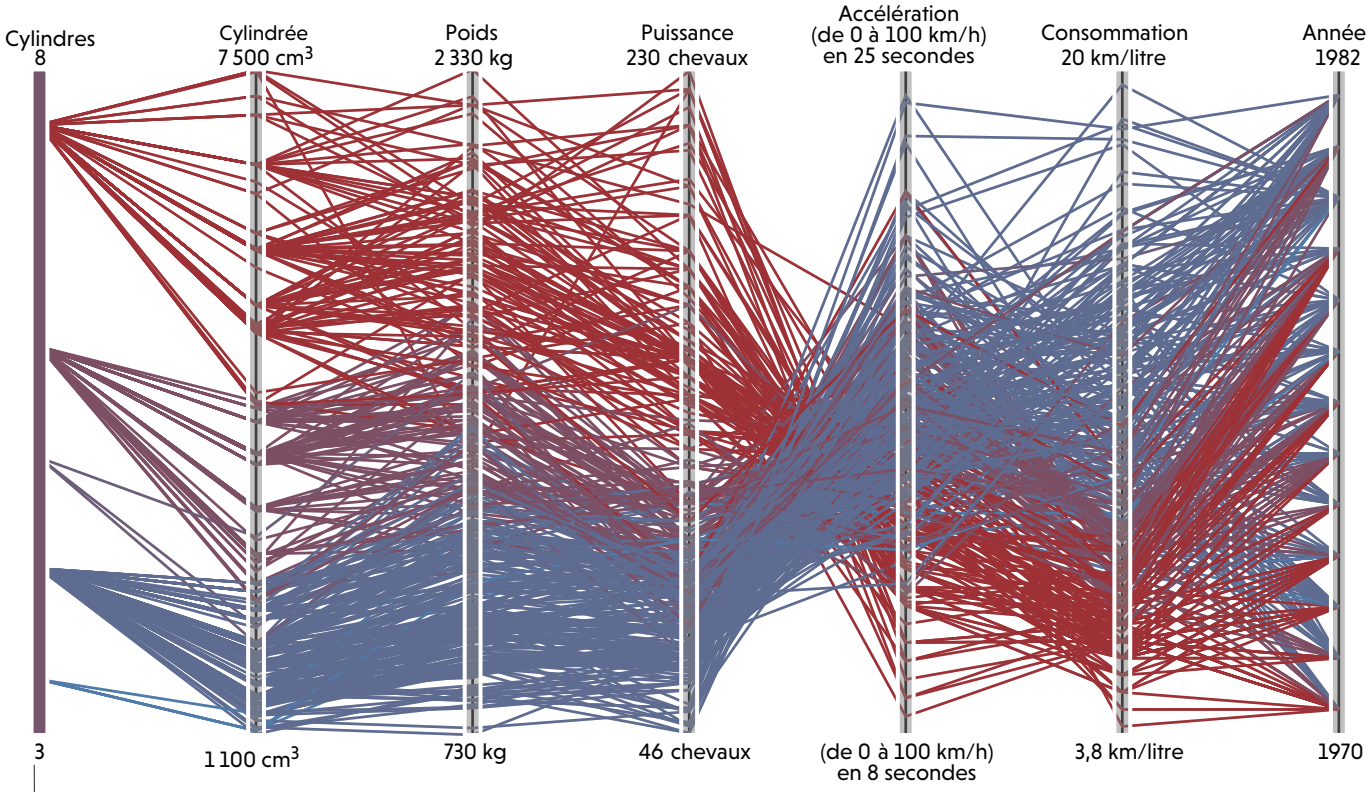


UN EXEMPLE HISTORIQUE

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.



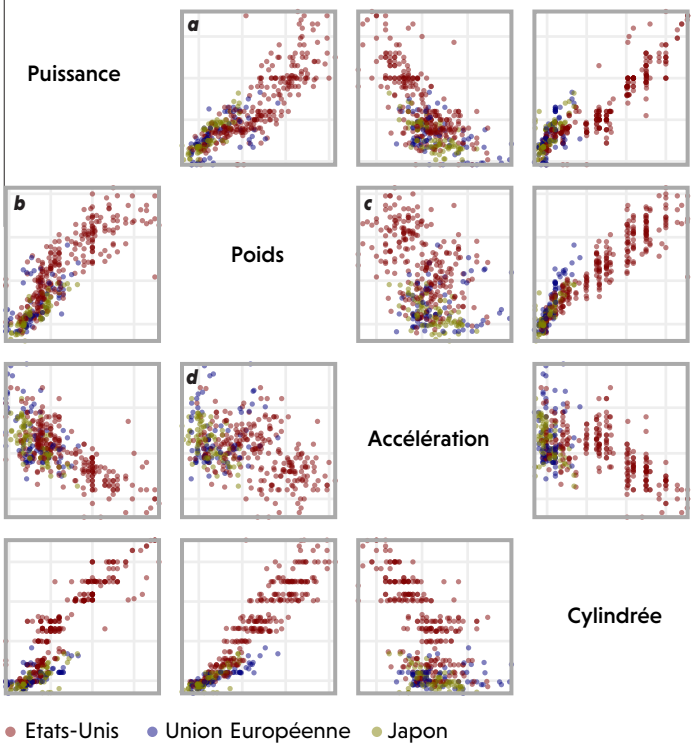
© Domaine public

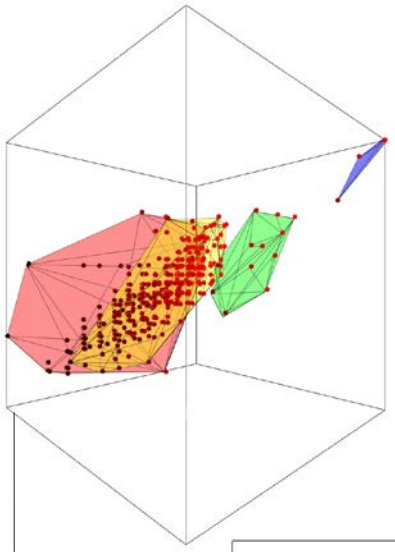


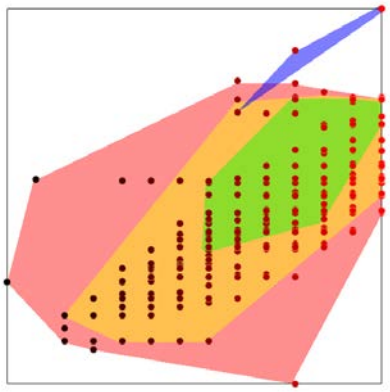
© J. M. Heer/ U. Washington

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*).





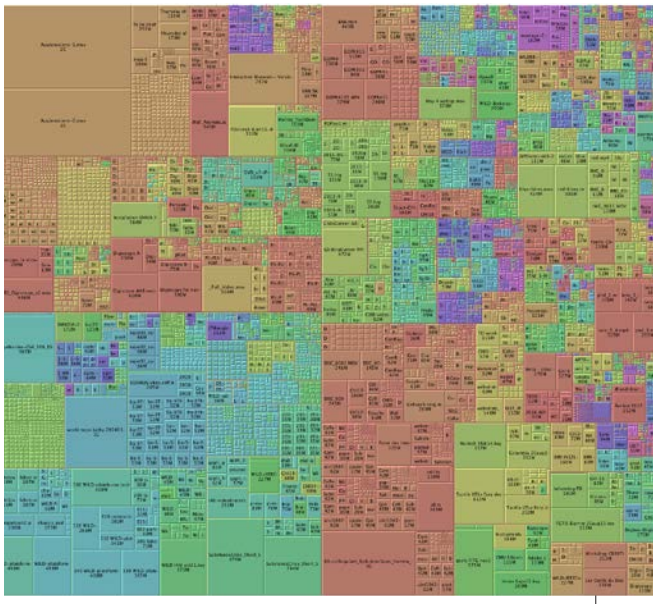


«ROLLING THE DICE»

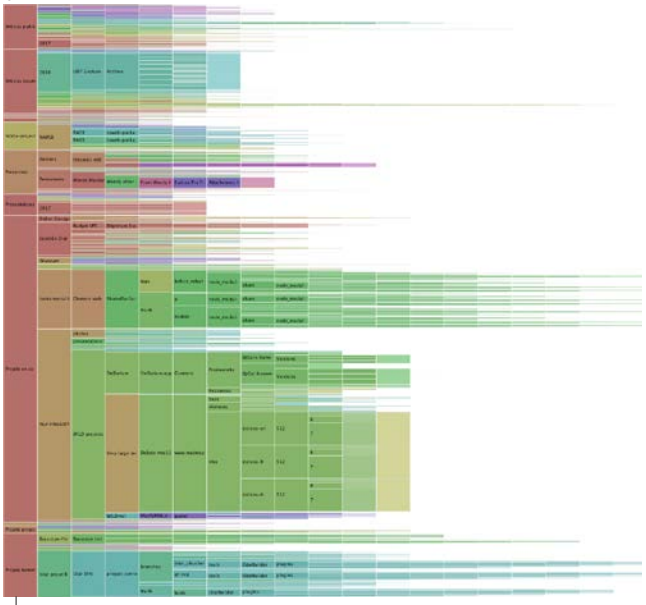
Avec cette technique d'animation 3D (littéralement « Faire rouler les dés »), on peut projeter dans l'espace les matrices de corrélations *(ci-contre)* et passer de l'une à l'autre afin de mieux comprendre leurs relations. On peut voir cette animation ici : http://bit.ly/RolDice

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

a



b

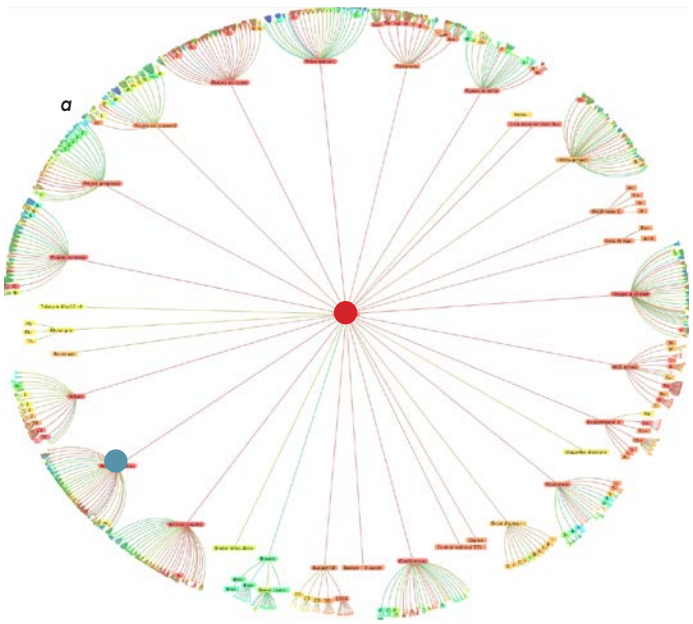


[source : http://www.randelshofer.ch/treeviz/]
[app Java libre disponible sur cette page]

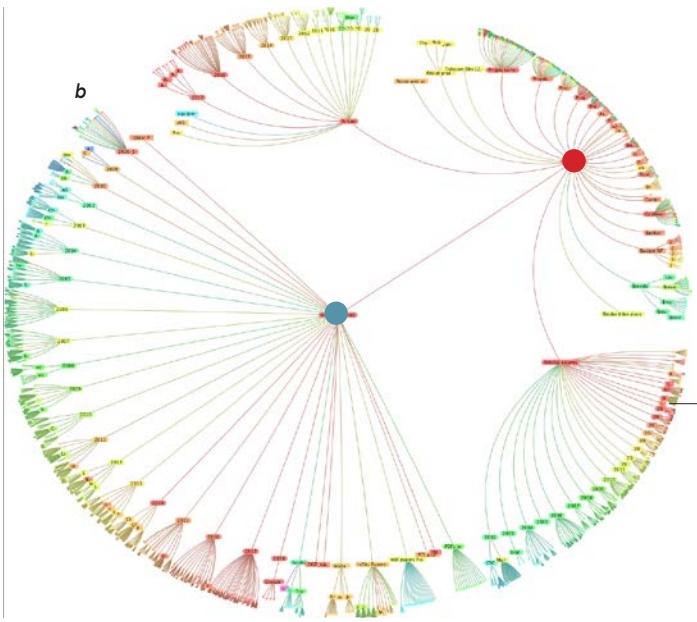
UNE FORÊT D'ARBRES

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.

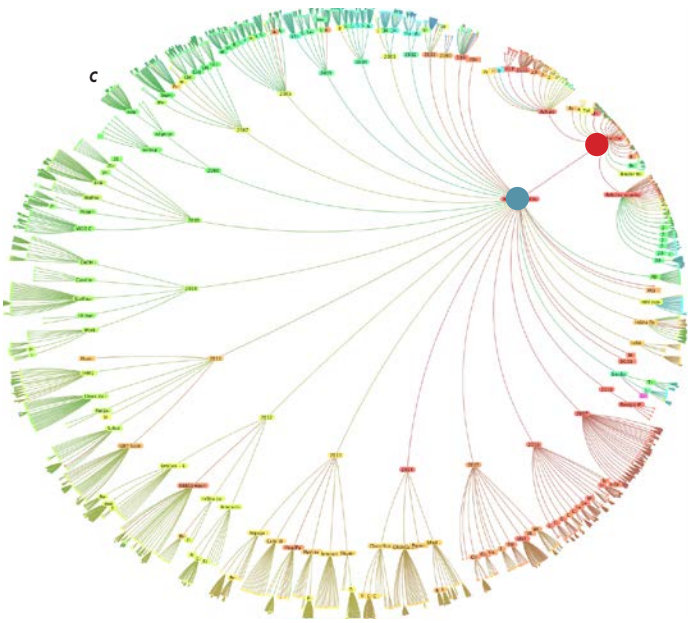
a



b



c



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.

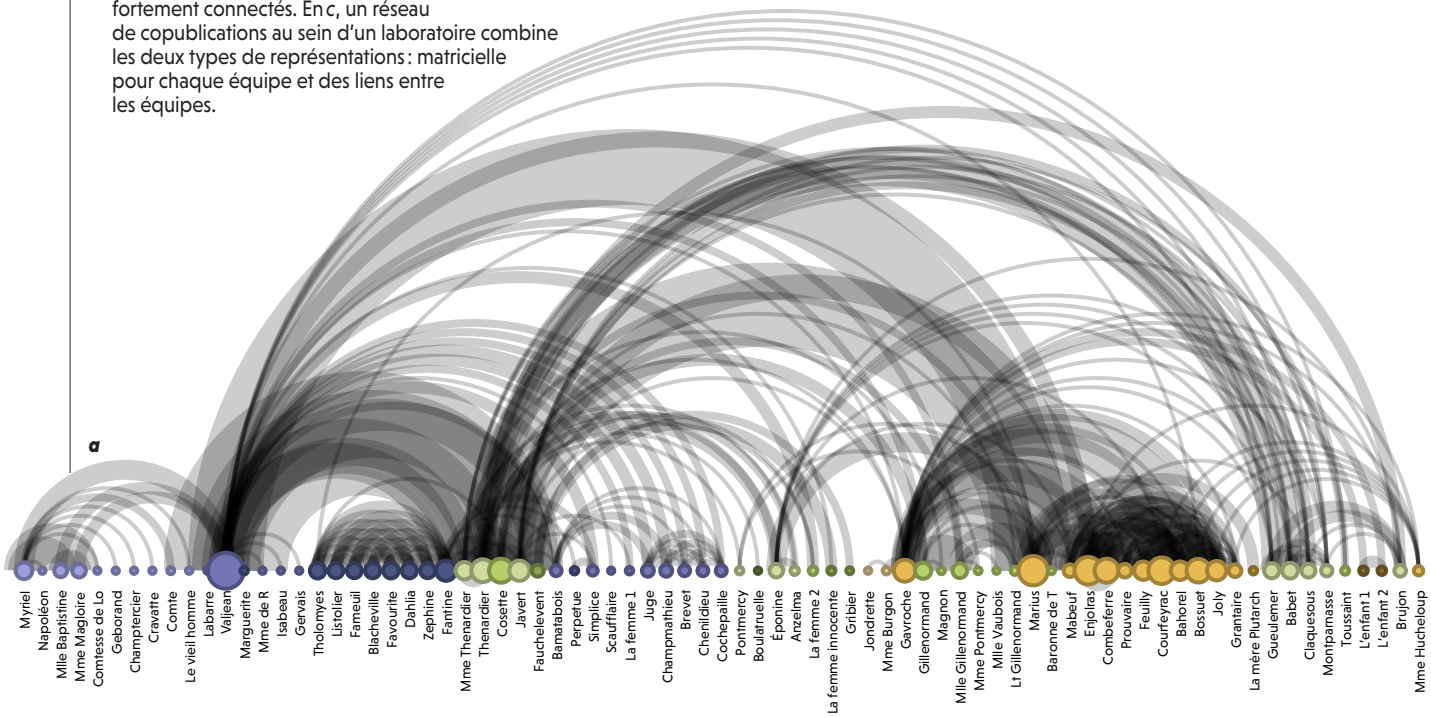
© Michel Beaudouin-Lafon

L'ARBRE DE LA VIE

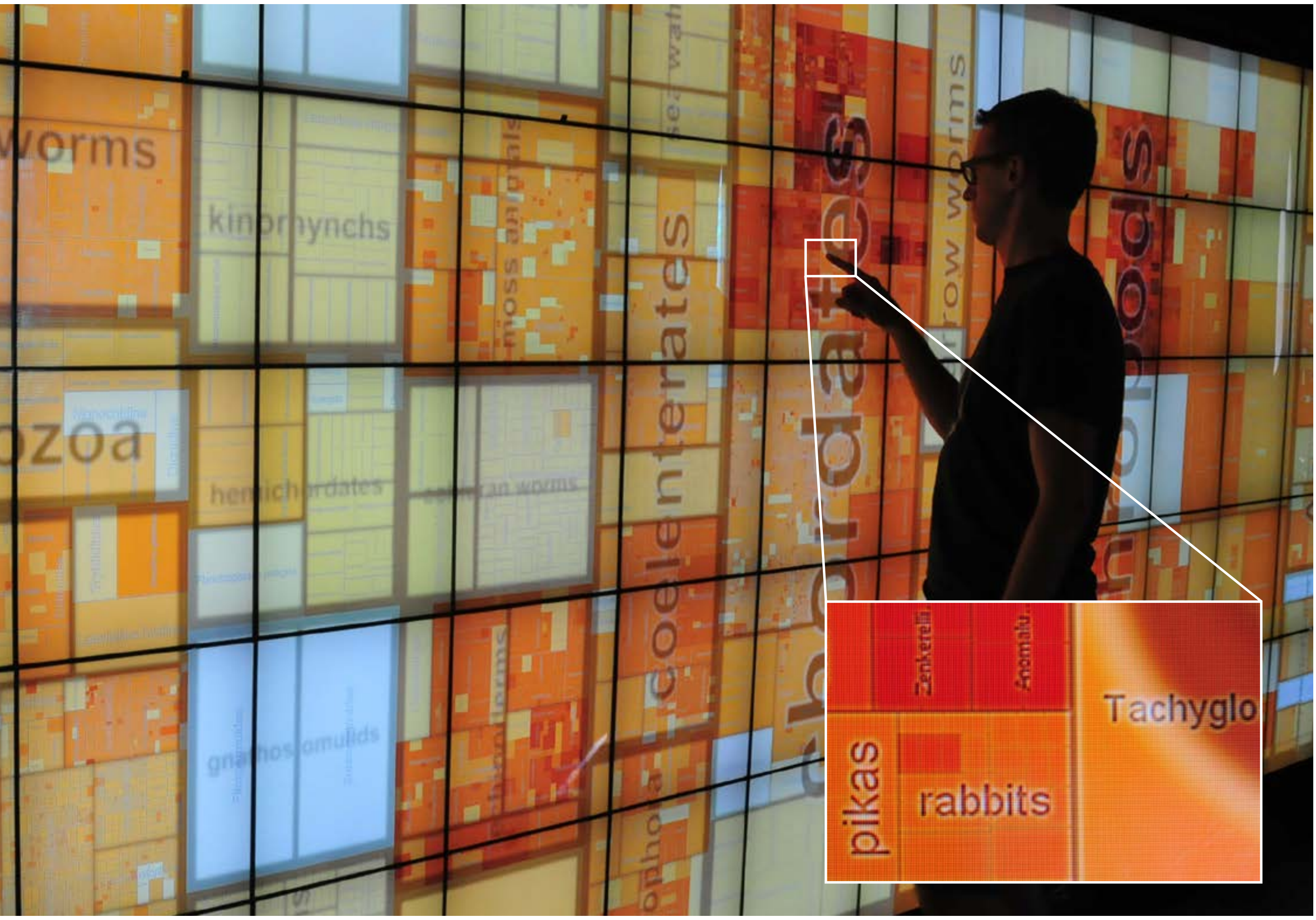
Sur le mur d'écrans Wilder du projet Digiscope (http://digiscope.fr), de six mètres de largeur et de 70 millions de pixels, s'affiche l'arbre planaire rassemblant près de 200 000 espèces d'organismes vivants. La technique des images hybrides permet de lire de loin les étiquettes des grandes catégories,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

NON AUX SPAGHETTIS !

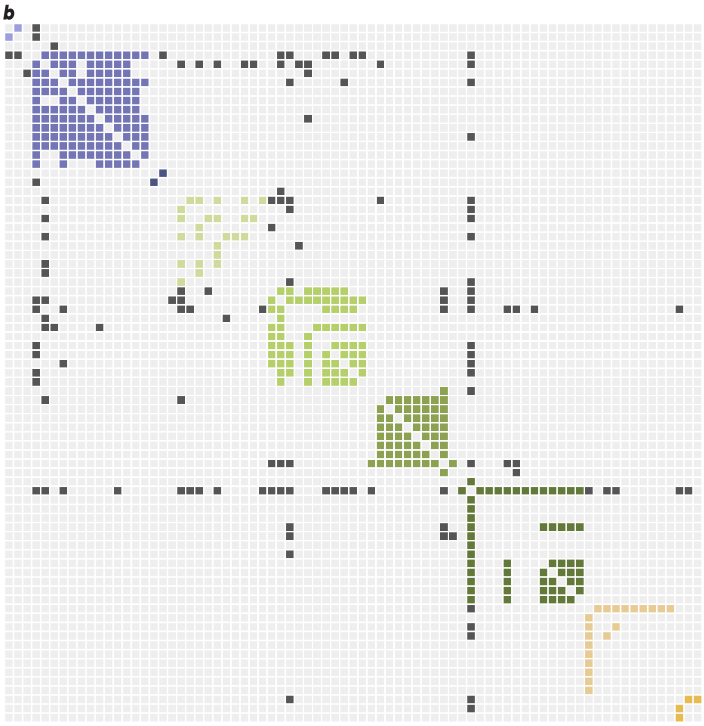
Les données en réseau sont difficiles à représenter car on obtient en général un « plat de spaghettis » illisible. Dans la représentation *a*, les arcs relient les personnages des *Misérables* qui apparaissent en même temps dans le livre. Ils ont été ordonnés et rassemblés par couleur de façon à rendre 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.



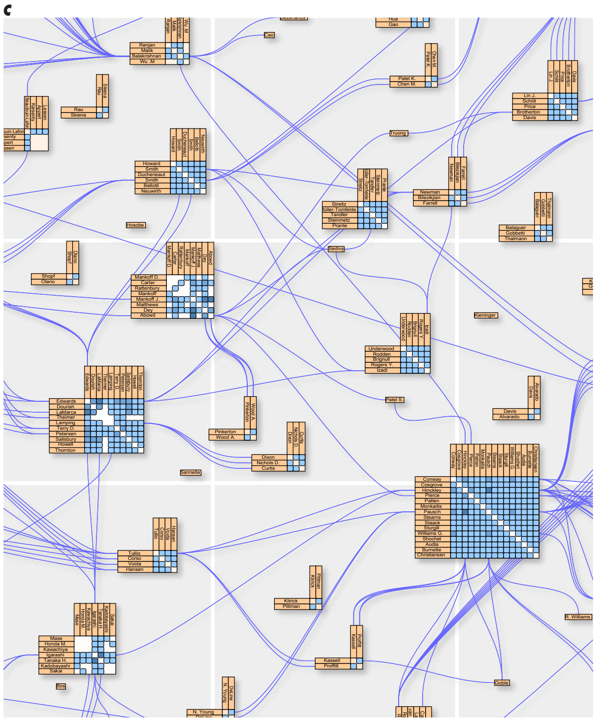
Images 1 et 2 : © J. M. Heer/ U. Washington
Image 3 : © 2007, N. Henry, P. Dragicevic, J.-D. Fekete et Inria

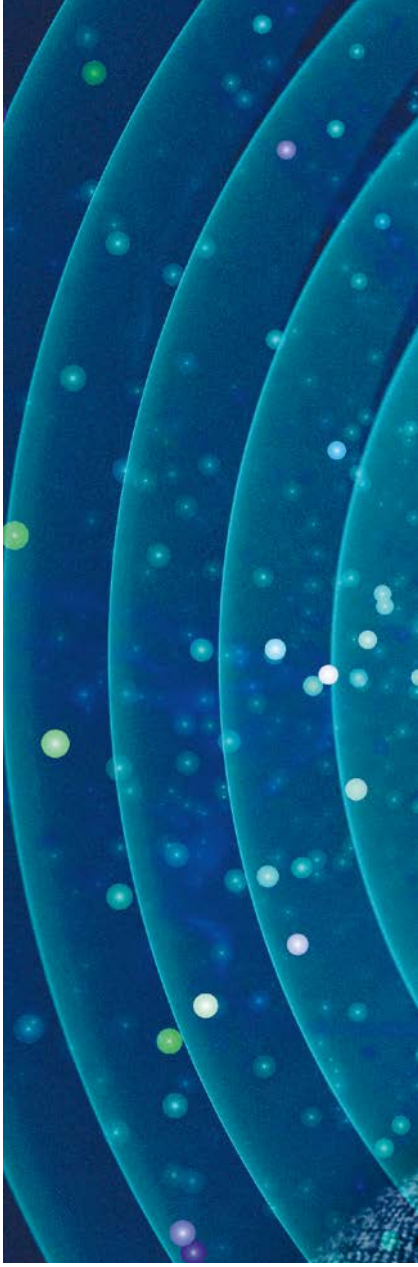

worms
kinorhynchs
zoa
hemichordates
coelenterates
Zenkerelli
Anomalu
Tachyglo
pikas
rabbits

b



c





L'ESSENTIEL

- La notion de cause a connu des fortunes diverses au cours des siècles et a cédé la place, en science, au principe de causalité.
- La science a continué de se bâtir sur des théories physiques et des lois universelles.
- Aujourd'hui, l'essor du *big data* s'accompagne d'une possible nouvelle science fondée sur les corrélations. Est-ce la fin de la théorie ?
- Si elle peut remporter quelques succès, cette voie ne conduira jamais à des théories du niveau de la relativité générale.

L'AUTEUR

ÉTIENNE KLEIN
est philosophe des sciences et directeur de recherche au CEA.

Vers une nouvelle SCIENCE ?

