

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

Révisseuse: 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 Bruguère 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 Floreke

Toutes demandes d'autorisation de reproduire, pour le public français ou francophone, les textes, les photos, les dessins ou les documents contenus dans la revue « Pour la Science », dans la revue « Scientific American », dans les livres édités par « Pour la Science » doivent être adressées par écrit à « Pour la Science S.A.R.L. », 162 rue du Faubourg Saint-Denis, 75010 Paris.

© Pour la Science S.A.R.L. Tous droits de reproduction, de traduction, d'adaptation et de représentation réservés pour tous les pays. La marque et le nom commercial « Scientific American » sont la propriété de Scientific American, Inc. Licence accordée à « Pour la Science S.A.R.L. ». En application de la loi du 11 mars 1957, il est interdit de reproduire intégralement ou partiellement la présente revue sans autorisation de l'éditeur ou du Centre français de l'exploitation du droit de copie (20 rue des Grands-Augustins, 75006 Paris).

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.



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-SÉRIE

N° 98
Février-Mars 2018

BIG DATA

VERS UNE RÉVOLUTION DE L'INTELLIGENCE ?

Constituez
votre collection
de *Hors-Séries*
Pour la science
Tous les numéros
depuis 1996

pourlascience.fr



© de la couverture
Shutterstock.com/Oleg Vyshnevskyy

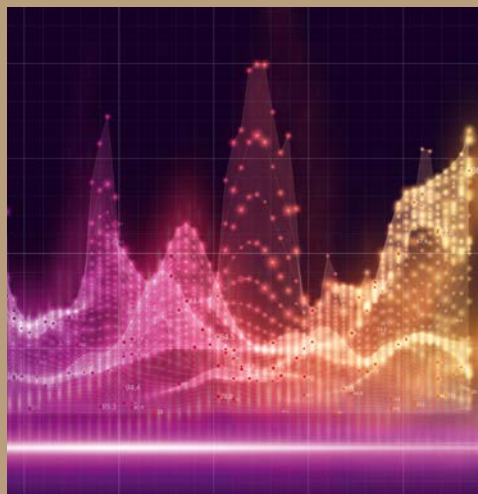
P. 6 Repères

L'indispensable pour apprécier ce numéro.

P. 8
Avant-propos
DANIEL COHEN
« Avec le *big data*,
la matérialité des corps
et des objets pourrait
disparaître. Comme dans
le film *Matrix*. »



P. 50 Cahier special Data Science
« Quand les “data
scientists”
nous simplifient la vie »
en partenariat avec



STATISTIQUES: MODE D'EMPLOI

P. 14 Déjouer les pièges des statistiques

Jean-Paul Delahaye

Les statistiques tendent des pièges:
la subtilité de certains nous ravit.

P. 20 Prévoir l'improbable

Rama Cont

Accident nucléaire, krach boursier...
Comment prévoir de tels événements rares?

P. 28 Le casse-tête des petits effets

Andrew Gelman et David Weakliem

On accorde parfois à de petites différences
un sens qu'elles n'ont pas.

P. 34 La malédiction de la valeur-*p*

Regina Nuzzo

La valeur-*p*, l'étalon-or des statistiques n'est
pas aussi fiable qu'on veut bien le croire.



LES DÉFIS DES BIG DATA

P. 42

La révolution de l'apprentissage profond

Yoshua Bengio

Pour créer une intelligence artificielle, pourquoi ne pas s'inspirer d'une intelligence naturelle?

P. 54

Traiter les big data avec raffinement

Mokrane Bouzeghoub et Anne Doucet

La gestion des *big data*, de la collecte jusqu'à l'analyse, requiert de nombreux algorithmes.

P. 62

Calculer plus vite, plus haut, plus fort

Jean-Laurent Philippe

Les besoins en calcul et en analyse des données explosent dans notre société du *big data*.

P. 70

Portfolio

Je vois, donc je comprends

Michel Beaudouin-Lafon

La visualisation des données est un complément indispensable au traitement des *big data*.

P. 76

Vers une nouvelle science ?

Étienne Klein

Peut-on confier à des machines, gavées de données, l'activité scientifique de demain?



RESTER MAÎTRE DU JEU

P. 84

Fake news : l'histoire secrète de leur succès

Walter Quattrociochi

Quels mécanismes expliquent la diffusion massive d'informations fausses?

P. 92

Santé : halte à la manipulation

Gerd Gigerenzer, Wolfgang Gaissmaier,

Elke Kurz-Milcke, Lisa Schwartz

et Steven Woloshin

La difficile interprétation des résultats médicaux et d'études de risques sanitaires.

P. 98

L'art de préserver l'anonymat

Tristan Allard, Benjamin Nguyen

et Philippe Pucheral

Les techniques des informaticiens pour protéger notre vie privée.

P. 104

Entretien

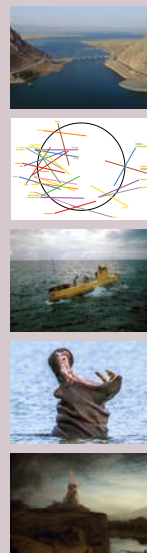
« Les algorithmes sont-ils transparentes et éthiques ?

Pour s'en assurer, nous avons
besoin d'outils adaptés. »

Nozha Boujemaa

P. 108

À lire en plus



RENDEZ-VOUS

par Loïc Mangin

P. 110

Rebondissements

Grand ménage autour de l'énergie sombre • Une nouvelle cité d'Alexandre? • L'enzyme bactérienne qui protège l'intestin • Des traces empreintes de mystère

P. 114

Données à voir

Un album de *Depeche Mode* a inspiré une datavisualisation sur l'évolution des pratiques sportives des Américains.

P. 116

Les incontournables

Des livres, des expositions, des sites internet, des vidéos, des podcasts... à ne pas manquer.

P. 118

Spécimen

Anthrax
et hippo furax

P. 120

Art & Science

Le Moulin, de Rembrandt, est-il un faux?

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.

1
OCTET

10³
KILO

10⁶
MÉGA

10⁹
GIGA

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.

**10¹²
TÉRA**

**10¹⁵
PÉTA**

**10¹⁸
EXA**

**10²¹
ZETTA**