Les maths ? Oui, ça sert !

Express

SOMMAIRE

Préface 3
Comité d'organisation du XXIe Salon Culture et Jeux Mathématiques

Introduction
Un cerveau naturellement conçu pour les mathématiques 5
Jean – Luc Berthier

Grands enjeux

L'électricité du numérique 11
Jean – Paul Delahaye

De la rationalité en matière environnementale et énergétique 17
Jean – Pierre Demailly

Les modèles mathématiques des épidémies 23
Étienne Pardoux

Émergence de résistance aux antimicrobiens 29
Lionel Roques

Les modèles de climat 35
Éric Blayo, Laurent Debreu et Christine Kazantsev

La fonte des calottes polaires 41
Jocelyne Erhel

Sciences de l'information

Les bases de données 47
Hervé Lehning

Les mathématiques de l'apprentissage 53
Clément Cartier

Réseaux de neurones et apprentissage 59
Gabriel Peyré

Générer ou reconnaître des images : les réseaux de neurones à la rescousse 65
Gabriel Peyré

L'intelligence artificielle sans les neurones 71
Enka Blanchard et Levi Gabasova

Maths et société

Des maths partout, même dans les jeux vidéo ! 77
Nicolas Nguyen

Et si on savait tous compter ? 83
Jean–Marie De Koninck

Les maths, ça sert… à être heureux ! 89
Emmanuël Houdart

Complément d'enquête sur l'Intelligence Artificielle

Il n'y a pas que les réseaux de neurones ! 95
Quentin Labernia

Ours 101

Préface du Comité d'organisation du XXIe Salon Culture et Jeux Mathématiques

Les maths, oui ça sert !

Lorsque nous réfléchissions au thème de ce 21ème Salon de la culture et des jeux mathématiques l'hiver dernier, c'est autour de l'idée qu'il fallait faire comprendre l'utilité des mathématiques que nous nous étions arrêtés. La raison en était simple : nous, mathématiciens, enseignant dans le primaire, le secondaire ou le supérieur, ou intervenant auprès du grand public, avons du mal à transmettre ce message de l'utilité de notre discipline.

Nous avons identifié quatre grands domaines où les mathématiques jouent un rôle très important : climat/environnement, santé, sciences du numérique et intelligence artificielle, et enfin sciences de l'ingénieur. Nous avons travaillé pour préparer un Salon où, à côté de la dimension purement ludique des mathématiques, en parallèle avec tout ce qui manifeste leur beauté, des stands, des conférences, des tables rondes porteraient sur leur utilité.

Mais ça, c'était avant. Avant Covid-19.

La pandémie qui frappe la France comme le reste du monde a bouleversé les vies, mis le pays à l'arrêt, fermé les établissements scolaires… et empêché la tenue d'événements publics. Qu'allions-nous pouvoir faire ? En mars et en avril, nous avons envisagé l'annulation pure et simple du Salon. Une idée s'est finalement imposée : organiser quand même le 21ème Salon, mais de manière entièrement « dématérialisée ».

Plus facile à dire qu'à faire : construire des stands est une chose, mettre en place les outils permettant la diffusion d'animations mathématiques sur Internet en est une autre. La mobilisation de tous a été formidable, et l'expérience menée par nécessité en 2020 ouvre des perspectives intéressantes pour l'avenir, avec la possibilité de promouvoir les mathématiques partout, depuis les centres des grandes villes jusqu'aux lieux les plus éloignés — sans pour autant renoncer aux rassemblements physiques, qui restent indispensables.

Douloureux paradoxe, l'épidémie de Covid-19 a démontré le rôle plus que jamais nécessaire des mathématiques pour comprendre notre monde. Le premier phénomène mathématique, pourtant très simple, qui a stupéfié tant l'opinion publique que les décideurs, a été de mesurer les conséquences de la progression exponentielle de l'épidémie : puisque le nombre de malades du Covid doublait tous les trois jours, en continuant à ce rythme, ce nombre serait multiplié par 1 000 en un mois, et par 1 000 000 en deux mois ! Ensuite, il a été question de graphes semi-logarithmiques, de probabilité de transmission, du mystérieux « R0 », de calcul du seuil d'immunité de groupes, de statistiques de prévalence… et plus récemment des questions de mathématiques et d'informatique liées à la mise en service éventuelle d'applications de traçage. Les épidémiologistes, avec leurs outils mathématiques et de science statistique, sont devenus les stars des médias.

Au-delà de l'actuelle crise sanitaire, les mathématiques seront au rendez-vous et seront indispensables pour faire face à tous les grands enjeux de demain, qu'il s'agisse de santé publique, de durabilité énergétique ou encore de climat, comme à ceux que nous ne connaissons pas encore.

Cette brochure Maths express, qui est disponible en téléchargement comme sous forme imprimée traditionnelle, donne un bel échantillon des multiples directions dans lesquelles les mathématiques sont utiles. Nous espérons qu'elle donnera envie aux jeunes de s'emparer des mathématiques pour agir sur le monde de demain.

Le comité d'organisation

Les organisateurs

Les vingt premiers Salons de la culture et des jeux mathématiques ont été organisés par le Comité international des jeux mathématiques (CIJM), sous la houlette de Marie José Pestel. C'est un formidable investissement dont nous sommes tous reconnaissants. En plein accord avec le CIJM, le Salon 2020 est pris en charge par un consortium des associations et entités suivantes :
Animath (coordination et pilotage), ***Association des professeurs de mathématiques de l'enseignement public***, ***Comité international des jeux mathématiques***, ***club Tangente***, ***Femmes et mathématiques***, ***fondation Blaise Pascal***, ***Fondation des sciences mathématiques de Paris***, ***Kangourou***, ***Science Ouverte***, ***Math.en.jeans***, ***Société de mathématiques appliquées et industrielles***, ***Société mathématique de France***, ***Société française de statistique.***

Un cerveau naturellement conçu pour les mathématiques

Jean-Luc Berthier

Spécialiste des sciences cognitives de l'apprentissage

Deux questions se posent autour de la relation entre les mécanismes cognitifs et les mathématiques, qui interrogent fortement leur enseignement dans le monde scolaire. Déjà, le cerveau humain est-il naturellement conçu pour manipuler les outils et traiter les concepts mathématiques ? Ensuite, que peut-on attendre des mathématiques pour le développement des fonctions cognitives ?

Des premiers éléments de réponses peuvent être avancés. On peut évoquer le ressenti croissant de difficulté, pouvant aller jusqu'à l'aversion, qu'éprouvent les élèves, depuis le cycle primaire (où la discipline est généralement appréciée) jusqu'à la fin du cycle lycée. Rares sont les adultes disant leur plaisir à avoir étudié les mathématiques ! On peut également s'interroger sur la place que peuvent occuper les mathématiques dans le développement des fonctions exécutives, qui sont le fondement de la cognition de tout humain.

Des prédispositions qui sont un cadeau de l'évolution !

L'idée clé désormais validée par la communauté des neuroscientifiques, qui ne l'était pas à l'époque du psychologue du développement Jean Piaget (1896—1980), est que le cerveau du bébé n'apprend pas de rien, telle une ardoise vierge. Il dispose à sa naissance de compétences le prédisposant à manipuler des concepts à caractère mathématique. C'est un cadeau de l'évolution humaine : le cerveau est prêt pour développer des compétences en numération et algèbre, analyse, géométrie, symbolisation, résolution de problème et abstraction.

Plasticité cérébrale, connectivité neuronale et modèles mentaux

Dès la vie utérine et au long des premières années, les neurones se créent à une vitesse fulgurante pour atteindre un nombre stabilisé d'une centaine de milliards, regroupés pour activer des fonctions dédiées, intégrer savoirs et compétences, agir, penser, apprendre. Les neurones se caractérisent par une connectivité évolutive qui va rendre possible l'évolution de la pensée et l'apprentissage. Les neurones modifient incessamment leur voisinage et leur structure, multipliant les liens entre eux, accroissant la vitesse de déplacement de l'information dans les axones, voire disparaissant. C'est la phase biologique de l'apprentissage, au cours de laquelle s'acquièrent les savoirs et compétences, se complexifie la pensée, se gèrent les erreurs.

Une connaissance, une idée, un concept, le sens d'un terme se représentent par des modèles mentaux, qui fourmillent et s'entremêlent dans notre esprit. L'apprentissage – on apprend continûment – peut s'interpréter comme une mise à jour et un accroissement permanent des modèles. Penser en mathématique, c'est reconfigurer, grâce à la plasticité, nos modèles mentaux, tous interconnectés. L'architecture mentale s'élabore tout au long de la vie, en particulier au cours de l'enfance et de la jeunesse, permettant par exemple de manipuler des symboles de plus en plus nombreux, de passer du monde concret au monde abstrait, d'élaborer des modèles mathématiques de plus en plus sophistiqués.

La manipulation des symboles et le passage à l'abstraction

Le monde mathématique fourmille de symboles, éléments de son langage. Le cerveau possède naturellement cette aptitude de pénétrer et évoluer dans le monde de l'abstrait en associant aux entités concrètes des symboles, puis en les manipulant pour s'en approprier de nouveaux afin d'embrasser un monde de plus en plus complexe. L'enfant se dote de chiffres, puis de nombres, de signes associés aux opérations mathématiques de base, de traits pour les fractions, du nombre π, des représentations géométriques… Plus grand, il se familiarise avec les parenthèses, les puissances, les logarithmes, l'infini, les nombres complexes, et le paysage ne cesse de s'enrichir. Parti de l'objet, de l'image, du concret, ses représentations s'abstractisent. Ce chemin progressif et naturel de la pensée est à la base de la « méthode de Singapour » : manipuler, verbaliser, abstractiser. L'enfant conçoit des quantités concrètes et visibles (trois boules, dix personnes, un dé, un cône…) puis, progressivement, l'imaginaire perd pied (cent personnes, mille grains de sable, des millions, des puissances de 10, le nombre d'Avogadro, l'infini…).

On peut comparer la progression vers l'abstrait à l'échelle d'un cerveau depuis la petite enfance, avec l'évolution des systèmes de numération à l'échelle collective : les petits traits gravés dans l'argile, la symbolisation par les lettres chez les Romains, puis les chiffres arabes discontinus, l'écriture décimale, les puissances, les logarithmes…

Le passage à l'abstrait devient de plus en plus difficile pour un nombre croissant d'élèves. C'est peut-être l'une des différences de fond qui distinguent les mathématiques d'autres disciplines.

Évoluer dans les limites de la mémoire de travail

Où, dans le cerveau, la cognition se déroule-t-elle, et avec quelles contraintes naturelles ? Nous plongeons dans deux mondes cérébraux intimement complémentaires : les *systèmes de la mémoire*, et les *fonctions exécutives* (mémoire de travail, attention, inhibition, flexibilité mentale, planification). Une fois les informations perçues et reconnues (*mémoires perceptives*), elles activent les populations neuronales associées à l'une de nos grandes fonctions exécutives, la *mémoire de travail*. Durant un temps limité (une poignée de secondes, voire de minutes), les informations sont comprises, traitées en relation avec la mémoire sémantique des savoirs et avec la mémoire procédurale des automatismes.

La mémoire de travail est quantitativement limitée en traitement d'informations. Plusieurs processus cognitifs se mettent en œuvre, fondamentaux en mathématique. Passons-les en revue.

Stanislas Dehaene, professeur au Collège de France, à l'occasion d'une conférence sur le Salon Culture et Jeux Mathématiques (mai 2017).
© É. Thomas, 2017

La mobilisation de mécanismes automatiques acquis. Avec eux, le cerveau fonctionne rapidement, inconsciemment, presque toujours sans erreur. C'est ce que l'on nomme le *système 1* de la pensée, ou des *heuristiques*. Il permet à la mémoire de travail de se libérer de la charge cognitive pour penser rationnellement, lentement et plus sûrement. C'est le fonctionnement en *système 2* ou *algorithmique*. Le mixage judicieux permettant d'optimaliser ce double fonctionnement intriqué s'opère grâce au *système 3*, dit système de *l'inhibition*, dont le rôle est de ne pas se laisser

entraîner dans des automatismes inopportuns tels qu'un raisonnement hâtif et erroné, des reflexes inadaptés, et plus généralement des impulsions excessives ou embarrassantes. C'est une forme de résistance cognitive régulatrice, indispensable aux raisonnements. L'utilisation d'opérations mathématiques relève du champ heuristique, qu'il convient d'entraîner, et d'entraîner encore, jusqu'à les automatiser en partie.

L'attention. Autre fonction exécutive directement associée à la mémoire de travail, l'attention est l'une des plus utiles chez l'humain. Elle améliore la mémorisation, la gestion des informations, la précision et la qualité du focus sur l'exécution de la tâche en cours. Nous sommes au cœur même de la compréhension, et de la production cognitive.

La manipulation d'un lexique de mots et de sens de concepts aussi riche et précis que possible. La compréhension de consignes, la communication, démarre chez le jeune par la désignation et l'explicitation. Plus ce vocabulaire de base est riche, meilleur (et rapide) est le fonctionnement de la mémoire de travail, pour laisser de l'espace au raisonnement.

La nécessité d'un bon entraînement à l'attention

L'insuffisance de la mémoire de travail entraîne de lourdes conséquences pour les élèves, relativement à l'attention, à la limitation des informations à traiter conjointement, à la planification du raisonnement, à l'inhibition des phases inappropriées d'un raisonnement. Peut-on améliorer à l'école ces capacités ? Oui, par des exercices d'entraînement à l'attention, par la mentalisation (activités de type Mathador, Réseau Canopé, 1999), par l'acquisition d'automatismes (opérations mathématiques, lecture d'un graphique, traitement des fractions, résolution d'équations…), par l'entraînement au contrôle de la pensée (mise au calme mental, préparation à la concentration).

Au fur et à mesure de l'apprentissage, l'élève progresse dans la gestion de la complexité, qui se traduit par le traitement d'un nombre croissant d'éléments en mémoire de travail, qui pourtant est limitée ; il mobilise un nombre plus important de situations de référence, qui nourrissent son intuition ; et il devient plus créatif, en se désemprisonnant des automatismes.

Une divergence opère entre ceux qui évoluent aisément dans le monde mathématique et les autres. C'est pour certains le début du décrochage.

Le cerveau est de nature prédictive et probabiliste

En toutes circonstances, y compris les plus anodines du quotidien, le cerveau se pose des questions, pour lesquelles il émet, consciemment ou inconsciemment, des hypothèses issues des modèles mentaux acquis. L'apprentissage a lieu lorsque les hypothèses sont confrontées à la réponse donnée par l'enseignant ou tout simplement à la vie. Ce processus d'émission repose sur un fonctionnement mental statistique (« une hypothèse semble plus probable que d'autres »). C'est le cas chez le bébé déjà.

On peut en tirer certaines conclusions : nous apprenons essentiellement par questionnement, par résolution de situations problèmes, par résolution d'erreurs dites pertinentes. C'est l'efficacité pédagogique du jeu, des groupes d'apprentissage, des énigmes, de l'implication active. On pressent en filigrane une forme efficace de pédagogie. Écouter, lire, imiter, ne suffisent pas.

Un sens inné de la quantité et du repérage spatial

L'estimation de la quantité (dire si l'une est plus grande ou plus petite que l'autre) fait partie des compétences innées. Chez l'humain, l'évaluation est d'abord approximative, avec la possibilité balbutiante d'amorcer les mécanismes de l'addition et de la soustraction. Puis la précision s'affine avec le comptage discret d'objets physiques. C'est sur cette capacité que va se construire la numératie. L'humain est prêt à la développer sans limite !

Citons également l'aptitude au repérage spatial, les notions de gauche, droite, haut, bas, arrière, avant, qui permettent dès les premiers mois de positionner les objets et les personnes, de les relier à d'autres. L'aptitude à la géométrie est là, même embryonnaire, soutenue par la compétence à faire pivoter les images mentales dans le cerveau.

L'appareil cognitif est donc conçu pour les mathématiques. Toute la question est la mise en œuvre de modalités pédagogiques subtilement adaptées pour les développer avec le moins possible d'empêchements pour chacun.

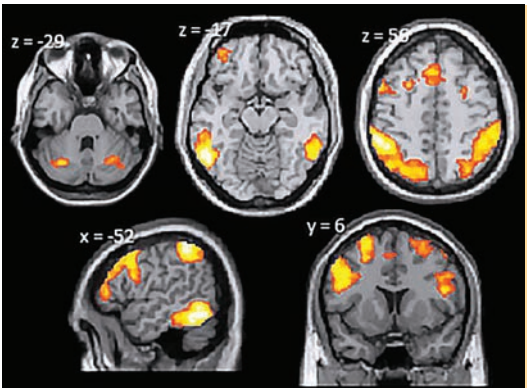
Les maths pour le développement des fonctions cognitives

Toute activité cognitive s'appuie sur les fonctions exécutives de base, les systèmes de la mémoire et la plasticité cérébrale. En cela, les mathéma-

tiques ne sont pas vraiment spécifiques. Mais elles sont un terrain magnifique d'opportunités de développement du cerveau : inhibition, rigueur et logique, étonnement et questionnement, symbolisation, planification.

C'est à partir d'une architecture innée, imparfaite et balbutiante que se structure toute pensée. Oui, le cerveau de l'humain est conçu pour évoluer dans le monde mathématique. Oui, l'activité mathématique dépassant largement la pratique des outils opérationnels de base permet aux fonctions exécutives de se développer efficacement, au service de tout acte de la vie, très au-delà de la seule culture mathématique.

J.–L. B.



Coupes montrant les zones du cerveau en activité lors d'une résolution de problème par un mathématicien expert.

Source : Origins of The Brain Networks For Advanced Mathematics In Expert Mathematicians, Proceedings Of The National Academy Of Sciences 113 (18), 2016

© *Marie Amalric et Stanislas Dehaene*

Pour en savoir (un peu) plus :

La bosse des maths, quinze ans après. Stanislas Dehaene, Odile Jacob, 2010.

Maths Langages Express. Comité international des jeux mathématiques, 2017.

Maths Jeux Culture Express. Comité international des jeux mathématiques, 2019.