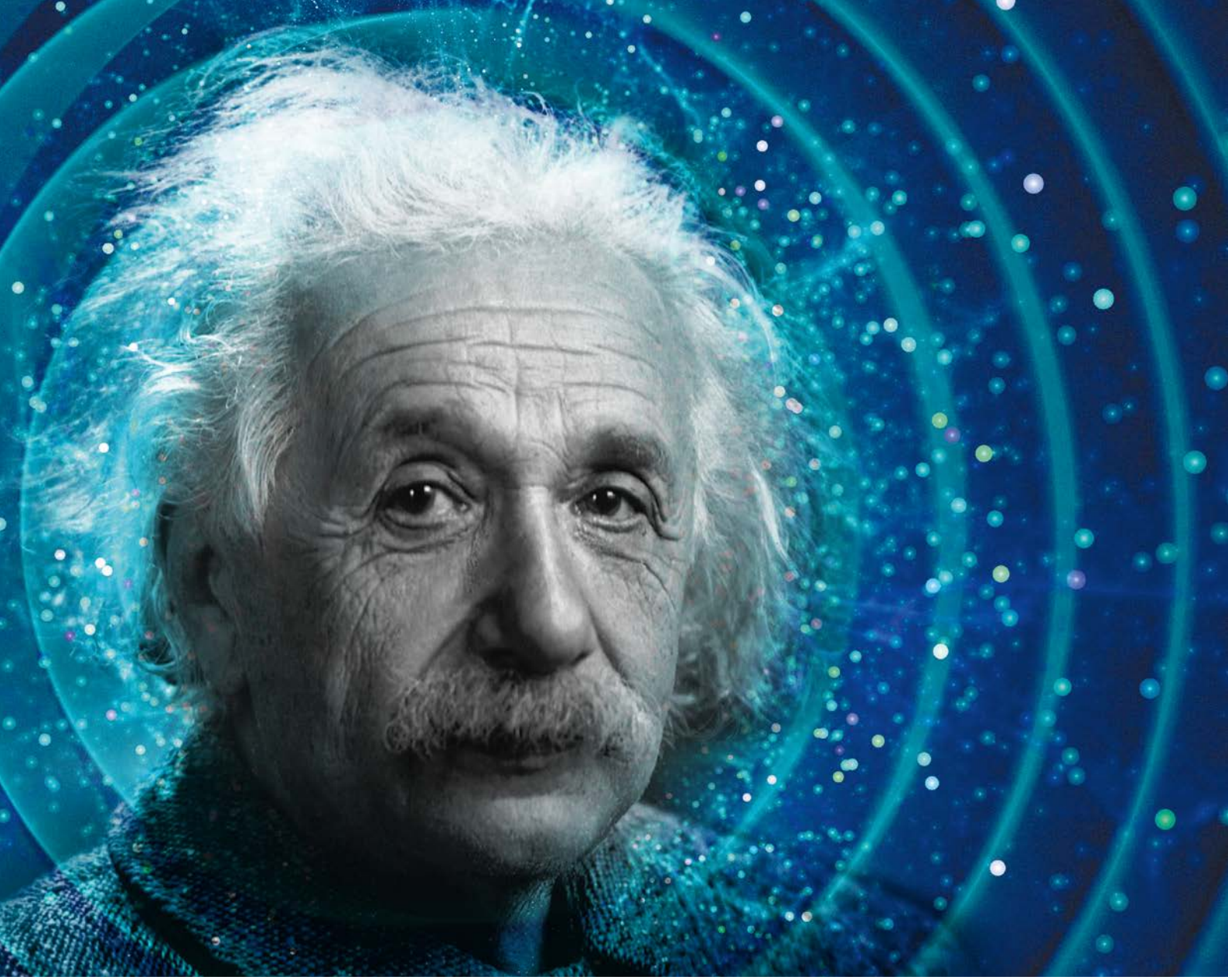
Aux échecs, au jeu de go, le silicium écrase parfois le neurone. Est-ce une raison pour confier à des machines, gavées de données, l'activité scientifique de demain ?

Cet article est extrait en partie du livre *Matière à contredire*, essai de philo-physique, à paraître le 7 février 2018 aux Éditions de l'Observatoire.

En juin 2008, Chris Anderson, rédacteur en chef du magazine *Wired*, publiait un article au titre provocateur: «*The end of theory: The data deluge makes the scientific method obsolete*» («**La fin de la théorie : le *big data* rend obsolète la méthode scientifique**»). Il y défendait l'idée que lorsque nous disposerons de suffisamment de données, les nombres parleront d'eux-mêmes et les corrélations qu'ils dévoileront remplaceront les relations de causalité que manifestent les lois théoriques. La science changerait alors de visage puisqu'elle pourrait se développer sans plus énoncer d'hypothèses, et sans plus s'appuyer sur des théories explicites. Prophétie ou délire techno-utopiste ?

DE LA MONARCHIE ANGLAISE

Pour trancher, ou au moins disposer d'éléments de réponse, nous devons d'abord revenir sur l'idée de causalité. On dit souvent des sciences dures, notamment de la physique, qu'elles visent à identifier les causes des phénomènes naturels. En réalité, la notion de cause, au sens fort du terme, a posé tant de problèmes aux physiciens qu'ils ont fini par quasiment l'abandonner. Ils ne l'invoquent en tout cas presque plus de façon explicite, même si le concept de cause demeure présent dans certains de leurs discours.



Pourrions-nous retrouver les équations d'Einstein, qui prévoient notamment les ondes gravitationnelles, à partir du seul *big data* ?

La cause en est (si l'on ose dire) celle identifiée par Bertrand Russell : « Il en va du concept de cause comme de la monarchie anglaise, à savoir qu'on ne l'a laissée survivre que parce qu'on suppose à tort qu'elle ne fait pas de dégâts. » Le mathématicien et philosophe britannique prenait l'exemple de la loi de gravitation afin de montrer comment celle-ci aplatit la notion de causalité. « Dans les mouvements des corps gravitant ensemble, il n'y a rien qui puisse être appelé une cause, et il n'y a rien qui puisse être appelé un effet : il y a là simplement une formule qui permet de calculer la configuration du système à n'importe quel instant. »

Après avoir joué un rôle essentiel dans la physique des XVII^e^ et XVIII^e^ siècles, l'idée de cause a en effet vu son importance décliner en deux étapes. Au XIX^e^ siècle d'abord, elle s'est effacée au profit de la notion de loi physique, et a pâti de l'assouplissement du déterminisme du fait de l'apparition des probabilités en physique statistique. Ensuite, au XX^e^ siècle, la physique quantique lui a porté le coup de grâce. De fait, l'usage que cette physique fait des probabilités interdit qu'on puisse parler, à propos des processus quantiques, de « cause » au sens strict du terme.

C'est ainsi que, progressivement, l'idée de cause semble avoir été évacuée des théories scientifiques pour se résorber dans la dynamique même des systèmes. Pourtant, le principe de causalité, lui, reste parfaitement vivace ! Il stipule que si un phénomène, nommé cause, produit un autre phénomène, nommé effet, alors l'effet ne peut précéder la cause. Désormais épuré de l'idée de cause (on cherche le suffisant et le nécessaire à un événement), le principe de causalité a permis l'élaboration des théories de la physique moderne (physique quantique et théories de la relativité), et il structure en profondeur celles qui sont à l'ébauche : il impose un ordre obligatoire et absolu entre divers phénomènes, sans que l'un puisse être présenté comme la cause de l'autre. >

En pratique, le principe de causalité se décline dans les différents formalismes de la physique : il s'adapte à chacun d'eux, y prend une forme qui dépend de la façon dont les événements et les phénomènes sont mathématiquement représentés. Ses conséquences sont toujours extrêmement contraignantes. Elles s'expriment sous la forme d'interdictions ou de prédictions, qui dépendent de façon cruciale de la théorie qu'on considère.

En physique newtonienne, la causalité brise le cercle du temps, celui-ci devenant linéaire et non cyclique. Cette ouverture topologique suffit à assurer qu'un effet ne peut pas rétroagir sur sa propre cause. En relativité restreinte, la causalité interdit qu'une particule puisse se propager plus vite que la lumière dans le vide, car alors elle pourrait voyager dans le passé et éventuellement y changer le cours d'événements qui se sont déjà produits…

En physique des particules, la prise en compte de la causalité a permis de prédire l'existence de l'antimatière, dûment confirmée ensuite par l'expérience.

AU NOM DE LA LOI, AU NOM DE LA CAUSE

Mais dès lors que la physique a renoncé aux causes proprement dites et a dissous l'idée même de causalité dans celle de loi physique, il nous incombe de clarifier cette dernière notion : quel lien peut-il exister entre l'univers physique et les lois universelles qui découlent des théories physiques ? En d'autres termes, quel est le statut des lois physiques et comment parviennent-elles à s'appliquer aux objets physiques ?

Une première piste consiste à considérer que les lois physiques seraient des produits de l'intellect, ou du moins qu'elles en seraient inséparables. Mais comment la pensée de l'homme, qui n'est qu'une partie micro-infime de l'Univers, parviendrait-elle à saisir la structure du tout qui la contient ? Autrement dit, grâce à quelle nature très spéciale les lois physiques pourraient-elles participer à la fois du monde qu'elles structurent et de la pensée qui comprend ce monde ? Y aurait-il une parenté de structure, voire une identité, entre l'esprit humain et l'Univers ?

On peut préférer considérer, derrière Platon, l'existence de deux mondes distincts ; d'une part, celui des formes intelligibles, constitué de réalités immuables et universelles, et faisant l'objet d'une connaissance et d'un discours vrais ; d'autre part, celui des choses sensibles, qui ne sont que des copies approximatives des formes pures. En langage moderne, cela reviendrait à dire que les équations mathématiques exprimant les lois physiques sont « transcendantes » et non pas immanentes, c'est-à-dire indépendantes de l'Univers empirique, et que ce dernier n'en constitue qu'une image mobile et imparfaite.

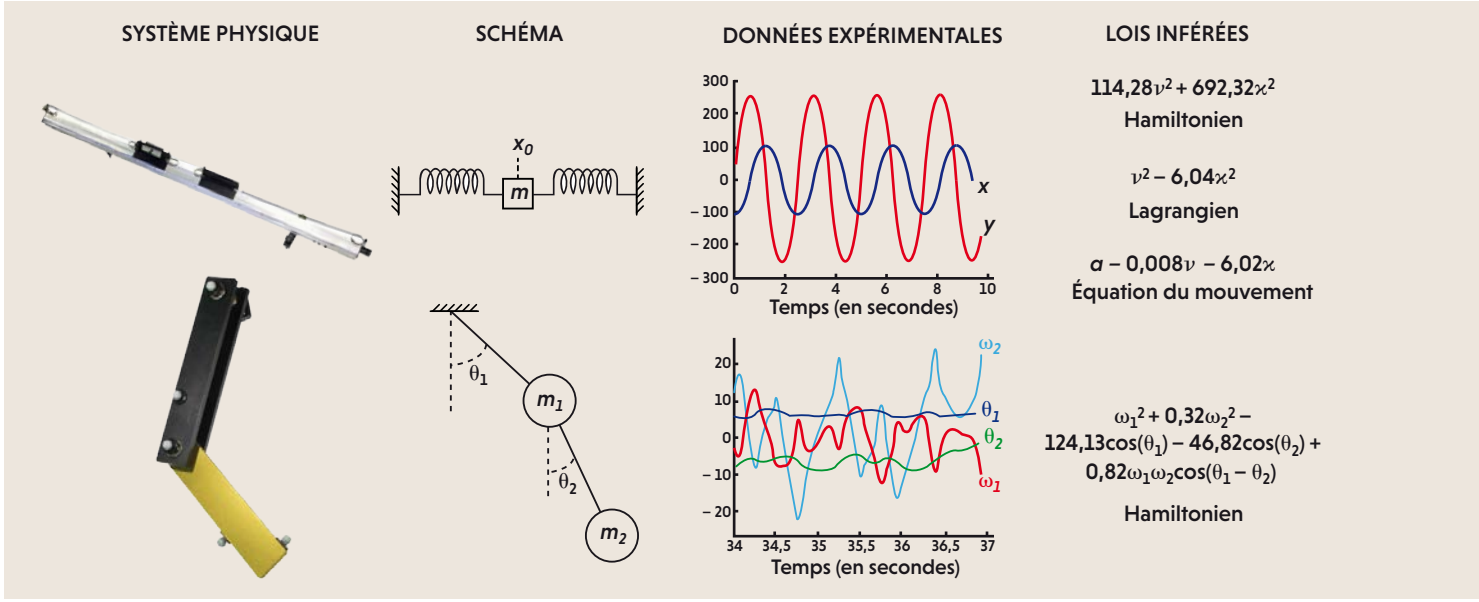
L'Univers serait en quelque sorte un écho physique dégradé de la pureté mathématique qui le tiendrait sous sa coupe. Mais si tel est le cas, comment le monde des équations parvient-il à structurer « à distance » le monde des phénomènes ? Ou, redit en langage platonicien, comment les formes intelligibles participent-elles aux formes sensibles ? En guise de réponse à cette question, Platon avait énoncé dans le *Timée* deux hypothèses donnant naissance à deux fictions : celle d'un démiurge, qui met en ordre l'Univers, et celle de la *khora*, le matériau sur lequel intervient le démiurge en s'inspirant du mieux qu'il peut, mais de manière infidèle, des formes intelligibles. Comment ces idées pourraient-elles s'adapter à la physique contemporaine, peu friande en démiurges ?

Un algorithme « naïf » a réussi à redécouvrir plusieurs lois physiques à partir de données

Ces embarras peuvent nous porter à suivre plutôt Spinoza, selon lequel il n'y a qu'un seul et même univers, mais qui se donne sous deux modalités différentes : d'un côté, un univers matériel et spatial, et, de l'autre, un univers législatif contenant des lois, des principes et des règles qui sont accessibles à la pensée. Mais là encore, comment ces deux modalités d'être de l'Univers communiquent-elles ? Par quelle médiation les lois de la matière, qui relèvent du deuxième mode, parviennent-elles à s'imposer à la matière, située, elle, dans le premier mode ?

On voit par là que les questions les plus vives de la cosmologie contemporaine font lointainement écho aux systèmes philosophiques les plus puissants. Mais parmi toutes ces options, ou d'autres également imaginables, comment choisir ?

La réponse s'annonce d'autant plus délicate qu'un bouleversement est en train de se



Un algorithme retrouve les lois de la mécanique *(à droite)* à partir des données *(au centre)* collectées sur des dispositifs mécaniques *(à gauche)*. En est-il pour autant un scientifique ?

produire, de plus en plus manifeste : l'arrivée du *big data*. Tout le monde en parle, et pour cause : couplé à l'intelligence artificielle (*voir La révolution de l'apprentissage profond, par Y. Bengio, page 42*), le *big data* a déjà, et aura plus encore dans l'avenir, un impact majeur sur les technologies, le travail, les emplois, l'économie (*voir l'entretien avec D. Cohen, page 8*), la médecine et les relations interpersonnelles. Et aussi, à l'évidence, sur les sciences, notamment sur la façon de mener les recherches, et peut-être aussi sur le statut même des lois physiques. De quelle façon ?

LES PILIERS DE LA SCIENCE

Galilée est à juste titre considéré comme l'inventeur de la science telle que nous la pratiquons depuis quatre siècles, dont les deux grands piliers ont été la théorisation et l'expérimentation : on élabore des hypothèses conduisant à la formulation de lois, qui sont ensuite testées grâce à des expériences spécifiques, souvent très éloignées de la simple observation. Par exemple, lorsque, en 1604, Galilée énonce que tous les corps, quelle que soit leur masse, tombent à la même vitesse, il formule une loi qui n'est ni conforme aux données empiriques de l'époque, ni directement observable (elle ne peut l'être que dans le vide) : elle ne sera expérimentalement vérifiée que bien plus tard.

Mais cette façon de faire de la science n'est-elle pas en train de changer sous nos yeux ? Avec l'arrivée du *big data*, ce modèle canonique demeurera-t-il toujours vrai ? Continuerons-nous d'honorer la pensée spéculative, d'incliner à « la gravité enjouée de l'enquête métaphysique » ? L'article de Chris Anderson cité en introduction répondait par la négative à ces questions.

Le *big data* est cet ensemble énorme de données brutes et silencieuses, dont le volume augmente à une vitesse prodigieuse. Ces données choisies mémorisent les « traces » de l'activité d'êtres vivants, de machines, d'objets, et leurs états successifs. Leur collecte est principalement automatique, puis ces données sont analysées par des algorithmes qui y détectent des régularités, par exemple dans le comportement des consommateurs, des machines, des indices économiques, du trafic routier…

À partir de ces régularités, ils infèrent des règles prédictives que nous avons tendance à considérer comme des normes, ou comme des lois générales, voire universelles, alors qu'elles ne sont que la condensation de ce qui a déjà eu lieu : dès lors que le futur qu'elles configurent n'est que du passé extrapolé, elles ne peuvent correctement prédire l'avenir qu'à la condition que celui-ci prolonge le passé, sans surprises ni ruptures ni inventions.

Privilégier de la sorte l'induction, n'est-ce pas faire exagérément confiance à une certaine uniformité du cours de la nature et trop croire que certaines choses n'arriveront pas ? L'idée de rupture n'est-elle pas en train de disparaître ?

Certes, le *big data* ouvre des perspectives fascinantes, notamment celle de redécouvrir des lois universelles déjà connues par la simple analyse de données massives. C'est ce qu'ont montré Michael Schmidt et Hod Lipson, de l'université Cornell, à Ithaca, aux États-Unis. Ils ont d'abord accumulé des données de positions, de vitesse, d'angles… enregistrées en certains points d'un dispositif mécanique en mouvement, notamment un double pendule au comportement chaotique (*voir la figure ci-dessus*). Ensuite, ces informations ont alimenté un algorithme « naïf », c'est-à-dire >

> dépourvu de toute connaissance en géométrie et en physique. Le programme a réussi à « découvrir » plusieurs lois (les Hamiltoniens, les Lagrangiens, la conservation du moment…). En d'autres termes, l'algorithme s'est mué en un physicien ! Dans ce cas, les lois identifiées étaient connues, mais pourquoi ne pas imaginer qu'un algorithme puisse identifier de nouvelles lois ?

Le *big data* peut également aider à la compréhension d'événements impliquant de très grands nombres de variables quantifiables, tels les phénomènes météorologiques ou climatiques, les comportements électoraux, l'usage des réseaux sociaux…

FUYEZ L'HÔPITAL !

Toutefois, ce transfert vers les *big data* n'est pas sans risque. De fait, il est aussi possible que nous nous égarions dans l'identification de multiples corrélations, pas forcément bien interprétées. Or une corrélation n'est pas une relation de cause à effet. Elle marque simplement le fait que deux grandeurs semblent dépendre l'une de l'autre, au sens où si l'une augmente, l'autre augmente ou décroît « de la même façon ».

Par exemple, il a été récemment démontré dans une université américaine que les étudiants qui sont les plus gros consommateurs d'alcool ont tendance à avoir des résultats scolaires moins bons que les autres. L'alcool est-il la cause directe de cette baisse de niveau ? N'allons pas trop vite en besogne. On peut certes imaginer que l'alcool rende stupide, mais d'autres hypothèses sont aussi envisageables : les étudiants qui boivent de l'alcool seraient à la base de moins bons étudiants ; ou bien ils seraient moins attentifs que les autres les lendemains de soirées arrosées, expliquant leurs déboires aux examens ; ou bien encore, ils boiraient plus que de raison afin de calmer leur peur d'échouer ou pour noyer des chagrins n'ayant rien à voir avec leurs études…

Ainsi, une corrélation manifeste n'est pas nécessairement la manifestation d'une relation de causalité : ce n'est pas parce qu'il y a des grenouilles après la pluie qu'on a le droit de dire qu'il a plu des grenouilles. Mais il arrive très souvent que nous confondions les deux notions, à la manière d'un Coluche conseillant de ne jamais aller à l'hôpital au motif qu'on y meurt en moyenne plus souvent que chez soi… Plusieurs sites recensent ce type de corrélations (*voir la figure page ci-contre*) relevant de l'effet cigogne : le taux de natalité augmentant comme le nombre de nids de cigognes, on en déduit que les cigognes apportent les bébés ! Il y a matière à sourire, mais le problème n'en demeure pas moins sérieux.

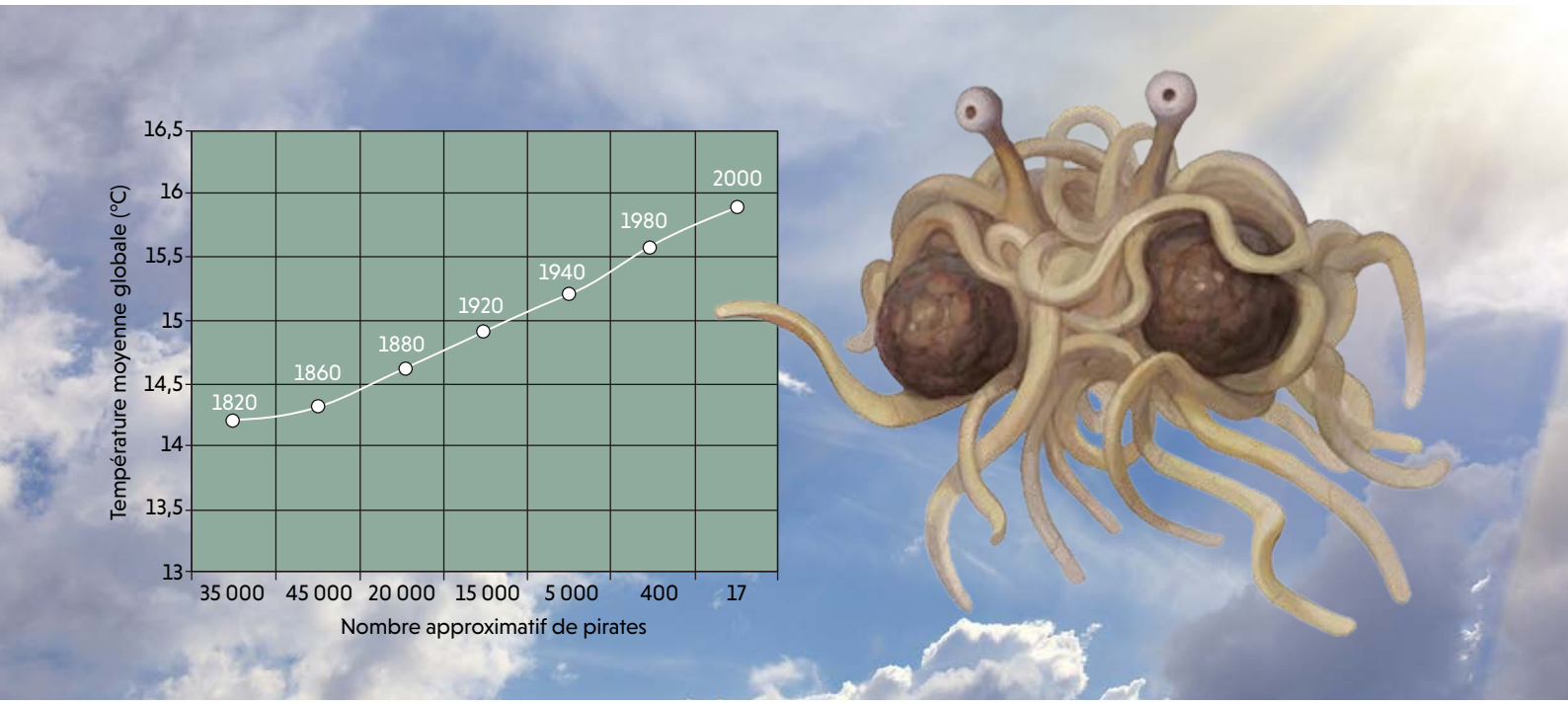
Il y a bien la possibilité qu'avec le *big data*, au lieu de théoriser, nous cédions aux facilités de l'induction et délaissions le « geste théorique » de type galiléen, celui qui consiste à énoncer des hypothèses portant bien au-delà des données disponibles.

À cet égard, le cas d'Albert Einstein est exemplaire. En 1915, il publiait la théorie de la relativité générale, une conception révolutionnaire de la gravitation, alors qu'il n'avait que très peu de données sur l'Univers à sa disposition : on ignorait, par exemple, pourquoi les étoiles brillent, que d'autres galaxies hormis la nôtre existent, que l'Univers est en expansion, etc. Cela n'a pas empêché les équations d'Einstein de parfaitement s'accommoder de la quantité gigantesque de données recueillies depuis un siècle par les télescopes et les satellites. En outre, elles constituent les fondations d'une véritable cosmologie scientifique, capable d'envisager l'Univers en tant qu'objet physique, doté de propriétés qui le caractérisent « lui-même ».

N'allez jamais à l'hôpital, on y meurt en moyenne plus souvent que chez soi

Elles ont même permis de prédire l'existence des ondes gravitationnelles un siècle avant leur première détection, en septembre 2016. Une théorie peut donc non seulement enrichir l'univers des données, mais également agir comme un « treuil ontologique » capable de faire apparaître de nouveaux éléments de réalité. En d'autres termes, la théorie en « dit plus » que les données, notamment par le fait qu'elle explicite des lois que les données n'illustrent jamais que de façon partielle.

Imaginons maintenant que les choses se soient passées dans l'ordre inverse, c'est-à-dire que nous ayons commencé avec toutes les données dont nous disposons aujourd'hui, mais sans avoir à notre disposition la théorie de la relativité générale. Pourrions-nous, par une sorte d'induction théorique permettant de passer des données aux lois, découvrir les



Selon le pastafarisme, une parodie de religion parfois considérée comme une version moderne de la théière de Russell (une métaphore qui illustre le fait que c'est au croyant de prouver les bases « invérifiables » de la religion), le déclin de la piraterie est la cause du réchauffement climatique, car la température moyenne sur la Terre est corrélée, de façon inverse, à la population de pirates.

équations d'Einstein? Rien n'est moins sûr, même si, nous l'avons vu, certaines lois physiques simples ont pu être récemment « redécouvertes » à partir de l'analyse d'un ensemble de données expérimentales.

La réponse d'Einstein à cette question aurait en tout cas été négative, du moins si l'on en croit la lettre qu'il écrivit un jour à son grand ami Maurice Solovine: « Aucune méthode inductive ne peut conduire aux concepts fondamentaux de la physique. L'incapacité à le comprendre est la plus grave erreur philosophique de nombreux penseurs du XIX^{e} siècle. »

Quelques années plus tard, lors d'une conférence qu'il donna à Oxford avant de quitter définitivement l'Europe pour l'Amérique, il précisera quelle était selon lui la base de l'invention des idées scientifiques: « L'expérience peut bien entendu nous guider dans notre choix des concepts mathématiques à utiliser, mais il n'est pas possible qu'elle soit la source d'où ils découlent. C'est dans les mathématiques que réside le principe vraiment créateur. En un certain sens, donc, je tiens pour vrai que la pensée pure est compétente pour comprendre le réel, ainsi que les Anciens l'avaient rêvé. »

En la matière, les prochaines décennies, toutes gorgées d'intelligence dite « artificielle » et de « *very big data* », viendront-elles contredire la pertinence de cet avis? Il est raisonnable d'en douter, pour au moins deux raisons.

La première tient à ce que l'intelligence artificielle porte bien son nom: elle n'est qu'une intelligence fabriquée au moyen de techniques informatiques, de sorte que rien ne permet de penser qu'elle soit elle-même « pensante ». Pour doter les machines d'une pensée autonome, analogue à la pensée humaine, il faudrait déjà comprendre en détail les mécanismes de cette dernière, ce qui est loin d'être acquis, et aussi en donner une définition qui soit à la fois opératoire et exhaustive. Classer des milliers de photographies mieux et plus vite que n'importe quel humain est une chose, mais inventer des concepts et les associer en est une tout autre...

APPELONS UN CHAT UN CHAT, MAIS COMMENT ?

La seconde tient en ce que les techniques de l'intelligence artificielle, tel le *deep learning* (*voir La révolution de l'apprentissage profond, par Y. Bengio, page 42*), fonctionnent comme des boîtes noires. Lorsque je vois un chat et reconnais aussitôt qu'il s'agit bien d'un chat, j'effectue une opération qui n'est pas de même nature que celle d'un logiciel d'intelligence artificielle: lui constate que l'objet qui lui est présenté ressemble à d'innombrables autres chats dont on lui a préalablement fourni les images pour « l'entraîner », mais sans que quiconque puisse déterminer comment il effectue ce constat, ni indiquer quelles sont les ressemblances qui font « tilt » pour lui.

Il est donc difficile d'imaginer que l'intelligence artificielle puisse élaborer des théories physiques, et encore moins, si d'aventure elle le pouvait, que celles-ci puissent nous être intelligibles. Jusqu'à preuve du contraire, physicien demeure un métier d'avenir. ■

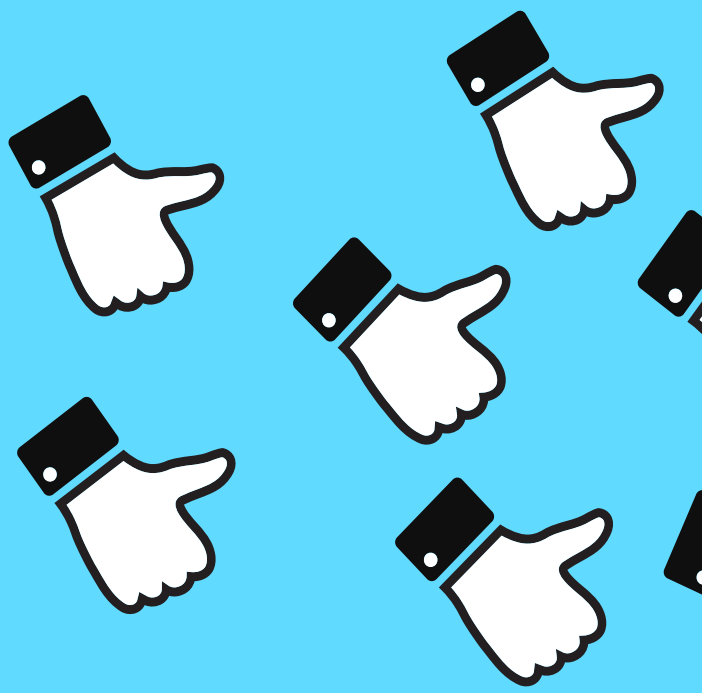
BIBLIOGRAPHIE

G. BERRY,
L'hyperpuissance de l'informatique : Algorithmes, données, machines, réseaux,
Odile Jacob, 2017

M. SCHMIDT ET H. LIPSON,
Distilling Free-Form Natural Laws from Experimental Data,
***Science*, vol. 324, pp. 81-85, 2009.**

FAKE NEWS

L'histoire secrète de leur succès



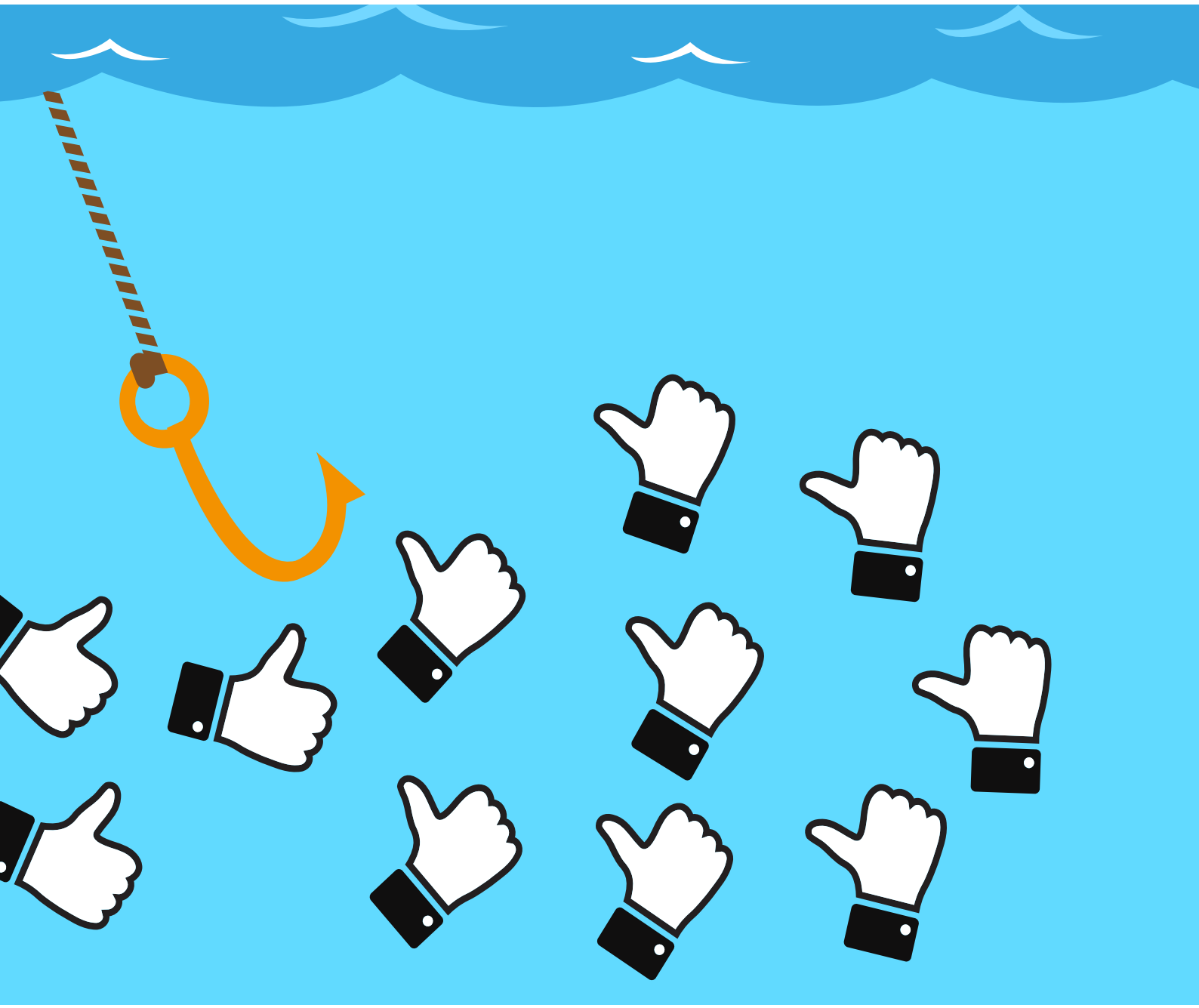
Les internautes diffusent massivement des informations fausses et des théories conspirationnistes farfelues. Quels mécanismes expliquent cet inquiétant phénomène ? Des études statistiques sur le réseau Facebook répondent.

L'ESSENTIEL

- Sur les réseaux sociaux tels que Facebook, de fausses informations et des théories fumeuses circulent facilement, avec parfois de graves conséquences.
- La masse de données disponibles permet d'étudier ces phénomènes de désinformation de façon quantitative.
- Les analyses montrent que le biais cognitif dit de confirmation joue un rôle majeur: les internautes forment des groupes solidaires qui autoentretiennent leurs opinions et préjugés.
- Il apparaît que les tenants des thèses conspirationnistes sont réfractaires à la démystification.

L'AUTEUR

WALTER QUATTROCIOCCHI coordonne le Laboratoire de sciences sociales computationnelles à l'école IMT des hautes études de Lucques, en Italie.



© Shutterstock.com/FGC

Le saviez-vous ? Le 11 octobre 2017 Google a racheté Apple. La Nasa exploite des enfants parqués dans des camps sur Mars. À Padoue, un restaurant chinois sert des pieds humains en hors-d'œuvre. Vous avez nécessairement vu passer ces informations. Faut-il préciser qu'il s'agit de rumeurs infondées… ?

Pourtant, elles ont été amplement colportées ou au moins commentées. Qu'est-ce qui a donc changé dans notre façon de nous informer, et donc de nous forger une opinion ? Quel rôle les médias sociaux tels que Facebook jouent-ils dans la diffusion de fausses informations ou de thèses conspirationnistes ? Quels sont les ressorts de cette mésinformation ou désinformation ? Est-il possible d'endiguer ces phénomènes ?

De nombreux sociologues se sont penchés sur les phénomènes sociaux liés à Internet et à ses médias, et notamment sur la « viralité » des informations infondées ou fausses. Ils ne sont pas les seuls. Depuis plusieurs années, des mathématiciens, des physiciens, des chercheurs en informatique se sont aussi intéressés à ces problématiques, en apportant leurs propres outils et méthodes d'analyse. Ainsi a émergé un nouveau champ de recherche : les « sciences sociales computationnelles ». >

> Grâce à l'analyse de grandes masses de données, cette discipline étudie les phénomènes sociaux de façon quantitative. Il s'agit d'exploiter les très nombreuses traces numériques que laissent les internautes sur les différents médias sociaux tels que Facebook, Twitter, YouTube, etc. lorsqu'ils sélectionnent, partagent ou commentent des informations. On peut ainsi étudier certains phénomènes sociaux à un niveau de précision sans précédent.

LES RACINES DE LA DÉSINFORMATION

Nos travaux s'inscrivent pleinement dans cette démarche. Notre groupe s'intéresse aux dynamiques de contagion sociale et à l'utilisation des contenus sur les différents réseaux sociaux d'Internet. Nous étudions en particulier la viralité des informations et la façon dont se forgent et se renforcent les opinions dans le cyberespace, une scène où les contenus sont mis en ligne et lus sans aucun intermédiaire ni contrôle.

Avant de présenter nos résultats sur la diffusion des informations et leur assimilation, sur la formation des opinions et sur la façon dont les personnes s'influencent mutuellement, commençons par souligner quelques traits généraux de la situation créée par Internet et ses réseaux sociaux apparus il y a une dizaine d'années.

Internet a modifié la façon dont les personnes s'informent, interagissent, trouvent des amis, des sujets et des intérêts communs, filtrent les informations et se forgent leurs propres opinions. Dans ce contexte, plusieurs facteurs contribuent au problème de la mésinformation ou de la désinformation.

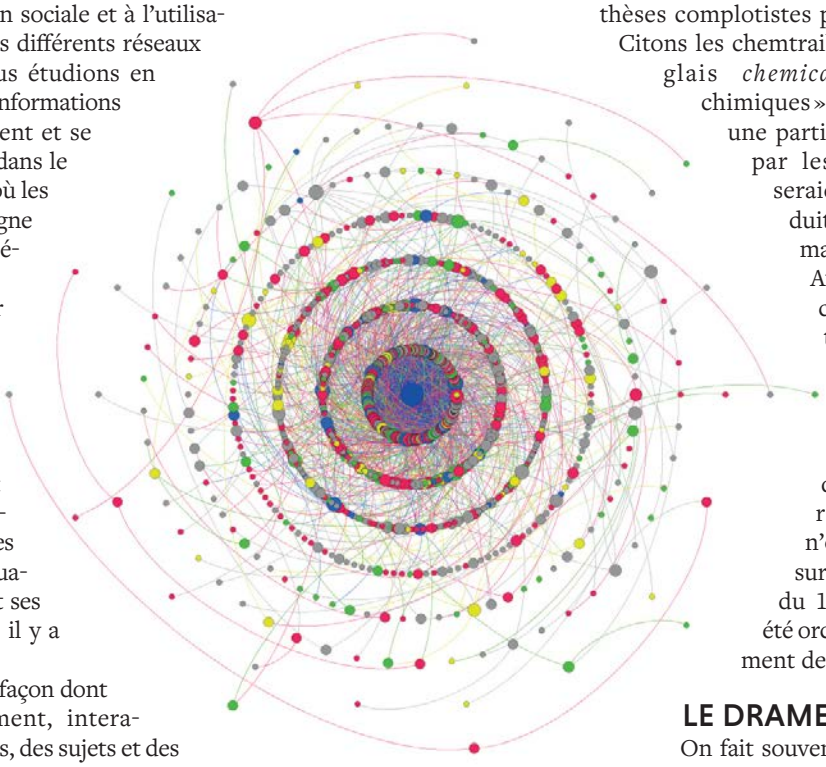
L'un est l'analphabétisme fonctionnel, c'est-à-dire l'incapacité à comprendre convenablement un texte; en France ou en Italie, cela concerne près de la moitié des personnes âgées de 16 à 65 ans, d'après les données de l'OCDE.

Un autre facteur est le «biais de confirmation» selon lequel chacun tend à privilégier les informations qui confirment ses opinions ou sa vision du monde, et à négliger ou ignorer celles qui les contredisent. Dans la masse d'informations de tous types véhiculées par Internet, chacun peut alors rechercher (et trouver...) ce qui le conforte dans ses préjugés et ses goûts, et délaisser le reste.

Un troisième facteur en jeu est le fait que, sur Internet, l'émission et la réception des contenus s'effectuent essentiellement sans intermédiaires. N'importe qui peut publier sa version des faits et ses opinions sur n'importe quoi, sans qu'aucune personne ou autorité n'ait au préalable contrôlé la véracité, ou au moins le fondement, de ce qui a été mis en ligne.

Pour ces raisons, on assiste à des phénomènes de masse impliquant de la mésinformation ou de la désinformation. D'ailleurs, en 2013, le Forum économique mondial, une fondation internationale indépendante qui débat des problèmes les plus urgents de la planète, a cité la diffusion massive de fausses informations comme l'une des menaces les plus graves auxquelles nos sociétés sont confrontées.

Et en effet, les fausses informations ou les thèses complotistes prolifèrent sur la Toile. Citons les chemtrails (contraction de l'anglais *chemical trails*, «traînées chimiques»), thèse selon laquelle une partie des traînées formées par les avions dans le ciel seraient composées de produits chimiques destinés à manipuler les populations. Autres exemples : les vaccins favoriseraient l'autisme ; les sources d'énergie alternatives et gratuites existent, mais sont tenues secrètes par les multinationales afin de protéger leurs intérêts ; les Américains n'ont jamais mis les pieds sur la Lune ; les attentats du 11 septembre 2001 ont été orchestrés par le gouvernement des États-Unis.

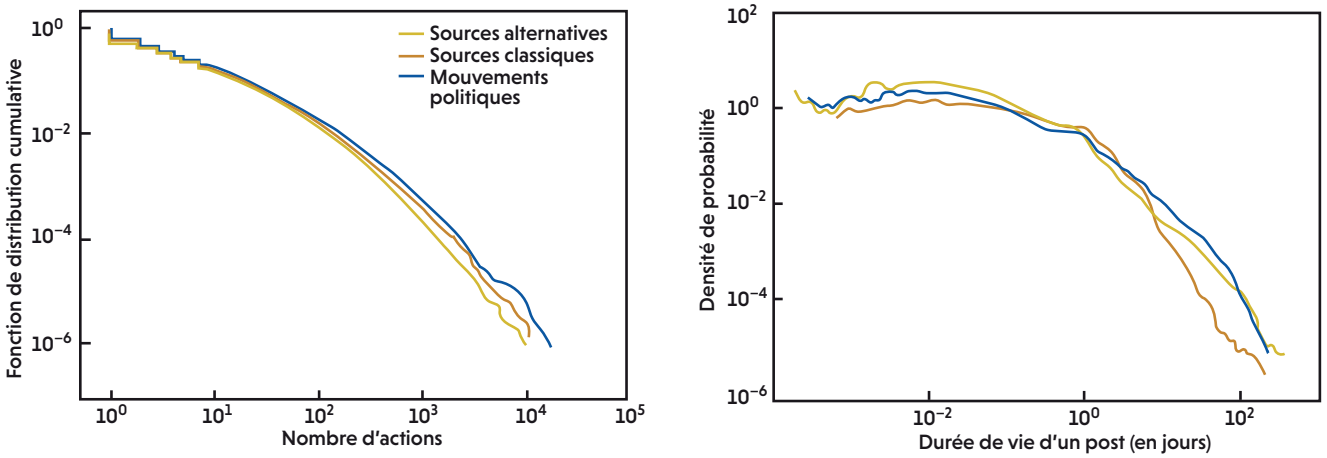


Un canular lancé par un post sur Facebook en 2012 affirmait qu'un certain sénateur italien Cirenga (qui n'existe pas) avait proposé un projet de loi visant à débloquer 134 milliards d'euros pour aider les parlementaires italiens à trouver du travail en cas de non-réélection. Ce canular s'est diffusé massivement sur Facebook en décembre 2012. Ce graphe visualise sa propagation. Les nœuds représentent les internautes, les arêtes représentent la relation de partage du post. Le post d'origine se trouve au centre. Les couleurs indiquent la polarisation de l'utilisateur, c'est-à-dire sa préférence pour un type de contenus : en jaune les utilisateurs qui suivent les sources classiques, en vert les discussions politiques, en rouge les sources alternatives, en bleu les trolls.

LE DRAME DU PIZZAGATE

On fait souvent l'hypothèse que l'être humain est rationnel, mais l'étude quantitative de ces phénomènes indique plutôt le contraire. Dans un environnement où les informations ne subissent aucun filtre, l'individu prend ce qui correspond à son schéma de pensée : c'est le biais de confirmation. Cela alimente les récits les plus disparates, appuyés par de piètres argumentations faisant appel davantage à des associations d'idées qu'à des raisonnements solides. Des récits se répandent alors massivement et exercent une forte influence sur la perception du public concernant des questions essentielles: santé, politique économique, géopolitique, réchauffement climatique...

Cela a parfois d'étranges, sinon de graves, conséquences. Ainsi en va-t-il du «Pizzagate» lancé par le tweet d'un avocat new-yorkais le 30 octobre 2016. Il faisait état d'une prétendue enquête de la police sur un grand réseau de pédophilie auquel Hillary Clinton aurait été liée. La rumeur a enflé sur Internet et a notamment



Le comportement des internautes est très similaire, qu'il s'agisse de posts sur des pages Facebook de sources alternatives d'actualités, sur des pages de sources classiques ou sur des pages de mouvements politiques. À gauche : la « fonction de distribution cumulative » du nombre d'actions (un « J'aime », un commentaire ou un « J'aime » sur un commentaire) des usagers selon le type de pages : pour une abscisse de valeur x, l'ordonnée est la probabilité qu'un post ait un nombre d'actions supérieur ou égal à x. À droite : la densité de probabilité p(x) que le temps écoulé entre le premier et le dernier commentaire d'un post soit égal à x.

mis en cause une pizzeria à Washington. Or le 4 décembre, Edgar Maddison Welch, un internaute pensant probablement faire justice lui-même, s'est rendu à cette pizzeria muni d'un fusil d'assaut, a menacé un employé et tiré dans l'établissement, heureusement sans faire de victimes, avant d'être arrêté par la police.

Afin de mieux comprendre ces phénomènes de désinformation, nous avons étudié en 2015 la consommation sur Internet, en Italie, d'informations qualitativement différentes : celles provenant de sources classiques, celles données par des sources alternatives et celles données par des mouvements politiques.

Les sources classiques désignent ici tous les journaux et agences qui couvrent l'information nationale. Les sources alternatives sont celles qui s'autoproclament promotrices de tout ce que les médias précédents cachent aux gens. Le troisième type de sources relève des mouvements et des groupes politiques qui utilisent Internet comme instrument de mobilisation politique.

Nous avons ainsi choisi 50 pages Facebook publiques, dont 8 sources classiques, 26 sources alternatives et 16 pages d'activisme politique. Puis nous avons téléchargé tous les messages mis en ligne (les posts) et les interactions de leurs utilisateurs respectifs sur une période de six mois, de septembre 2012 à février 2013, qui était une période de campagne électorale. Nous avons pris en compte les posts eux-mêmes, les « J'aime », les commentaires, les partages et les « J'aime » sur les commentaires. Nous avons traité ces données, qui sont publiques, sous une forme agrégée et anonyme. Leur analyse donne un aperçu du comportement de plus de 2,3 millions d'internautes.

Les résultats montrent que, malgré leurs différences qualitatives, les trois types d'informations présentent, dans leur propagation, des propriétés statistiques similaires. Par exemple, la loi statistique qui décrit le nombre de réactions de la part des usagers de Facebook (un « J'aime », un commentaire ou un « J'aime » sur un commentaire) est très similaire pour les trois types de sources (*voir la figure ci-dessus*).

Il en est de même des lois statistiques portant sur l'écho que suscitent les posts. En particulier, pour chacune des trois catégories de sources, la durée moyenne de l'attention prêtée à un post est d'environ 24 heures.

LA CHASSE AUX TROLLS

Lors de cette étude, nous avons aussi examiné sur Facebook le comportement vis-à-vis des trolls, ce terme désignant initialement un message ou un élément de discussion sur Internet dont l'objectif est de perturber cette dernière et de créer une polémique. Stimulée par l'énorme hétérogénéité des groupes et des intérêts qui ont envahi Internet, cette figure a évolué en quelque chose de plus structuré. Là où une dynamique sociale emporte les foules et conduit à des opinions extrêmes sur un sujet donné, il est fréquent qu'une contrepartie apparaisse sous forme de troll et en fasse la parodie. On trouve ainsi des pages qui singent le comportement des membres de mouvements politiques locaux, des pages qui publient toujours la même photo de chanteurs célèbres pour accompagner un quelconque contenu massivement diffusé sur le Web…

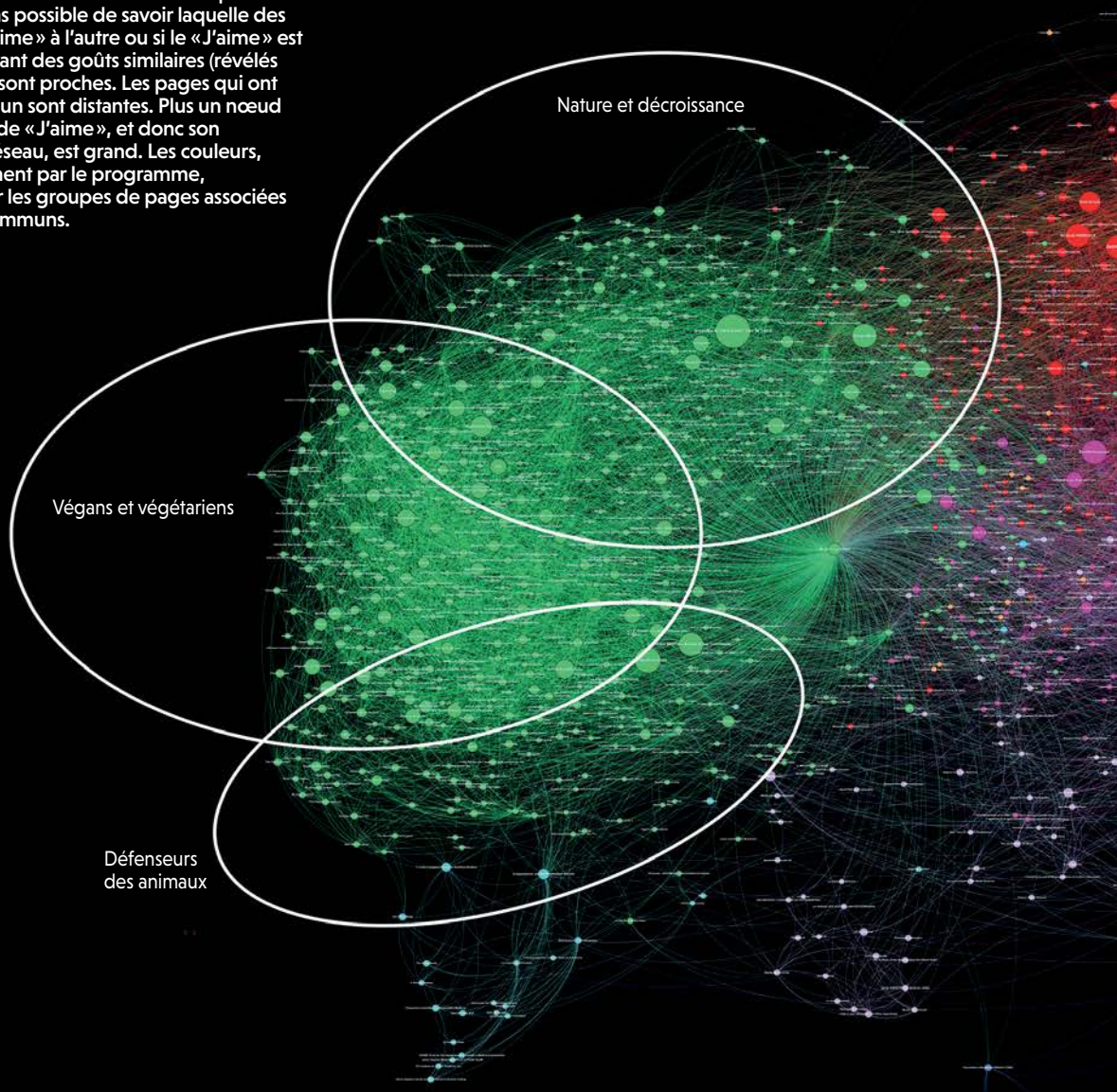
Les théories du complot font aussi partie des divers sujets de moquerie. On a ainsi des trolls qui parlent d'abolir les lois de la thermodynamique au Parlement, ou selon lesquels une récente analyse de la composition chimique des chemtrails prouve la présence de citrate de sildénafil, c'est-à-dire le principe actif du Viagra…