L'électricité du numérique

Jean-Paul Delahaye

Professeur émérite à l'Université de Lille

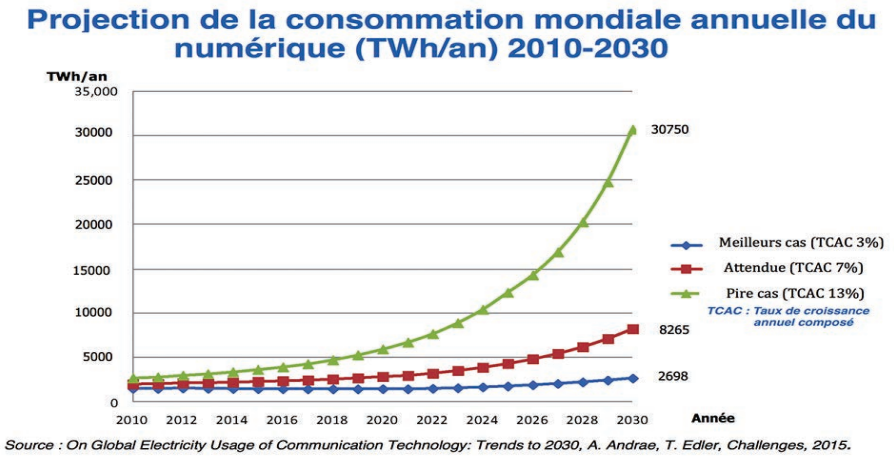
Les dispositifs numériques partout présents dans le monde consomment massivement de l'électricité. Il s'agit de nos smartphones, de nos ordinateurs, de nos écrans, des serveurs qui font fonctionner les réseaux et permettent aux données qu'on envoie de circuler, et aux données qu'on veut consulter de nous parvenir. Il s'agit bien sûr aussi des data-centers, ces centres de données qui collectent, gardent et trient toutes sortes d'informations : moteurs de recherche, médias d'actualité, centres de stockage d'articles scientifiques en ligne, équipements informatiques pour mettre à jour et faire fonctionner les GPS, centres de calcul météo ou de recherche…

Une part croissante de l'énergie dépensée par le numérique

Dans un rapport publié en 2016, le conseiller scientifique Joël Hamelin écrit : « *En 2012, la consommation électrique du numérique était de 930 TWh/an, soit 4,7 % de la consommation mondiale […]. Ce ratio était de 4 % en 2007. La progression de l'énergie électrique dépensée par le numérique est estimée en 2016 à 6,6 % par an et devrait évoluer au même rythme dans les prochaines années. La consommation électrique mondiale [totale] a progressé durant les quarante dernières années de 3,2 % par an. En 2020, la consommation électrique mondiale [du numérique] devrait être de 1 550 TWh/an et de 2 940 TWh/an en 2030, soit respectivement 5,8 % et 8 % de la consommation globale.* » (*Énergie et numérique*, indicateurs France Stratégie, 2016)

Devons-nous nous inquiéter de cela ? Devons-nous tenter de limiter cette part croissante de l'énergie électrique dépensée par le numérique ?

La réponse est bien évidemment oui. Autant que nous le pouvons, nous devons limiter cette dépense, de la même façon que nous devons limiter toutes les dépenses énergétiques car toutes ont un impact environnemental que l'on doit réussir à contrôler.



Prévisions jusqu'en 2030 de la dépense du numérique selon diverses hypothèses.

Source : On Global Electricity Usage of Communication Technology : Trends to 2030. Anders Andrae, Tomas Edler, *Challenges*, 2015

Pour autant, il ne faut pas se tromper d'ennemi. Lorsque l'on utilise un ordinateur pour consulter un article de journal (ou les références proposées en dernière page de ce texte), cela nous évite probablement un déplacement physique qui aurait coûté en énergie fossile, et cela nous évite aussi de consommer le papier qui aurait servi à l'imprimer. Le bilan précis permettant de savoir si l'on est « gagnant » avec la version numérique de l'article numérique est très délicat à calculer car il dépend de la taille de l'article lu, de l'éloignement du site Internet utilisé pour télécharger l'article, de la taille de l'écran pour le visualiser, et bien sûr dans l'hypothèse de l'achat du journal de la distance parcourue pour aller l'acheter et de la façon dont il est produit et distribué.

Il est cependant certain que le numérique est plus économique dans la très grande majorité des situations. Cette remarque vaut aussi pour les courriers électroniques comparés aux envois de courriers postaux, à condition bien évidemment de ne pas faire imprimer ses courriels.

Il est important de ne pas se tromper d'ennemi !

Deux autres points ne doivent pas être oubliés. Déjà, envoyer un courrier électronique sans fichier attaché demande un transfert de données bien inférieur à ce qui est induit par la consultation d'un article sur un média Internet car ce dernier comportera sans doute des photos ou des images qui

transiteront sur le réseau : on passe de quelques milliers d'octets de données à quelques centaines de milliers d'octets, ou même plus. Cette consultation d'article elle-même demandera beaucoup moins de ressources réseaux (et donc de consommation d'énergie au final) que l'écoute d'un morceau de musique en *streaming*, ou pire le visionnage d'une vidéo ou d'un film : on passe cette fois de quelques centaines de milliers d'octets à quelques dizaines de millions d'octets. Si vous voulez limiter les dépenses électriques liées à votre usage du numérique, ne vous préoccupez donc pas trop de vos courriels, mais plutôt des téléchargements de fichiers volumineux (provoqués par le « surf » sur Internet) et surtout prenez garde au *streaming* audio et vidéo : un épisode de série télé équivaut à plusieurs milliers ou dizaines de milliers de courriers électroniques.

Ensuite, ce que l'on pouvait dire il y a quelques années concernant par exemple le coût d'une requête sur un moteur de recherche (ou de tout autre usage du numérique) évolue rapidement à la baisse. La loi de Gordon Moore indique que les coûts des opérations numériques sont divisés par 2 tous les deux ans. Cette loi est en train lentement de s'épuiser et doit être réévaluée (les deux ans devenant trois, et sans doute bientôt quatre). Pourtant, actuellement, il se produit toujours une baisse du coût énergétique d'une opération donnée. Si la consommation du numérique croît au rythme de 6 % ou 7 % par an, c'est que l'on est de plus en plus nombreux à l'utiliser, que chacun l'utilise beaucoup plus et que l'on se permet des opérations de plus en plus coûteuses en calcul (le *streaming* par exemple), ce qui, malgré la baisse des coûts de chaque opération, engendre globalement une augmentation.

La taille des écrans des smartphones et ordinateurs est un exemple de ce confort que l'on se permet et qui demande, pour fonctionner, des quantités de calculs croissantes, qui heureusement (grâce notamment à la loi de Moore) ne se traduisent pas proportionnellement en dépenses électriques.

Les chaînes de blocs et la consommation électrique

Les centres de données tentent, pour des raisons économiques évidentes, de dépenser moins d'énergie ; ils y parviennent grâce aux progrès de leur conception. D'une manière générale, toutes les opérations de calcul

qu'engendre l'usage du numérique continueront de baisser par opération, que ce soit l'envoi d'un courrier électronique, une requête informatique, l'écoute d'une chanson en ligne… En outre, les opérations que l'on effectue grâce au numérique font souvent économiser d'autres dépenses énergétiques (transports, papier…). Il faut être vigilant, attentif et progresser encore, mais ne pas considérer que le numérique est mauvais en soi pour l'environnement et qu'il faut absolument en limiter les usages, car on risquerait d'aboutir à l'inverse de ce que l'on souhaite !

Il existe cependant un domaine où la dépense électrique du numérique augmente follement, et cela sans que cela puisse être sérieusement justifié. Il s'agit des opérations de « minage » des crypto-monnaies de type Bitcoin.

Le minage du Bitcoin. Une usine de la firme chinoise Bitmain qui mine du Bitcoin. De dizaines de bâtiments contiennent des milliers de machines qui calculent en continu la fonction SHA256 révisions jusqu'en 2030 de la dépense du numérique selon diverses hypothèses.

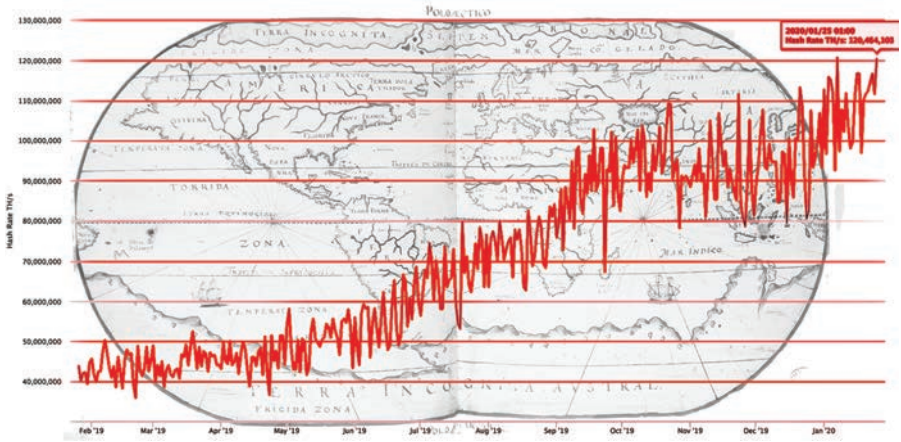
© *Shofon.net*

Sans entrer dans les détails faute de place, expliquons le problème avec le minage. Le fonctionnement des monnaies cryptographiques sans autorité centrale de contrôle exige que les ordinateurs validateurs des opérations exécutées sur le réseau (transactions d'un compte vers un autre, ajout de nouvelles pages à la mémoire de chaque ordinateur validateur) se coordonnent et s'accordent.

Cette recherche de consensus exige que soit désigné périodiquement un *nœud validateur principal* – toutes les dix minutes environ dans le cas du Bitcoin – sans que l'on puisse anticiper qui jouera ce rôle. Cette désignation tournante du nœud validateur principal peut se faire selon plusieurs méthodes. Les trois principales sont (a) la *preuve de travail*, (b) la *preuve d'enjeu*, et (c) le *fonctionnement tournant* pour les systèmes de consortium, où les nœuds validateurs ne sont pas anonymes et en nombre limité.

Mauvais choix technologique et électricité gaspillée

La *preuve de travail*, qui est la méthode utilisée par le Bitcoin, consiste à résoudre un problème mathématique. Le réseau formule automatiquement, toutes les dix minutes environ, un problème exigeant par nature une grande quantité de calculs pour être résolu. L'énoncé est tel que résoudre le problème demande « seulement » de disposer d'une certaine capacité à calculer la fonction SHA256 (il s'agit d'une fonction cryptographique de hachage standard). Si un nœud dispose d'une capacité à calculer ce SHA256 équivalent à 10 % par exemple de la capacité totale du réseau Bitcoin à calculer ce SHA256, alors ce nœud validateur sera désigné dans 10 % des cas comme validateur principal, ce qui lui rapportera à chaque fois une certaine somme car 12,5 bitcoins sont distribués au « gagnant » toutes les dix minutes (en février 2020, et deux fois moins à partir de mai 2020). Ce concours qui, au total, a distribué l'équivalent de plus de cinq milliards de dollars en 2019, a conduit ceux qui y participent à s'équiper en matériels spécialisés et optimisés pour calculer le plus rapidement possibles des SHA256.



La fonction SHA256 est calculée 120 × 10^{18} fois par seconde, par le réseau Bitcoin (janvier 2020). Ce calcul provoque une dépense électrique équivalente à ce que produisent environ cinq réacteurs nucléaires.

© Trading View, 2020

Ces dispositifs dépensent de l'électricité ! Depuis 2009, la date de mise en marche du Bitcoin, on a assisté à une course folle entre nœuds validateurs, qui a abouti aujourd'hui à ce que le réseau Bitcoin calcule plus de 10^{20} évaluations de la fonction SHA256 par seconde. Ces calculs consomment 40 TWh par an d'électricité au minimum, ce qui est environ le double de ce

que produisent toutes les éoliennes de France (24 TWh/an) et correspond, en ordre de grandeur, à la consommation électrique de pays comme la Suisse ou l'Autriche.

L'absurdité de la situation provient de ce que (b) et (c), les méthodes de consensus concurrentes de la preuve de travail (a), donnent une sécurité au réseau équivalente, sans provoquer de dépense sensible d'électricité. On peut donc dire aujourd'hui que l'on brûle pour rien plus de 40 TWh/an d'électricité à cause de ce mauvais choix technologique au cœur de Bitcoin ! Il faudrait changer la méthode de consensus du Bitcoin, et c'est d'ailleurs ce que va faire le réseau de crypto-monnaie Ethereum (le deuxième en importance derrière Bitcoin). Pour des raisons de gouvernance mal conçue dans le cas de Bitcoin, cela semble impossible pour Bitcoin, et rien aujourd'hui ne semble aller dans ce sens pour cette crypto-monnaie.

Une autre source de gâchis indirectement liée au numérique résulte des appareils mis en veille. Rien qu'en France, la dépense de ces systèmes de veille atteindrait de 8 à 13 TWh/an. Même si tous ces systèmes de mises en veille ne sont pas totalement inutiles, il se trouve sans doute là une possibilité d'économie non négligeable.

Oui, il y a des économies à faire sur les calculs du numérique. Mais ce n'est pas en vous retenant d'écrire des courriels ou même en limitant le nombre de requêtes à votre moteur de recherche préféré que vous parviendrez à quoi que ce soit de sensible. Concentrons-nous sur les plus gros gâchis et laissons tranquilles les usages du numérique vraiment importants, qui sont aussi sources d'économies (en papier, en déplacements…). N'oublions pas aussi, lorsqu'on jette des ordinateurs, des tablettes, des *smartphones* pour les remplacer par des neufs, que cela a un impact environnemental important. Faire durer les matériels est aussi un moyen très sérieux d'économiser la planète.

J.-P. D.

Pour en savoir (un peu) plus :

Empreinte environnementale du numérique mondial. Fréderic Bordage, 2019, disponible en ligne.

La face cachée du numérique. Agence de l'environnement et de la maîtrise de l'énergie (ADEME), 2019, disponible en ligne.

L'insoutenable usage de la vidéo en ligne. Rapport du Shift Project, 2019, disponible en ligne.

Pour une sobriété numérique. Rapport du Shift Project, 2018, disponible en ligne.

La folie électrique du Bitcoin. Jean-Paul Delahaye, *Pour la science 294*, février 2018, disponible en ligne sur le site de l'auteur.

Coinshare research. The Bitcoin Mining Network, 2019, disponible en ligne.

De la rationalité en matière environnementale et énergétique

Jean-Pierre Demailly

Professeur à l'Université Grenoble-Alpes,
Mathématicien, membre de l'Académie des sciences

L'accès à l'énergie est un élément essentiel de prospérité de l'humanité. Si l'on parvient aujourd'hui à nourrir 7,6 milliards d'êtres humains, c'est en grande partie grâce à l'accroissement considérable des rendements agricoles depuis le milieu du XX^{e} siècle, dans un facteur 5 à 8, permis par la mécanisation et l'emploi d'engrais chimiques, et donc par l'énergie. Même si des politiques de sobriété sont mises en place, le développement des pays du Sud, tout comme les besoins accrus de recyclage des ressources dans les pays développés, nécessiteront de disposer de davantage d'énergie. Sinon, la régression économique pourrait mener à un effondrement, à la famine et à la guerre. Le remarquable rapport Meadows, *les Limites à la croissance*, publié en 1972, soulignait déjà avec force les limites naturelles de la planète, par exemple sur le plan des ressources minérales ou alimentaires. En même temps, la lutte contre le réchauffement climatique réclame de se tourner vers des sources d'énergie décarbonées. Ce n'est malheureusement pas l'évolution constatée aujourd'hui dans le monde. Nombre d'autres scientifiques – voir par exemple les analyses de Jean-Marc Jancovici – considèrent que la politique énergétique menée par la plupart des pays est déficiente sous de nombreux aspects. En Europe, la France ne tire son épingle du jeu que par un héritage nucléaire qu'elle s'évertue, hélas, à dilapider, faute de le maintenir à niveau et d'investir dans les technologies les plus prometteuses.

Pour des décisions fondées sur la science et sur le calcul

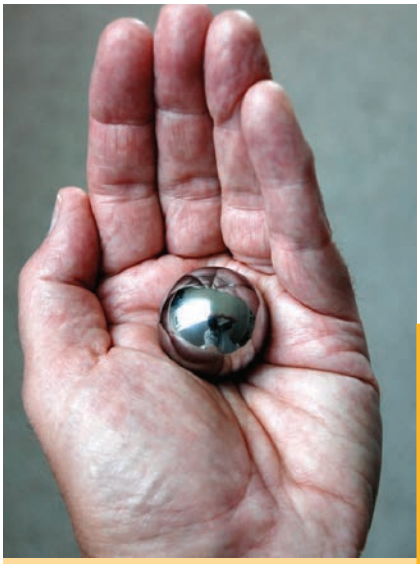
Les décisions en matière énergétique devraient être fondées de manière rationnelle sur la science et le calcul, à partir des données physiques fondamentales. Le recours historique aux combustibles fossiles tient au fait qu'il s'agit d'énergies facilement accessibles, bon marché, ayant une densité énergétique importante. Ceci explique l'usage massif des hydrocarbures dans les transports.

Fissile, vous avez dit « fissile » ?

En physique nucléaire et en chimie, on appelle isotopes (d'un élément chimique donné) les noyaux d'atomes partageant le même nombre de protons (caractéristique de cet élément chimique) mais possédant un nombre de neutrons différent. Un isotope est fissile si son noyau peut subir une fission nucléaire sous l'effet d'un bombardement par des neutrons (désintégration enclenchant une réaction en chaîne).

Le seul isotope fissile présent en quantité non négligeable sur Terre est l'uranium 235. D'autres isotopes fissiles peuvent être produits artificiellement.

De son côté, l'énergie nucléaire est, de très loin, la championne de la densité énergétique : 1 kg de combustible fissile comme l'uranium 235 (^{235}U) peut fournir théoriquement la même quantité d'énergie que deux mille quatre cents tonnes de charbon, ou que mille six cents tonnes de pétrole, soit un gain massique d'environ 2×10^6 par rapport à l'énergie chimique. On pourrait ainsi tenir dans la main les quelques centaines de grammes nécessaires pour assurer l'approvisionnement énergétique d'un Européen pendant toute sa vie ! Une consommation annuelle de cinquante tonnes d'éléments fissiles par un parc de soixante réacteurs nucléaires de $1{,}35 \times 10^9$ W (avec un facteur de charge typique de 80 %) suffit à assurer la totalité de la production électrique de la France, soit 550×10^{12} Wh par an. Un autre fait important est que les réactions nucléaires n'émettent pas de CO_2, et ne participent donc pas au réchauffement climatique.



Uranium ou thorium : l'énergie de toute une vie dans une main !

© *Flibe Energy*

Puissance (W) et énergie dépensée (Wh)

La puissance électrique (mesurée en watts, ou W) indique, à un instant donné, la consommation en énergie. C'est en fait une énergie par seconde. Une consommation ou une production annuelle est ainsi souvent exprimée en watts-heures (Wh), qui correspondent à l'énergie dépensée (à savoir 1 W pendant une heure). Cette dernière se mesure également en joules (J), avec l'équivalence 1 Wh = 3,6 kJ.

Le solaire photovoltaïque et l'éolien ne sont pas la solution

Les nouvelles énergies renouvelables comme le solaire photovoltaïque et l'éolien ont évidemment leur place, par exemple pour des installations autonomes ou dans des lieux reculés. Cependant, si elles devaient prendre une part importante de la production électrique, leur intermittence exigerait des capacités de stockage actuellement hors de portée, et leur faible densité énergétique se traduirait par un impact environnemental exorbitant.

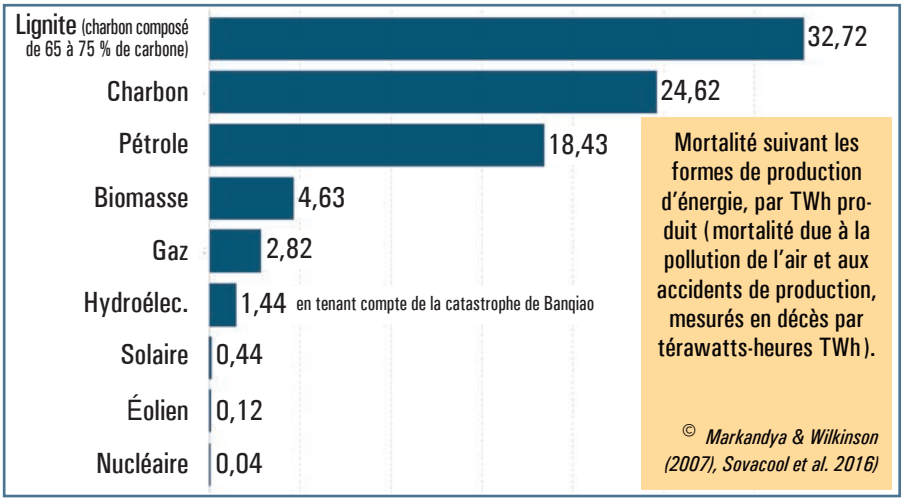
Sachant qu'un panneau photovoltaïque de puissance de crête égale à 300 W (pour 1,5 m² et 18 kg) a un facteur de charge de l'ordre de 15 % en moyenne, on calcule que la production de 550×10^{12} Wh photovoltaïques demanderait $550 \times 10^{12} / (300 \times 0{,}15 \times 24 \times 365)$, soit plus de $1{,}39 \times 10^{9}$ panneaux. Cela représente une masse totale de $2{,}5 \times 10^{7}$ tonnes de matériaux de haute technologie, requérant des traitements physico-chimiques élaborés et polluants, pour une quantité de matériaux bruts au moins dix fois supérieure, à renouveler tous les trente ans environ. Ces estimations ne tiennent pas compte des dispositifs de stockage à ajouter, ni de l'énergie considérable nécessaire pour la production des panneaux et leur recyclage, qui devrait idéalement être renouvelable, elle aussi ! Elle est à comparer aux mille cinq cents tonnes d'éléments fissiles nécessaires pendant les mêmes trente années, ce qui, au moyen de réacteurs surgénérateurs (voir plus loin), nécessiterait de brasser « seulement » deux cent mille tonnes de minerai, même pour des minerais pauvres dont la teneur avoisine 0,75 %.

Le nucléaire actuel ne sera pas soutenable très longtemps !

Les reproches faits à l'énergie nucléaire, telle qu'elle existe aujourd'hui, tiennent à son caractère non renouvelable et à une production significative de déchets radioactifs, bien que ceux-ci représentent *in fine* des volumes relativement faibles pouvant faire l'objet d'un stockage géologique. Il y a également une sûreté perçue comme insuffisante, les accidents de Tchernobyl et de Fukushima ayant beaucoup frappé les esprits. Si l'on s'en tient aux estimations « objectives » (factuelles) du nombre de décès par TWh produit (1 T = un téra = 10^{12}), on constate que l'énergie nucléaire est tout de même la plus sûre des énergies actuellement disponibles.

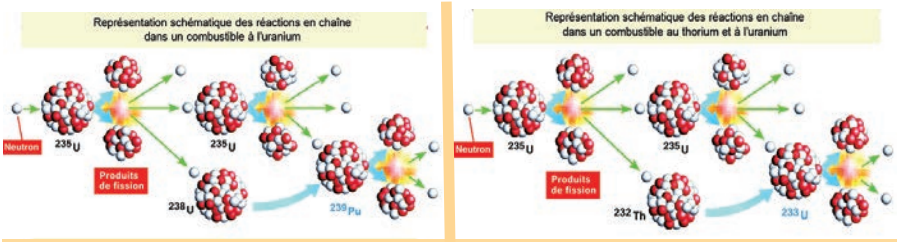
Mais il convient de se demander si le nucléaire reste une option viable dans l'hypothèse où l'on passerait d'un parc mondial de quatre cent cinquante réacteurs à une flotte de dix mille à vingt mille réacteurs de puissance – la fourchette haute permettant de couvrir la plus grande part des besoins énergétiques mondiaux. Ceci inclurait, par exemple, la production d'hydrogène

par électrolyse, dans le but de remplacer les combustibles fossiles. Dans ce cas, la technologie actuelle des réacteurs à eau pressurisée (REP) n'est plus soutenable, car l'horizon des ressources en uranium ne serait plus que de l'ordre de cinq à dix ans : les REP ne consomment en gros que la fraction ^{235}U de l'uranium naturel, qui comporte 99,3 % de ^{238}U (qui n'est pas fissile) et seulement 0,7 % de ^{235}U fissile.



La bonne nouvelle est que les physiciens ont dans leurs cartons des *réacteurs à neutrons rapides* (RNR), qui peuvent, par capture neutronique, fertiliser ^{238}U en plutonium ^{239}Pu fissile, et également le thorium ^{232}Th, élément naturel assez abondant, lequel se trouve transmuté en un autre isotope fissile de l'uranium, à savoir ^{233}U. Lorsque les pertes de neutrons sont suffisamment faibles, ce qui impose de ne pas les ralentir dans un milieu modérateur comme l'eau, on obtient ainsi des réacteurs qui peuvent exploiter la totalité des éléments fertiles ^{238}U et ^{232}Th, et produisent même un peu plus d'éléments fissiles qu'ils n'en consomment, d'où le nom de « surgénérateurs ».

Ainsi, deux cycles de réactions en chaîne sont possibles : le cycle $^{238}U - ^{239}Pu$ et le cycle $^{232}Th - ^{233}U$. Dans cette situation, le combustible disponible devient de trois cents à quatre cents fois plus abondant que le seul ^{235}U présent dans la nature, et l'horizon des ressources repasse à plusieurs milliers d'années. Un gros réacteur électrogène d'une puissance électrique de $1{,}5 \times 10^9$ W pourrait ainsi ne consommer annuellement qu'à peine plus d'une tonne d'uranium naturel ou de thorium. Avec un stock de trois cent cinquante mille tonnes d'uranium appauvri et de dix mille tonnes de thorium, cela permettrait à la France d'attendre des milliers d'années avant de reprendre toute extraction minière !

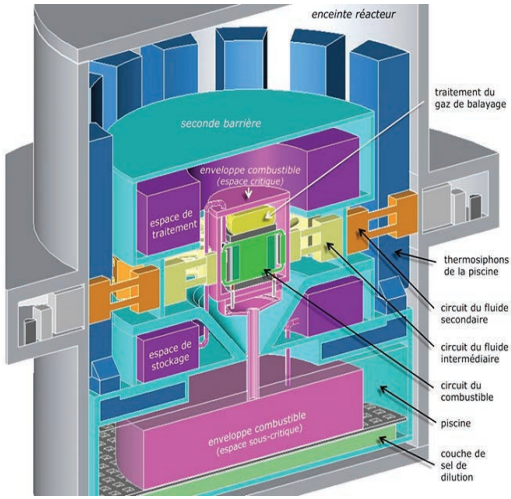


Deux cycles de réactions en chaîne.

© *Corinne Beurtey, CEA*

Schéma conceptuel du MSFR. Seules les fonctions sont représentées, les détails techniques n'étant pas définis, et les proportions relatives ne sont que spéculatives.

© *CNRS/LPSC, Grenoble*



Initiatives et prototypes en Russie, en Chine et en France

En réalité, les ressources disponibles sont encore bien plus grandes, car l'efficacité supérieure des RNR permettrait d'exploiter des minerais de basse teneur, voire l'uranium dissous dans l'eau de mer, dont la masse totale est estimée à plus de 4×10^9 tonnes.

La filière la mieux testée est celle des RNR caloportés au sodium, avec les réacteurs BN-600, BN-800 (en cours d'exploitation en Russie) et CFR-600 (en Chine). Dans cette direction, la France avait construit les prototypes Phenix et Superphenix dans les années 1973–2010, et le Commissariat à l'énergie atomique et aux énergies alternatives (CEA) a récemment achevé le projet d'études Astrid pour la réalisation d'un réacteur industriel d'une puissance électrique d'environ 600 MW.

À Grenoble (Isère), le Centre national de la recherche scientifique étudie de son côté un modèle de réacteur rapide à sels fondus, le Molten Salt Fast Reactor (MSFR), qui, à échéance d'une vingtaine d'années, pourrait offrir des performances encore supérieures, et une sûreté optimale. L'usage d'un combustible liquide, sous forme de chlorures ou de fluorures portés à 750° C, permet en effet de procéder au retraitement pyrochimique en ligne du combustible. Le réacteur est rechargé en continu, sans qu'il y ait besoin de procéder à des plans de chargement tous les quatre ou cinq ans. Le coefficient de contre-réaction thermique, très négatif en raison de la dilatation des sels, rend la réaction en chaîne parfaitement stable, même en l'absence de barres de contrôle, tandis que la cuve est à l'abri des accidents de fusion grâce à un refroidissement passif par simple convection. Enfin, on peut atteindre des puissances électriques élevées, de l'ordre de $1{,}35 \times 10^9$ W, pour seulement 18 m^3 de sels fondus, et un combustible (Th ou U) qui n'a plus besoin d'être enrichi. Cerise sur le gâteau, de tels réacteurs produisent peu de déchets, la durée de vie de ceux-ci n'excédant pas quelques siècles, et ils peuvent même incinérer la partie gênante des déchets des réacteurs actuels, à savoir les actinides mineurs à longue vie.

La simplicité de conception et le fonctionnement du MSFR à basse pression devraient, selon certains experts, le rendre à terme très compétitif avec les énergies fossiles. Toutes ces promesses nécessiteront encore beaucoup d'efforts, notamment sous forme de modélisations mathématiques et numériques, avant d'aboutir à un premier démonstrateur. Comme la Chine ou l'Amérique du Nord, il est à souhaiter que la France reprenne en toute première priorité le financement inconsidérément interrompu de la recherche et développement sur les réacteurs de quatrième génération, tels que les RNR-sodium et le MSFR.

J.–P. D.

Pour en savoir (un peu) plus :

« Le MSFR. » Daniel Heuer, université d'été *« Sauvons le climat »*, 23 septembre 2016, document de présentation disponible en ligne.

The limits to growth. (Les limites à la croissance). Donella Meadows, Dennis Meadows, Jørgen Randers et William Behrens, Universe books, 1972.

«Vers quoi l'Allemagne transite-t-elle exactement ?» Jean-Marc Jancovici, 2013, chronique disponible en ligne.

Fermeture de Fessenheim : une forfaiture. Yves Brechet, Progressistes, 2020, disponible en ligne.

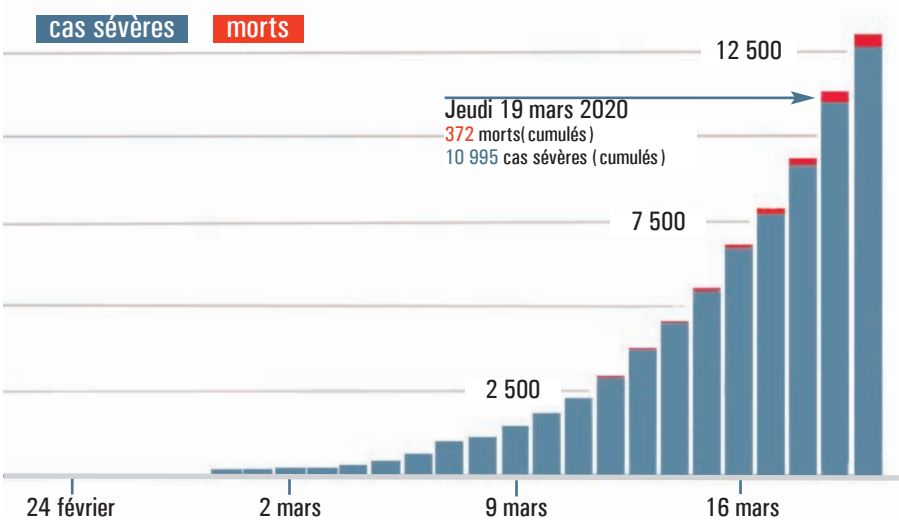
Thorium: energy cheaper than coal. Robert Hargraves, CreateSpace Independant Publishing Platform, 2012.

Les modèles mathématiques des épidémies*

Étienne Pardoux**

Professeur émérite à Aix-Marseille Université

L'année 2020 restera marquée par une pandémie mondiale inédite, qui s'est répandue à partir de la Chine dans le monde entier par le transport aérien, et se transforme en épidémies nationales une fois les frontières verrouillées. Les modèles mathématiques des épidémies permettent de répondre à des questions du type : jusqu'où la courbe des infectés et des décès va-t-elle grimper ? quelle est l'efficacité des mesures de confinement décidées dans la plupart des pays ?



La propagation du Covid-19 en France lors des premières semaines.

© *CIJM, d'après* Le Monde

* Cet article est une version complétée et remise à jour de la vidéo de l'auteur *Modèles mathématiques des épidémies comme celle du Covid-19*, disponible en ligne (**https://youtu.be/fu92X74MS_M**), elle-même inspirée par celle (en anglais) de Tom Britton, professeur à l'université de Stockholm (**https://www.youtube.com/watch?v=gSqIwXl6IjQ**). C'est également une adaptation de l'article de l'auteur *Modèles mathématiques des épidémies* paru dans le journal en ligne *The Conversation*.

**L'auteur remercie le professeur Tran Viet Chi, de l'université Paris-Est-Marne-la-Vallée, pour l'avoir aidé à s'initier en accéléré à l'utilisation du logiciel R pour les simulations.

Le modèle SIR et le nombre de reproduction de base R_0

L'un des plus vieux modèles mathématiques des épidémies est celui de Reed–Frost (1929). Il est simpliste mais il permet d'introduire des notions essentielles et d'obtenir une formule mathématique importante. Les individus sont de trois types : S (comme « susceptibles » d'être infectés), I (comme « infectés et infectieux », capables d'infecter un individu susceptible), et R (comme « remis » ou « retiré », soit guéri, soit mort). Dans ce modèle, le temps t est discret et représente par exemple un nombre de semaines. La taille n de la population est supposée « grande ». Au début du modèle, la population comporte $n-1$ individus de type S, 1 de type I et 0 de type R. Un individu qui est infecté une semaine infecte chaque susceptible avec la probabilité p la semaine suivante, puis guérit. L'épidémie se poursuit tant qu'il y a des infectés, puis elle s'arrête. Pour simplifier, on néglige la phase d'incubation, et on suppose qu'un individu de type R, s'il n'est pas mort (ce qui heureusement est le cas de l'immense majorité des R), est immunisé.

Le *nombre de reproduction de base* R_0 est le nombre moyen de susceptibles qu'un individu I infecte « au début de l'épidémie », lorsque presque toute la population est susceptible. Combien vaut R_0 ? Prenons $n=1\,000$ et $p=0{,}0025$. Le premier infecté a, autour de lui, $n-1$ (soit environ 1 000) individus susceptibles. Puisqu'il infecte chacun d'eux avec la probabilité p, R_0 vaut (environ) $n \times p$, soit 2,5.

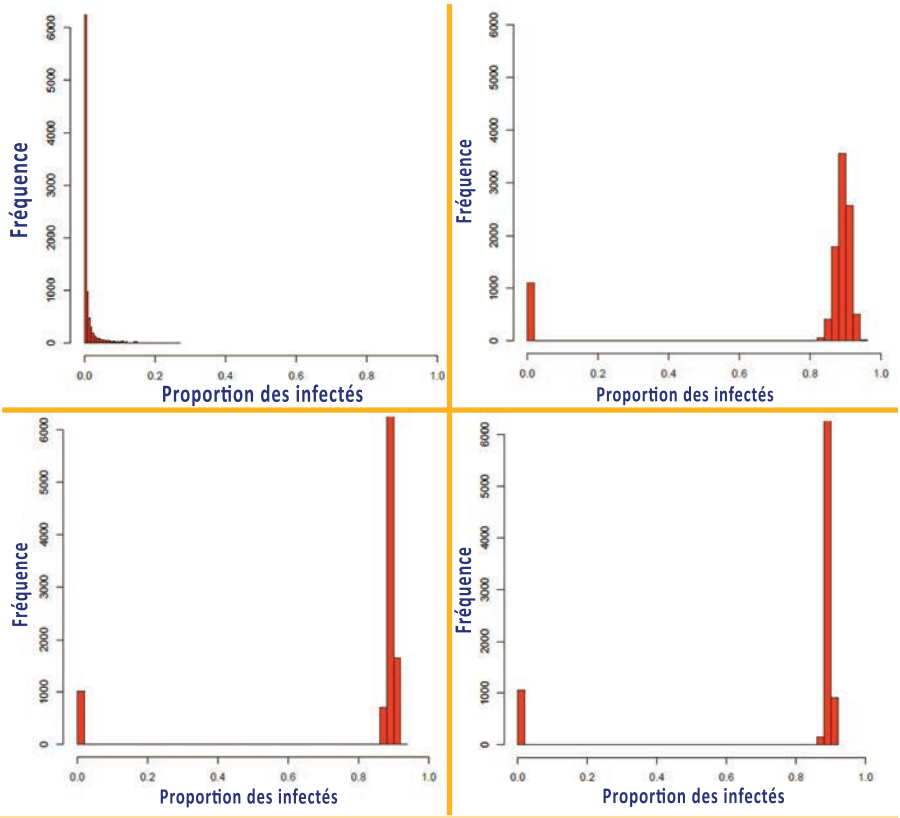
Si $R_0<1$, il n'y aura pas d'épidémie majeure avec un petit nombre d'infectés initiaux. De même si $R_0=1$. Par contre, si $R_0>1$, un seul infecté initial peut déclencher une épidémie majeure (qui touche une fraction importante de la population).