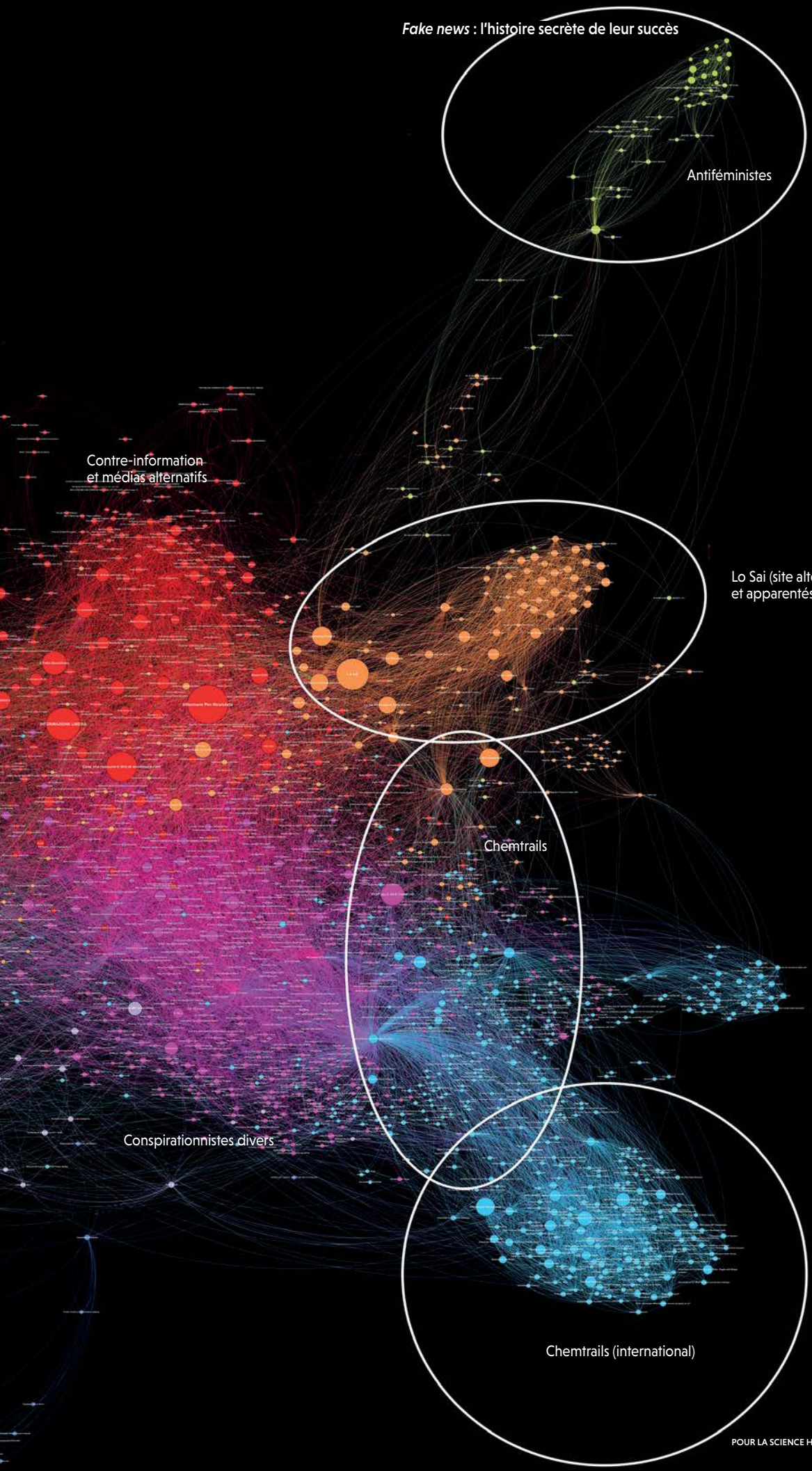
Ces contenus parodiques et caricaturaux ont été essentiels dans notre étude, parce qu'ils nous ont permis de mesurer les capacités de vérification de l'information (le *fact-checking*) des internautes. Étant conçus à des fins parodiques, les trolls sont intentionnellement faux et véhiculent des contenus paradoxaux ; ils révèlent jusqu'à quel point le biais de confirmation est déterminant dans le choix des contenus fait par l'internaute. >

UNE NÉBULEUSE CONSPIRATIONNISTE

Cette carte est un graphe, c'est-à-dire un ensemble de nœuds raccordés par des arêtes. Élaboré avec un programme de visualisation de données, il représente l'immense réseau des pages Facebook italiennes promouvant des théories conspirationnistes. J'ai commencé par un groupe de 17 pages dont l'appartenance au monde conspirationniste et de la désinformation est difficilement contestable, comme *Stop alle scie chimiche* («*Stop aux chemtrails*») ou l'association COMILVA qui lutte contre la vaccination. Des logiciels ont ensuite récupéré le nom de toutes les pages Facebook qui adhèrent à ces premières pages *via* le bouton «J'aime» et, de même, les noms des pages reliées à ces nouvelles pages. Les données récoltées correspondent, dans ce graphe, à 2 612 pages Facebook (les nœuds) et 22 879 liens «J'aime» (les arêtes).
Dans cette version, la direction de la relation n'est pas indiquée. Il n'est donc pas possible de savoir laquelle des deux pages a mis un «J'aime» à l'autre ou si le «J'aime» est réciproque. Les pages ayant des goûts similaires (révélés par les pages «aimées») sont proches. Les pages qui ont peu, voire rien, en commun sont distantes. Plus un nœud est gros, plus le nombre de «J'aime», et donc son «influence» au sein du réseau, est grand. Les couleurs, attribuées automatiquement par le programme, permettent de distinguer les groupes de pages associées à des centres d'intérêt communs.

Les pages à caractère franchement complotiste sont presque toutes au centre (la région fuchsia). En s'éloignant vers l'extérieur, le lien avec les théories du complot tend à disparaître : c'est le cas de plusieurs journaux crédibles (*La Repubblica*, le *Corriere della Sera*...) et d'organisations telles qu'Emergency, qui figurent sur ce graphe uniquement parce qu'ils sont suivis par d'autres pages du réseau.
On observe surtout qu'il n'existe pas de ligne de démarcation nette entre les pages dédiées aux théories du complot, les pages qui publient des canulars et de fausses et tendancieuses nouvelles *(la région rouge)* et l'information véritable et authentique. On retrouve une caractéristique fondamentale des conspirationnistes : leur incapacité à distinguer les sources fiables de celles qui ne le sont pas.




Antiféministes
Contre-information
et médias alternatifs
Lo Sai (site alternatif)
et apparentés
Chemtrails
Conspirationnistes divers
Chemtrails (international)

> Plus précisément, nous avons d'abord classé les utilisateurs de Facebook selon le type d'informations qu'ils préfèrent à partir des mentions « J'aime » qu'ils donnent à chaque type d'informations. Puis nous avons mesuré comment ils réagissaient à un ensemble déterminé d'environ 2800 trolls.

Sans surprise, les adeptes des sources d'informations alternatives sont les plus enclins à réagir aux trolls par un « J'aime » et à les partager, exactement comme ils le font avec les autres informations alternatives. Plus précisément, parmi 1279 utilisateurs ayant une orientation bien définie, 55 % de ceux qui ont cliqué « J'aime » sur les trolls considérés sont des amateurs de sources alternatives, contre 23 % et 22 % respectivement pour les amateurs de sources classiques et ceux de mouvements politiques.

Ce résultat est intéressant, car il met en évidence le « paradoxe de la conspiration » : les internautes les plus attentifs à la prétendue manipulation perpétrée par les médias orthodoxes sont les plus enclins à interagir avec des sources d'informations intentionnellement fausses. En d'autres termes, les plus méfiants vis-à-vis des médias classiques sont aussi les plus enclins à être manipulés !

Comment expliquer que des informations de types différents se propagent avec des lois statistiques semblables ? On peut émettre l'hypothèse qu'il existe des groupes dont l'intérêt se focalise sur des contenus spécifiques, mais que le comportement des internautes est universel.

Cette idée s'accorde bien avec la notion d'exposition sélective aux informations, due au biais de confirmation, et avec l'idée qu'Internet, en facilitant la connexion entre les personnes et l'accès aux contenus, accentue la formation de chambres d'écho. Cette expression renvoie à des communautés d'individus aux intérêts similaires qui sélectionnent le même type d'informations, ne discutent qu'entre eux et renforcent ainsi leurs propres croyances autour d'un récit commun.

Dans un second temps, nous avons comparé le comportement d'internautes s'intéressant à des sources d'informations scientifiques avec celui d'internautes qui suivent les sources d'informations alternatives et conspirationnistes.

Il s'agit de sources de types très différents. Les informations scientifiques ont notamment un auteur bien identifié, un responsable du message. L'information scientifique fait référence à des travaux le plus souvent publiés en détail dans des revues scientifiques professionnelles, dont on connaît les auteurs, leurs institutions… Tout le contraire des informations conspirationnistes, qui se réfèrent à une quelconque machination secrète et fomentée par des individus puissants, des groupes ou des États rarement identifiés.

Une autre différence substantielle, indépendamment de la véracité des informations rapportées, est que les récits sont aux antipodes. Le premier type repose sur un paradigme rationnel qui, presque toujours, recherche des preuves empiriques. Le deuxième se fonde sur des croyances en des liens de causalité d'après lesquels les événements ou phénomènes considérés sont le résultat d'une intention humaine.

La pensée conspirationniste traduit une incapacité à attribuer des événements à des causes aléatoires ou complexes. Martin Bauer, de l'École d'économie et de sciences politiques de Londres, y voit une façon « quasi religieuse » de penser le monde. À l'aube de l'humanité, nos ancêtres attribuaient aux tempêtes une origine divine, aujourd'hui les conspirationnistes font de même en cherchant une explication simple à des phénomènes complexes qui les chagrinent.

Dans notre étude comparative des deux types d'internautes, nous avons d'abord déterminé un ensemble des pages Facebook à explorer et jeté notre dévolu sur 39 conspirationnistes et 34 scientifiques. Notre analyse a porté sur 1,2 million d'utilisateurs sur une période de cinq ans, entre 2010 et 2014.

DES INTERNAUTES POLARISÉS

Que montrent les résultats ? En Italie, les utilisateurs qui suivent les pages Facebook d'informations complotistes sont trois fois plus nombreux que ceux qui suivent des sources d'informations scientifiques. De plus, les deux populations sont très polarisées : ceux qui suivent les deux types de sources sont peu nombreux. Autrement dit, les utilisateurs sortent rarement de leur chambre d'écho.

Les informations, qu'elles soient fondées ou non, sont consommées de façon similaire, mais mutuellement exclusive. Cette caractéristique de l'interaction sociale sur Facebook semble avoir un rôle déterminant dans la diffusion des fausses rumeurs. En examinant 4 709 informations visant à imiter parodiquement ou sarcastiquement les théories du complot et ayant un caractère manifestement absurde, nous avons constaté que ceux qui interagissent avec ces contenus (un « J'aime » ou un commentaire) sont, à environ 80 %, des utilisateurs d'informations conspirationnistes.

Autre constat intéressant : les utilisateurs principalement centrés sur les informations complotistes ont une plus grande propension à partager ces informations que ceux intéressés par les informations de nature scientifique (avec ces mêmes informations).

Cette forte polarisation des utilisateurs se reflète-t-elle aussi dans les « amitiés » virtuelles ? En étudiant de plus près les chambres d'écho, nous avons reconstitué les réseaux sociaux des deux groupes et avons découvert une régularité statistique surprenante : à

EN CHIFFRES (POUR 2016)

Internet

Dans le monde : plus de 3,4 milliards d'internautes, soit 46 % de la population.

En Europe : 616 millions d'internautes, soit 73 % de la population.

En France : 55,4 millions d'internautes, soit 86 % de la population. En moyenne, chacun passe 4,6 heures par jour en ligne (3,5 au Japon, 9,1 au Brésil).

En janvier 2016, 38,6 % des pages web consultées l'ont été *via* des téléphones mobiles contre 28,9 % en janvier 2014. En moyenne, le trafic de données représente 1,4 gigaoctet par mois et par téléphone.

Réseaux sociaux

Dans le monde : plus de 2,3 milliards (31 % de la population mondiale). Leur nombre croît de 10 % par an.

En France : en moyenne, un usager des médias sociaux y consacre 1,3 heure par jour (0,3 au Japon, 3,3 au Brésil).

En France, les plateformes principales sont Facebook (43 %), Facebook Messenger (22 %), Google+ (11 %) et Twitter (11 %).

Parmi les 32 millions d'utilisateurs de Facebook en France, la moitié ont entre 20 et 40 ans.

mesure que le nombre de « J'aime » augmente sur un type de récit spécifique, la probabilité d'avoir un réseau social virtuel composé uniquement d'« amis » ayant le même profil augmente aussi, de façon proportionnelle. Ainsi, plus un internaute polarisé est actif, plus ses amis ont le même profil.

Cette division des réseaux sociaux en groupes homogènes selon le type de contenu consommé aide ainsi à comprendre la viralité des phénomènes sur la Toile. Chaque groupe a tendance à exclure tout ce qui n'est pas en cohérence avec sa vision du monde. Il s'agit ainsi d'une structure collective qui amplifie la sélection des contenus par le biais de confirmation à l'œuvre à l'échelle de chaque individu.

Nous avons ensuite poursuivi nos travaux en examinant les effets des campagnes de désintoxication visant à corriger la diffusion des fausses informations sur les médias sociaux. Nous avons ainsi comparé, parmi les utilisateurs généralement exposés à des sources conspirationnistes, ceux qui ont été exposés à des posts de démystification avec ceux qui ne l'ont pas été. Nous avons notamment mesuré la persistance, c'est-à-dire la probabilité de continuer à accorder des « J'aime » à un type de contenu spécifique dans le temps, pour les utilisateurs exposés ou pas à des campagnes de décodage des fausses informations.

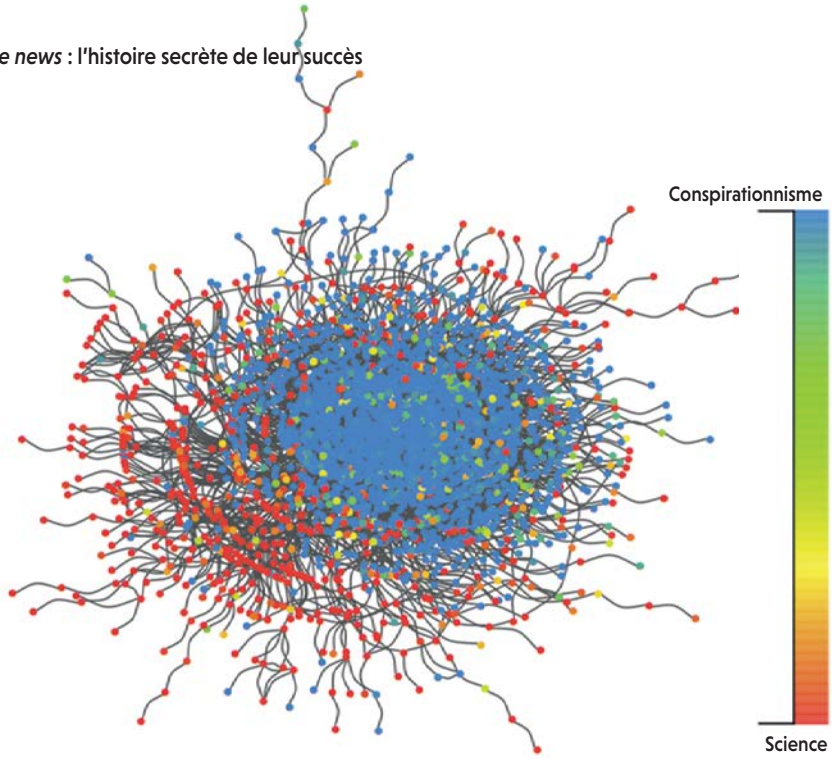
Pour ces utilisateurs exposés à la démystification, la probabilité de continuer à interagir avec des informations conspirationnistes est supérieure d'environ 30 % que pour les autres ! Essayer de convaincre un fervent défenseur de la théorie des chemtrails de son erreur renforce ses croyances et augmente ses interactions avec les sources d'informations de type complotiste.

Pour écarter toute particularité liée au contexte italien, nous avons mené une analyse similaire avec le Facebook américain. On retrouve les mêmes tendances…

DES DISCUSSIONS QUI DÉGÉNÈRENT

Un autre résultat intéressant a été obtenu au moyen d'algorithmes qui, bien entraînés, sont en mesure de fournir avec une bonne approximation le sentiment exprimé par les utilisateurs dans les commentaires des posts. Plus la discussion entre les internautes est longue, plus on se dirige vers un sentiment négatif. Et cela vaut tant pour les informations conspirationnistes que pour les informations scientifiques, même si les internautes conspirationnistes ont tendance à être plus négatifs. Dans tous les cas, une discussion prolongée dégénère.

Les dynamiques sociales qui émergent de nos études révèlent les problématiques relatives à la formation et à la propagation massive, sur les réseaux sociaux d'Internet, de récits potentiellement erronés.



Les internautes qui suivent habituellement les sources d'informations scientifiques ont relativement peu de liens d'« amitié » (au sens de Facebook) avec ceux qui suivent les sources conspirationnistes, comme le montre ce réseau de liens d'amitié reconstitué par l'auteur et ses collègues (en rouge, les utilisateurs plutôt exposés à des sources scientifiques, en bleu, les utilisateurs exposés à des sources conspirationnistes). On visualise ainsi l'effet « chambre d'écho » : les communautés d'internautes ont peu de contacts entre elles.

BIBLIOGRAPHIE

M. DEL VICARIO *ET AL.*, The spreading of misinformation online, *PNAS*, vol. 113(3), pp. 554-559, 2016.

D. MOCANU *ET AL.*, Collective attention in the age of (mis)information, *Computers in Human Behavior*, vol. 51, pp. 1198-1204, 2015.

A. BESSI *ET AL.*, Science vs conspiracy: Collective narratives in the age of misinformation, *PLoS One*, vol. 10(2), e0118093, 2015.

A. BESSI *ET AL.*, Trend of narratives in the age of misinformation, *PLoS One*, vol. 10(8), e0134641, 2015.

A. BESSI *ET AL.*, Social determinants of content selection in the age of (mis)information, Social Informatics, Lecture Notes in Computer Science, vol. 8851, pp 259-268, 2014.

La sélection des contenus par l'internaute s'effectue sous l'influence du biais de confirmation, ce qui conduit à la formation de groupes relativement homogènes de personnes qui s'intéressent aux mêmes thèmes et aux mêmes récits. Parallèlement au renforcement de cette focalisation, les membres de chaque groupe tendent à ignorer tout ce qui ne conforte pas leurs préjugés. Et quand discussion il y a avec des personnes ayant une opinion différente, elle dégénère souvent, ce qui encourage la polarisation.

Dans ce contexte, il devient difficile d'informer correctement, et arrêter la propagation d'une nouvelle infondée ou fausse devient impossible (*voir la figure page 86*). Le problème de la désinformation sur les réseaux sociaux est devenu si prégnant que les opérateurs de ces réseaux sont maintenant souvent accusés de ne pas prendre leurs responsabilités et sont incités à réagir de façon à limiter la propagation des fausses rumeurs et théories. Mais comment ? La question est complexe et les réponses ne font pas encore consensus.

Une partie des solutions pourrait être d'adapter les algorithmes qui gèrent les réseaux sociaux. Par exemple, Facebook a introduit pour les utilisateurs la possibilité de signaler les fausses informations, tandis que Google étudie le moyen de tenir compte de la fiabilité des pages dans le classement des résultats affichés à la suite d'une requête par mots-clés. Mais nos études ne nous invitent guère à l'optimisme.

Nous entendrons probablement pour quelque temps encore parler du complot mondial orchestré par les reptiliens, des extraterrestres d'apparence humaine qui seraient dissimulés parmi nous (et auxquels croient des millions d'Américains). Nous ne sommes pas dans l'ère de l'information mais bien dans celle de la crédulité ! ■

L'ESSENTIEL

- Préoccupés par notre santé, nous sommes prêts à subir de nombreux tests médicaux... qui n'apportent aucune certitude.
- Pour bien interpréter les résultats, médecins, malades et non-malades devraient se familiariser avec les statistiques.
- En apprenant à distinguer risque absolu et risque relatif, à utiliser la fréquence naturelle d'une maladie, les taux de mortalité plutôt que la survie à cinq ans, nous ne nous inquiéterons plus inutilement.

LES AUTEURS

GERD GIGERENZER travaille à l'institut Max-Planck de Berlin.
WOLFGANG GAISSMAIER est professeur à l'université de Constance.
ELKE KURZ-MILCKE travaille à l'Institut de mathématiques et d'informatique, à Ludwigsbourg.
LISA SCHWARTZ et **STEVEN WOLOSHIN** sont professeurs à l'école de médecine de Dartmouth, à Hanovre, aux États-Unis.

SANTÉ : halte à la manipulation

Interpréter les résultats d'examens médicaux et d'études de risques sanitaires est un art difficile auquel peu de personnes sont initiées. Quelques conseils pour mieux comprendre leur signification ne sont donc pas inutiles.

En 1938, dans son essai *World Brain*, l'écrivain britannique H. G. Wells, auteur de *L'Homme invisible* et de *La Guerre des mondes*, prédisait que penser en termes statistiques serait aussi indispensable aux citoyens éduqués d'une démocratie moderne que lire et écrire. Qu'en est-il presque un siècle plus tard, en ce début du XXIe siècle? Presque tous ceux qui vivent dans les sociétés industrielles savent lire et écrire, mais peu savent interpréter correctement les statistiques et comprendre les notions de risque et d'incertitude. C'est aussi le lot de nombreux médecins, journalistes et hommes politiques qui, en conséquence, répandent de fausses idées dans le public.

L'inculture statistique n'est pas due à des déficits intellectuels particuliers – le «gène des statistiques» n'existe pas! – mais à divers facteurs sociaux et psychologiques: dans le domaine de la médecine, la nature paternaliste de la relation médecin-malade, l'illusion que la médecine offre des certitudes, que les interventions médicales sont toujours bénéfiques. L'anxiété et les espoirs des citoyens peuvent être facilement manipulés

Les statistiques médicales sont sujettes à diverses interprétations. Bien les comprendre éviterait des frayeurs inutiles.

pour des raisons politiques et commerciales. Avec des conséquences médicales et psychiques parfois redoutables.

Bonne nouvelle, on peut éviter certaines manipulations statistiques en médecine, donner du sens à des données chiffrées parfois peu claires, et utiliser cette information pour prendre les bonnes décisions. Et surtout, on peut tous sensibiliser les enfants aux statistiques de façon à les aider à résoudre des problèmes concrets.

Cela fait longtemps que la médecine se méfie des statistiques. Pendant des siècles, les thérapies se fondaient sur une confiance mutuelle plutôt que sur des données chiffrées, auxquelles on reprochait d'être impersonnelles ou peu pertinentes. Aujourd'hui encore, certains médecins se fient davantage à leur intuition et à leur propre jugement qu'aux statistiques. De leur côté, nombre de patients préfèrent faire confiance à leur médecin plutôt que d'analyser eux-mêmes les résultats qui les concernent.

Les gens n'aiment pas les statistiques parce qu'ils ont besoin de certitude face à la maladie, tandis que les statistiques obligent à prendre des décisions sans certitude. Ainsi, une étude menée auprès de 1000 Allemands majeurs, en 2006, suggère que la plupart des gens considèrent que les tests de dépistage du VIH et les tests génétiques sont fiables à 100%, ce qui est faux.

De même, alors que la mammographie a réduit le risque de décès par cancer du sein des >

femmes cinquantenaires d'environ 5 à 4 pour 1000 en treize ans, 60% d'un échantillon aléatoire de femmes américaines pensaient que le bénéfice était 80 fois plus élevé.

Les citoyens des sociétés où la technologie est omniprésente sont confrontés à de nombreux dilemmes médicaux. Une femme enceinte âgée de 35 ans doit-elle subir une amniocentèse pour dépister une éventuelle anomalie chromosomique du fœtus, alors que cette procédure présente un risque (de l'ordre de 1%) d'entraîner une fausse couche? Doit-on vacciner les filles contre les papillomavirus humains afin de les protéger contre le cancer du col de l'utérus, alors que quelques complications ont été signalées, notamment un risque potentiel de paralysie?

Pour prendre des décisions éclairées, nous devons comprendre les statistiques médicales. En particulier, nous devons distinguer un risque absolu d'un risque relatif, et interpréter correctement la fréquence naturelle d'une maladie pour en déduire la probabilité d'en être atteints en cas de test positif. Nous devons aussi davantage nous fier aux taux de mortalité plutôt qu'aux statistiques, trompeuses, de survie à cinq ans.

RISQUES ABSOLU ET RELATIF

En octobre 1995, l'Agence de sécurité sanitaire du Royaume-Uni émit un avis selon lequel les pilules contraceptives de troisième génération doublaient le risque de phlébite potentiellement mortelle dans les jambes ou les poumons (un caillot sanguin obstrue une veine); ce risque augmentait donc de 100%. Cette information fut transmise par courrier à 190000 médecins généralistes, pharmaciens et directeurs de services médicaux, et sous forme de messages d'alerte dans les médias. La nouvelle émut tout le pays, et beaucoup de femmes cessèrent de prendre la pilule; l'année suivante, on compta 13000 avortements supplémentaires en Angleterre et au pays de Galles; quelque 800 jeunes filles de moins de 16 ans eurent un enfant. Pourtant, les avortements et les grossesses augmentent le risque de thrombose dans des proportions bien supérieures à celle possiblement liée à la pilule de troisième génération.

Une telle panique aurait pu être évitée avec une meilleure information. En réalité, les données montraient qu'environ 1 femme sur 7 000 prenant une pilule de deuxième génération avait une thrombose; ce chiffre passait à 2 pour 7000 avec la pilule de troisième génération. Ainsi, l'augmentation du risque absolu (*voir les Repères, page 6*) n'était que de 1 pour 7000, alors que celle du risque relatif était, effectivement, de 100%.

Annoncer des risques relatifs peut provoquer des espoirs infondés, aussi bien que des inquiétudes inutiles. Nombre de patients et de médecins évaluent plus favorablement un traitement ou un test si les bénéfices correspondent à une diminution du risque relatif. En 2007, les travaux de Judith Covey, de l'université de Durham, en Angleterre, le montraient: lorsque le bénéfice d'un médicament était présenté sous forme d'une réduction du risque relatif, 91% des généralistes danois le recommandaient à leurs patients. Mais lorsque l'information était présentée sous forme de réduction du risque absolu, seuls 63% recommandaient ce même médicament.

Les brochures d'information, les médecins, les revues médicales et les médias continuent à informer le public en termes de changements relatifs, en partie parce que les chiffres élevés attirent davantage l'attention. La confusion est encore plus grande quand on conjugue bénéfices et risques. Ainsi une publicité a affirmé que le traitement hormonal substitutif recommandé pour compenser le déficit en œstrogènes chez les femmes ménopausées «protège les femmes contre le cancer colorectal (jusqu'à plus de 50%)», tandis que le risque de cancer du sein «pourrait augmenter de 0,6%». En fait, le bénéfice relatif de 50% correspond à un nombre absolu inférieur à 6 pour 1000. En d'autres termes, moins de 6 femmes sur 1000 sont protégées du cancer colorectal par le traitement. Cela signifie que cette thérapie engendre au total plus de cancers qu'elle n'en prévient. Néanmoins, selon une étude de 2003 dans laquelle on distribuait cette brochure à 80 femmes âgées de 41 à 69 ans, 60 en avaient conclu le contraire.

Le risque absolu est plus informatif parce qu'il intègre l'information sur les proportions réelles à partir desquelles sont effectués les calculs. Du risque absolu on peut déduire le risque relatif, mais l'inverse n'est pas vrai. Une réduction de 50% du risque relatif peut décrire une diminution importante de 200 à 100 pour 10000, ou bien minime, de 2 à 1 pour 100000. En médecine, les résultats apportés

L'annonce de risques relatifs peut susciter des espoirs infondés ou des inquiétudes inutiles

LA CERTITUDE N'EXISTE PAS

Les questions à se poser quand on parle d'un risque:

De quel risque s'agit-il? À quoi se réfère le risque: est-ce le risque de mourir d'une maladie, d'attraper une maladie ou d'en présenter les symptômes? Parle-t-on de risque absolu (la probabilité de développer une maladie durant une période de temps donnée) ou de risque relatif (le rapport entre le risque dans le groupe concerné et le risque dans un groupe témoin).