La fraction de la population qui sera touchée

Une question importante est d'estimer la fraction de la population totale qui sera touchée par l'épidémie. Si l'on admet qu'une personne guérie est immunisée, l'épidémie va s'arrêter tôt ou tard, au pire quand tout le monde aura été contaminé, et sera guéri et immunisé (un petit pourcentage étant mort). Mais dans la réalité tout le monde n'est pas touché. Lorsqu'une fraction de la population est immunisée, pour savoir combien de susceptibles un individu I infecte en moyenne, on multiplie p par le nombre de susceptibles. Bien avant que ce nombre ne s'annule, ce produit passe en dessous de 1, et alors l'épidémie s'arrête.

Effectuons dix mille simulations du modèle de Reed-Frost, dans les cas $R_0=0{,}95$ (lorsque $R_0<1$, aucune épidémie majeure n'a lieu) et $R_0=2{,}5$. La hauteur de chaque barre indique le nombre de simulations qui ont conduit à la proportion d'infectés indiquée sur l'axe des abscisses. Dans le cas $R_0=2{,}5$, une

certaine fraction des simulations (qui ne dépend pas de n) n'aboutissent pas à une épidémie majeure, tandis que la proportion d'individus infectés dans le cas d'une épidémie majeure se concentre quand n augmente autour d'une certaine valeur (laquelle augmente avec R_0).



Résultat des simulations dans les cas $R_0 = 0{,}95$ (en haut à gauche), $R_0 = 2{,}5$ et $n = 500$ (en haut à droite), $R_0 = 2{,}5$ et $n = 2\,500$ (en bas à gauche) et enfin $R_0 = 2{,}5$ et $n = 5\,000$ (en bas à droite).

© *É. Pardoux, 2020*

Supposons donc $R_0 > 1$. Dans la réalité, il n'y a pas un seul infecté initial, il y en a eu suffisamment dans chaque pays pour que l'épidémie majeure soit inévitable. Supposons que l'épidémie touche au total une fraction τ de la population. Alors la fraction $1-\tau$ de la population qui n'est pas touchée par l'épidémie est égale à la probabilité qu'un individu ne soit infecté par aucun des infectieux, car tous les individus dans la population ont la même probabilité d'être infecté. Ce qui est égal, par définition, à $(1-p)^{n\tau}$, ou encore

à $(1-R_0/n)^{n\tau}$, car les « choix des victimes » par les divers infectés sont indépendants. Cette quantité vaut approximativement $\exp(-R_0\tau)$ pour n « grand ». De même que la proportion de « piles » dans un « grand » nombre de lancers successifs d'une pièce équilibrée tend vers 1/2, la proportion aléatoire τ tend vers une valeur donnée τ^* quand n tend vers l'infini. Cette valeur τ^* est donc la solution de l'équation $\tau = 1-\exp(R_0\tau)$.

0 est toujours solution de cette équation, et c'est la seule solution si $R_0 \leq 1$. Par contre, si $R_0 > 1$, alors il y a une seconde solution $0 < \tau^* < 1$, qui ne dépend que de R_0. C'est elle qui nous intéresse.

Un ordinateur fournit une valeur approchée de τ^* pour chaque valeur donnée de R_0. On trouve en particulier, pour $R_0 = 1{,}5$, 2, 2,5 et 3, les valeurs respectives de τ suivantes : 58 %, 79 %, 89 % et 94 %.

Trois hypothèses essentielles

Les résultats obtenus avec le modèle de Reed-Frost peuvent en fait être établis dans un cadre plus général, respectant les trois hypothèses suivantes.

(H1) Pas d'immunité au départ (naturelle ou par vaccination). (H1) est vraie pour la maladie à coronavirus 2019 (Covid-19), pas pour la grippe saisonnière.

(H2) Une communauté homogène. (H2) n'est pas vraie pour le Covid-19, ni pour aucune maladie ! Les hétérogénéités qu'il convient de prendre en compte dépendent de chaque maladie et de son mode de transmission. En situation réelle, on considère qu'il faut réduire le nombre d'infectés d'environ 10 % à 20 % par rapport aux prédictions du modèle homogène.

(H3) Le comportement des individus ne change pas au cours de l'épidémie. Les mesures prises par les autorités (fermeture des écoles et des lieux publics, confinement...) changent évidemment la donne.

Réduire R_0 : confinement, mesures barrières, vaccination…

Comment réduire R_0 ? Factorisons ce nombre ainsi pour voir comment il est construit : $R_0 = p \times c \times u$, avec p la probabilité qu'un contact produise une infection, c le nombre moyen de « contacts » par jour et u la durée (en nombre de jours) de la période d'infection. Les mesures de prévention visent à réduire R_0 en diminuant p (port de masques, lavage des mains…), c (confinement, interdiction des regroupements…) et u (diagnostic rapide, mise à l'isolement des infectés…).

Maintenant, si une proportion v de la population est vaccinée, alors R_0 est remplacé par $R_v = (1-v)R_0$ car chaque « tentative d'infecter » ne réussit qu'avec la probabilité $1-v$. On en déduit que $R_v < 1$ dès que $v > 1 - 1/R_0$. Ainsi, R_0 nous indique la fraction de la population qu'il faut vacciner si l'on veut être sûr qu'il n'y aura pas d'épidémie majeure (si $R_0 = 2{,}5$, il faut vacciner 60 % de la population).

Estimer R_0 est vraiment crucial ! Mais, observer la courbe d'incidence (la première figure de l'article) ne suffit pas. Le taux de croissance r de la courbe dépend en effet de deux facteurs : R_0, et le *temps de génération* G (la durée entre le moment où l'on est infecté et celui où l'on infecte). Le taux r croît avec R_0, décroît avec G. Les mathématiciens ont établi des formules qui relient r, R_0 et la loi de probabilité de G. Par une méthode de *suivi des contacts* (recherche de qui a infecté qui et quand), on peut obtenir des informations sur la loi de G. En les combinant avec la lecture de r sur les courbes d'incidence, on peut estimer R_0 !

Un modèle déterministe pour la dynamique de l'épidémie

Pour décrire l'évolution de l'épidémie, dans le cas d'une « grande » population, et partant d'une situation où une fraction significative de cette population est déjà touchée, on va utiliser un modèle déterministe.

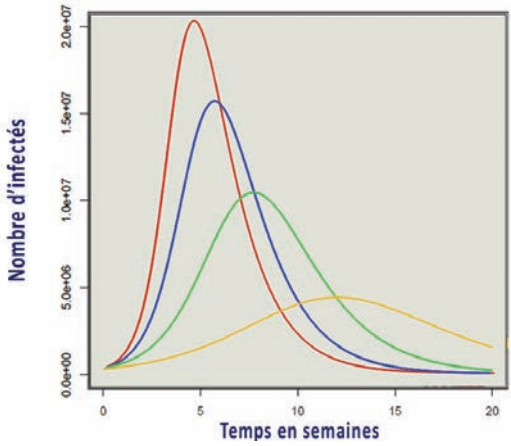
Le modèle d'évolution

Le modèle déterministe SIR pour décrire l'évolution de l'épidémie est le suivant. On appelle $s(t)$ la fraction d'individus susceptibles à l'instant t, et $i(t)$ celle des infectés. Le modèle s'écrit sous la forme d'un système d'équations différentielles :

$$\begin{cases} \dfrac{ds}{dt}(t) = -\alpha\, s(t)\, i(t), \\ \dfrac{di}{dt}(t) = \alpha\, s(t)\, i(t) - \beta\, i(t), \end{cases}$$

où α est la force de l'infection et β l'inverse de la durée moyenne d'infection. Au début de l'épidémie, $s(t)$ est « proche » de 1, et chaque individu I infecte au taux α pendant une durée de moyenne $1/\beta$. Donc $R_0 = \alpha/\beta$.

Le système d'équations différentielles qui régit le modèle est non linéaire. Si l'on fixe la condition initiale, ce système a une solution unique. Sur les quatre premières courbes ci-dessous, plus R_0 est grand, plus la vague des infectés arrive tôt et monte haut. En outre, lorsque l'on réduit R_0, la réduction de la hauteur du pic est plus marquée que la réduction de la taille totale de l'épidémie.



L'évolution du nombre d'infectés pour des valeurs de R_0 égales à 3 (rouge), 2,5 (bleu), 2 (vert) et 1,5 (jaune).

© *É. Pardoux, 2020*

Concernant le Covid-19 en France, alors même que l'épidémie était déjà à un stade très avancé, beaucoup d'incertitudes subsistaient quant à la valeur de R_0 : les estimations variaient entre 2,5 et 5. Si le gouvernement n'avait pas pris de mesures fortes, entre 60 et 80 % de la population aurait probablement été touchée. Surtout, la « vague » serait arrivée très vite et montée très haut, submergeant le système de santé. D'autres facteurs ont contribué à rendre les modèles particulièrement délicats avec le Covid-19 : on ne dispose pas à ce jour d'une bonne estimation du nombre de personnes en France qui ont été touchées par le virus. Cette maladie est en effet bénigne, et même indécelable, dans beaucoup de cas. Il est donc difficile de savoir quelle fraction de la population est immunisée !

Les modèles mathématiques sont indispensables pour les décideurs. Il convient de poursuivre leur étude pour obtenir, à l'avenir, des modèles plus fins, plus proches des conditions dans lesquelles une nouvelle épidémie pourrait survenir. Voici quelques problèmes sur lesquels travaillent les mathématiciens et statisticiens qui s'intéressent aux modèles des épidémies : prise en compte de l'hétérogénéité de la population (maisonnées et lieux de travail, répartition spatiale, voyages…) et analyser ses effets ; comparer l'intérêt des modèles les plus « réalistes » et des modèles plus « simples » ; mesurer l'effet de diverses mesures préventives ; délimiter précisément ce que les données disponibles permettent d'obtenir comme conclusions ; identifier les données supplémentaires qui pourraient être utiles… Le chantier est vaste. Donc si vous voulez sauver des vies, vous pouvez aussi faire des maths !

É. P.

Émergence de résistance aux antimicrobiens

Lionel Roques

Directeur de recherche à l'Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE)

Résistance aux traitements : bactéries pathogènes, mais également plantes et animaux

D'après l'Organisation mondiale de la santé (OMS), l'émergence de résistances aux antimicrobiens constitue l'une des plus grandes menaces sur la santé humaine. Pour survivre, certains microbes parviennent à développer des mécanismes de résistance à l'action toxique des antimicrobiens. Ainsi, en 2017 déjà, l'OMS avait publié une liste de douze familles de bactéries les plus menaçantes pour la santé humaine, car résistantes à la plupart des antibiotiques. Il existe même des bactéries résistantes à quasiment tous les antibiotiques connus (bactéries *toto-résistantes*). Ces phénomènes d'émergence de résistance aux traitements, bien connus du grand public lorsqu'il s'agit de bactéries pathogènes pour l'homme, concernent également les plantes et les animaux.

Darwin, la sélection naturelle et le sauvetage évolutif

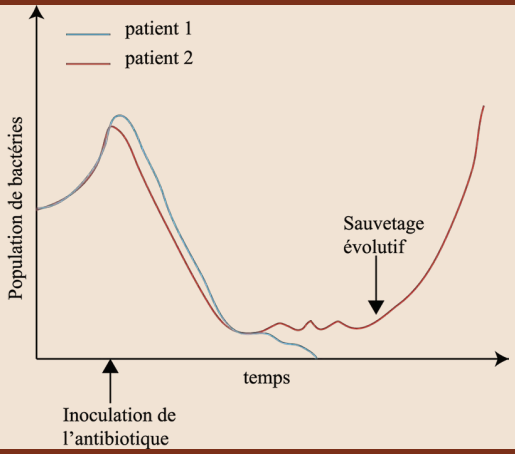
Un des principaux mécanismes biologiques permettant d'expliquer l'émergence de ces résistances est la *sélection naturelle*, principe initialement décrit par Darwin, et correspondant au fait que certains individus dans une population (ici des microbes) ont tendance à avoir plus de descendants car ils sont mieux adaptés à leur environnement. Au cours des générations, des mutations peuvent apparaître. La plupart sont délétères, et conduisent à des individus moins bien adaptés à leur environnement que leurs parents, mais certaines sont bénéfiques : les individus porteurs de ces mutations seront mieux adaptés à leur environnement que leurs parents.

Ces individus auront donc tendance à avoir plus de descendants, porteurs de ces mêmes mutations … et ainsi de suite, jusqu'à l'émergence d'individus bien adaptés à l'environnement.

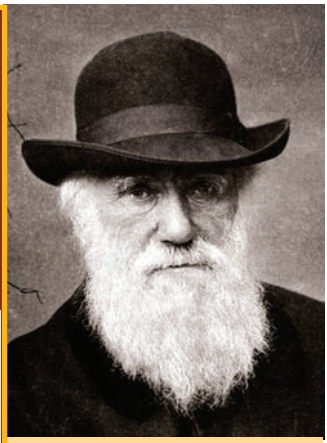
Le sauvetage évolutif correspond au phénomène suivant : une population initialement déclinante, car soumise à un environnement hostile (comme une population de bactéries en présence d'un antibiotique), parvient à s'adapter à cet environnement et finalement à s'y multiplier librement.

Populations de bactéries soumises à un antibiotique

Un antibiotique est inoculé à deux patients infectés par une même bactérie. Dans le cas du patient 1, le traitement est efficace et conduit à l'éradication de la bactérie. En revanche, un phénomène de sauvetage évolutif se produit chez le patient 2, rendant le traitement antibiotique inefficace.



© L. Roques, 2020



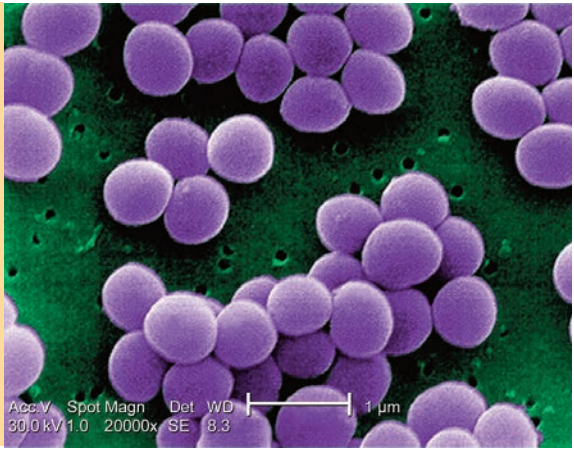
Charles Darwin (1809–1882) est l'auteur de *l'Origine des espèces* en 1859, ouvrage dans lequel il introduit la notion de sélection naturelle.

© Elliott & Fry, 2009

En temps normal, notre flore microbienne est faite de millions de souches, plus ou moins sensibles aux antibiotiques. Lorsqu'un patient est malade et prend un traitement antibiotique, même si celui-ci se révèle efficace pour traiter le microbe responsable de ses symptômes, la pression de sélection qu'il exerce sur l'ensemble des microbes présents dans le corps a tendance à sélectionner ceux qui sont les plus résistants, et peut conduire à des phénomènes de sauvetage évolutif, créant ainsi des microbes résistant à cet antibiotique, qui pourront par la suite envahir l'environnement autour du patient… et finir par faire le tour de la planète.

Le staphylocoque doré (ou *staphylococcus aureus*) est une bactérie présente chez 30 à 50 % des sujets sains et souvent impliquée dans les infections nosocomiales en milieu hospitalier, sous sa forme multirésistante SARM (*staphylococcus aureus* résistant à la méticilline).

© *Janice Haney Carr, 2009*



Le concept de sauvetage évolutif ne concerne pas uniquement les bactéries pathogènes ; il peut également être mis en jeu, par exemple, lors de l'adaptation d'espèces vivantes à un changement du climat.

Émergence d'une résistance : les maths à la rescousse

Des modèles mathématiques peuvent être construits pour décrire l'évolution de populations de microbes en présence d'un antibiotique. Cela permet de comprendre, en limitant le nombre d'expériences, et pour des catégories génériques de microbes, le rôle des différents paramètres sur la probabilité d'émergence d'une résistance. Un outil mathématique simple, la notion de suite récurrente, permet déjà de construire un premier modèle démographique (sans évolution).

Considérons d'abord une population $N(t)$ de microbes à un instant $t \geq 0$. Au temps $t+1$, la population est de $N(t+1)=N(t)+R$ individus, avec R la différence entre le nombre de naissances et le nombre de morts parmi les microbes pendant une unité de temps (disons une génération). L'hypothèse la plus naturelle est de supposer que R est proportionnel à $N(t)$; on peut alors écrire $R=(r-1)N(t)$, avec r le taux de croissance de la population. On a donc $N(t+1)=rN(t)$.

Dans ce modèle très simple, la croissance est dite *malthusienne*, en référence à l'économiste britannique Thomas Robert Malthus (1766–1834). La solution est simplement égale à $N(t)=N_0 r^t$, pour tout instant $t \geq 0$, avec N_0 la population initiale. Donc si $r>1$, la population croît exponentiellement, tandis que si $r<1$, la population tend vers 0 (donc s'éteint).

Ces modèles sont dits *déterministes* : une fois que les paramètres N_0 et r sont fixés, il n'y a qu'une issue possible, qui dépend ici uniquement de r. Pourtant, on sait expérimentalement que des conditions comparables peuvent engendrer des résultats très différents (comme dans l'exemple de l'encadré page 28). C'est en particulier vrai quand l'expérience fait intervenir de « petites » tailles de populations : dans le cas du sauvetage évolutif, la population de microbes prend des « petites » valeurs avant de rebondir.

Une façon de prendre cette variabilité en compte consiste à considérer des approches *probabilistes*. Imaginons un dé à quatre faces (un tétraèdre), marquées –1, 0, 0, 1. À chaque pas de temps, on tire ce dé, et on note la valeur obtenue $\varepsilon(t)$. Un exemple de modèle d'évolution de population probabiliste serait alors $N(t+1) = r\,N(t) + \varepsilon(t)\sqrt{N(t)}$. Même avec $r > 1$, ce modèle peut conduire à une extinction de la population.

Comment prendre en compte les mutations

Sous l'effet de la sélection naturelle, le terme de croissance r devrait changer au cours du temps. L'émergence d'une résistance correspondrait à un facteur r initialement plus petit que 1 (à cause de l'antibiotique), qui deviendrait par la suite plus grand que 1. Pour décrire le mécanisme de sélection naturelle, il convient de prendre en compte le fait que certains individus ont tendance à avoir plus de descendants.
Ainsi, supposons que la population $N(t)$ soit potentiellement composée de deux types : un type 1, sensible à l'antibiotique ($r_1 < 1$, population N_1) et un type 2, résistant ($r_2 > 1$, population N_2). Les tailles de populations correspondantes sont notées $N_1(t)$ et $N_2(t)$. Pour décrire la mutation, on fait l'hypothèse qu'à chaque pas de temps, une partie des individus N_1 deviennent des N_2, et inversement. Notons u (compris entre 0 et 1) ce taux de mutation. On peut calculer $N_1(t)$ et $N_2(t)$ de la façon décrite précédemment, avec des suites récurrentes :

$$N_1(t+1) = r_1\,N_1(t) + u\,N_2(t) - u\,N_1(t) + \varepsilon_1(t)\sqrt{N_1(t)}$$

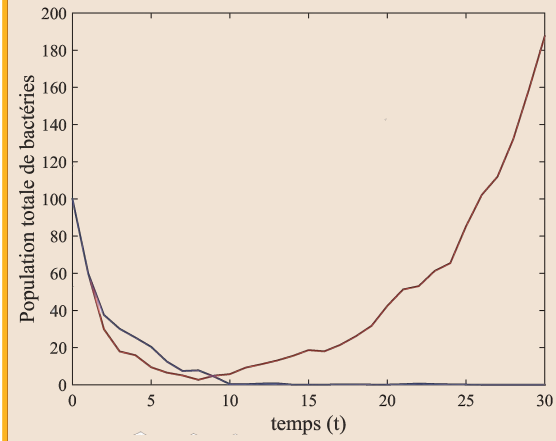
et

$$N_2(t+1) = r_2\,N_2(t) + u\,N_1(t) - u\,N_2(t) + \varepsilon_2(t)\sqrt{N_2(t)}$$

En additionnant les deux équations, on voit que le taux de croissance global de la population $N(t) = N_1(t) + N_2(t)$ est :

$$r(t) = \frac{r_1\,N_1(t) + r_2\,N_2(t)}{N_1(t) + N_2(t)}.$$

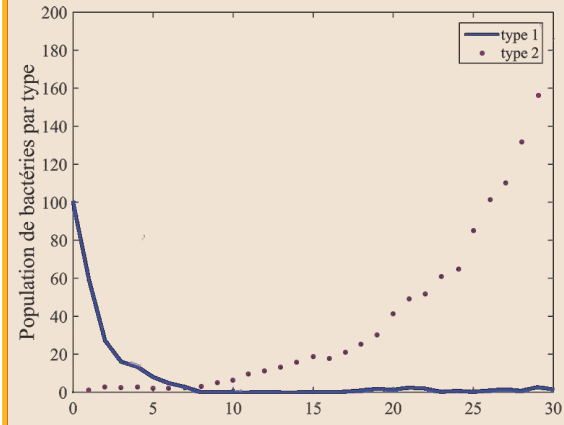
Le travail du mathématicien est avant tout d'établir des propriétés du modèle, sans nécessairement avoir à le simuler avec un programme informatique. Mais cela nécessite des outils mathématiques plus avancés (relevant ici de l'analyse stochastique). Contentons-nous donc de simulations informatiques. Le modèle étant très simple, une calculatrice programmable suffit !



Simulations numériques

On part d'une population clonale de bactéries : seul le type 1 (sensible) est présent initialement, avec un effectif de cent bactéries. La trajectoire bleue correspond à un cas où la bactérie est éradiquée. Ici, les valeurs des paramètres sont $r_1 = 0{,}6$, $r_2 = 1{,}2$ et $u = 0{,}01$.

© L. Roques, 2020



Simulations numériques

Avec les mêmes paramètres du modèle, une autre simulation conduit à la trajectoire rouge (sauvetage évolutif). Dans ce cas, on est tenté de regarder la composition de la population. Le type 1 (sensible) décline rapidement, mais les mutations font apparaître le type 2 (résistant), qui, malgré de faibles effectifs, évite l'extinction et finit par croître rapidement.

© L. Roques, 2020

Les résultats de simulations (en encadré) laissent apparaître le phénomène de sauvetage évolutif ! Le modèle arrive donc bien à reproduire des phénomènes observés expérimentalement.

Utilisons-le maintenant pour améliorer notre compréhension du phénomène d'émergence de résistance. Pour chaque valeur du taux de mutation u, simulons un grand nombre de fois le modèle. En faisant le rapport entre le nombre de cas où un sauvetage évolutif a lieu et le nombre total de simulations, on en déduit une probabilité d'émergence de résistance. En répétant ceci pour des valeurs du taux de mutation u dans l'intervalle $[0, 1]$, on obtient le graphique suivant. On observe qu'un taux de mutation nul ($u = 0$) conduit à l'impossibilité d'émergence d'une résistance (car seul le type 1, sensible à l'antibiotique, est initialement présent). Puis la probabilité augmente et atteint des valeurs très proches de 1, avant de diminuer à nouveau quand le taux de mutation u augmente. Ce dernier phénomène, connu des biologistes, s'appelle la *mutagénèse létale*. Quand le taux de mutation u devient très grand, tout se passe comme si on avait un unique type de bactérie, avec un taux de reproduction égal à la moyenne de r_1 et r_2. Ici, cette moyenne étant plus petite que 1, l'émergence d'une résistance n'est pas possible.

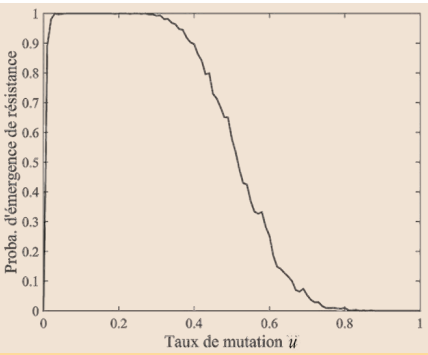


Illustration du phénomène de mutagénèse létale. La probabilité d'émergence de résistance, en fonction du taux de mutation u, est calculée en simulant un grand nombre de fois le modèle.

©*L. Roques, 2020*

Des modèles plus complexes pour tester des combinaisons

Grâce aux mathématiques, il est possible de tester une multitude de scénarios, dont tous ne seraient pas envisageables expérimentalement, souvent pour des raisons pratiques ou financières, parfois pour des raisons éthiques. On peut ainsi tester des combinaisons d'antibiotiques, faire varier les durées de traitement, leur intensité… et concevoir des protocoles thérapeutiques optimisés, qui resteront bien sûr à évaluer médicalement *in fine*. Parmi les pistes envisagées pour éviter l'émergence de résistance, certaines font appel à la phagothérapie (destruction de bactéries en utilisant des virus bactériens, les bactériophages). Pour décrire ce type de traitement, il faudrait non seulement décrire l'évolution de la population de bactéries, mais également l'évolution de la population des bactériophages. Le modèle construit ici est simplement basé sur la notion de suite récurrente : le temps t change de façon discrète, par sauts successifs. Les modèles continus en temps sont souvent plus réalistes et plus simples à analyser mathématiquement. Ils nécessitent toutefois de mobiliser des outils mathématiques différents (analyse et équations différentielles).

L. R.

Les modèles de climat*

Éric Blayo, Laurent Debreu et Christine Kazantsev

Université Grenoble-Alpes et Inria

Montée du niveau des océans, modification des régimes de pluie, intensification de phénomènes météorologiques extrêmes, fonte des glaces, conséquences sur la faune et la flore : évaluer l'ampleur et les conséquences du changement climatique en cours, et déterminer l'impact de futurs choix politiques et économiques, sont des enjeux actuels majeurs. Il est nécessaire pour cela de décrire et comprendre le « fonctionnement » du climat, notamment par l'étude des climats passés et par la mise au point de modèles permettant d'effectuer des projections dans le futur. La complexité du système climatique et la diversité des questions le concernant amènent pour cela à faire appel à de nombreuses branches des mathématiques.

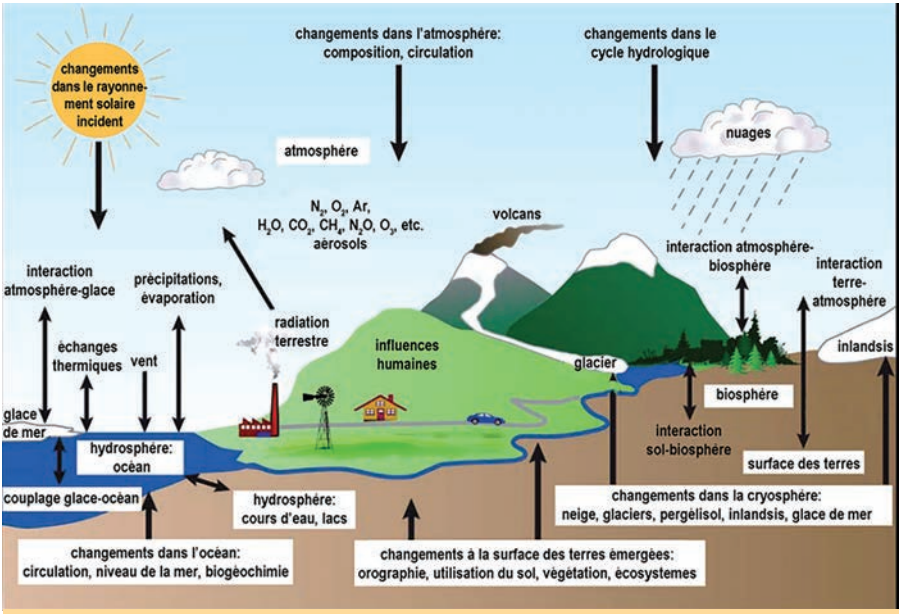
Des mathématiques pour décrire la nature

Décrire le climat, c'est tout d'abord expliciter le fonctionnement de ses principales composantes : l'atmosphère et l'océan, bien sûr, qui sont au cœur du système, mais aussi la cryosphère (calottes polaires, banquises, glaciers, surfaces enneigées), les cours d'eau, les sols, la biosphère… Ceci est réalisé au travers de modèles, c'est-à-dire d'équations mathématiques traduisant les principes qui régissent l'évolution du système : lois de conservation (par exemple masse ou énergie), transformations chimiques, relations entre espèces… Certaines sont connues depuis les travaux de grands scientifiques du passé, comme Leonhard Euler (1707–1783), Jean Baptiste Joseph Fourier (1768–1830), Claude Louis Marie Henri Navier (1785–1836), George Gabriel Stokes (1819–1903), Gustave Gaspard Coriolis (1792–1843) ou Joseph Valentin Boussinesq (1842–1929).

Aujourd'hui, les lois de comportements de l'atmosphère et de l'océan sont assez bien connues. C'est par contre moins vrai pour celles régissant la glace, la biogéochimie marine ou encore la végétation terrestre, même si la connaissance progresse sans cesse. De plus, au-delà de la description de chacun de ces « compartiments », il faut aussi expliciter leurs interactions :

* Cet article est une version mise à jour d'un texte publié en décembre 2013 sur le site **interstices.info**

échanges d'énergie, d'eau, de carbone… Ces interactions sont souvent très complexes, et il reste encore pour beaucoup d'entre elles une grande marge d'incertitude quant à leur mise en équations.



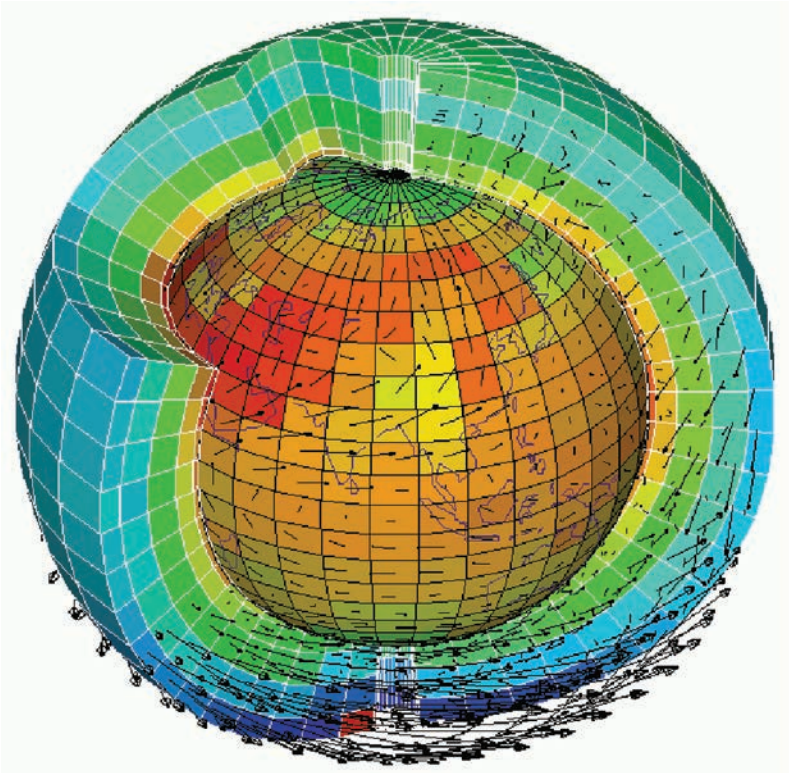
Le système couplé atmosphère – hydrosphère – cryosphère – lithosphère – biosphère (dont anthroposphère) de la planète Terre et dont l'océan est une composante.

© *Intergovernmental Panel on Climate Change, 1999*

Le climat, un système dynamique sophistiqué à l'extrême

De nombreux modèles ont été développés pour décrire le climat. D'un point de vue mathématique, il s'agit de *systèmes dynamiques non linéaires*, domaine dont la théorie a été initiée notamment par Henri Poincaré (1854–1912). À vrai dire, on ne sait analyser les propriétés mathématiques des modèles de climat que pour des cas extrêmement simplifiés, qui n'ont que peu de choses à voir avec la réalité. Il est cependant généralement admis que le climat est un système dynamique présentant plusieurs *bassins d'attraction* (régimes d'équilibre ou de quasi-équilibre), potentiellement très différents les uns des autres, et entre lesquels des transitions abruptes seraient possibles. Le risque d'un tel basculement du régime actuel vers un autre régime inconnu est souvent évoqué lorsque l'on envisage des scénarios catastrophes, comme par exemple l'arrêt du Gulf Stream sous l'effet d'importants apports d'eau douce en mer du Labrador dus notamment à la fonte des glaces du Groenland.

Les modèles mathématiques de climat sont donc constitués de systèmes d'équations traduisant le comportement de chaque « compartiment » (atmosphère, océan…) ainsi que les interactions entre ces compartiments. La complexité de ces modèles est telle qu'il est hors de question de les résoudre de façon exacte. Par contre, on peut en chercher des solutions approchées (mais pas approximatives !) grâce aux techniques de *simulation numérique* sur ordinateur. L'idée de base consiste à définir un *maillage* de chaque composante (par exemple, découper l'atmosphère en « un grand nombre de briques » de quelques kilomètres de côté) et à approcher les équations de départ à l'intérieur de chaque maille en ne faisant plus intervenir que les quantités physiques dans les mailles voisines (on parle de *discrétisation*). Ainsi, la dérivée selon la direction x de la température au centre de la maille numéro i sera remplacée par un taux d'accroissement calculé à partir des températures dans les mailles de gauche ($i-1$) et de droite ($i+1$), par exemple.



La Terre maillée : maillage tridimensionnel de la composante atmosphérique d'un modèle de climat. Les couleurs représentent la température, et les flèches, le vent.

© *Vincent Landrin, d'après Laurent Fairhead/LMD/CNRS*

On remplace donc les équations continues initiales, très complexes, et contenant de nombreuses dérivées, par un grand système d'équations dont les inconnues sont « seulement » les variables physiques (température, vitesse, pression, concentrations chimiques…), et qui peut être résolu par ordinateur. Les mathématiques appliquées interviennent ici à plusieurs niveaux : pour faire en sorte que les solutions des équations discrétisées au sein de chaque composante soient une « bonne » approximation des solutions des équations de départ, pour coupler de façon cohérente ces composantes, ou encore pour résoudre efficacement le très grand système discrétisé.

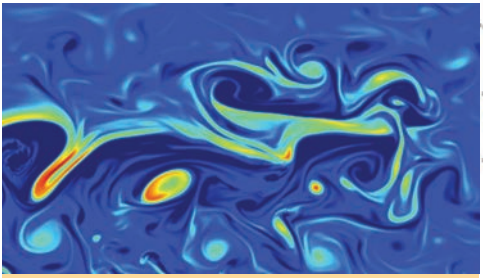
La mise en œuvre informatique de tels modèles est un véritable défi de calcul scientifique. Le comportement du système climatique, même à grande échelle, est en effet largement influencé par les phénomènes de plus petite échelle. Il est donc théoriquement nécessaire d'utiliser les modèles avec une résolution spatiale suffisamment fine. Idéalement, celle-ci devrait être de l'ordre de quelques kilomètres pour l'atmosphère, et être encore plus fine pour l'océan. Mais la taille des systèmes à résoudre, et donc le nombre d'opérations à effectuer, dépasserait alors très largement les capacités des plus puissants supercalculateurs actuels. La résolution des modèles climatiques est donc fixée pour l'instant, et le sera encore pendant de nombreuses années, en fonction de la puissance de calcul disponible. Elle demeure à l'heure actuelle relativement faible : de l'ordre de 50 à 100 km sur l'horizontale pour l'atmosphère et l'océan, et quelques dizaines à centaines de mètres sur la verticale. Les plus fines échelles spatiales ne sont donc pas présentes dans les modèles, et leurs effets sur le climat à plus grande échelle ne peut alors qu'être simulé de façon imparfaite, au moyen de termes *ad hoc* artificiellement ajoutés aux équations.

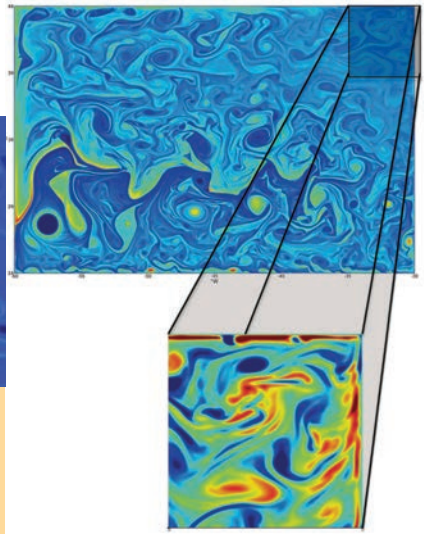
Mais même avec cette résolution insuffisante, les modèles doivent calculer l'évolution de plusieurs dizaines (voire centaines) de millions de variables, sur des durées de plusieurs dizaines d'années, avec une fréquence de l'ordre de quelques minutes ou dizaines de minutes, ce qui représente un volume de calcul gigantesque.