Quelle est l'échelle de temps? On se représente plus concrètement les risques à l'horizon d'une dizaine d'années plutôt que de la vie entière.

Quelle est l'importance du risque? Les nombres devraient être exprimés en termes absolus, ou bien sous une forme comparative, liant le risque à d'autres.

Ce risque me concerne-t-il? Le calcul du risque repose-t-il sur l'étude d'individus au profil similaire (âge, sexe, santé...)?

Une mammographie positive peut provoquer une angoisse considérable. Elle serait pourtant bien moins alarmante si l'on annonçait qu'avec ce résultat, la probabilité d'avoir un cancer du sein est bien inférieure à 100 %.

par les essais cliniques sont particulièrement fiables, mais s'ils sont exprimés de façon inadéquate, le public n'a aucune chance de les interpréter correctement.

DÉMÊLER LE VRAI DU FAUX POSITIF

Prenons le cas d'une femme qui, face au résultat positif d'une mammographie, demande à son médecin si elle a vraiment un cancer du sein, ou quelle est la probabilité qu'elle en soit vraiment atteinte. En 2007, l'un de nous (G. Gigerenzer) a demandé à 160 spécialistes de répondre à une telle question, en tenant compte des données suivantes : la prévalence de cancer du sein dans la région de la patiente est de 1 % ; si une femme a un cancer du sein, la probabilité que le test soit positif (sensibilité) est de 90 % ; si une femme n'a pas de cancer du sein, la probabilité qu'un test soit quand même positif (faux positif) est de 9 %.

Quelle est, parmi les propositions suivantes, la meilleure réponse à donner à la patiente ? (*a*) la probabilité qu'elle ait un cancer du sein est d'environ 81 % ; (*b*) sur 10 femmes ayant une mammographie positive, environ 9 ont un cancer du sein ; (*c*) sur 10 femmes ayant une mammographie positive, environ 1 a un cancer du sein ; (*d*) la probabilité qu'elle ait un cancer du sein est d'environ 1 %.

La meilleure réponse est la troisième : en moyenne, sur 10 femmes dont les résultats sont positifs au dépistage mammographique, une seule environ a effectivement un cancer du sein. Les 9 autres sont alarmées inutilement. Seuls 21 % des spécialistes interrogés avaient choisi la bonne réponse…

Nombre de médecins ne connaissent pas la probabilité qu'une personne soit effectivement malade en cas de test de dépistage positif, c'est-à-dire la valeur prédictive positive de ce test. Ils sont aussi incapables de l'estimer à partir de probabilités dites conditionnelles telles que la sensibilité du test (la probabilité d'un test positif en présence de la maladie) et la spécificité du test (le taux de faux positifs). De telles lacunes risquent d'entretenir des frayeurs inutiles. Or plusieurs mois après avoir reçu un résultat faux positif de mammographie, 1 femme sur 2 signale une anxiété importante liée à ce résultat, et 1 sur 4 rapporte que cette anxiété a affecté son humeur et sa vie quotidienne.

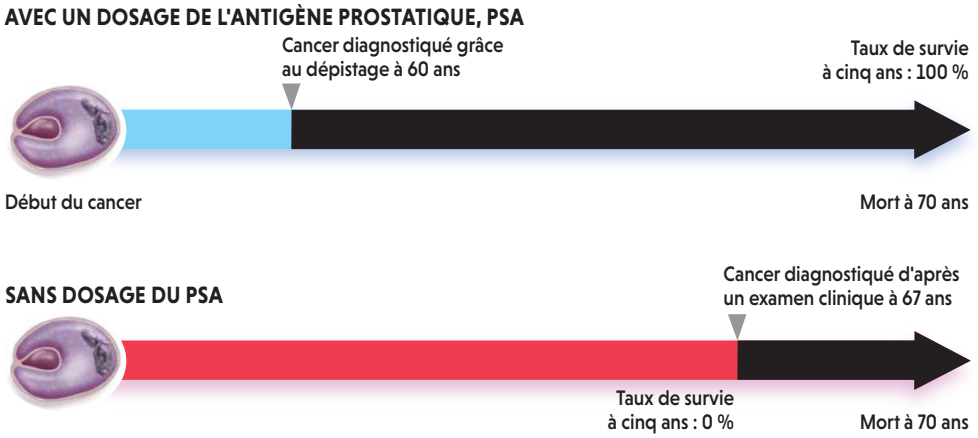
Les médecins seraient plus à même de déduire les probabilités correctes avec des statistiques relatives aux tests présentées sous forme de fréquences naturelles. Par exemple, avec les données de la mammographie évoquées plus haut : 10 femmes sur 1 000 ont un cancer du sein ; sur ces 10 femmes, 9 ont un test positif ; sur les 990 femmes non atteintes, environ 89 ont quand même un résultat positif. Ainsi, 98 patientes (89 + 9) ont un test positif, mais seulement 9 ont un cancer.

Les médecins devraient toujours informer leurs patients qu'aucun test n'est parfait, et que les résultats doivent être interprétés avec prudence, ou qu'ils doivent être réitérés pour voir si l'on obtient à nouveau le même résultat. Toutes les femmes qui passent une mammographie devraient savoir que les résultats indiquant une suspicion peuvent être de fausses alertes.

Une incertitude similaire existe pour tous les tests de dépistage, y compris celui du VIH. Bien que le test détecte effectivement 99,9 % des infections réelles, et que 99,99 % des résultats négatifs soient effectivement négatifs, le très faible risque chez les hommes hétérosexuels implique que le risque qu'ils ont d'être infectés n'excède pas 50 % même quand un test est positif (*voir l'encadré page 97*). Cependant, lorsque le risque intrinsèque est supérieur, comme chez les homosexuels qui ont des rapports sexuels non protégés ou les toxicomanes qui partagent leurs seringues, la probabilité que le sujet soit infecté par le virus en cas de test positif est presque de 100 %. Ainsi, le risque intrinsèque à un groupe donné détermine la signification d'un test positif.

UN INDICE TROMPEUR

Dans un spot télévisé de campagne électorale, en 2007, l'ancien maire de New York Rudy Giuliani annonçait : « J'ai eu un cancer de la prostate, il y a cinq ans. Mes chances de survie étaient de 82 % aux États-Unis. En Angleterre, où le système de santé est socialisé, elles n'auraient été que de 42 %. Dieu merci, j'ai guéri. » L'ex-édile sous-entendait qu'il avait eu de la chance de vivre à New York. Cette déclaration a fait les gros titres de la presse américaine. Pourtant, les >



Les statistiques de survie dépendent du moment du diagnostic, ce qui rend trompeuses les données statistiques. Dans le cas fictif présenté ici, un diagnostic de cancer de la prostate à l'âge de 60 ans *(en haut)* peut faire augmenter le taux de survie à cinq ans par rapport à un diagnostic posé sept ans plus tard *(en bas)*. Pourtant, dans les deux cas, l'âge du décès est le même : 70 ans.

> chiffres qu'il avait donnés trahissaient une grossière erreur d'interprétation !

Selon les données de l'année 2000 qu'il a apparemment utilisées, sur les 49 Britanniques sur 100 000 chez qui un cancer de la prostate avait été diagnostiqué, 28 étaient décédés au bout de cinq ans, ce qui correspondait à un taux de survie à cinq ans d'environ 43 %. Le taux correspondant aux États-Unis était de 82 %, ce qui suggérait que les Américains avaient deux fois plus de chances de survivre que les Britanniques à un cancer de la prostate. Mais cette déduction était fausse, parce que les statistiques de survie reflètent davantage des différences de diagnostic entre les deux pays que des traitements de meilleure qualité.

ÊTRE MALADE QUELQUE PART

Pour le comprendre, imaginons un groupe fictif de patients souffrant d'un cancer de la prostate diagnostiqué, d'après leurs symptômes, à l'âge de 67 ans au Royaume-Uni, et qui meurent tous à 70 ans. Chaque malade n'a survécu que trois ans, et en conséquence le taux de survie à cinq ans de ce groupe est égal à zéro. Prenons un groupe comparable aux États-Unis, où les médecins détectent la plupart des cancers de la prostate en dosant un antigène spécifique (le PSA), alors que ce dosage n'est pas réalisé en routine au Royaume-Uni. Ces patients sont diagnostiqués plus tôt, vers 60 ans, mais ils meurent quand même à 70 ans. Ils ont tous survécu dix ans, et en conséquence le taux de survie à cinq ans est de 100 %. Bien que les taux de survie soient radicalement différents, l'âge du décès est le même dans les deux groupes. Ainsi, en fixant le moment du diagnostic plus tôt, on augmente les taux de survie (biais d'avance au diagnostic) bien qu'aucune vie n'ait été prolongée ou sauvée (*voir la figure ci-dessus*).

Des taux de survie artificiellement élevés peuvent aussi résulter d'un « surdiagnostic », par exemple la détection d'anomalies qui sont des cancers, mais qui n'évolueront jamais assez pour menacer la vie du patient qui décédera d'une autre cause. Supposons 1000 hommes atteints d'un cancer de la prostate évolutif ne bénéficiant pas d'un test de dépistage. Au bout de cinq ans, 440 seront toujours en vie, ce qui représente un taux de survie à cinq ans de 44 %. Dans un autre groupe de 3000 hommes, le dosage du PSA révèle que 1000 ont un cancer évolutif et 2000 un cancer non évolutif (ils ne mourront pas de ce cancer dans les cinq ans). En ajoutant ces 2000 cas aux 440 qui ont survécu au cancer évolutif, on aboutit à un taux de survie à cinq ans gonflé, de 81 %, alors que la mortalité n'a pas été réduite.

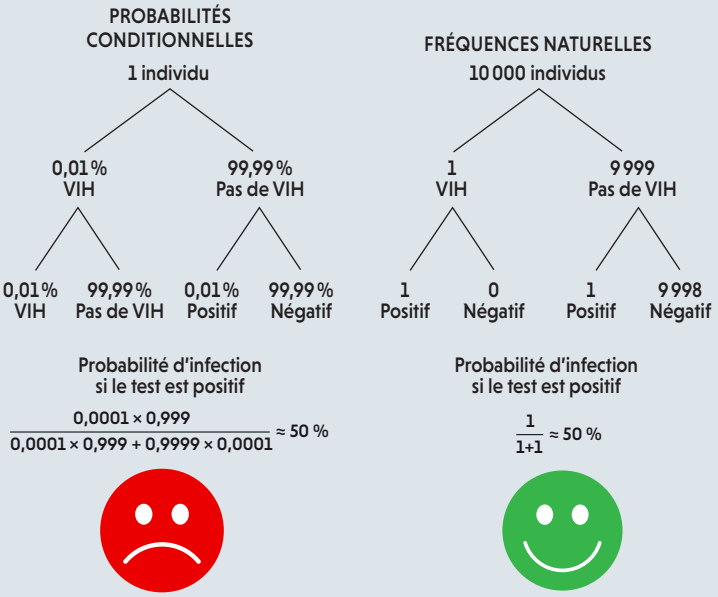
Aux États-Unis, le dépistage du cancer de la prostate par dosage du PSA a conduit à la fin des années 1980 à une explosion du nombre de nouveaux cancers diagnostiqués. Au Royaume-Uni, l'effet a été bien plus faible à cause d'une utilisation moins systématique du dosage de cet antigène. Cette disparité dans les diagnostics explique en grande partie pourquoi le taux de survie à cinq ans pour les cancers de la prostate est plus élevé outre-Atlantique.

Lorsque les pratiques diagnostiques diffèrent d'un pays à l'autre, la différence des taux de mortalité à cinq ans ne reflète pas de façon fiable la différence des taux de mortalité. Et pourtant, nombre d'agences officielles continuent de faire état de taux de survie à cinq ans. Un rapport du Bureau des statistiques du Royaume-Uni notait que le taux de survie à cinq ans du cancer du côlon était de 60 % aux États-Unis contre 35 % au Royaume-Uni. Les experts qualifiaient ce résultat de « scandaleux » et appelaient à un doublement des dépenses gouvernementales pour le traitement du cancer du côlon. En fait, les taux de mortalité dus à ce cancer sont à peu près les mêmes dans les deux pays.

Plus curieux encore, en 2001, une brochure publicitaire du Centre Anderson, au Texas, mélangeait les taux de survie avec les taux de mortalité : « Tandis que le taux de mortalité du cancer de la prostate a fluctué entre 1960 et 1990, le taux de survie au cancer de la prostate des patients du Centre Anderson a continué de progresser. »

PRÉVOIR UNE INFECTION

Si votre test de dépistage du VIH est positif et que vous êtes un homme à faible risque d'infection (hétérosexuel, non drogué…), quel est le risque que vous soyez effectivement porteur du virus ? Les probabilités conditionnelles *(à gauche)* proposent un calcul compliqué. En revanche, en se fondant sur les fréquences naturelles *(à droite)*, on obtient facilement la réponse : sur 10 000 hommes, on s'attend à ce qu'un seul soit infecté et que son test soit donc positif ; parmi les 9 999 qui ne sont pas infectés, 1 devrait aussi avoir un résultat positif. En conséquence, on a deux tests positifs, mais un seul individu infecté. Donc la probabilité d'infection donnée par un test positif n'est pas de 100 %, mais de 50 %.



Les taux de mortalité sont des indicateurs plus fiables de la valeur des programmes de dépistage que les taux de survie à cinq ans. Si l'on se fie à ces taux, un homme doit-il faire systématiquement un dosage de l'antigène PSA ? Un fumeur doit-il passer systématiquement un scanner des poumons ? Il est vrai que ces deux examens détectent plus de cancers à un stade précoce ; mais aucun des deux ne permet de réduire la mortalité.

LES ÉPIDÉMIES DE DIAGNOSTICS

Les gens considèrent souvent les tests de dépistage comme des garants de leur santé. Toutefois, des examens supplémentaires peuvent conduire à des interventions médicales inutiles dont les effets sont parfois délétères. Et pour les nombreux patients inutilement diagnostiqués, le traitement a forcément des conséquences indésirables. Une épidémie de diagnostics peut être aussi dangereuse pour la santé que la maladie.

Les erreurs d'interprétation des statistiques seraient moins fréquentes si les chercheurs, les médecins et les médias utilisaient des données chiffrées directes au lieu de nombres qui prêtent à confusion : le risque absolu au lieu du risque relatif, les fréquences naturelles à la place des probabilités conditionnelles, et les taux de mortalité plutôt que les taux de survie à cinq ans. En outre, nous devons mieux éduquer les jeunes à la science du risque et de l'incertitude.

Comme le suggérait H. G. Wells, les statistiques devraient être enseignées en même temps que la lecture et l'écriture. De fait, aux États-Unis, l'Association nationale des enseignants de mathématiques insiste pour que l'enseignement des statistiques et des probabilités commence à l'école primaire. Si les enfants apprenaient que le monde n'est pas fait de certitudes et ce d'une façon ludique, les statistiques seraient mieux comprises.

Au lieu d'apprendre aux étudiants comment appliquer des formules de probabilité pour résoudre des problèmes virtuels, les professeurs devraient leur montrer comment utiliser les statistiques pour résoudre des problèmes concrets. Par exemple, ils pourraient leur enseigner l'usage des statistiques et des probabilités quand il s'agit de décider comment se comporter face aux drogues, à la consommation d'alcool, à la conduite automobile, aux biotechnologies et à d'autres questions importantes pour la vie quotidienne.

Un livre scolaire américain du secondaire raconte l'histoire vraie d'une mère célibataire de 26 ans. À la suite d'un test positif de dépistage du VIH, elle perd son travail, déménage dans un foyer hébergeant d'autres personnes séropositives, a des relations sexuelles non protégées avec l'un d'entre eux, et attrape une bronchite. Son médecin lui prescrit alors un nouveau test de dépistage. Le résultat est négatif, tout comme celui de son échantillon sanguin précédent, qui a été réanalysé. Cette femme a vécu un cauchemar parce que ses médecins n'ont pas compris qu'un résultat positif à ce test n'était pas un verdict définitif, mais qu'il signifiait que cette femme avait une probabilité d'être infectée de 50 %, étant donné son appartenance à un groupe à faible risque.

L'éducation statistique peut changer des vies, aider les gens à prendre de meilleures décisions personnelles, à reconnaître les messages trompeurs et à développer une attitude plus sereine envers leur santé. Comme le recommandait le philosophe Emmanuel Kant : « Osez savoir ! » ■

BIBLIOGRAPHIE

WOLOSHIN *ET AL.*, *Know your chances : Understanding health statistics*, University of California Press, 2008.

A. PLEASANT, Communiquer sur les statistiques et le risque, *SciDev Net*, 15 décembre 2008.

A. FAGERLIN *ET AL.*, Making numbers matter : present and future research in risk communication, *Am. J. of Health and Behav.*, vol. 31, pp. S47-S51, 2007.

G. GIGERENZER, *Calculated risk : how to know when numbers deceive you*, Simon & Schuster, 2003.

B. FALISSARD, *Comprendre et utiliser les statistiques dans les sciences de la vie*, Masson, Abrégés, 3e éd., 2005.

L'ESSENTIEL

- Cachés derrière un pseudonyme, nous laissons de nombreuses données personnelles sur Internet ou à divers organismes.
- C'est insuffisant pour garantir l'anonymat de ces données !
- Diverses techniques sont développées pour protéger les données sensibles sans empêcher leur exploitation.

LES AUTEURS

TRISTAN ALLARD est postdoctorant à l'Inria (Institut national de recherche en informatique et en automatique).

BENJAMIN NGUYEN est chercheur à l'Inria et maître de conférences à l'Université de Versailles-Saint-Quentin-en-Yvelines.

PHILIPPE PUCHERAL est chercheur à l'Inria et professeur à l'Université de Versailles-Saint-Quentin-en-Yvelines.

L'art de préserver L'ANONYMAT

Presque tous nos comportements laissent des empreintes numériques. Les informaticiens conçoivent et développent des techniques pour que l'analyse de ces gisements de données ne compromette pas la vie privée.

Santé, déplacements, achats, appels téléphoniques, réseaux sociaux, recherches d'information: toutes ces facettes de nos vies laissent des empreintes numériques qui sont stockées et classées dans des bases de données gigantesques, telles celles des moteurs de recherche, qui gardent l'historique des requêtes associées à une adresse IP (la carte d'identité d'un appareil connecté) pendant des années. Ces masses d'informations constituent une manne sans précédent pour les sociétés humaines. Dans le domaine de la santé, par exemple, on peut analyser les caractéristiques individuelles de chaque patient, afin de mieux le prendre en charge, ou de mener des études de santé publique. Si l'analyse de données sur les individus est pratiquée depuis des millénaires, notamment lors des recensements, la quantité et la diversité des informations aujourd'hui disponibles en multiplient l'intérêt potentiel… et les dangers.

En effet, ces myriades d'informations font planer une menace, elle aussi sans précédent, sur la vie privée des individus. La plupart du temps, les données ne sont pas publiques, mais elles circulent néanmoins entre différents acteurs : des ensembles de données médicales sont transmis à des organismes de recherche, les sites en ligne peuvent vendre une partie des données personnelles de leurs utilisateurs… Certaines données sont même accessibles à tous sur Internet : par exemple, le site Netflix a publié des données de ses utilisateurs à l'occasion d'un concours visant à optimiser les algorithmes qui déterminent les films à recommander.

Or de nombreuses études montrent qu'il est souvent possible d'identifier la personne associée à un jeu de données, même quand celui-ci ne contient ni son nom, ni ses coordonnées. Des pans entiers de la vie privée peuvent être dévoilés, avec des préjudices multiples : discrimination au crédit bancaire ou à l'assurance selon l'état de santé, discrimination à l'emploi selon l'orientation sexuelle ou le groupe ethnique… Protéger les données personnelles sans empêcher leur exploitation est devenu un enjeu majeur à l'ère du tout-numérique. La diffusion de ces données est un art de l'équilibre, où l'on recherche le meilleur compromis entre utilité des données et protection des individus.

UN BESOIN D'ÉQUILIBRE

Le souci de la protection des données s'est accru pendant la seconde moitié du XX^e^ siècle, à mesure que l'informatique se généralisait. Il a fait naître un domaine de recherche spécifique dans les années 1970, visant à garantir un anonymat plus ou moins complet aux titulaires des données. Le phénomène s'est encore accéléré depuis le début des années 2000, avec l'essor d'Internet, qui permet de consulter et de croiser instantanément ou presque de multiples sources d'information (réseaux sociaux, annuaires…). Lors de la publication de données, il est essentiel de prendre en compte les connaissances annexes accessibles à un individu mal intentionné. En outre, les analystes se contentent de moins en moins de statistiques sur les données personnelles et demandent d'accéder directement à celles-ci, afin d'augmenter la précision de leurs études.

La conjonction de ces deux facteurs a exacerbé la nécessité d'une protection robuste. Au cours des dix dernières années, de nombreux chercheurs ont étudié la façon de publier des données tout en préservant la vie privée des individus concernés. Ils ont développé des méthodes dites d'assainissement des données, qui consistent à dégrader leur précision de façon contrôlée, afin de réduire la probabilité que l'on puisse réidentifier la personne correspondante ou accéder à des informations sensibles la concernant. >

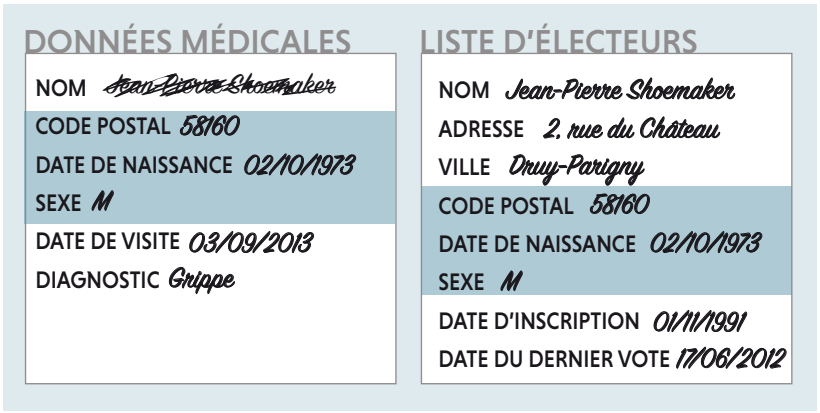
> Comment rendre anonyme un jeu de données? Le moyen le plus ancien et le plus intuitif est de remplacer le nom de l'individu concerné par un pseudonyme ou un nombre – on parle de pseudonymisation. C'est la méthode qu'a utilisée l'entreprise AOL, en août 2006, lorsqu'elle a mis en ligne 20 millions de requêtes adressées à son moteur de recherche par 650000 utilisateurs sur une période de trois mois (pour les rendre disponibles aux chercheurs). Les noms des utilisateurs étaient remplacés par des nombres aléatoires, mais les mots-clés des recherches étaient laissés en clair. Quelques jours plus tard, des journalistes du *New York Times* ont réussi à identifier la personne portant le pseudonyme 4417749 en croisant les mots-clés de ses requêtes avec des données disponibles sur le Web.

ASSAINIR LES DONNÉES

De nombreux cas similaires illustrent l'échec de la pseudonymisation à rendre les données anonymes. Remplacer les seuls champs directement identifiants, tels que le nom et le prénom ou le numéro de sécurité sociale ne prémunit pas contre une réidentification ultérieure (*voir la figure ci-dessus*). Pourtant, la pseudonymisation reste aux yeux de la loi une forme acceptable d'anonymisation, sans doute parce que ses failles sont mal comprises par les législateurs.

Au début des années 2000, Latanya Sweeney, de l'université Carnegie-Mellon, aux États-Unis, a proposé une méthode, nommée *k*-anonymat, pour prévenir les réidentifications *via* des croisements de données. Elle a d'abord montré qu'il était parfois possible de retrouver le titulaire d'un jeu de données médicales pseudonymisé à partir des champs {Sexe, Date de naissance, Code postal}. En effet, ces champs sont aussi présents dans d'autres jeux de données comprenant le nom du titulaire. En outre, la combinaison des renseignements correspondants est unique pour la plupart des gens, selon plusieurs études récentes.

Pour empêcher les réidentifications, Latanya Sweeney a proposé de scinder les champs du jeu de données en deux catégories: les champs quasi identifiants (notés QID), dont la combinaison de valeurs peut être unique, et les champs sensibles (SD), qui doivent rester privés (par exemple, un diagnostic médical). Les valeurs des quasi-identifiants d'un individu sont ensuite rendues identiques à celles d'au moins $(k - 1)$ autres individus dans le jeu de données, pour un entier k convenablement choisi. En d'autres termes, le *k*-anonymat dissimule chaque individu dans une foule d'au moins k personnes impossibles à distinguer par leurs quasi-identifiants. Tout croisement avec une autre source de données aura ainsi une précision inférieure à $1/k$.



Le titulaire d'un dossier médical dépourvu de coordonnées précises *(à gauche)* peut souvent être retrouvé en croisant plusieurs jeux de données. En effet, la plupart du temps, la combinaison {Code postal, Date de naissance, Sexe} *(en rouge)* est unique. Il suffit alors de trouver un autre jeu de données, par exemple une liste électorale *(à droite)*, qui la contient et qui renferme aussi les coordonnées de la personne pour identifier cette dernière.

Les algorithmes du *k*-anonymat se fondent sur le principe de généralisation, qui consiste à remplacer les valeurs précises des quasi-identifiants par des ensembles de valeurs. Ceux-ci peuvent être des intervalles numériques ou des catégories. Il serait simple d'élaborer un algorithme qui parcoure les quasi-identifiants et forme un ensemble les incluant tous, tel {âge de 0 à 150 ans, habitant sur la Terre}. Mais les données seraient trop dégradées (c'est-à-dire trop imprécises) pour être utiles aux analystes. On a alors développé des algorithmes dits par partitionnement: ils forment des ensembles d'au moins k individus en regroupant les quasi-identifiants voisins, qu'ils remplacent ensuite par un intervalle ou une catégorie les incluant tous. Le plus utilisé est l'algorithme dit de Mondrian (*voir l'encadré page 102*).

Cependant, le *k*-anonymat ne résout pas tous les problèmes. Considérons la situation suivante: un attaquant dispose d'un jeu de données *k*-anonymes et recherche la valeur sensible (tel le diagnostic médical) d'un individu cible dont il connaît le quasi-identifiant. Il peut retrouver le groupe auquel appartient sa cible, et donc l'ensemble des valeurs sensibles de ce groupe. La protection apportée par le *k*-anonymat est d'autant plus faible que le nombre de ces valeurs est petit. Dans le cas extrême, supposons que les k individus du groupe partagent la même valeur sensible, par exemple un même cancer. L'attaquant sait alors que sa cible souffre de ce cancer: la valeur sensible n'est pas protégée.

BROUILLER LES PISTES

De multiples techniques ont succédé au *k*-anonymat. Elles cherchent à imposer une diversité minimale aux valeurs sensibles de chaque groupe, afin de limiter ce que peut apprendre un attaquant muni de divers renseignements sur sa cible. Certaines tentent d'estimer ces renseignements, afin de prendre en compte des attaques potentielles réalistes.

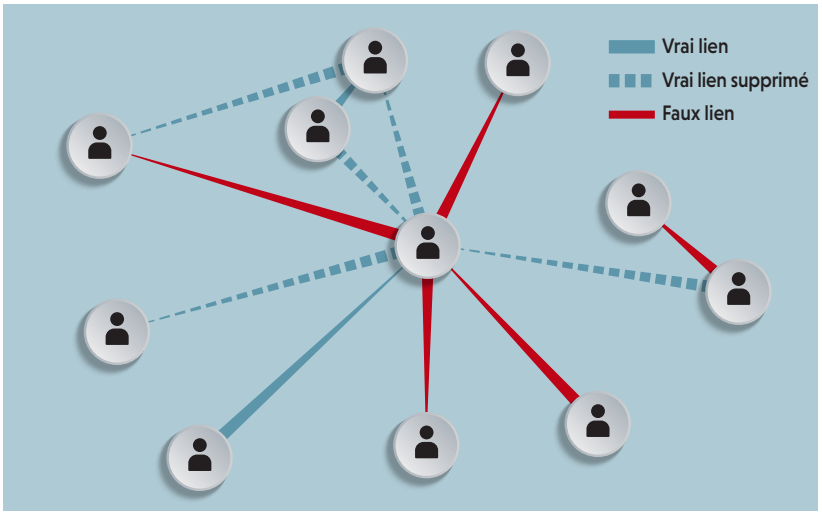
Ainsi, le modèle de Confidentialité bayésienne optimale, proposé en 2006 par Ashwin Machanavajjhala, de l'université Cornell, à New

York, et ses collègues, quantifie la connaissance préalable qu'a l'attaquant de sa cible par la probabilité de trouver la valeur convoitée avant l'observation du jeu de données: si l'attaquant recherche le diagnostic médical d'une personne nommée Dupont, qu'il ne connaît de sa cible que son nom et qu'il sait que 10% des Dupont ont un cancer, il a 1 chance sur 10 de ne pas se tromper en affirmant que sa cible a un cancer. Cette connaissance est dite *a priori*. Après la découverte des données, l'attaquant a plus de chances de trouver la donnée qu'il cherche; la connaissance de sa cible qui en résulte, connaissance dite *a posteriori*, est quantifiée par la probabilité de lui associer la bonne valeur sensible.

C'est ici que la théorie bayésienne intervient, car elle permet de calculer des probabilités conditionnelles: les probabilités de déduire telle ou telle information d'un jeu de données sachant qu'on a telle ou telle connaissance préalable. La confidentialité des données publiées est caractérisée par la différence entre connaissance *a posteriori* et connaissance *a priori*, autrement dit par la connaissance supplémentaire sur la cible apportée par les données. Assurer un niveau adéquat de protection revient à limiter cette valeur. Cependant, cette technique est difficile à mettre en œuvre, car elle suppose de connaître exactement ce que l'attaquant sait de sa cible.

La l-diversité, méthode conçue la même année et par les mêmes chercheurs, est davantage applicable. Elle consiste à imposer la présence de l valeurs sensibles distinctes dans chaque groupe (*voir l'encadré page 102*). Un attaquant qui ne connaît de sa cible que le groupe auquel elle appartient ne peut trouver sa valeur sensible avec une probabilité supérieure à $1/l$.

De nombreuses autres techniques ont été élaborées depuis, afin de s'adapter à la diversité des données, des attaques et des attaquants. La «t-proximité» vise à créer des groupes au sein desquels la distribution des données est à peu près la même que dans la population globale.



Les graphes de réseau social représentent les individus par des points et leurs liens par des traits. Pour qu'ils ne dévoilent pas la vie privée des individus, on peut leur appliquer un algorithme dit de confidentialité différentielle, consistant, par exemple, à supprimer de nombreux vrais liens et à les remplacer par autant de faux liens.

La «(c, k)-sûreté» prend en compte la capacité de l'attaquant à formuler k déductions logiques du type: «Si Adrien a la grippe, alors Bettie aussi.» La «3D-confidentialité» considère que l'attaquant connaît *a priori* trois types de données: l valeurs sensibles que sa cible n'a pas, k valeurs sensibles que d'autres individus ont et m déductions logiques entre individus. La «m-invariance» protège contre les comparaisons entre les publications successives d'un même jeu de données, telles certaines informations statistiques sur un hôpital, dont les variations reflètent les arrivées, les départs et les évolutions des patients.

L'application de ces techniques commence souvent par celle de l'algorithme de Mondrian. Il sert à construire des groupes k-anonymes, au sein desquels on vérifie la conformité de la distribution des données sensibles vis-à-vis du modèle choisi (*voir l'encadré page 102*).

Volontairement ou non, tous ces modèles sont des exemples du paradigme de non-information, formulé en 1977 par le statisticien suédois Tore Dalenius: selon ce paradigme, un jeu de données est d'autant plus confidentiel qu'il renseigne peu l'attaquant sur sa cible. En 2006, Cynthia Dwork, alors chercheuse à Microsoft, a proposé une nouvelle façon de définir la confidentialité, qualifiée de confidentialité différentielle: un jeu de données serait confidentiel s'il n'est presque pas modifié par l'ajout des données d'un individu, quelles qu'elles soient.

NOYER LES DONNÉES

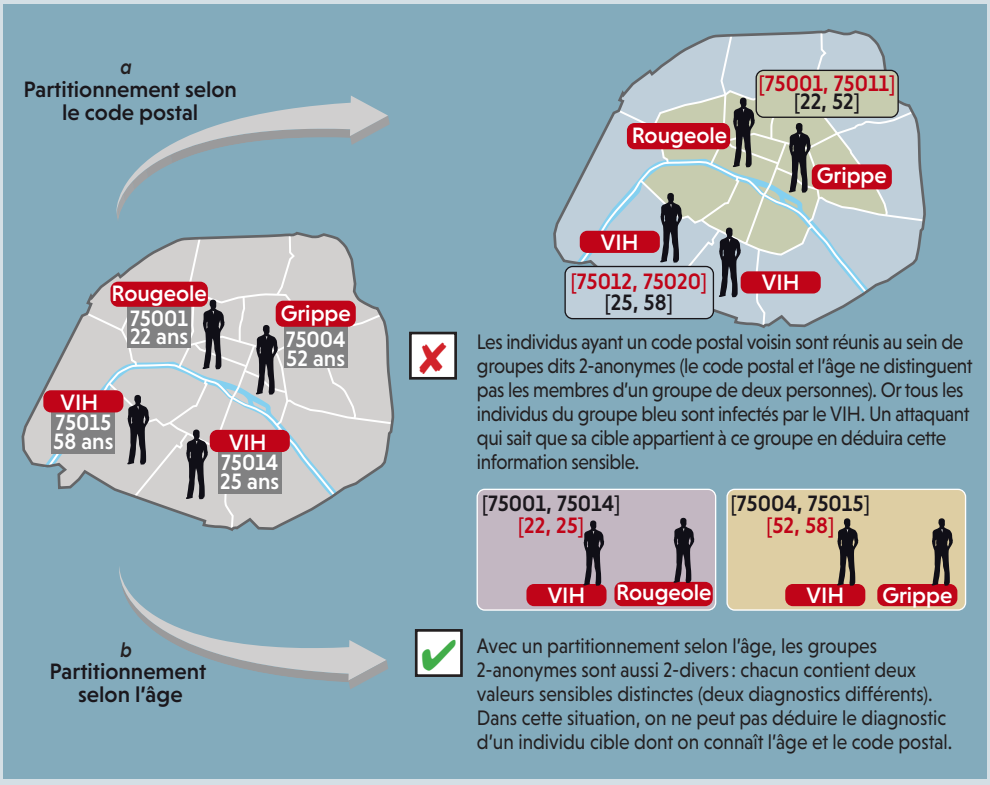
L'assainissement consiste alors à perturber les données de sorte que celles de chaque individu se trouvent noyées dans la perturbation (et non plus dans la foule, comme dans le k-anonymat). On introduit, par exemple, un grand nombre de fausses données (typiquement 100 fois plus que de vraies), afin que le jeu de données qui contient la contribution d'un individu particulier ressemble beaucoup à celui qui ne la contient pas. Certains algorithmes suppriment aussi de vraies données. La protection est assurée par l'impossibilité de distinguer les fausses données des vraies dans le jeu de données assaini, dont on ne peut même pas garantir qu'il contient la contribution d'un individu précis.

L'idée de la confidentialité différentielle est en plein essor. Elle est désormais applicable à des types variés de données, tels les graphes de réseaux sociaux. Ces graphes représentent les individus par des nœuds, connectés par des liens. La perturbation peut alors consister à supprimer de vrais liens et à en introduire de faux. Des informations tel le nombre de liens sont toujours extractibles du graphe, sans que l'on puisse déterminer les connexions précises d'un individu.

Le succès de la confidentialité différentielle s'explique en partie par sa simplicité: l'attaquant >

MONDRIAN ET L'ANONYMISATION

Pour compliquer voire empêcher l'identification d'un individu à partir de ses données, une des meilleures solutions est l'algorithme dit de Mondrian. Pourquoi une telle référence au peintre néerlandais ?



Pour rendre le titulaire d'un jeu de données impossible à identifier, on se contente souvent de remplacer les champs tels que le nom ou le numéro de sécurité sociale par un pseudonyme. C'est insuffisant. D'autres techniques ont alors été développées, dont celle du *k*-anonymat, la plus employée aujourd'hui. Elle consiste à brouiller certains champs, dits quasi-identifiants parce que leur combinaison est parfois unique : l'âge, le sexe, le code postal… On remplace leurs valeurs précises par des intervalles de valeurs ou des catégories, de façon à ce que la combinaison résultante soit partagée par au moins *k* personnes.

Pour appliquer cette technique, on utilise souvent l'algorithme de Mondrian, élaboré en 2006. Son nom renvoie au partitionnement de l'espace cher au peintre hollandais Piet Mondrian. De même, l'algorithme partitionne l'espace des quasi-identifiants, où ces derniers sont indiqués sur des axes et où les individus sont repérés par des points. L'algorithme commence par choisir un quasi-identifiant, selon lequel il divise les individus en deux groupes, par exemple ceux dont le code postal est inférieur ou égal à 75011, et les autres *(a)*. Si *k* est supérieur à 2, les groupes résultants sont de nouveau divisés en deux pour former quatre groupes. Ce découpage continue jusqu'à ce qu'aucun groupe ne puisse plus être divisé sans englober moins de *k* individus. La dernière étape consiste à remplacer les quasi-identifiants de chacun par l'ensemble de valeurs couvert par son groupe.

Le problème est que les données personnelles sensibles à protéger, tel le diagnostic médical, peuvent être identiques au sein d'un groupe. Un attaquant qui sait que sa cible appartient à ce groupe a alors accès à sa valeur sensible. Plusieurs techniques visent à y remédier. La *l*-diversité, par exemple, consiste à construire des groupes ayant au moins *l* valeurs sensibles. On choisit la valeur de *l* en fonction de ce que l'attaquant sait de sa cible, c'est-à-dire du nombre maximal de valeurs sensibles qu'on l'estime capable d'éliminer. Plus il sait de choses, plus *l* sera élevé, plus il faut inclure d'individus dans chaque groupe afin d'avoir une diversité suffisante, et moins les statistiques calculées à partir du jeu de données assaini sont fines. Le choix de *l* traduit un compromis entre utilité et protection.

En pratique, on construit souvent les groupes grâce à l'algorithme de Mondrian. À chaque division d'un groupe en deux, on vérifie que les nouveaux groupes contiennent au moins *l* valeurs sensibles distinctes. Dans le cas contraire, on revient en arrière et on partitionne le groupe selon un autre quasi-identifiant *(b)*.

> n'apparaissant pas dans les algorithmes, on évite les difficultés liées à l'estimation de ce qu'il sait de sa cible, et la distinction entre quasi-identifiants et données sensibles, parfois peu évidente, n'est pas nécessaire. Mais cette simplicité est à double tranchant: des travaux publiés en 2011 ont montré que la non-prise en compte des connaissances annexes de l'attaquant et des relations potentielles entre les individus d'un jeu de données ouvre des failles dans la protection.

Un modèle d'assainissement universel des données reste donc chimérique. Chaque modèle a ses défenseurs et ses détracteurs, et est plus ou moins efficace selon la situation.

Outre le modèle d'assainissement, l'architecture de gestion des données a une importance. Les données nominatives sont souvent extraites du système informatique assurant leur usage quotidien (tel le serveur d'un centre de soin), puis copiées sous une forme pseudonymisée dans un entrepôt de données. C'est à partir de cet entrepôt que seront produits à la demande des jeux de données assainis pour différents destinataires. En France et en Angleterre, où le système de dossier médical personnel national fonctionne ainsi, les épidémiologistes peuvent par exemple recevoir des jeux de données plus précis que les industriels pharmaceutiques.

Toutefois, ce principe d'assainissement centralisé nécessite une fiabilité totale du gestionnaire de l'entrepôt de données. Or on constate régulièrement des fuites d'information sur de nombreux serveurs, dues à des négligences ou à des attaques. Même si les données stockées dans les entrepôts sont pseudonymisées, nous avons vu que cela n'apporte pas une protection suffisante. Il est donc légitime de s'interroger sur les risques de la centralisation.

L'ASSAINISSEMENT DISTRIBUÉ

Les statisticiens ont proposé un mécanisme d'assainissement décentralisé dès les années 1960, pour parer aux réticences à leur confier des données personnelles sensibles. Le principe est de perturber les données de chacun au moment de leur collecte, ce qui les protège avant tout enregistrement.

Cependant, on sait aujourd'hui qu'une perturbation indépendante de chaque donnée ne permet pas d'atteindre le niveau de qualité d'un assainissement réalisé sur le jeu de données dans son ensemble. La centralisation des données personnelles est-elle alors nécessaire à un assainissement de qualité? Non, car il est aussi possible d'élaborer des mécanismes décentralisés où les perturbations ne sont pas indépendantes. En d'autres termes, la perturbation à appliquer n'est pas décidée localement, mais par une entité centrale. Celle-ci possède des informations sur toutes les réponses, sans connaître les réponses elles-mêmes.

Des algorithmes regroupés sous le nom de Secure Multi-Party Computation, qui font souvent intervenir des techniques de cryptographie, visent à assurer la confidentialité de l'assainissement distribué (où les calculs sont répartis entre plusieurs acteurs).

Les mécanismes distribués constituent un pas majeur vers la sécurisation du processus d'assainissement, mais leur mise en œuvre pose des problèmes de passage à l'échelle, car ils nécessitent des calculs importants et une forte connectivité des participants. Ils sont pour l'instant relégués au traitement de jeux de données peu volumineux.

Face à ce problème, notre équipe a développé une architecture de gestion de données personnelles fondée sur des dispositifs individuels, telles des clés USB, sécurisés. Nous avons montré que les calculs nécessaires aux algorithmes d'assainissement peuvent être répartis entre ces dispositifs et une entité centrale sans que celle-ci ne dispose des données personnelles non chiffrées. La puissance de calcul de l'entité centrale assure l'applicabilité à grande échelle et, grâce au fait qu'elle coordonne les dispositifs personnels, ceux-ci n'ont pas besoin d'être interconnectés en permanence. Cette architecture est capable d'appliquer des algorithmes d'assainissement à des jeux de données massifs, correspondant à plusieurs millions d'individus.

Une telle architecture décentralisée pourrait assurer la gestion des dossiers médicaux, des factures ou de tout autre type de dossier personnel. En pratique, les données seraient stockées localement sur les appareils de l'utilisateur et ne seraient jamais exportées sous une forme non perturbée ou non chiffrée. Aucun serveur ne les regrouperait toutes, mais les échanges entre l'entité centrale et les appareils des utilisateurs permettraient tout de même de collecter certaines informations, en respectant la vie privée.

Pendant la dernière décennie, une grande diversité de modèles et d'algorithmes d'assainissement de données ont été élaborés, afin de mieux prendre en compte les capacités de l'attaquant. Aujourd'hui, la pseudonymisation reste massivement utilisée, alors qu'elle ne garantit pas une protection suffisante; dans le reste des cas, la *k*-anonymisation est le plus souvent appliquée, et les techniques plus complexes sont encore exotiques. Ces techniques doivent donc continuer à se répandre et à se perfectionner.

Pour autant, l'assainissement reste le meilleur compromis entre utilité et confidentialité des données, dont la protection absolue est illusoire. L'analyse des nouveaux gisements de données personnelles pouvant apporter des bénéfices notables, l'assainissement de données est une question sociétale avant d'être un défi scientifique. ■

BIBLIOGRAPHIE

T. ALLARD *ET AL.*, METAP: Revisiting privacy-preserving data publishing using secure devices, *Distributed and Parallel Databases*, pp. 1-54, 2013 (à paraître).

B. C. CHEN *ET AL.*, Privacy-preserving data publishing, *Found. and Trends in Databases*, vol. 2, n° 1-2, pp. 1-167, 2009.

A. GREENFIELD, *Everyware. La révolution de l'ubimédia*, Pearson, 2007.

C. DWORK, Differential privacy, Proceedings of the 33rd international conference on Automata, *Languages and Programming*, vol. 2, pp. 1-12, 2006.

L. SWEENEY, *k*-Anonymity: A model for protecting privacy, *Int. J. Uncertain, Fuzziness Knowl.-Based Syst.*, vol. 10(5), pp.557-570, 2002.

L'Open Security Foundation, organisation non gouvernementale américaine, recense les cas les plus significatifs sur son site Internet DataLossDB.org

NOZHA BOUJEMAA

« Les algorithmes sont-ils transparents et éthiques ? Pour nous en assurer, nous avons besoin d'outils adaptés. »

Un algorithme peut-il être juste ?

Nozha Boujemaa : Pour répondre, un référentiel est indispensable. C'est précisément une des difficultés que l'on rencontre lorsqu'on s'intéresse à la qualité des algorithmes. La question à se poser est de savoir : « Pour qui ? ». La notion d'algorithme juste, ou loyal (d'une façon générale, on parlera de transparence), est différente selon que l'on se place du point de vue du consommateur du service adossé à l'algorithme ou du producteur de ce service.

Le consommateur individuel est un citoyen, et de fait les questions de transparence algorithmique ont souvent été soulevées par des affaires de libertés individuelles et de droits civiques. Quand l'utilisateur est une entreprise ou bien même les services de l'État, les problèmes possibles sont ceux, d'une part, de la concurrence déloyale, des abus de position dominante et, d'autre part, de souveraineté nationale.

BIO EXPRESS

17 JUILLET 1963
Naissance

2010-2015
directrice du centre de recherche Inria Saclay Ile-de-France.

2017
Directrice du projet *TransAlgo*, de l'Inria et de l'institut Dataia.

La question de la transparence est essentiellement celle de l'asymétrie informationnelle, entre celui qui propose un service logiciel et celui qui l'utilise, le second en sachant beaucoup moins que le premier. La réduction de cette asymétrie est cruciale : on doit savoir comment nos données sont calculées et par où elles transitent. Notons que la discrimination ou les biais potentiels ne sont pas toujours intentionnels.

Ces problématiques sont le cœur du projet *TransAlgo* que vous coordonnez. Comment est-il né ?

Nozha Boujemaa : Cette initiative fait suite à la loi pour une République numérique (promulguée le 7 octobre 2016) proposée par l'ancienne secrétaire d'État au numérique Axelle Lemaire, ainsi qu'au rapport commandé au Conseil général de l'économie (CGE) sur la gouvernance des algorithmes. La loi prévoit

notamment le droit à l'explication pour tout citoyen quant aux décisions algorithmiques prises par les services de l'État. C'est une avancée notable qui fait de la France un précurseur sur ces sujets aux niveaux européen et international.

En termes de protection des libertés individuelles, elle complète le Règlement général sur la protection des données (RGPD) adopté par le Parlement européen le 14 avril 2016. Ce dernier porte surtout sur les données personnelles, la portabilité, alors que la loi numérique s'intéresse à l'explication des décisions, celles-ci ne s'appuyant pas nécessairement sur des données personnelles. Le droit à l'explication est tout à fait original et anticipe bien l'évolution de la société où les algorithmes vont être de plus en plus amenés à prendre des décisions.

C'est une question de confiance dans l'État, et ses services, au moment où il entame sa transformation numérique. Il s'engage à cette occasion sur la transparence et l'explication des décisions algorithmiques. Il pourra alors répondre à des questions du type : « Pourquoi mon voisin et moi avons-nous été traités différemment alors que nous sommes dans la même situation ? »

Très tôt dans la genèse de la loi numérique, la nécessité de l'accompagner par un volet scientifique s'est imposée. En d'autres termes, on a pris conscience des difficultés techniques qui allaient se dresser devant le législateur. De fait, l'explicabilité des algorithmes (expliquer le cheminement de la prise de décision et sa traçabilité) n'est pas une question résolue, tant s'en faut. De plus, une des recommandations phare, du CGE était la création d'une plateforme scientifique dédiée à la régulation et à la gouvernance des algorithmes. C'est là qu'entre en scène Inria, sollicité pour aider à la montée en compétence des services de l'État en ce domaine.

Mais il a d'abord fallu se mettre d'accord sur les objectifs ?

Nozha Boujemaa : De notre point de vue, il est préférable de ne pas parler de gouvernance ni de régulation des algorithmes, parce que c'est un peu utopique… Ce n'est pas le bon objectif à se donner. Je crois que nous avons été compris, même si chez nos interlocuteurs la tentation a été grande, je peux vous l'assurer, d'imaginer une autorité des algorithmes, un système de certification de transparence, avec des labels…

Ce n'est pas notre rôle en tant qu'organisme de recherche : chacun son métier et la régulation n'est pas le nôtre. En revanche, nous sommes compétents pour répondre à des questions difficiles sur l'explicabilité, l'auditabilité et la loyauté des algorithmes, sur les biais possibles à la fois des données et des algorithmes, les deux étant liés dans ce qu'il vaudrait mieux appeler un système algorithmique. Cette dualité est essentielle. C'est sur ces questions que se penche *TransAlgo*.

Il s'agit donc d'ouvrir la boîte noire à laquelle on compare souvent les algorithmes ?

Nozha Boujemaa : Nous réfléchissons à des outils dits d'auditabilité pour essayer de comprendre le fonctionnement d'un algorithme. Pour ce faire, on n'a pas nécessairement besoin du code source. L'une des approches est la rétro-ingénierie, ou *reverse engineering* : avec des jeux de données tests bien calibrés et dont on maîtrise les variations, on étudie le comportement d'un algorithme afin d'en comprendre le mieux possible les rouages. Par exemple, on peut étudier un algorithme en ne faisant varier que les informations concernant la géolocalisation. On voit alors dans quelle mesure elle influe sur les résultats de l'algorithme.

La rétro-ingénierie est particulièrement utile pour un type de boîte noire, les algorithmes dont le fonctionnement est bien connu du concepteur, mais que celui-ci ne communique pas.

On doit bien maîtriser le protocole d'expérimentation et la méthodologie. Par exemple, avec un algorithme de classement, tel un moteur de recherche, les résultats peuvent varier simplement parce qu'au cours du temps, le corpus sur lequel opère l'algorithme a changé. Les précautions qui s'imposent parfois rendent indispensable le travail interdisciplinaire : des mathématiciens, des informaticiens, mais aussi des juristes, des sociologues, des économistes collaborent sur ces questions-là et encadrent ensemble les expérimentations.

L'autre type de boîte noire concerne les algorithmes dont même le concepteur ne maîtrise pas tout le fonctionnement. C'est le cas de ceux issus de l'apprentissage profond, tels les réseaux de neurones, où les paramètres sont ajustés automatiquement, à mesure des exemples, pour remplir la tâche assignée. On doit alors se concentrer sur l'explicabilité. Des travaux de recherche sont indispensables pour rendre explicable ce qui ne l'est pas aujourd'hui.

Insistons sur le fait que tous les algorithmes de l'intelligence artificielle ne sont pas inexplicables, seuls ceux d'une famille particulière, relevant du *deep learning*, le sont.

Quelles approches sont explorées ?

Nozha Boujemaa : Dans cet élan, les approches d'*accountability* sont très intéressantes. Il s'agit de l'idée de rendre compte du comportement des algorithmes, d'évaluer leur loyauté en se penchant par exemple sur les données qui les ont alimentés et de repérer celles qui conduisent à des discriminations.

Il ne faut pas les éliminer, mais les identifier, les prendre en compte et maîtriser le processus de décision de sorte qu'elles n'influent pas dessus. On parle d'*Equity by design* («équité par construction»).

L'élaboration de nouvelles démarches d'apprentissage fait aussi partie de nos objectifs, afin que les questions éthiques et juridiques soient prises en compte dès le départ : c'est la *transparency by design*.

Les constructeurs d'algorithmes sont-ils prêts à vous suivre ?

Nozha Boujemaa : Au départ, plusieurs industriels n'ont pas très bien pris notre démarche. Pour eux, elle s'apparentait à un manque de confiance en leurs produits. Ils pensaient qu'on allait traquer les failles, les pointer du doigt. Mais nous ne sommes pas des gendarmes. Notre rôle est de mettre à disposition des outils logiciels au >

« Le droit à l'explication anticipe l'évolution d'une société où les algorithmes vont prendre plus de place »

service de tout le monde, citoyen, industriel, autorités de régulation…

Après un long travail de pédagogie, nous avons assisté à un changement des mentalités à notre égard. Les industriels ont compris qu'on leur proposait un moyen d'établir un lien de confiance avec le consommateur. Or la confiance incite à l'achat…

On peut faire le parallèle avec l'agroalimentaire ou l'industrie pharmaceutique, deux domaines où l'importance de ce lien de confiance est reconnue depuis longtemps. D'ailleurs on m'a un jour fait remarquer que les médicaments nécessitent une autorisation de mise sur le marché, pourquoi n'en serait-il pas de même avec les algorithmes?

Cette remarque est pertinente, car en effet, on met en circulation des algorithmes dont on sait peu de choses en fin de compte. C'est pour ça que l'État s'est posé la question de la régulation, et que nous avons proposé de travailler sur les outils de la transparence.

Pour certaines questions, des travaux de recherche sont encore nécessaires et peuvent réclamer du temps. Mais du chemin a été parcouru: des interlocuteurs hier méfiants ont compris l'intérêt de ces approches et proposent aujourd'hui de financer des thèses sur le sujet.

Comment faire admettre ce genre d'idée aux Gafam?

Nozha Boujemaa: Il se trouve que ces Gafam (Google, Amazon, Facebook, Apple et Microsoft), et d'autres encore, sont à l'origine de Partnership on AI, une initiative visant à établir de bonnes pratiques dans les systèmes d'intelligence artificielle et à éduquer le public à ce sujet. Ils ont compris la vague citoyenne mondiale autour des questions d'éthique, de transparence, de discrimination, et ont pris les devants après la médiatisation de certains abus.

Il y a clairement une part de marketing, mais il y va aussi de leur survie commerciale. Aux États-Unis, de nombreux rapports et livres ont pointé du doigt les risques et certaines pratiques douteuses. Leur business menacé, ces grandes entreprises se devaient de réagir.

La clé de voûte de *TransAlgo* n'est-elle pas le volet pédagogique?

Nozha Boujemaa: C'est en effet un point très important dans nos objectifs. Aujourd'hui, des groupes de travail thématiques ont commencé à réfléchir sur la définition des concepts que nous avons évoqués (juste, loyal, éthique, responsable…) et à identifier ce qui est objectivable, invariable. Cela pose la question de la relativité des concepts. Un site dédié est en cours de production, avec en page d'accueil cet aspect pédagogique essentiel.