Comportement statistique du temps et projections climatiques

Au-delà de la simulation des climats passés, un intérêt majeur de ces véritables monstres numériques que sont les modèles climatiques est bien évidemment la prévision du climat futur. Il est intrinsèquement impossible de prévoir la météo au-delà d'une échéance de l'ordre de quinze jours. C'est le fameux *effet papillon*, rendu célèbre par Edward Norton Lorenz (1917 – 2008), qui illustre le caractère fondamentalement chaotique de la dynamique de l'atmosphère. Il en est de même d'ailleurs pour la dynamique de l'océan, sur des échéances toutefois un peu plus longues.

Cela étant dit, le comportement statistique (c'est-à-dire le climat) est, lui, tout à fait prévisible plusieurs dizaines d'années à l'avance. Dans l'idée, cela est du même ordre que prévoir précisément l'évolution de la température moyenne dans une casserole d'eau sur une cuisinière à gaz selon le réglage de la flamme, alors que l'on est incapable d'anticiper le comportement exact de chaque molécule.





Images de turbulences :
simulation numérique de tourbillons océaniques
(résolution de 1 km).

© Pierre-Antoine Bouttier, 2014, thèse de doctorat de l'université de Grenoble

Être capable de prévoir le climat futur suppose au préalable d'être capable de reproduire le climat passé. Les modèles numériques sont donc réglés et validés en comparant leurs résultats aux observations disponibles, notamment celles, nombreuses, des décennies récentes. Une fois mis au point, ils peuvent alors fournir des projections sur les évolutions possibles du climat actuel dans les décennies, voire les siècles, à venir.

Ces études sont coordonnées au niveau international par le GIEC (Groupe d'experts intergouvernemental sur l'évolution du climat). Créé en 1988 par l'ONU, il fournit régulièrement un rapport de synthèse sur les connaissances scientifiques concernant le changement climatique (cinq rapports ont été publiés entre 1990 et 2014, le sixième est prévu pour 2021–2022). Plusieurs « futurs possibles » des sociétés humaines (développement économique et démographique, choix énergétiques, évolution des comportements individuels) sont traduits en termes de scénarios d'émissions de gaz à effet de serre. Les centres de recherche de par le monde (plus de vingt pour le prochain

rapport) simulent alors avec leurs différents modèles les évolutions climatiques qui en résulteraient. Une synthèse globale est ensuite réalisée afin de dégager des tendances fiables, et de quantifier les incertitudes autour de ces tendances. Cette synthèse requiert d'ailleurs une méthodologie statistique sophistiquée : comment synthétiser de manière cohérente les résultats de plusieurs dizaines de simulations climatiques ?

À l'heure actuelle, il y a globalement convergence entre ces modèles concernant l'évolution à l'échelle planétaire. Par contre, préciser les futurs changements climatiques à l'échelle d'un pays ou d'une région demeure extrêmement difficile, du fait de l'influence locale du caractère chaotique de la météorologie.

Extrêmement active à l'échelle internationale, la recherche en sciences du climat a conduit à des progrès remarquables ces dernières années. Les travaux, souvent fondamentalement pluridisciplinaires, font appel à une large palette de mathématiques : théorie des systèmes dynamiques, méthodes statistiques, analyse numérique, calcul scientifique… Pour les années à venir, de grandes questions demeurent : projections climatiques à l'échelle régionale et à échéance de dix ou vingt ans, modélisation des impacts du changement sur les espèces vivantes ou sur l'économie… La contribution de scientifiques d'horizons très divers est absolument nécessaire pour relever ces défis, et les mathématiciens ont un rôle à y jouer.

É. B. – L. D.– C. K.

Pour en savoir (un peu) plus :

Chaos. Aurélien Alvarez, Étienne Ghys et Jos Leys, 120 mn, 2013, film disponible en ligne.

Le climat à découvert : outils et méthodes en recherche climatique. Sous la direction de Catherine Jeandel et Rémy Mosseri, CNRS Éditions, 2011.

Le climat en équations. Éric Blayo, *Interstices*, 2013

« Tout savoir sur la météo, le climat et Météo-France. » **meteofrance.fr/climat-passe-et-futur.**

« The Intergovernmental Pannel on Climate Change. » **https://www.ipcc.ch** (en anglais).

La fonte des calottes polaires*

Jocelyne Erhel

Institut national de recherche
en informatique et automatique (INRIA)
Institut de recherche mathématique de Rennes

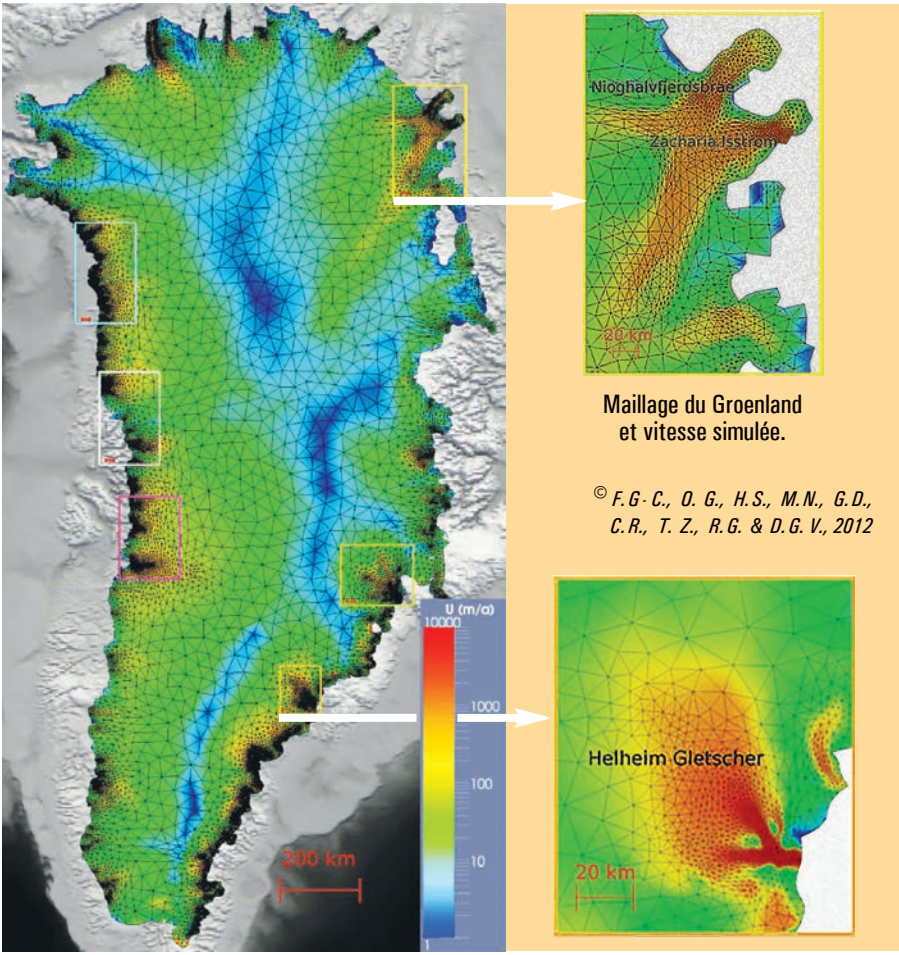
Le niveau des mers monte, pour différentes raisons liées au réchauffement climatique. Les glaciers de l'Antarctique et du Groenland, que l'on appelle *calottes polaires* ou *inlandsis*, jouent un rôle majeur dans l'évolution du niveau des mers. Peut-on prévoir l'évolution future de ces calottes polaires et en particulier le vêlage (perte de fragments) d'icebergs dans l'océan ?

Des compétences scientifiques issues de toutes les disciplines

Trois phénomènes contribuent à la montée du niveau des mers. Le premier est la modification en surface des précipitations neigeuses et de la fonte de glace, à cause du changement climatique et du changement d'altitude du glacier. Le deuxième est l'amincissement des plateformes flottantes, qui fondent par en-dessous à cause du réchauffement de l'océan. Le dernier est la décharge de glace dans l'océan, lorsque la glace s'écoule vers la mer puis se brise en larguant des icebergs. Ces processus physiques sont étudiés par des glaciologues, en collaboration avec des climatologues, océanographes, mathématiciens et informaticiens. Un modèle numérique, basé sur un modèle mathématique décrivant la physique du système, permet de simuler l'évolution des calottes polaires, notamment le vêlage d'icebergs. Pour construire le modèle mathématique, on adopte la démarche générique présentée dans l'article précédent. On considère la calotte polaire comme un volume délimité par des frontières, puis, grâce aux lois de la physique (mécanique des fluides, lois de conservation…), on construit un système d'équations qui traduit ce qui se passe à l'intérieur du glacier et à ses frontières. Pour calculer la vitesse et l'altitude de surface en chaque point de la calotte polaire et à chaque instant, on a recours à un modèle numérique ; la solution numérique approchée ainsi obtenue est assortie d'une erreur que l'on sait contrôler mathématiquement.

La première étape est la discrétisation spatiale : la surface du glacier est découpée en petits triangles, et l'épaisseur du glacier en tranches, ce qui aboutit à un découpage du volume glaciaire en prismes (les mailles).

* Ce texte est une version actualisée de deux articles parus dans la revue de culture scientifique en ligne *Interstices*, créée par des chercheurs pour vous inviter à explorer les sciences du numérique (**https://interstices.info**).



Maillage du Groenland et vitesse simulée.

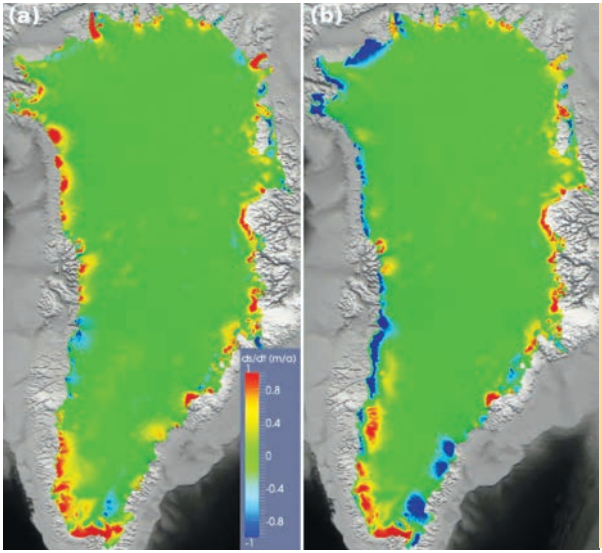
© *F. G - C., O. G., H. S., M. N., G. D., C. R., T. Z., R. G. & D. G. V., 2012*

Pour calculer la vitesse et la pression à l'intérieur de la calotte polaire, connaissant l'altitude de surface, il faut résoudre les équations de Stokes discrétisées. Pour déterminer l'altitude de surface, il faut en plus discrétiser le temps et calculer l'altitude à des instants réguliers (tous les ans par exemple). Pour cela, il faut recourir à des mesures et actualiser le bilan de masse de surface (BMS, différence entre l'accumulation et la fonte de glace liées aux conditions météorologiques et au déplacement du glacier) : on calcule le BMS pour l'année 0 et on en déduit l'altitude de surface pour l'année 1 ; on calcule alors la vitesse pour l'année 1, puis le BMS, puis l'altitude de surface pour l'année 2, *etc*.

On combine ainsi des modèles et des observations.

Pour résoudre les équations de Stokes discrétisées, il y a deux difficultés : les équations sont non linéaires, et le système contient beaucoup d'équations (environ quatre fois le nombre de mailles). Comme on ne sait pas résoudre de façon exacte un tel système, on utilise un algorithme itératif, qui procède par approximations successives. On contrôle mathématiquement l'erreur et on arrête l'algorithme lorsque l'erreur est jugée « assez petite ».

Un code informatique adéquat permet enfin de simuler l'évolution de la calotte polaire selon divers scénarios socio-économiques.



Simulation de l'évolution de l'altitude de surface en dix ans : en rouge, l'altitude augmente ; en bleu, elle diminue ; en vert, elle ne bouge pas.

© *F.G-C., O. G., H.S., M.N., G.D., C.R., T. Z., R.G. & D.G. V., 2012*

Les figures ci-dessus montrent des résultats de simulation pour une période de dix ans, obtenus avec le modèle Elmer/Ice. Les conditions à la base restent identiques. Dans (a), les conditions climatiques actuelles restent constantes pendant dix ans (ce qui n'est pas réaliste en raison du réchauffement actuel). Dans (b), on utilise la moyenne de dix-huit modèles utilisant le scénario médian d'émissions de CO_2. Les bords du Groenland s'épaississent pour (a) et s'amincissent pour (b), même en seulement dix ans.

Des calculs qui prennent encore beaucoup trop de temps !

Le modèle a été utilisé pour simuler l'évolution du Groenland pendant un siècle. Il est également bien adapté pour simuler l'évolution des glaciers terrestres. Mais même avec l'aide de supercalculateurs, les calculs prennent trop de temps : le temps de calcul augmente avec le nombre d'équations, donc

avec le nombre de mailles. Or, pour avoir une bonne précision, il faut disposer de beaucoup de mailles, surtout dans les zones de forte vitesse. Les glaciologues et les mathématiciens ne renoncent pas pour autant. En utilisant le fait que l'épaisseur de glace est « très faible » par rapport à l'étendue horizontale de la calotte, ils ont proposé un modèle simplifié, dit *modèle couche mince*, un petit peu moins précis que le précédent mais beaucoup plus rapide !

L'évolution d'une calotte polaire peut ainsi être simulée pendant une période donnée, ce qui permet de réaliser des prévisions, en combinant le modèle mathématique et les mesures physiques. Ce couplage se fait de manière indirecte, en remontant par exemple aux conditions physiques à la base de la calotte polaire, grâce à des mesures à la surface de celle-ci. Hélas, les prévisions obtenues sont très incertaines, à cause de la sensibilité de plusieurs paramètres. Des outils mathématiques puissants, les « méthodes inverses » (à la base de nos prévisions météos quotidiennes !), permettent de pallier ce problème et d'effectuer des prévisions plus fiables en prenant en compte, en plus du modèle d'évolution, les mesures physiques disponibles.

Paramètres sensibles mais mal connus et méthodes inverses

Les entrées du modèle mathématique sont des paramètres qui dépendent de la calotte polaire étudiée et du scénario climatique envisagé : il s'agit de la géométrie et de la température de la calotte, du flux géothermique, des coefficients mécaniques, des précipitation. Leurs valeurs sont déterminées par les processus physiques ou choisies par l'utilisateur. En pratique, les valeurs de ces paramètres sont entachées d'erreurs, qui vont se répercuter dans les équations du modèle et produire une erreur sur la sortie. Un paramètre est *sensible* si une « petite » erreur à l'entrée se traduit par une « grosse » erreur à la sortie ; il est *peu sensible* si la « petite » erreur en entrée reste « petite » en sortie. Pour ces derniers paramètres, on n'a pas besoin d'être extrêmement précis, une bonne approximation suffit. Pour les paramètres sensibles, en revanche, il est crucial de réduire l'erreur autant que possible en vue d'effectuer des prévisions fiables.

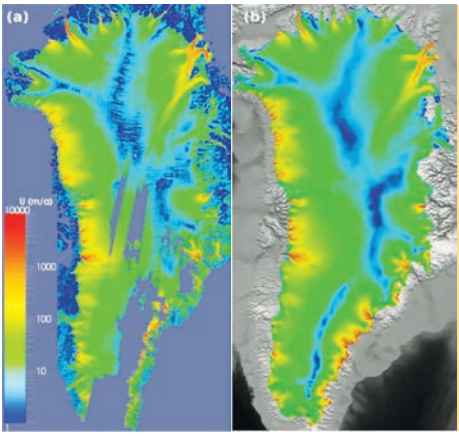
La perte de masse des calottes polaires par vêlage d'icebergs est contrôlée par un petit nombre de fleuves de glace ou de glaciers côtiers appelés *glaciers émissaires* (la glace à proximité des limites de la calotte glaciaire peut se retrouver emprisonnée dans des vallées taillées dans le roc, formant ce type particulier de glacier). La présence ou non de fleuves de glace est liée à la nature des conditions à la base de la calotte, qui permettent ou non un fort glissement et donc de grandes vitesses d'écoulement (jusqu'à une dizaine de kilomètres par an).

Un autre paramètre essentiel est l'altitude du socle rocheux (toujours à la base de la calotte), notamment dans les zones côtières. En effet, la glace se déforme sous l'effet de son propre poids, qui est proportionnel à l'épaisseur de glace, donc à la différence entre l'altitude de surface et l'altitude du socle. Les conditions basales que sont le coefficient de frottement et l'altitude du socle sont des paramètres sensibles : la moindre incertitude provoque une grande variation de la vitesse d'écoulement, donc potentiellement une grande variation dans le volume de glace perdu par la calotte. Or, les observations pour les conditions de frottement à la base sont très limitées. En pratique, on ne dispose donc pas d'une valeur précise. Pour la profondeur du socle, on dispose là aussi d'un nombre très limité de mesures, faites par avion. Des méthodes mathématiques permettent d'estimer l'altitude dans les zones non observées mais l'incertitude peut alors atteindre plusieurs centaines de mètres. Le coefficient de frottement et l'altitude du socle sont donc des paramètres à la fois sensibles et mal connus.

Par contre, il est possible de disposer d'autres observations bien plus précises : des mesures par satellite fournissent l'altitude de la surface et la vitesse de la glace en surface. On dispose de données d'observation annuelles depuis environ 2005. Peut-on les utiliser pour en déduire les conditions basales ? L'idéal serait de pouvoir « inverser le modèle », autrement dit de pouvoir calculer les paramètres du modèle à partir de ces observations. Malheureusement, mathématiquement parlant, le modèle n'est pas inversible : il n'existe pas de façon simple de passer des observations aux paramètres d'entrée inconnus. Pour autant, les mathématiciens ont développé des méthodes inverses, qui permettent malgré tout de répondre au problème (estimer « au mieux » les paramètres mal connus, en utilisant à la fois les observations et le modèle). Car même si l'on ne peut pas écrire formellement un « modèle inverse », on peut encore utiliser le modèle direct. En combinant les quelques observations et l'expertise des glaciologues, on dispose d'une première estimation des paramètres d'entrée. L'idée est alors de fournir au modèle ces paramètres d'entrée approchés et d'exécuter le code associé pour obtenir des simulations qui vont du passé vers le présent. Les sorties de ce modèle direct sont *a priori* inexactes, mais fournissent une altitude de surface et un champ de vitesses de surface que l'on peut comparer aux observations. La méthode consiste alors, grâce à l'écart entre les observations et les sorties simulées, à corriger les paramètres d'entrée, afin de réduire cet écart.