Nous avons commencé à rédiger des textes synthétiques sur ces concepts et à sélectionner des cas d'usage emblématiques, par exemple sur la variabilité des prix, pour interpeller un large public. En complément, nous ouvrirons un espace de ressources avec des articles, des rapports, des vidéos, des logiciels en open source.

Et sur le plan scientifique?

Nozha Boujemaa: En parallèle à ce volet pédagogique, nous lançons des groupes de travail scientifiques thématiques sur les algorithmes de classement, de recommandation, d'apprentissage, le contrôle de l'usage des données, les biais des données et des algorithmes. L'idée est de constituer une communauté scientifique de recherche sur le sujet réunissant plusieurs disciplines scientifiques (les télécommunications, la gestion de bases de données, l'apprentissage…), mais aussi des juristes, des économistes, des sociologues… Ce que produira cette plateforme scientifique sera ouvert à tous, vérifiable, testable.

Plusieurs autorités de régulation sont aussi intéressées pour se joindre à cette plateforme et l'accompagner. Le Conseil national du numérique (CNNum) est également très impliqué. Le secrétariat d'État au numérique suit *TransAlgo* de près.

Mentionnons également Dataia, l'institut interdisciplinaire en sciences de données, intelligence et société que nous lançons (un des Instituts convergences financés par l'ANR et les Investissements d'avenir). Les sciences des données y seront étudiées sans oublier leurs interfaces avec les sciences humaines et sociales. Une des priorités définies porte sur la transparence et l'éthique des systèmes algorithmiques. Le projet, prévu pour huit ans et demi et financé à hauteur de dix millions d'euros, rassemblera cent trente enseignants-chercheurs autour de quatorze partenaires académiques (Inria, CEA, École polytechnique, HEC, Ensae, Télécom Paris Tech, Centrale, Supélec, université Paris Sud, CNRS…).

Le numérique n'ayant pas de frontières, qu'en est-il de la dimension internationale?

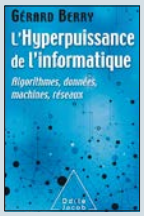
Nozha Boujemaa: Elle est fondamentale! Nous ne pouvons pas nous contenter d'avoir notre propre définition de la transparence, de l'équité… Si c'était le cas, ce serait un coup d'épée dans l'eau. Aussi, dès le début du projet, nous avons discuté avec nos homologues du Data Science Institute, à l'université Columbia, aux États-Unis, avec la National Science Foundation, aux États-Unis, et avec le programme Big Data Crest, de la Japan Science and Technology agency, au Japon, ainsi qu'avec l'institut Alan Turing, en Angleterre. L'idée de transparence algorithmique fait son chemin, malgré les différences marquées des sphères culturelles. Avec mes collègues américains, nous travaillons à ce que ce domaine bénéficie de crédits de recherche.

Nous avons aussi noué des contacts avec l'agence qui conseille le premier ministre italien sur l'agenda numérique, avec leurs homologues auprès du ministère du numérique espagnol… Il importe d'avancer ensemble. La mise en route est difficile, mais un mouvement international est indéniablement lancé. ■

« Les Gafam ont compris la vague citoyenne mondiale portant sur les questions d'éthique, de transparence… »

PROPOS RECUEILLIS PAR LOÏC MANGIN

À LIRE EN PLUS



L'Hyperpuissance de l'informatique Algorithmes, données, machines...

GÉRARD BERRY

ODILE JACOB, 2017

(**512 PAGES, 35 EUROS**)

Depuis peu, on se rend compte à quel point l'informatique est en passe de transformer notre société, et même de la bouleverser. L'auteur, professeur au Collège de France, nous aide à nous repérer dans cette société où la science et la technologie informatiques mettent l'information au cœur de l'action. Pour ce faire, il passe en revue cinq domaines de transformations massives: les télécommunications, Internet, la photographie et la cartographie, la médecine. Pour finir, il donne sa vision de l'évolution de l'informatique.

Les Big Data à découvert

MOKRANE BOUZEGHOUB ET RÉMY MOSSERI (DIR.)

CNRS ÉDITIONS, 2017

(**368 PAGES, 39 EUROS**)

Depuis les tablettes mésopotamiennes jusqu'à l'intelligence artificielle, l'humanité n'a eu de cesse de collecter des données. Mais le phénomène s'est amplifié depuis l'essor de l'informatique et d'Internet qui nous a conduits vers l'ère des *big data*. Dans cet ouvrage, un large panel d'experts présente, à travers 150 articles synthétiques, l'état de l'art sur cette révolution: les défis soulevés, les bienfaits attendus, les dérives à éviter, les risques à maîtriser... Un livre indispensable pour saisir tous les enjeux et participer aux débats de société engagés.

Le Temps des algorithmes

GILLES DOWEK ET SERGE ABITEBOUL

LE POMMIER, 2017

(**192 PAGES, 17 EUROS**)

Les algorithmes ont envahi le débat public et les discussions. C'est sans doute parce qu'ils inquiètent autant qu'ils fascinent. Nous nous réjouissons qu'ils nous facilitent la vie, mais redoutons qu'ils nous asservissent... Les auteurs, spécialistes reconnus, nous invitent à sortir de cette vision manichéenne et proposent un nouveau regard sur ce qui, après tout, n'est que créations de l'esprit humain: les algorithmes sont ce que nous avons voulu qu'ils soient. L'ouvrage nous enjoint à ne plus les subir, mais à chercher à les comprendre. C'est ainsi que nous pourrons être maîtres de notre destinée.

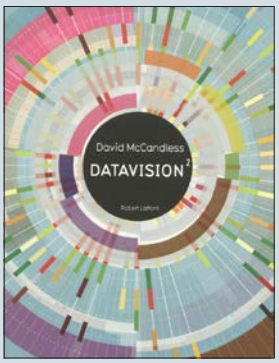
Les Statistiques en BD

LARRY GONICK

LAROUSSE, 2016

(**256 PAGES, 17,99 EUROS**)

L'auteur, dessinateur, professeur et mathématicien fait le pari de vous faire aimer les statistiques. Avec un humour décalé, il dédramatise l'univers complexe de cette branche des mathématiques aujourd'hui omniprésente: économie, finance, technologie, informatique, biologie, communication. Nous sommes entourés de statistiques, alors autant ne pas se laisser piéger par leurs chiffres et acquérir les grands principes de base pour les... déchiffrer, surtout si c'est en s'amusant! Les lois de Bernoulli s'offrent à vous, ne résistez pas.



BEAU LIVRE

Datavision². Le savoir est un art

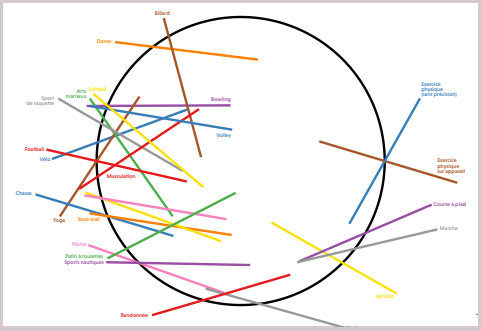
DAVID MCCANDLESS

ROBERT LAFFONT, 2014

(**224 PAGES, 23,50 EUROS**)

L'auteur, designer et pionnier de la datavisualisation, propose ici ses plus belles infographies. Elles démontrent que les images offrent une approche plus directe, mieux hiérarchisée, plus facilement mémorisable et plus séduisante du savoir universel. L'ouvrage offre une réponse graphique et esthétique à toutes sortes de questions. Au hasard: une guerre nucléaire totale est-elle envisageable? L'huile empêche-t-elle les pâtes de coller? Quels sont les mots de passe les plus utilisés sur le Web? Quelles villes seront sous l'eau en 2100?

RENDEZ-VOUS



Hors-Série 97 : Big Bang

Grand ménage autour de l'énergie sombre

La fusion de deux étoiles à neutrons a émis des ondes gravitationnelles, mais aussi un flash « lumineux ». L'observation couplée de ces événements permet de mettre au rencard plusieurs théories sur l'énergie sombre.

Souvenez-vous l'été dernier, le 17 août les collaborations *Ligo*/*Virgo*, avec d'autres équipes, annonçaient avoir détecté les ondes gravitationnelles émises par la fusion de deux étoiles à neutrons. Les ondes gravitationnelles, évoquées dans le *Hors-Série* n° 97 : « Et si le Big Bang n'avait pas existé ? », sont ces vibrations de l'espace-temps, prédites par Albert Einstein au début du xx^{e} siècle. Les premières avaient été repérées en septembre 2015, mais elles résultaient alors de la fusion de deux trous noirs. Quelle différence ?

Dans le cas de deux étoiles à neutrons, la coalescence est suivie, ici après 1,7 seconde, d'un « flash » électromagnétique. Or selon Thomas Kitching, de l'*University College*, à Londres, cet événement aide à y voir plus clair sur un phénomène apparemment sans rapport, l'accélération de l'expansion de l'Univers.

Cette accélération fut proposée à la fin des années 1990 sur la base d'observations de supernovæ de type Ia. Pour l'expliquer, les cosmologistes ont invoqué une mystérieuse forme d'énergie – l'énergie sombre – dont les propriétés gravitationnelles sont opposées à celles de la matière ordinaire : la pression exercée est négative.

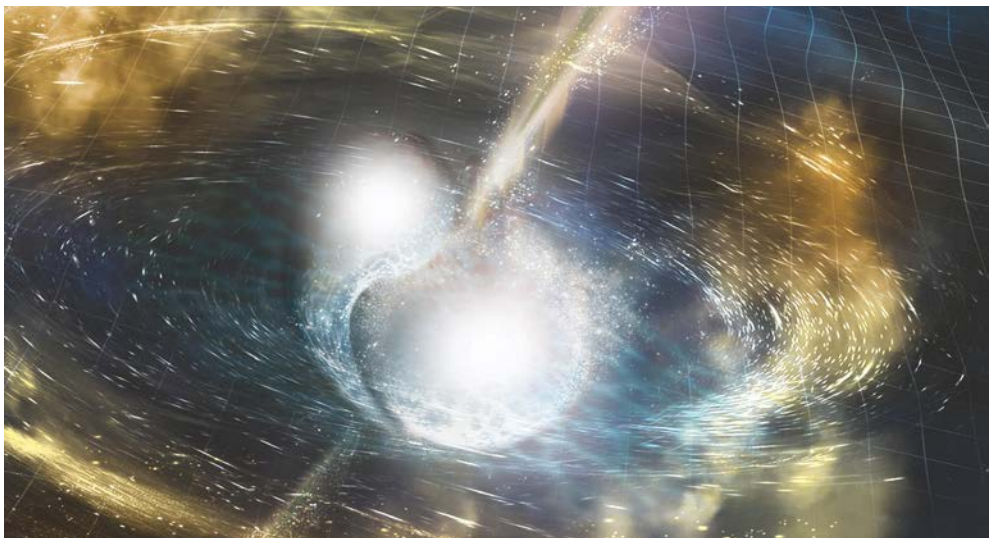
De nombreuses théories ont été échafaudées pour étendre voire remplacer la théorie de la relativité générale d'Einstein de façon à rendre compte de cette accélération de l'expansion. Certaines prévoyaient que le signal lumineux de la coalescence des étoiles à neutrons arriverait bien après les ondes gravitationnelles. L'observation d'août a définitivement mis au rebut plusieurs de ces théories candidates.

Les théories rescapées de cet élagage sont les plus simples. L'une d'elles, qui fait désormais figure de favorite, fait intervenir l'énergie du vide, même si, aujourd'hui, la valeur que l'on calcule pour la densité d'énergie du vide et celle requise pour l'énergie sombre sont incompatibles. L'autre théorie qui reste en lice est fondée sur l'existence d'un champ « scalaire », semblable au champ de Higgs associé au boson éponyme. Ce champ pourrait conduire à une accélération de l'expansion cosmique.

Pour trancher, ou au moins encore réduire le champ des possibles, les spécialistes comptent désormais sur de nouveaux instruments, notamment les télescopes des projets Euclid, LSST, SKA…

T. KITCHING, *THE CONVERSATION*, 2017.
HTTP://BIT.LY/CONV-DE

La fusion de deux étoiles à neutrons *(vue d'artiste)* émet des ondes gravitationnelles et un rayonnement électromagnétique.



© NSF/LIGO/Sonoma State University/A. Simonnet

Un petit ancêtre

Le *Hors-Série* n° 94 : « Évolution. La saga de l'humanité » retraçait l'histoire de notre espèce et celle de nos cousins les grands singes. Mais à quoi ressemblait notre ancêtre commun ? Pour en savoir plus, Mark Grabowsli, de l'université de Tübingen, en Allemagne, et William Jungers, de l'université Stony Brook, aux États-Unis, se sont livrés à une analyse phylogénétique portant sur la masse des espèces. Leur étude inclut des humains, des hominines fossiles (les genres *Homo*, *Australopithecus*…) et des grands singes européens, africains et asiatiques du Miocène (23 à 5 millions d'années). Ils ont ainsi pu estimer la masse de l'ancêtre commun à toutes ces espèces. Conclusion ? Alors qu'on l'imaginait de l'ordre de celle d'un chimpanzé (environ 50 kilogrammes), elle aurait été plus proche de celle d'un gibbon (une dizaine de kilogrammes seulement).

M. GRABOWSKI ET W. JUNGERS, *NATURE COMMUNICATIONS*, PRÉPUBLICATION EN LIGNE, 2017

Alexandre et son demi-frère

Parti de Macédoine, Alexandre le Grand a parcouru le monde à travers un long périple, décrit dans le *Hors-Série* n° 96 : « Alexandre le Grand. Quand l'archéologie bouscule le mythe », qui l'a notamment mené en Égypte, en 332 avant notre ère. Sa famille aussi a voyagé. En effet, la mission archéologique égyptienne a récemment annoncé avoir découvert dans l'ancien temple de Kom Ombo, au sud de Louxor, un bas-relief sur lequel apparaît Philippe III Arrhidée, le demi-frère du conquérant. On y distingue son visage orné de la couronne rouge de basse Égypte et son nom, en hiéroglyphes.

Hors-Série 95 : Intestin et microbiote

L'enzyme bactérienne qui protège l'intestin

Les liens entre le microbiote et la bonne santé du système immunitaire étaient détaillés dans le *Hors-Série* n° 95 : « Intestin, l'organe qui révolutionne la médecine ». Plusieurs mécanismes ont déjà été identifiés, mais Pere Santamaria, de l'université de Calgary, au Canada, et ses collègues viennent d'en ajouter un à la liste. Ils ont découvert le rôle protecteur d'une protéine d'origine microbienne contre les maladies inflammatoires chroniques de l'intestin (Mici).

Il s'agit plus précisément d'une enzyme nommée intégrase produite par plusieurs espèces de bactéries intestinales du genre *Bacteroides*. Des études sur des souris axéniques, c'est-à-dire dépourvues de microbiote, dont on a colonisé le tube digestif avec différentes souches de *Bacteroides* (produisant ou non des intégrases) ont révélé que l'enzyme recrute dans l'intestin des lymphocytes T $CD8^+$. Là, ces cellules cytotoxiques ciblent les cellules présentatrices d'antigènes activées par les bactéries productrices d'intégrase qui, sinon, déclencheraient des réactions inflammatoires. En d'autres termes, l'intégrase empêche le déclenchement des Mici en recrutant rapidement les globules blancs qui vont détruire les cellules immunitaires responsables de l'inflammation.

Qu'en est-il chez les êtres humains ? Les lymphocytes T circulants reconnaissent également l'intégrase des *Bacteroides* et pourraient freiner l'inflammation. Les études restent à mener. Les biologistes en concluent que les lymphocytes T cytotoxiques pourraient devenir des alliés dans le contrôle des maladies inflammatoires chroniques de l'intestin.

R. NANJUNDAPPA *ET AL.*, *CELL*, VOL. 171, PP. 655–667, 2017.

Hors-Série 96 : Alexandre le Grand

Une nouvelle cité d'Alexandre ?

Des images « déclassifiées » d'un satellite américain avaient révélé dans les années 1990 l'emplacement d'une cité fondée par Alexandre le Grand. Les fouilles ont commencé en septembre 2017…

Le site de Qalatga Darband *(au-delà du pont, à droite)* et un fragment de tuile *(en haut)*.

Sur les traces du conquérant Macédonien, le *Hors-Série* n° 96 : « Alexandre le Grand. Quand l'archéologie bouscule le mythe » recensait les différentes villes qu'il avait fondées. Alexandrie, en Égypte, bien sûr, mais aussi Alexandrie de Margiane, celle du Caucase, de l'Oxus, d'Arachosie… Une autre aurait été récemment découverte, en Iraq.

L'histoire commence par l'observation de photographies prises par des satellites espions américains dans les années 1960 et libérées du secret en 1996. On repère rapidement ce qui pourrait être un site intéressant, mais les fouilles sont longtemps restées impossibles pour des raisons de sécurité. La situation s'est améliorée depuis peu, rendant l'endroit accessible. Exploré depuis septembre 2017, il offre l'occasion de former des archéologues iraquiens à la sauvegarde d'un patrimoine endommagé par l'État islamique, notamment dans les anciennes villes de Ninive, Nimroud et Hatra. Le programme est dirigé par John MacGinnis, du British Museum, à Londres.

Le site, dont l'importance a été confirmée par des images prises depuis un drone, est celui de Qalatga Darband, près du lac Dukan, à proximité de la ville de Ranya, dans le Kurdistan iraquien. Les archéologues y ont découvert les restes d'un petit bourg fortifié, probablement construit en 331 avant notre ère, alors qu'Alexandre poursuivait le Perse Darius III. Des statues, des pièces de monnaie et des tuiles en terre cuite indiquent une indéniable influence grecque. Les vestiges de matériel de vinification ont aussi été retrouvés. Qalatga Darband fut sans doute une cité marchande prospère, située sur une route importante.

Parmi les trouvailles remarquables, on peut mentionner deux statues, l'une serait celle de Perséphone, déesse du retour de la végétation, l'autre représenterait Adonis, symbole de fertilité.

Selon les auteurs Plutarque et Arrien, Alexandre le Grand aurait fondé quelque soixante-dix villes. Après Qalatga Darband, il n'en reste plus qu'une bonne cinquantaine à découvrir…

LE SITE DU BRITISH MUSEUM DÉDIÉ À L'ÉQUIPE : HTTP://BIT.LY/BM-IRAK

Hors-Série 94 : Évolution

Des traces empreintes de mystère

En Crète, la découverte de traces laissées il y a 5,7 millions d'années laisse perplexes les paléoanthropologues. Elles sont résolument humaines, alors que l'on croyait les représentants de cette lignée cantonnés en Afrique.

Jusqu'où nos cousins de la lignée humaine, les australopithèques, sont-ils allés? Le *Hors-Série* n° 94: «Évolution. La saga de l'humanité» faisait état de l'idée dominante chez les paléoanthropologues selon laquelle les australopithèques ne sont pas sortis d'Afrique. De fait, les découvertes de fossiles se sont succédé durant le xx^{e} siècle, d'*Australopithecus africanus*, par Raymond Dart en 1924 à *Australopithecus sediba*, par Lee Berger en 2010 en passant par Lucy, une *Australopithecus afarensis* mise au jour par Yves Coppens en 1974, mais toutes ont été faites en Afrique, dans le sud et l'est du continent.

Cette hypothèse est remise en cause par la découverte d'une équipe emmenée par Erik Ahlberg, de l'université d'Uppsala, en Suède. Ils ont découvert des empreintes étrangement humaines… en Crète, plus précisément près de Trachilos, dans l'ouest de l'île. Ces empreintes rassemblées en une piste évoquent celles mises au jour en 1978 par Mary Leakey, près de Laetoli, en Tanzanie, et datées de 3,6 à 3,8 millions d'années. Celles de Crète sont plus anciennes encore, puisqu'elles sont marquées depuis la fin du Miocène, il y a quelque 5,7 millions d'années! Qui les a laissées là?

L'affaire est délicate. Nombre de caractéristiques anatomiques, notamment une longue plante des pieds, cinq orteils courts sans griffes, et surtout un gros orteil (le hallux) plus gros que les autres et très semblable à celui des humains modernes par sa taille, sa position et sa forme, plaident pour un hominine (la lignée humaine qui réunit les genres *Homo*, *Australopithecus*…).

Pourtant, les traces sont 1 million d'années plus vieilles qu'*Ardipithecus ramidus*, un hominine africain doté d'un pied de grand singe et en qui certains voient l'ancêtre direct des hominines plus récents. L'écheveau est très emmêlé.

Une piste peut-être? Cette année, les fossiles de *Graecopithecus* ont été réinterprétés et sont passés du statut de grand singe à celui d'hominine. Or ils vivaient en Grèce et en Bulgarie il y a 7,2 millions d'années… Ont-ils foulé le sol crétois? C'est possible, d'autant plus qu'à la fin du Miocène, ce n'était pas encore une île, mais là encore, difficile de trancher, car on ne connaît *Graecopithecus* que par des fragments de dents et de mâchoires… L'enquête continue et elle va sans doute bousculer des idées que l'on croyait acquises. ■

G. GIERLIŃSKI *ET AL.*, *PROCEEDINGS OF THE GEOLOGISTS' ASSOCIATION*, VOL. 128, PP. 697-710, 2017.

Cette trace de pas a été laissée en Crète il y a 5,7 millions d'années, mais par qui?

Le microbiote en mouvement

L'intestin et le microbiote ont tissé des liens étroits que le *Hors-Série* n° 95: «Intestin, l'organe qui révolutionne la médecine» détaillait. Andrea Murillo-Rincón, de l'université de Kiel, en Allemagne, et ses collègues en ont découvert un pour le moins inattendu. Ils ont montré que le microbiote pourrait intervenir dans le contrôle des mouvements péristaltiques de l'intestin, ces contractions musculaires qui favorisent la progression du contenu de l'organe. Ce rôle a été mis en évidence chez l'hydre, un organisme aquatique peu évolué: sans bactérie, ses mouvements péristaltiques étaient moins intenses et irréguliers. Des investigations chez l'humain s'imposent, et on peut imaginer qu'un tel mécanisme ait été conservé dans le règne animal au cours de l'évolution. De fait, les maladies inflammatoires chroniques de l'intestin (Mici) sont associées à des perturbations du péristaltisme.

A. MURILLO-RINCÓN *ET AL.*, *SCIENTIFIC REPORTS*, VOL. 7, ART. 15937, 2017.

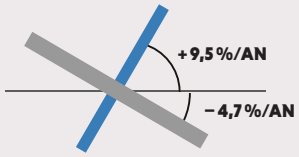
Un trou noir embarrassant

Branle-bas de combat chez les astrophysiciens, en particulier ceux qui se penchent sur les débuts de l'Univers, le thème du *Hors-Série* n° 97: «Et si le Big Bang n'avait pas existé?» Ils se perdent en conjectures devant un objet, un trou noir de 800 millions de masses solaires et âgé de 13,1 milliards d'années. Aucune théorie ne prévoit qu'un trou noir si gros ait pu se former seulement 690 millions d'années après le Big Bang, pendant la réionisation de l'Univers. On doit revoir ce que l'on croyait savoir de la croissance de tels monstres et supposer dorénavant qu'elle pouvait être plus rapide qu'on ne le pensait.

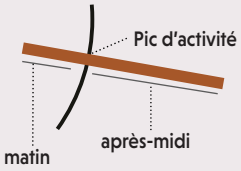
E. BAÑADOS ET *AL.*, *NATURE*, PRÉPUBLICATION EN LIGNE, 2017.

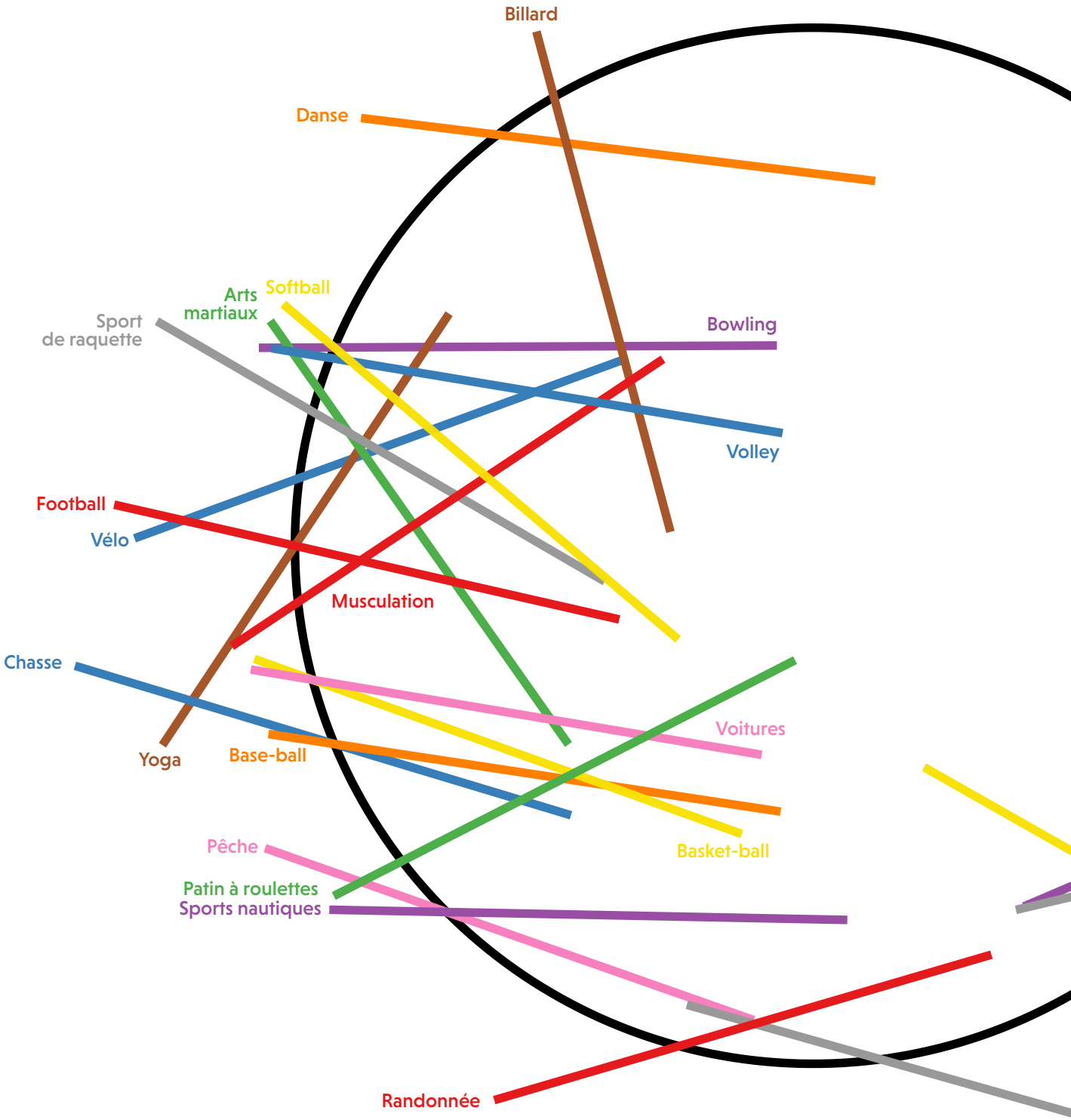
HEURES DE LA JOURNÉE

ÉVOLUTION ENTRE 2003 ET 2015



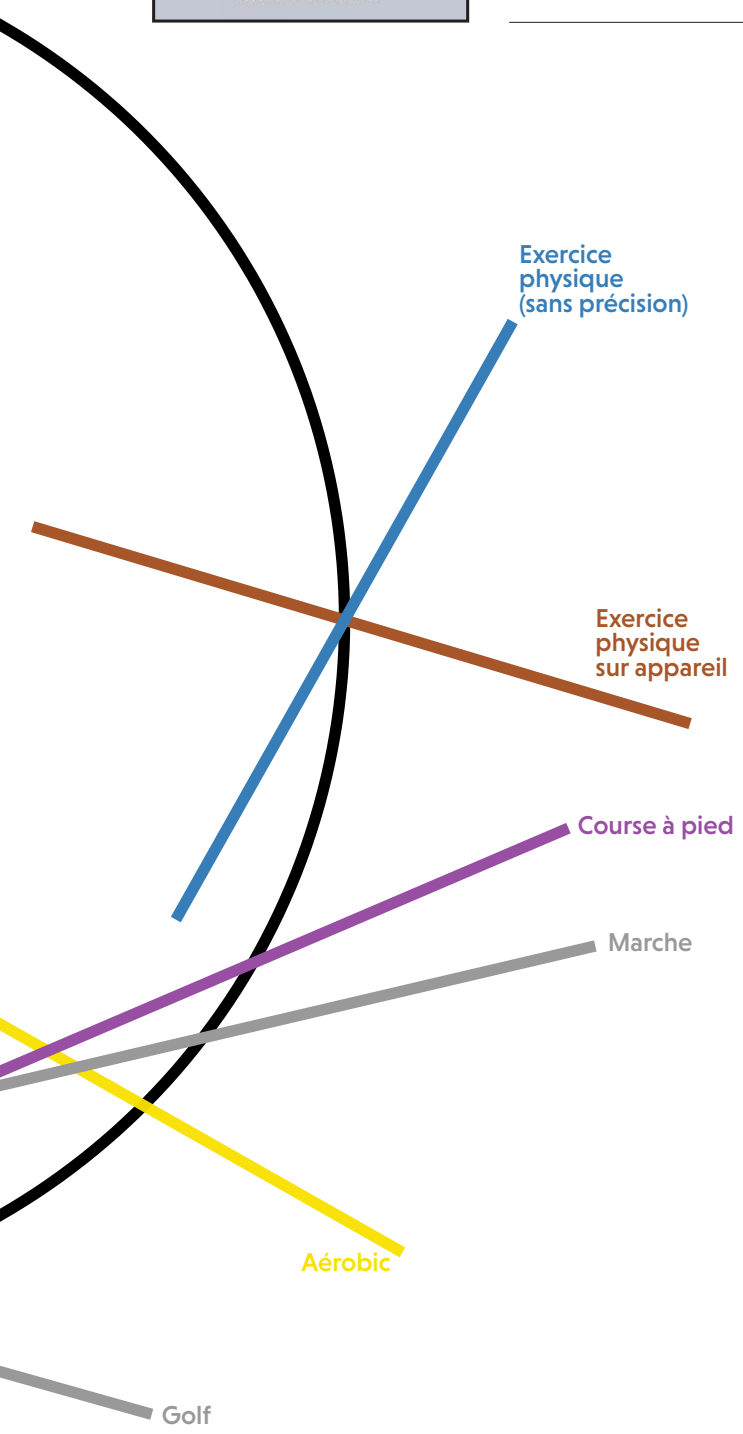
PRATIQUE JOURNALIÈRE





Sports of the Universe

Le graphisme de l'album *Sounds of the Universe*, du groupe Depeche Mode, a inspiré une datavisualisation récapitulant l'évolution des pratiques sportives des Américains entre 2003 et 2015.



De toute part, des messages gouvernementaux ou non nous enjoignent à faire du sport pour améliorer notre santé. Ces recommandations sont-elles suivies? La pratique sportive se développe-t-elle? La datavisualisation peut aider à répondre, notamment celle concoctée par le Suédois Henrik Lindberg, de la société Studentvikarie. Elle montre l'évolution des habitudes sportives aux États-Unis entre 2003 et 2015. Comment lire ce graphique?

D'abord, chaque activité est représentée par une ligne colorée. Le point d'intersection de cette ligne avec le cercle indique l'heure à laquelle elle est le plus pratiquée (la circonférence correspond à 24 heures). Par exemple, on constate que les Américains pratiquent le billard surtout le soir, le pic de fréquentation étant situé vers 22 heures. Les golfeurs et les randonneurs sont les plus nombreux en milieu de journée. Enfin, les sports sur appareil sont essentiellement pratiqués tôt le matin, vers 6 heures.

Deuxième type d'informations, la longueur de la ligne à gauche de cette intersection indique la part du temps consacré à cette activité avant midi, celle à sa droite correspondant à l'après-midi. En d'autres termes, plus la partie droite de la ligne est longue, plus l'activité est pratiquée entre midi et minuit. Le base-ball et le basket-ball sont indubitablement des sports d'après-midi. L'aérobic et la chasse sont surtout des activités du matin, même si le pic principal de la chasse est vers 16h30.

Enfin, l'orientation de la ligne indique l'évolution de la pratique de l'activité correspondante entre 2003 et 2015. L'angle est proportionnel à l'augmentation ou à la diminution, en pourcent, du nombre de minutes consacrées par jour par l'ensemble de la population. On voit que le billard ne fait plus vraiment recette, alors que le yoga, et dans une moindre mesure le patin à roulettes, sont des pratiques en vogue. Les sports nautiques et le bowling sont restés constants.

Si Depeche Mode ne vous convient pas, Henrik Lindberg propose une autre version de ces données, inspirée cette fois de l'album *Unknown Pleasures*, de Joy Division...

Retrouvez les travaux d'Henrik Lindberg ici : https://github.com/halhen/
Et suivez-le sur Twitter : @hnrklndbrg

@hnrklndbrg | Source: American Time Use Survey

À lire

DirtyBiology
La grande aventure du sexe

LEO ET COLAS GRASSET

DELCOURT, 2017

(182 PAGES, 19,50 EUROS)

Le saviez-vous ? Le champignon *Schizophyllum commune* a la particularité d'avoir… 28 000 sexes différents ! Chez *Osedax mucofloris* – un ver zombie ainsi nommé parce qu'il fore les os des cadavres à l'acide pour extraire sa nourriture –, la femelle porte sur elle les mâles microscopiques, restés à l'état larvaire. Les gobies (*Gobiodon* sp.) passent alternativement de mâle à femelle plusieurs fois au cours de leur vie. La vie est foisonnante, et la sexualité des espèces vivantes bien plus encore.

Cette bande dessinée est une adaptation de la chaîne YouTube *Dirty Biology* dont on retrouve l'esprit. L'humour omniprésent n'empêche pas, bien au contraire, de rentrer dans les détails des explications ici d'un rituel amoureux étrange, là d'une anatomie complexe qui ne se comprennent qu'à l'aune de l'évolution des espèces. Sans qu'il s'en rende compte, le lecteur est entraîné dans un riche manuel de biologie qui, sous une forme plus classique, l'aurait peut-être rebuté. Là, même les concepts les plus ardus de la grande aventure du sexe deviennent accessibles grâce à un graphisme simple et clair. La vulgarisation scientifique trouve ici une alliée parfaite, la bande dessinée. Une fois ce livre terminé, la sexualité des êtres humains vous semblera bien terne.

À vivre

Nous ne sommes pas le nombre que nous croyons être

LES 2 ET 3 FÉVRIER 2018

CITÉ INTERNATIONALE DES ARTS
18 RUE DE L'HÔTEL-DE-VILLE
75004 PARIS

Que diriez-vous de passer 36 heures non-stop en compagnie d'artistes, de chercheurs, de penseurs pour réfléchir à notre futur ? C'est l'ambitieuse proposition de l'événement « Nous ne sommes pas le nombre que nous croyons être » qui se tiendra cet hiver à la Cité internationale des arts, à Paris.

Plus de 300 participants internationaux iront à la rencontre du grand public à travers des studios-ateliers-laboratoires, des conférences, des expositions d'œuvres artistiques, des performances interactives… pour s'interroger sur le présent et esquisser ensemble des voies d'avenir. Le fil conducteur de l'événement sera l'univers romanesque qui entoure *Les Quatre Vents du désir* (*The Compass Rose*), un recueil de nouvelles fantastiques de l'Américaine Ursula Le Guin, paru en 1982. Parmi les temps forts prévus, citons l'échange qu'auront l'anthropologue Bruno Latour et le philosophe Pierre-Damien Huyghe autour des différences et des similitudes entre arts, sciences et technologies, sur la place de l'artiste et du designer, ainsi que de leurs productions, dans notre société. Sous la houlette du généticien Jonathan Weitzman, de l'université Paris-Diderot, et de son Académie vivante, des étudiants s'empareront de l'idée de « paysage épigénétique » et aideront à l'élaboration d'œuvres d'art. Les artistes Samuel Bianchini et Jochen Dehn, l'astrophysicien Stavros Katsanevas, les physiciens Jean-Marc Chomaz et Camille Duprat seront aussi de la partie. Cet événement est proposé par la fondation Daniel et Nina Carasso, en partenariat avec la chaire « arts & sciences » qu'elle porte avec l'École polytechnique et l'Ensad/PSL.

www.chaire-arts-sciences.org

À visiter

La Galerie des Amériques

Au dernier étage du Muséum d'histoire naturelle de Rouen, s'étend sur toute la longueur du bâtiment, un grand espace : bienvenue à la Galerie des Continents ! En la parcourant, le visiteur fait le tour du monde au gré des sections dédiées aux différents continents de cet espace permanent. Après l'Océanie en 2011 et l'Asie en 2014, celle consacrée aux Amériques (du Nord et du Sud) a été récemment inaugurée. L'originalité du projet ? L'élaboration des vitrines est confiée à des membres des peuples concernés, en l'occurrence les Osages de l'Oklahoma, aux États-Unis, et les Kayapos de la forêt amazonienne du Brésil. Ainsi, l'Osage Joe Don Brave et les Kayapo Bepkamrek et Nhakti ont choisi parmi les collections du Muséum les objets qui leur paraissent importants au regard de leur propre culture et les ont mis en scène.

Parallèlement à l'ouverture de la galerie des Amériques, la Réunion des musées métropolitains Rouen-Normandie a puisé dans les collections des musées de l'agglomération rouennaise pour proposer l'exposition « Potluck » (ce mot désigne dans la tradition nord-américaine un repas partagé auquel chacun contribue), jusqu'au 21 janvier 2018. Peintures, dessins, sculptures, céramiques, textiles, enseignes, photographies, affiches… témoignent des liens qui unissent depuis longtemps la région et les Amériques.

http://museumderouen.fr/

À cliquer

Le potager idéal

Vous disposez d'un espace, même petit, où vous aimeriez faire pousser quelques plantes potagères, mais vous ne vous y connaissez guère ? L'application Permapp est faite pour vous ! Sur le site, vous dressez d'abord le plan de votre parcelle et prévoyez le nombre de rangées et de colonnes. Ensuite, selon les principes de la permaculture, vous choisissez les espèces et leurs voisines en fonction de leurs affinités. Par exemple, les tomates s'accommodent bien des aubergines, de la bourrache, des oignons, des poireaux… mais il est déconseillé de planter des blettes, des betteraves, du fenouil, des pommes de terre… à côté d'elles. Le site vous indique également quelles espèces privilégier pour les rotations des cultures. Que planter après les tomates ? Pourquoi pas des carottes ?
Autres services proposés : d'abord, un calendrier des semis, plantations, floraison, récoltes… et ensuite un agenda des tâches à accomplir au gré des mois en fonction des espèces choisies. En janvier, vous pouvez déjà semer des carottes…

https://www.permapp.fr/

À voir

Un voyage au long cours

Embarquez avec Jeffrey Tsang, marin et photographe. Lors d'un périple de 30 jours sur un porte-conteneurs qui l'a mené de la mer Rouge, *via* des escales au Sri Lanka et à Singapour, jusqu'à Hong Kong, il a pris près de 80 000 clichés qu'il a ensuite réunis en une vidéo *timelapse* d'une dizaine de minutes. Ciels étoilés ou déchirés par des éclairs, couchers de soleil, tempêtes, bateaux de pêcheurs de poulpes, clairs de lune, séquences frénétiques et multicolores de déchargement se succèdent. Le résultat est hypnotique.

https://youtu.be/AHrCI9eSJGQ

À visiter

Explorateurs des abysses

À Cherbourg-en-Cotentin, la Cité de la mer rend hommage aux plus illustres explorateurs des océans. Le musée donne l'occasion unique de plonger avec eux.

L'effervescence et l'émotion règnent ce 12 octobre 2017 à la Cité de la mer. C'est qu'il y a du beau monde ! Américains, Russe, Chinois, Japonais, Français… le gratin de l'exploration des abysses est réuni à l'initiative du musée et de son directeur, Bernard Cauvin, qui inaugure ce jour-là un mur des célébrités (un *wall of fame*) recensant les records de plongée : un Hollywood boulevard des profondeurs ! Parmi les seize honorés, les Français qui ont embarqué à bord de l'*Archimède* (record 9545 mètres pour Henri-Germain Delauze en 1962), ainsi que les recordmen absolus, le Suisse Jacques Piccard et l'Américain Don Walsh, descendus à 10916 mètres de profondeur, à bord du *Trieste*, le 23 janvier 1960, dans la fosse des Mariannes. Il s'en est fallu de peu qu'ils soient détrônés par le réalisateur canadien James Cameron, également convié, qui dans son *Deepsea Challenger*, en 2012, s'est arrêté à… 10908 mètres ! Citons aussi l'Américaine Sylvia Earle (la Jeanne d'Arc des océans selon Cameron), seule femme de ce panthéon. Ce dernier se dresse désormais dans la Grande Galerie des engins et des hommes (l'ancienne gare maritime transatlantique de 280 mètres de longueur), aux côtés de douze submersibles emblématiques de la plongée profonde (l'équipe du musée espère un jour accueillir un engin venu de Chine, un pays très en pointe dans le domaine aujourd'hui). Cette proximité entre humains et machines souligne davantage encore les exploits des « océanautes » tant certains des appareils semblent parfois fragiles. Pourtant, ils ont fait la preuve de leur fiabilité et de leur robustesse. L'aventure n'est pas finie, il reste encore de la place sur le mur.

Des vocations pourraient d'ailleurs bien naître d'une visite à la Cité de la mer, tant le lieu, résolument tourné vers la transmission aux plus jeunes, donne à contempler et à comprendre les richesses (patrimoniales, naturelles, scientifiques…) du monde marin. Quelques exemples ? On peut visiter *Le Redoutable*, premier sous-marin lanceur d'engins, construit en 1967… à Cherbourg. Le pôle océan abrite l'aquarium abyssal le plus profond d'Europe, ainsi que seize autres bassins plus petits à vertu pédagogique. Ils démontrent les trésors que recèle l'océan, par exemple en termes de médicaments. L'attraction « On a marché sous la mer » offre l'occasion, virtuelle, d'explorer les abysses. Une exposition invite à revivre les grandes heures des navires transatlantiques, dont le plus célèbre, le *Titanic*, fit escale à Cherbourg le 10 avril 1912.

Les océans couvrent près de 70 % de la planète et restent pourtant inconnus à 95 %, selon Gabriel Gorsky, du laboratoire d'océanographie de Villefranche-sur-Mer. Ils nous réservent à coup sûr encore bien des surprises. Pour s'y préparer, la première escale est à Cherbourg.

https://www.citedelamer.com/

Anthrax et hippo furax

Ne jamais se fier à un hippopotame ! Sous son air débonnaire et sa silhouette pataude se cache l'un des animaux les plus dangereux d'Afrique. Farouchement territorial, il cause la mort de quelque 300 individus chaque année en défendant son pré carré. Ce n'est pas le seul risque qu'il fait courir aux humains.

Melissa Marx, du Centre de contrôle et de prévention des maladies (CDC) de Lusaka, en Zambie, et ses collègues ont récemment incriminé ces animaux (*Hippopotamus amphibious*) dans le déclenchement d'une épidémie de maladie du charbon, ou anthrax. En 2011, 511 personnes ont été contaminées par *Bacillus anthracis*, et cinq en sont mortes. L'enquête a montré que 84 % des malades interrogés avaient mangé de la viande d'hippopotames trouvés morts (tués par l'anthrax également). Que faire face à ce mode de transmission ? En premier lieu, améliorer les conditions de vie des villageois, car dans un contexte d'insécurité alimentaire, près d'un quart seraient prêts à prélever de nouveau de la viande sur des dépouilles d'hippopotames malgré les risques d'infection. ■

Cette photographie est extraite du blog Best of Bestioles : http://bit.ly/PLS-BOB

M. Lehman *et al.*, Role of food insecurity in outbreak of anthrax infections among humans and hippopotamuses living in a game reserve area, rural zambia, *Emerg. Infect. Dis.*, vol. 23(9), pp. 1471-1477, 2017.

Meunier, tu dors, ton MOULIN va à l'envers

Un Rembrandt exposé aux États-Unis serait un faux. Pourquoi ? Les ailes du moulin ne seraient pas conformes à la réalité, ce qui est impensable de la part d'un peintre dont le père était meunier.

À la National Gallery of Art, à Washington, aux États-Unis, le tableau *Le Moulin* est à l'honneur et trône au centre d'un mur. L'œuvre en question (*voir page ci-contre*) est un Rembrandt. Vraiment ? Certains en doutent… En effet, l'attribution au maître néerlandais est remise en question par un de ses compatriotes, le sculpteur Rinus Roelofs, à qui l'on doit déjà la détection d'une erreur dans la représentation d'un rhombicuboctaèdre par Léonard de Vinci.

Sur la base de quels arguments remet-il en cause l'authenticité du tableau ? L'orientation des ailes du moulin. Pour comprendre, penchons-nous sur l'anatomie de ce symbole des Pays-Bas. Fichées dans l'axe principal (l'arbre) relié aux meules, les quatre ailes sont constituées d'un long axe, la verge, hérissé de barreaux transversaux. Ils sont parfois plantés selon une orientation variable qui confère à l'aile une forme hélicoïdale.

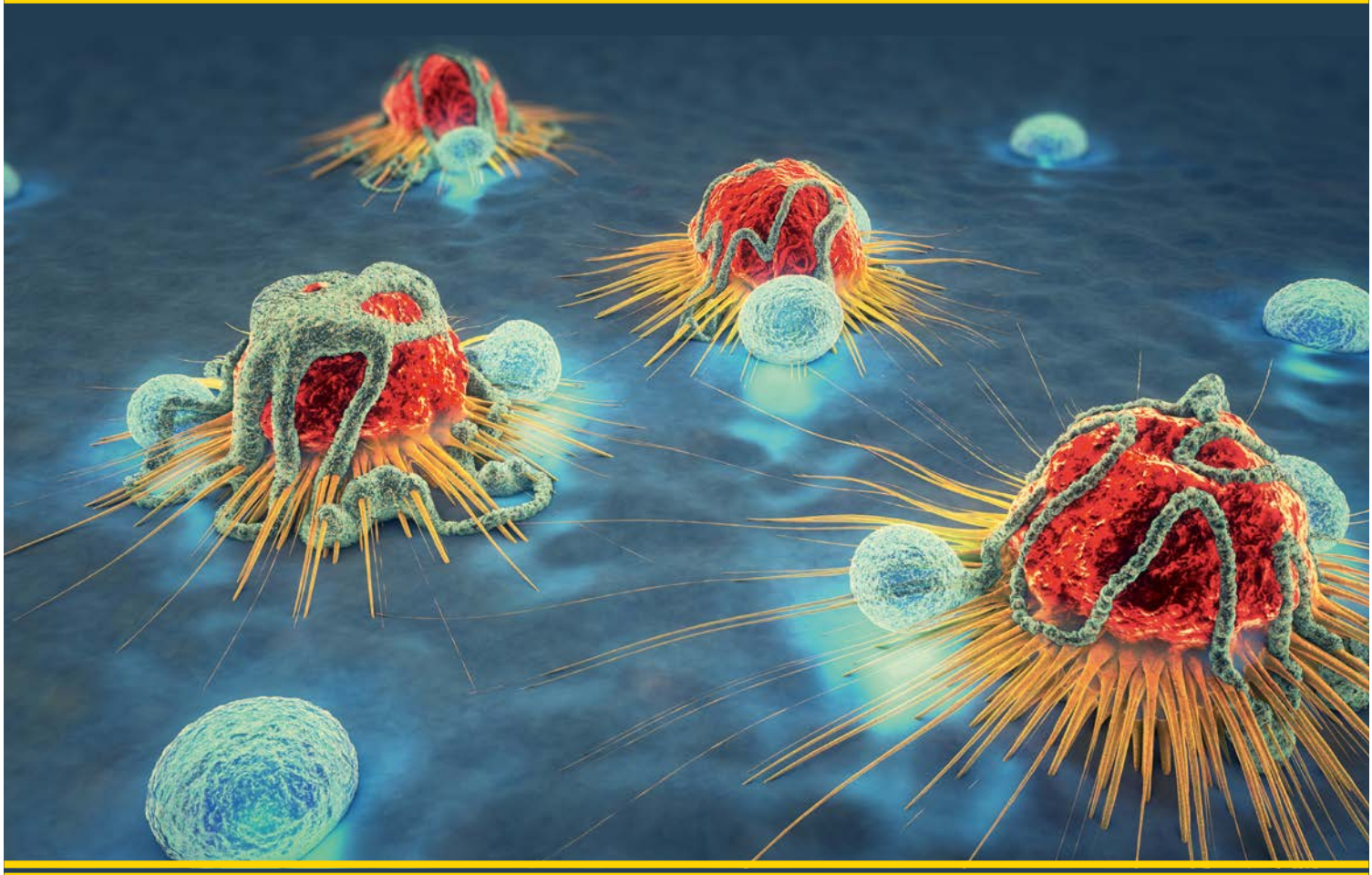
En France et en Europe du Sud, la plupart des ailes sont symétriques, c'est-à-dire que la verge est en leur milieu. En revanche, dans le Nord de l'Europe, et particulièrement aux Pays-Bas, où l'on a compté jusqu'à 10000 moulins, les ailes sont asymétriques : les barreaux sont d'un seul côté de la verge, en l'occurrence à gauche (quand l'aile est en bas). Ces ailes tournent donc dans le sens inverse des aiguilles d'une montre. Pourquoi toujours le même sens ? Parce que, tout au moins en Europe du Nord, les arbres dans lesquels on a débité les… arbres du moulin présentent tous la même torsion (les fibres du bois s'enroulent en hélice). Aussi, pour ne pas « contrarier » cette torsion, les ailes tournent dans le même sens.

Maintenant, regardez le tableau… Les ailes sont orientées dans le mauvais sens ! Elles tournent dans le sens des aiguilles d'une montre. Or on peut difficilement imaginer que Rembrandt se soit trompé. Il est le fils d'un meunier de Leyde, et a donc passé son enfance sous les ailes d'un moulin. Selon certains, le modèle du tableau serait le moulin de la ville de Kilder. Or tout visiteur peut constater que celui-ci est bien orienté. Comment expliquer cette erreur ?

Le mathématicien flamand Dirk Huylebrouck a mené l'enquête. Sur une gravure de Rembrandt représentant le moulin de Kilder, les ailes sont mal orientées, mais c'est normal, une gravure étant une image inversée résultant de l'impression sur papier à partir d'une plaque gravée. L'engouement pour les œuvres du peintre après sa mort en 1669 a-t-il incité un de ses apprentis à réaliser un tableau à partir de la gravure ? C'est une hypothèse que l'on ne peut pas exclure, même si la National Gallery se refuse à l'envisager, et continue de voir dans *Le Moulin* l'un des chefs-d'œuvre de Rembrandt.

PROCHAIN HORS-SÉRIE

en kiosque le 11 avril 2018



Cancer

REPENSER LA MALADIE POUR DE NOUVEAUX ESPOIRS

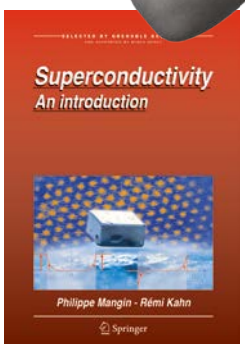
Influence du microbiote, mécanobiologie, rôle du hasard, épigénétique, cellules souches... La recherche sur le cancer explore des aspects inédits pour mieux comprendre l'apparition de la maladie, son développement et sa dissémination. Au final, ce sont de nouvelles pistes thérapeutiques qui émergent.

Achevé d'imprimer chez Roto Aisne (02) - N° d'imprimeur N° 17/11/0034 - N° d'édition : M0770698-01 - Dépôt légal : janvier 2018.
Commission paritaire n° 0917K82079 du 19-09-02-Distribution : Presstalis - ISSN 1 246-7685 - Directrice de la publication et gérante : Sylvie Marcé.

UGA Editions

Collection
Grenoble Sciences

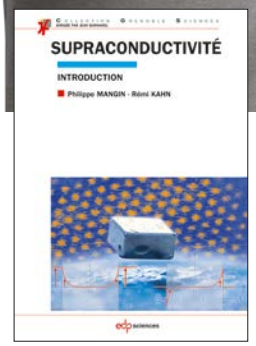
POUR TOUT COMPRENDRE SUR LA SUPRACONDUCTIVITÉ



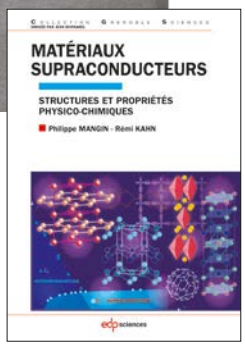
ISBN 978-3-319-50525-1
Selected by Grenoble Sciences

Une introduction complète et pédagogique pour un public varié.

Environ 250 illustrations, version française 388 pages, version anglaise 380 pages.



ISBN 978-2-7598-0858-8



ISBN 978-2-7598-2216-4

Une vision globale sur les familles et sous-familles de matériaux.

Environ 150 illustrations, 232 pages.

ISBN 978-2-7598-2137-2

Fils et câbles supraconducteurs : élaboration et applications.

Environ 235 illustrations, 322 pages.

À PARAÎTRE

Dispositifs supraconducteurs en micro et nanotechnologie : mise en œuvre, utilisations, phénomènes physiques.

Environ 120 illustrations, 220 pages.
Parution premier semestre 2018.

EN VENTE EN LIBRAIRIE ET SUR
laboutique.edpsciences.fr
springer.com

UGA Éditions - Photo : Pixabay

Springer edp sciences

www.uga-editions.com

APPEL À PROJETS 2018

LA PRÉVENTION DES RISQUES VIA

L'INTELLIGENCE ARTIFICIELLE ET LE BIG DATA

Déposez votre dossier jusqu'au 23 février 2018

Rendez-vous sur le site **www.fondation-maif.fr**

A la **Fondation MAIF,** nous finançons des projets de recherche afin de mieux comprendre et prévenir les risques liés à la mobilité, les risques de la vie quotidienne, les risques numériques et les risques naturels. Reconnue d'utilité publique, la Fondation MAIF a soutenu **150 projets** de recherche depuis sa création en 1989 et plusieurs se sont concrétisés par des innovations qui apportent plus de sécurité et une meilleure qualité de vie pour tous.

Soutenir la recherche pour prévenir les risques

MAIF FONDATION