Pour réduire l'écart à chaque étape, on utilise une méthode de descente : une randonneuse en montagne qui souhaite descendre dans la vallée avec le moins d'étapes possible suit la ligne de plus grande pente pour corriger sa position. À chaque étape, elle recalcule sa position et la ligne de plus grande pente, qu'elle emprunte alors pour effectuer son nouveau déplacement. Elle s'arrête lorsqu'elle est dans la vallée, quand la pente est « proche » de zéro.

On procède de façon similaire avec l'écart entre sorties et observations, qui correspond à l'altitude sur la montagne. On corrige les paramètres d'entrée (qui correspondent à la position sur la montagne) grâce à une ligne de plus grande pente, définie mathématiquement.



Vitesses à la surface du Groenland. En (a), les vitesses observées par satellite ; en (b), les vitesses reconstituées après détermination du coefficient de frottement basal. Les principaux glaciers émissaires (vitesses supérieures à 10 km/an) sont bien reproduits.

©AIRSEA & LGGE, 2015

Dans notre cas, la résolution du problème inverse a permis d'obtenir un coefficient de frottement plus précis, et de faire des simulations à l'échelle du siècle pour estimer la contribution du Groenland à la montée du niveau des mers.

J. E.

Pour en savoir (un peu) plus :

Modéliser et simuler la fonte des calottes polaires. Maëlle Nodet et Jocelyne Erhel, *Interstices,* 2015.

Des outils mathématiques pour prévoir la fonte des calottes polaires. Maëlle Nodet et Jocelyne Erhel, *Interstices,* 2015.

« Simuler la fonte des calottes polaires. » Maëlle Nodet et Jocelyne Erhel, *Imaginary*, module primé lors de la compétition internationale des Mathématiques de la planète Terre, 2017, disponible en ligne.

« Le futur des glaciers ». Guillaume Jouvet, 5 mn, vidéo pédagogique disponible en ligne.

De la glace à la mer. Maëlle Nodet, *Matapli* 100, 2013, disponible en ligne.

L'évolution des glaciers, modélisation et prédiction. Guillaume Jouvet, *Accromath* 8.2, 2013, disponible en ligne.

Greenland ice sheet contribution to sea-level rise from a new-generation ice-sheet model. Fabien Gillet-Chaulet, Olivier Gagliardini, Hakime Seddik, Maëlle Nodet, Gael Durand, Catherine Ritz, Thomas Zwinger, Ralf Greve et David Glyn Vaughan, *The Cryosphere* 6, 2012, disponible en ligne.

Elmer/Ice. Elmer/Ice Project, 2018, logiciel libre disponible en ligne.

Les bases de données

Hervé Lehning

Agrégé de mathématiques, écrivain scientifique,
membre de l'ARCSI et commandant de réserve

Nous assistons de nos jours à un déluge de données venant des réseaux sociaux, des sites qu'on visite comme des objets connectés qu'on utilise. En France, cette collecte des données, qui fait le *big data*, est encadrée par la loi. La CNIL (Commission nationale de l'informatique et des libertés) est chargée de la faire appliquer. Les médias nous montrent chaque jour les conséquences dévastatrices occasionnées par des données tombées dans de mauvaises mains.

Recherche dans une base de données : des méthodes ingénieuses !

Imaginez. Vous entrez dans une bibliothèque pour chercher le livre conseillé par un ami. Il vous a donné son titre et le nom de l'auteur. Vous vous renseignez. On vous envoie à deux fichiers, l'un par thème, l'autre par nom d'auteur. Vous choisissez celui-ci et, très vite, vous trouvez votre auteur, puis la référence du livre que vous cherchez, un numéro de classement dans les rayonnages. Vous les parcourez rapidement et, au numéro dit, trouvez votre livre. Sans le savoir, vous venez de consulter une base de données. Toutes peuvent se représenter comme les livres d'une bibliothèque. De façon générale, on peut l'imaginer comme un tableau dont les lignes représentent les données (les livres de notre bibliothèque) et les colonnes les renseignements les concernant (numéro de classement dans la base, nom de l'auteur, thème, photographie…).

On peut de même se représenter un dictionnaire, le fichier des clients d'une entreprise, celui de ses fournisseurs… Une base de données est donc avant tout un tableau, une liste de fiches de plusieurs natures, quantitatives ou qualitatives. Si elle est « grande », disons qu'elle comporte un million de fiches, sans index, il est difficile d'y trouver ce que l'on cherche. La seule solution serait de parcourir les fiches de la première à la dernière. Si vous avez de la chance, celle que vous cherchez est la première, mais elle peut aussi bien être la dernière. En moyenne, il vous faut consulter la moitié des fiches pour trouver celle que vous cherchez. Une telle recherche est dite

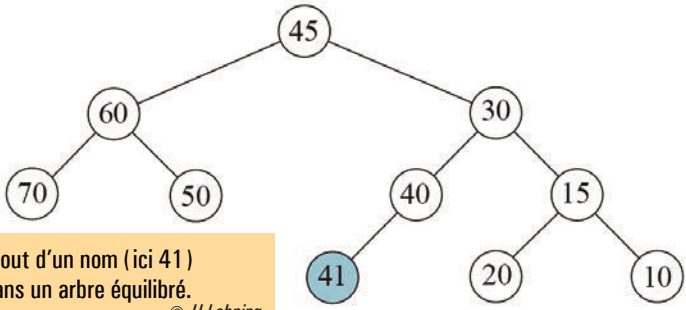
séquentielle. Même à l'aide d'un ordinateur, elle n'est praticable que pour de « petites » bases de données car, en moyenne, elle utilise un temps proportionnel au nombre d'éléments de la base.

Pour simplifier la consultation d'une base de données, la première idée est d'utiliser des index classés dans l'ordre suivant un critère, le nom par exemple. Imaginons que nous cherchions le mot « CIJM » parmi un million de fiches classées ainsi. Essayons d'abord celle du milieu ! Si l'on tombe sur le mot cherché, la recherche s'arrête là. Si la fiche porte un nom antérieur à celui recherché, la bonne fiche (si elle existe) se trouve entre les fiches numéro 500 001 et numéro 1 000 0000, sinon entre les numéros 1 et 499 999. On recommence de même. À chaque étape, la longueur de l'intervalle de recherche est divisée par 2. Dans le pire des cas, la consultation sera terminée après vingt essais puisque $2^{20} = 1\,048\,576$.

Mise à jour d'une base : les atouts d'une structure arborescente

Pour une base qui n'évolue pas (ou peu), tout ceci est parfait. Mais que faire quand on veut ajouter, modifier ou supprimer un élément ? Prenons le cas d'un ajout. Aucune difficulté pour la base elle-même : on place la nouvelle fiche à la suite des autres. Pour les index, c'est plus difficile. Si vous voulez insérer une nouvelle fiche, il faut la placer au bon endroit. Le trouver est facile, il suffit de le chercher comme précédemment. Mais, pour l'insérer, il faut décaler ceux qui le suivent pour lui faire une place. En moyenne, le temps d'insertion est proportionnel à la taille de la base. Il en est de même pour une suppression. Seule une modification ne prend que le temps de la recherche.

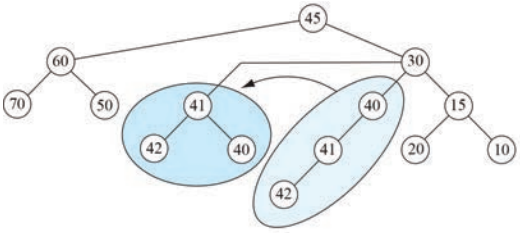
Pour simplifier la gestion des bases de données, une solution est l'utilisation d'une *structure d'arbre*, analogue à celle des arbres généalogiques. Plus précisément, elle est composée de *nœuds* portant le critère de recherche en *étiquette*, chaque nœud ayant un *père* (sauf le premier, la racine) et deux *fils* (sauf les derniers, les *feuilles*). Pour chaque nœud, l'étiquette du fils droit est antérieure et celle du fils gauche, postérieure à celle de son père. De plus, tout le long de l'arbre, les hauteurs des fils d'un même père diffèrent d'au plus une unité. Un tel arbre est appelé un *arbre de recherche équilibré*. Imaginons que l'index d'une base de données soit structuré ainsi. Pour y chercher un nom, partons de la racine. Si ce nom est antérieur à la racine, il se trouve (s'il existe effectivement) dans son fils droit, sinon dans son fils gauche. La recherche se poursuit ainsi en parcourant les branches de l'arbre. En moyenne, son temps d'exécution est proportionnel à la hauteur de l'arbre, et non pas à son nombre d'éléments.



Ajout d'un nom (ici 41)
dans un arbre équilibré.
© *H.Lehning*

Les éléments d'un arbre équilibré peuvent être des nombres ou des objets de toute nature, du moment qu'il est possible de les comparer deux à deux. Pour chercher l'élément 40, on le compare à la racine (45). Il est strictement inférieur donc, s'il figure effectivement dans l'arbre, il appartient à son fils droit. On le compare alors à la racine du fils droit (30). Il est strictement supérieur, donc appartient à son fils gauche. Et ainsi de suite. Si maintenant on souhaite ajouter l'élément 41, on opère de même (en bleu sur la figure). Il se trouve qu'ici l'arbre reste équilibré. Ce n'est pas forcément le cas.

Comment ajouter un élément dans un tel index ? Si l'index est vide, aucune difficulté. La belle affaire, direz-vous. Méfiez-vous, l'idée est plus subtile qu'il n'y paraît ! Si l'arbre n'est pas vide, il a une racine. Si l'élément à ajouter est antérieur à cette racine, il convient de l'ajouter à son fils droit, sinon à son fils gauche. Et on recommence avec ce fils. Cette seule idée suffit pour insérer le nouvel élément. Un tel algorithme est dit *récursif*. Quel est son temps d'exécution ? Le même que celui d'une recherche ! Il est proportionnel à la hauteur de l'arbre. Bien entendu, il reste un problème délicat : ce nouvel arbre n'est plus forcément équilibré. Il faut le rééquilibrer. Ceci se fait par des « rotations » des parties déséquilibrées. Cet algorithme d'ajout d'un élément dans la base clôt la question : on a décrit les deux opérations essentielles liées aux bases de données (la consultation et la mise à jour).



Rééquilibrage d'un arbre.
L'arbre se déséquilibre lorsque l'on ajoute
les deux éléments 41 et 42. Pour le rééquilibrer,
il suffit d'opérer une rotation sur son élément médian (41).
La partie bleu clair est donc remplacée par la partie bleu foncé.
© *H.Lehning*

Le partitionnement de documents : classer, résumer, affecter…

Retournons dans notre bibliothèque. Elle est partitionnée en plusieurs rayons : romans, livres pratiques, sciences… De même, on peut vouloir partitionner un lot de photos en paysages, bâtiments, animaux, humains… Si la taille de la collection de photos est grande, il est important de le faire automatiquement, de trouver un algorithme réalisant l'opération pour nous. Pour cela, on définit une distance euclidienne sur l'espace des photos, ce que l'on peut faire en comparant les pixels des images. Pour d'autres types de documents, la première étape du partitionnement est le choix de cette distance euclidienne, du nombre *k* de classes et d'un représentant de chaque classe, donc ici d'une photo de paysage, d'une photo de bâtiment…
La première étape de l'algorithme est d'attribuer une classe à chaque document : celle dont le représentant est « le plus proche ». On recommence alors en remplaçant le représentant de chaque classe par le centre de gravité (la moyenne) de la classe… et on recommence ainsi jusqu'à obtenir un partitionnement stable. Cette procédure est appelée *algorithme des k-moyennes*. Elle est utile pour analyser des données sophistiquées en les résumant à *k* exemples. Elle permet aussi à une équipe de ventes de définir des cibles pour leurs campagnes de marketing.

Les nuages (ou *clouds*) sont indispensables aux grandes bases de données. Leurs avantages sont nombreux. Plus besoin de disposer de tous les logiciels sur son ordinateur, il suffit d'utiliser ceux se trouvant dans le nuage qui, de plus, sont toujours à jour. De même, vous disposez de toute la mémoire nécessaire pour vos besoins quotidiens. Bien des gens le font en déposant des vidéos sur des sites comme YouTube (Google, 2006), par exemple.

Malgré ces avantages, l'*informatique en nuage* (ou *cloud computing*) possède deux gros inconvénients. Le premier est de dépendre d'Internet. En cas de coupure de réseau, vous n'avez plus accès à ce que vous avez externalisé. Le second est plus grave, il s'agit du manque de confidentialité. Vous ne savez plus où transitent vos données. Pourquoi pas chez vos pires ennemis ? Pour le comprendre, il est bon de savoir où elles se trouvent physiquement. Dans quel pays, sous quelles menaces (légales ou illégales) ? En fait, un nuage correspond à un certain nombre de centres de données (*data centers* en anglais), dont il est bon de savoir où, et sous quelle législation, ils se trouvent. Ils sont grands consommateurs d'énergie et grands producteurs de chaleur, ce qui milite pour les placer dans des endroits froids, proches d'agglomérations dont ils peuvent fournir une partie du chauffage urbain.

Un centre de données, élément d'un nuage. Où se trouve-t-il ?

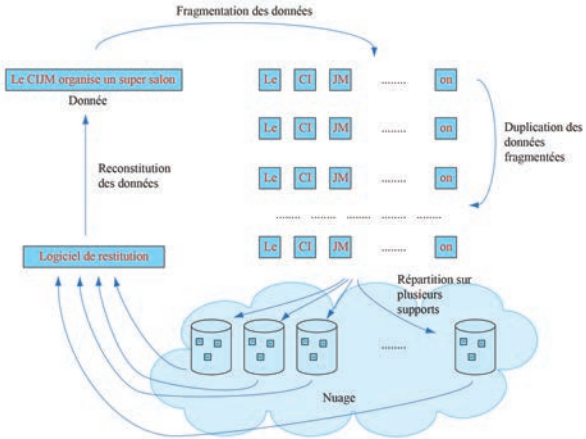
© *Más grande del mundo, 2018*

On parle de « nuages » pour signifier que l'on ne sait pas bien où sont situées les données que l'on y place car elles sont en fait fragmentées entre plusieurs centres de données qui, contrairement au *cloud*, sont des lieux physiques.

Plus précisément : un *cloud* est toujours un espace virtuel ; les données contenues dans un cloud sont fragmentées ; les fragments de données sont toujours dupliqués et répartis sur un ou plusieurs supports physiques ; un *cloud* possède une fonction de restitution des données permettant de reconstituer les données qui ont été fragmentées.

Fonctionnement d'un nuage.

© *H.Lehning*



Le chiffrement des données : la clé de la confidentialité

Si on ignore les algorithmes utilisés, il est en principe impossible de savoir où se trouve l'information, et on ne peut avoir accès qu'à des fragments qui ne permettent pas de retrouver le sens des données. Ces impossibilités

sont cependant fondées sur le secret de la méthode utilisée, ce qui est une grave faiblesse face à l'espionnage ou à l'intelligence économique. Pour garantir une véritable confidentialité, il est prudent de chiffrer les données.

Il existe un système permettant d'assurer une confidentialité correcte, à base d'un chiffre symétrique dont la clef est transmise par un chiffre asymétrique. Ce procédé de chiffrement est en principe solide, mais il n'est pas utilisable sur les «petits» objets connectés, en particulier ceux que l'on implante dans le corps humain comme les pacemakers ou les pompes à insuline. L'utilisation d'un chiffre asymétrique, très gourmand en énergie, ferait chauffer l'objet, ce qui serait source d'insécurité ! Il en résulte que les «petits» objets connectés sont difficiles à protéger. Ce sont donc des cibles faciles pour les pirates, pour se constituer des réseaux de zombies (ou *botnets*), c'est-à-dire d'ordinateurs dont ils se sont rendus maîtres pour lancer des actions malveillantes, telles des opérations de hameçonnage (le tristement fameux *phishing*). Cela peut sembler de la science-fiction. Pourtant, cela s'est déjà produit ! Les objets connectés se comptant par dizaines de milliards, le marché est immense… pour des personnes mal intentionnées.

H.L.

Les systèmes de chiffrement

Il existe deux grands systèmes de chiffrement. Premièrement, les *systèmes symétriques*, pour lesquels savoir chiffrer implique savoir déchiffrer. La clef doit être tenue secrète. Le chiffrement est rapide, le temps étant de l'ordre de celui d'une addition. Le plus connu est le système AES (*advanced encryption standard*, 1997).

Deuxièmement, les *systèmes asymétriques*, pour lesquels savoir chiffrer n'implique pas de savoir déchiffrer. La clef de chiffrement est publique, celle de déchiffrement est secrète. Le plus connu est le système RSA (Rivest, Shamir, Adleman, 1977), où la clef publique est liée au produit $N = p \times q$ de deux «grands» nombres premiers p et q (idéalement de plus de deux cents chiffres), et la clef secrète est liée à p et q. La difficulté de la factorisation de N si on ne connaît ni p, ni q explique ce mystère. Ce chiffrement est lent et gourmand en énergie, le temps étant de l'ordre d'une exponentiation.

Pour en savoir (un peu) plus :

Toutes les mathématiques du monde. Hervé Lehning, Flammarion, 2017.

La bible des codes secrets. Hervé Lehning, Flammarion, 2019.

Les mathématiques de l'apprentissage

Clément Cartier

Mathématicien

L'intelligence artificielle (IA), en particulier celle capable d'apprendre par elle-même, n'est pas une idée nouvelle : elle est déjà évoquée par le mathématicien britannique Alan Mathison Turing (1912–1954) dans un célèbre article de 1950, et la théorie mathématique de l'apprentissage automatique est posée dans les années 1960–1970 par le mathématicien russe Vladimir Naumovitch Vapnik (né en 1936). Mais c'est surtout aujourd'hui qu'elle occupe une place croissante dans le monde des applications informatiques, et soulève un certain nombre de débats. Quelles réalités recouvre donc cet apprentissage automatique ?

De façon un peu intuitive, on peut le définir comme un processus qui permet d'améliorer notre capacité à réaliser une action, à accomplir une tâche, ce qui passe par la réalisation d'exercices, et par l'assimilation d'exemples. D'un point de vue plus formel et mathématique, on peut traduire cette idée d'apprentissage en termes de recherche d'une fonction. On va en effet essayer de trouver une façon d'associer, à un objet quelconque X, par exemple une valeur numérique, un vecteur, une matrice (c'est-à-dire un tableau de nombres), une image (qu'on peut représenter comme une matrice), un son…, un autre objet tout aussi quelconque Y. X peut ainsi représenter l'information captée par une caméra installée dans une voiture, et Y la force à appliquer sur la pédale de frein.

Faire passer une droite ou une courbe par des points

L'apprentissage est une méthode qui permet de déterminer cette fonction en prenant un nombre N d'objets x_1, $x_2 \dots x_N$ (par exemple, tout un tas d'images prises sur la route) pour lesquels on connaît déjà les N solutions correspondantes $y_1, y_2 \dots y_N$ (comme la force qui devait être appliquée sur le frein pour rester en sécurité au moment où les images correspondantes ont été prises).

Différents types d'apprentissage supervisé

En fonction de la nature de l'objet Y que l'on souhaite déterminer, on va pouvoir distinguer :

– des *algorithmes de classification*, si Y est une donnée qualitative (typiquement, lorsque l'on souhaite déterminer si une image représente un visage, une radiographie du poumon, une tumeur..., ou dès que l'on recherche le coup « le plus efficace » à jouer dans une partie d'échecs). On peut alors utiliser des techniques d'*apprentissage supervisé*, où les données d'apprentissage ont un caractère discret (elles peuvent être indexées sur l'ensemble des entiers naturels) et permettent d'identifier les différentes classes possibles;

– des *algorithmes de régression*, si Y est une donnée quantitative (comme la probabilité qu'un coup à jouer dans une partie d'échecs conduise à la victoire, ou la force à appliquer sur le frein d'une voiture).

Les algorithmes de classification et de régression peuvent parfois être très similaires, notamment lorsque l'on veut déterminer la probabilité que l'objet X appartienne à chacune des classes Y (*problème de régression*) pour ensuite choisir la classe « la plus probable » (*problème de classification*).

Les points (x_1, y_1), $(x_2, y_2)\ldots$ (x_N, y_N) sont alors autant d'*exemples d'apprentissage*, à partir desquels on va chercher à interpoler un fonction f qui, non seulement nous fournit bien $f(x_1)=y_1, f(x_2)=y_2$ $f(x_N)=y_N$ (ou qui reste « le plus proche possible » de ces valeurs), mais qui sera aussi capable de trouver une solution Y pour de nouveaux éléments X qui n'étaient pas dans nos exemples d'apprentissage.

Dans le cas $N=2$, où il n'y a que deux exemples d'apprentissage (x_1, y_1) et (x_2, y_2) avec x_1 différent de x_2, on peut faire une simple interpolation linéaire, qui nous donnera une fonction affine, c'est-à-dire une simple droite, dont on étudie les équations en fin de collège :

$$y = \frac{y_2 - y_1}{x_1 - x_2} x + \frac{x_2 y_1 - x_1 y_2}{x_2 - x_1}.$$

Cependant, dès que $N>2$, il y a de fortes chances que les points ne soient pas alignés, et que donc l'interpolation linéaire ne produise pas de bons résultats. La phase d'apprentissage va donc nécessiter d'explorer un ensemble de fonctions pour trouver celle qui « s'approche le plus » des exemples d'apprentissage, et de déterminer les paramètres de cette fonction.

Mettre en œuvre l'apprentissage

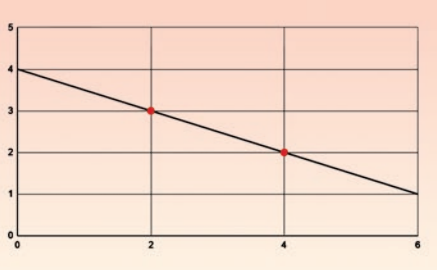
Pour mettre en œuvre ces différents types d'apprentissage, outre les méthodes décrites dans le précédent encadré, on peut également utiliser :

– des algorithmes d'*apprentissage par renforcement*, où l'on ne connaît pas à l'avance les solutions y_i pour les exemples d'apprentissage, mais où l'on est capable de tester les solutions proposées par l'algorithme dans la réalité pour voir si ces solutions sont de bonne qualité ou non ;

– des algorithmes d'*apprentissage par transfert*, où l'algorithme sait déjà réaliser certaines tâches et cherche à identifier les similitudes entre ces différentes tâches pour le généraliser à de nouvelles.

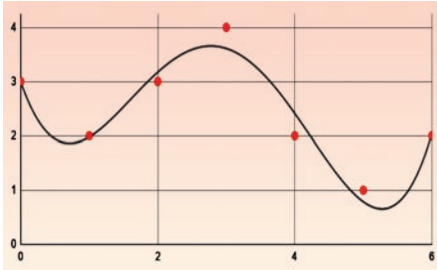
Mettre en œuvre ces différents algorithmes nécessite d'avoir accès à une grande quantité de données pour produire les exemples d'apprentissage, mais aussi d'avoir un minimum de compréhension de la façon dont fonctionne la tâche que l'on veut enseigner à la machine : la seule connaissance mathématique des algorithmes d'apprentissage ne suffit pas, il faut également des personnes expertes dans le domaine d'application pour aider à déterminer quels algorithmes seront « les plus efficaces » compte tenu de la nature du problème, et pour sélectionner les données « les plus pertinentes » pour l'apprentissage.

En pratique, la première difficulté consiste à déterminer intelligemment l'ensemble des fonctions que l'on veut explorer : si cet ensemble est « trop petit », on risque de passer à côté de solutions intéressantes, tandis que s'il est « trop grand », l'exploration risque de prendre beaucoup trop de temps. En pratique, il est donc nécessaire, avant de commencer le processus d'apprentissage, de déterminer le niveau de précision, le temps de calcul, les conditions que doivent vérifier la solution…



Exemple d'interpolation linéaire avec deux points (en rouge) ; la droite obtenue a pour équation $y = -0{,}5x + 4$.

© C.C.



Exemple d'interpolation polynomiale (non linéaire) avec six points.

© C.C.

Parfois, au cours du processus d'apprentissage, on peut essayer de proposer à l'algorithme un nouvel exemple (x_{N+1}, y_{N+1}). Il est alors possible que la fonction qui avait été trouvée précédemment par l'algorithme tombe « très proche » des N points précédents, mais soit « très loin » du nouveau point. On va donc vouloir que l'algorithme revoie sa copie, et se mette à la recherche d'une « meilleure » fonction, qui passe toujours par les points (x_1, y_1), $(x_2, y_2)\ldots(x_N, y_N)$, mais aussi par le nouveau point (x_{N+1}, y_{N+1}).

Dans ce scénario, où de nouveaux exemples peuvent être ajoutés au fur et à mesure, la fonction deviendra « de plus en plus précise », jusqu'à un moment où il devient possible de lui « faire suffisamment confiance » pour prédire le Y correspondant à n'importe quel X, même s'il ne fait pas partie des exemples d'apprentissage. L'algorithme est alors prêt à accomplir sa tâche.

L'apprentissage automatique, en 2020, c'est donc principalement collecter les données qui constituent les exemples d'apprentissage, et mettre en place la recherche de cette fonction optimale à partir des exemples d'apprentissage. La grande question est alors de savoir quel crédit accorder aux résultats que produit la fonction ainsi obtenue. C'est exactement la problématique à laquelle Vladimir Vapnik a répondu dans les années 1960 – 1990, à partir d'expériences menées les décennies précédentes sur des machines américaines capables d'apprentissages, comme le Perceptron (Frank Rosenblatt, 1957).

Quelle est la probabilité que la fonction nous induise en erreur ? Pour le savoir, on définit deux variables : l'*erreur d'apprentissage*, ou *risque empirique* R_{emp}, qui correspond à l'erreur commise par la fonction sur les exemples d'apprentissage ; et l'*erreur de généralisation*, ou *erreur de test* R, qui correspond à l'erreur commise par la fonction sur des données qui ne font pas partie des exemples d'apprentissage.

L'un des problèmes est de chercher à majorer (et réduire au maximum) R. Une approche courante sur des problèmes de régression, issue de la statistique bayésienne, peut être tentée pour déterminer R en connaissant R_{emp}. Hélas, le calcul des risques *a priori* demande beaucoup de calculs, ce que Vapnik considère comme un inconvénient majeur. Avec l'aide d'Alexey Yacovlevitch Chervonenkis (1938 – 2014), Vapnik a réussi à démontrer de façon directe des résultats importants et fondateurs en théorie de l'apprentissage automatique et de l'apprentissage statistique.

Le Perceptron

Le Perceptron, construit à l'université Cornell, est un ordinateur analogique capable d'apprentissage. Conçu pour reconnaître des formes, il est doté d'une « rétine » (captant des images d'une résolution d'environ 20 pixels par 20 pixels). Il renvoie une suite de nombres $z_1, z_2 \ldots z_m$ (voltages et tensions), fournie par un calculateur analogique (le *classifieur*).

Il en fait une somme pondérée, de la forme $y = \mathrm{sgn}\left(\sum_{i=1}^{m} w_i z_i + b\right)$ où sgn(u) renvoie 1 si $u \geq 0$

et 0 sinon, où $w_1, w_2 \ldots w_m$ sont des poids (entiers relatifs), et où b est un seuil fixé. La sortie y vaut donc 0 ou 1.

L'ensemble d'entraînement est $(z_1, y_1), (z_2, y_2) \ldots (z_m, y_m)$. Le but du Perceptron est d'entraîner les coefficients $w_1, w_2 \ldots w_m$ de sorte que la sortie vaille 1 pour les objets que l'on souhaite reconnaître et 0 pour ceux que l'on souhaite exclure (apprentissage supervisé).

La règle d'entraînement est simple. On initialise les poids à 0. Chaque poids w_i à l'instant $t + 1$ est égal au poids à l'instant courant t auquel on ajoute une erreur entre la sortie désirée et la sortie produite par le Perceptron. Sur l'exemple d'apprentissage (z_i, y_i), si la sortie y_i attendue vaut 1 mais que le Perceptron a répondu 0, on augmente de 1 les poids w associés à des valeurs de x négatives. On fait l'inverse si la sortie attendue était 0 mais que le Perceptron a répondu 1. On ne fait rien si la sortie attendue et la sortie obtenue coïncident.

L'idée générale de la théorie de Vapnik – Chervonenkis est que pour que R soit la plus petite possible, il faut que la complexité du modèle soit adaptée au nombre d'exemples d'apprentissage (une formule explicite de majoration de R est obtenue). Ainsi, si la complexité du problème est très grande (ou qu'elle tend vers l'infini), alors R sera toujours importante, même si R_{emp} est faible : la généralisation n'est alors pas possible !

Cela pose les limites des algorithmes d'apprentissage automatique tels qu'on les connaît aujourd'hui. Pour autant, les techniques utilisées pour mettre en œuvre l'apprentissage repoussent peu à peu ces limites : dans les années 1995 – 2005, les algorithmes privilégiés (des machines à vecteurs de support, introduits par Vapnik et Isabelle Guyon) n'étaient capables de traiter que de modèles très simples, souvent limités à des problèmes de classification linéaire. Aujourd'hui, les nouvelles technologies et le développement de techniques d'apprentissage profond (le *deep learning*) permettent de s'intéresser à des modèles bien plus sophistiqués.

C'est encore une question ouverte de déterminer exactement jusqu'où les limites de l'apprentissage automatique pourront être poussées, mais des limites sont fixées par les principes fondamentaux de la statistique, qui ne dépendent pas de la nature des machines qui apprennent.

Des questions mathématiques, mais aussi sociales et éthiques

Le développement de ces nouvelles technologies pose cependant un certain nombre de questions éthiques, sociales et économiques, au premier rang desquelles se trouve la question de propriété des données : il faut en effet des bases de données extrêmement importantes pour mettre en place de telles démarches d'apprentissage automatique, et une façon « facile » de faire est de collecter… les données des utilisateurs d'applications sur Internet.

La question est d'autant plus délicate lorsqu'il s'agit d'utiliser l'IA dans la pratique médicale, où les données impliquées sont des données de santé. Par ailleurs, des questions se posent quant aux conséquences en cas d'erreur commise par l'IA : l'erreur de test restera toujours positive, et des applications malheureuses sont donc inévitables. Dès lors, qui tenir pour responsable d'un accident ?

En outre, les décisions prises par une IA issue d'un apprentissage automatique dépendent directement des choix mis en œuvre dans la conception des algorithmes et dans la sélection des données d'apprentissage (deux éléments sur lesquels de simples utilisateurs ont peu de chances d'avoir un avis éclairé au moment de décider du crédit qu'ils accordent à l'IA, alors que ces éléments peuvent causer le renforcement de stéréotypes qui auraient été reproduits dans la sélection des données).

Les enjeux sont donc d'améliorer la vitesse d'exécution et le coût en données des algorithmes d'apprentissage, de trouver des méthodes pour s'attaquer à de nouveaux problèmes, mais aussi d'augmenter la fiabilité des algorithmes, en essayant de réduire les risques de biais dans la sélection des données.

C. C.

Pour en savoir (un peu) plus :

La théorie de l'apprentissage de Vapnik et les progrès de l'intelligence artificielle. Yann LeCun, conférence à la Bibliothèque nationale de France, 75 mn, mercredi 4 avril 2018, disponible en ligne.

Computer machinery and intelligence. Alan Turing, *Mind* 59, 1950.

Statistical learning theory. Vladimir Vapnik, Wiley–Blackwell, 1998.

Neural networks and deep learning. Michael Nielsen, 2019, disponible en ligne.

Réseaux de neurones et apprentissage

Gabriel Peyré

CNRS & DMA
PSL, École normale supérieure

Depuis 2012, les réseaux de neurones profonds ont révolutionné l'apprentissage automatique. Bien que relativement ancienne, cette technique a permis ces dernières années des avancées spectaculaires pour la reconnaissance de textes, de sons, d'images et de vidéos. Comprendre les enjeux de ces méthodes soulève des questions à l'interface entre les mathématiques et l'algorithmique.

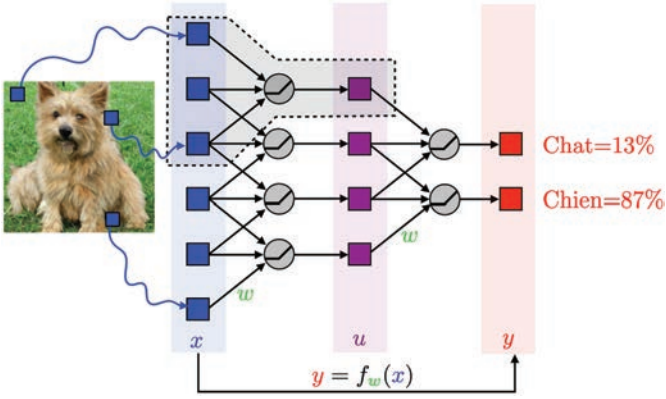
L'algorithmique et les mathématiques de l'apprentissage

Les réseaux de neurones sont des algorithmes, qui permettent à partir d'une entrée x (par exemple une image) de calculer une sortie y. Cette sortie est le plus souvent un ensemble de probabilités : sur la figure, la première sortie est la probabilité que l'image contienne un chat (plus ce nombre est proche de 100 %, plus cela signifie que l'algorithme est « sûr de lui »), la deuxième est la probabilité que l'image contienne un chien.

Pour simplifier, on ne considèrera que deux classes, les chats et les chiens, mais en pratique on peut considérer une sortie y avec plusieurs milliers de classes. On se restreint également à l'exemple des images, mais les réseaux de neurones sont aussi très performants pour reconnaître des textes ou des vidéos.

Exemple d'un réseau de neurones discriminatif s avec deux couches.

© G. Peyré



Mathématiquement, un tel algorithme définit une fonction f_w. Ainsi, on a $y = f_w(x)$. Le programme informatique qui permet de calculer cette fonction est très simple : il est composé d'un enchaînement de plusieurs étapes, et chaque étape effectue des calculs élémentaires (des additions, des multiplications, et l'évaluation d'un maximum entre plusieurs nombres). En comparaison, les programmes informatiques que l'on trouve dans le système d'exploitation d'un ordinateur sont beaucoup plus élaborés. Mais ce qui fait l'énorme différence entre un algorithme « classique » et un réseau de neurones, c'est que ce dernier dépend de *paramètres*, qui sont les *poids* des neurones. Avant d'utiliser un réseau de neurones, il faut modifier ces poids pour que l'algorithme puisse résoudre le mieux possible la tâche demandée. C'est ce que l'on appelle *entraîner* un réseau de neurones, et cela nécessite beaucoup de temps, de calculs machine et d'énergie.

Utiliser à bon escient de tels algorithmes nécessite donc des compétences en informatique et en mathématiques. Il faut ainsi manipuler les concepts clefs de l'algorithmique (méthodes itératives, temps de calcul, espace mémoire, implémentation efficace…) et des mathématiques (algèbre linéaire, optimisation, statistiques…).

De la biologie au modèle mathématique abstrait

Un réseau de neurones artificiel est construit autour d'une métaphore biologique. On connaît relativement bien la structure du cortex visuel primaire, et la découverte en 1962 de l'organisation des neurones dans les premières couches a valu le prix Nobel en physiologie à David Hubel et Torsten Wiesel. Ainsi, dans une vision extrêmement simplifiée du fonctionnement du cerveau, les neurones sont organisés en couches, chaque neurone récupère de l'information d'une couche précédente, effectue un calcul très simple, et communique son résultat à des neurones de la couche suivante.

Il ne s'agit cependant que d'une métaphore et d'une source d'inspiration : les réseaux biologiques ont des connexions beaucoup plus sophistiquées et les équations mathématiques qui les décrivent sont également très élaborées (elles ont été découvertes en 1952 par Alan Hodgkin et Sir Andrew Huxley, qui ont eux aussi reçu le prix Nobel). Il reste ainsi difficile de mettre précisément en relation les performances parfois surprenantes des neurones artificiels et les capacités cognitives du cerveau. Par exemple, les techniques d'entraînement des réseaux artificiels sont très différentes de la façon dont un enfant apprend.

La première figure détaille un exemple d'un réseau artificiel. Pour simplifier, il est ici composé de seulement deux couches de neurones (la première couche entre x et u, la seconde entre u et y), mais les réseaux actuels les plus performants peuvent comporter plusieurs dizaines de couches ; on dit qu'ils

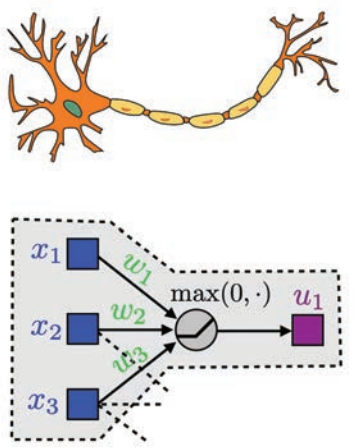
sont *plus profonds*. Ici, les entrées x sont les pixels d'une image. Une image contient typiquement des millions de pixels, et la figure n'en représente volontairement qu'un petit nombre (un réseau réaliste est très complexe). De plus, chaque pixel qui compose x est en fait constitué de trois valeurs (une pour chaque couleur primaire rouge, vert et bleu).

Le passage d'une couche (par exemple la couche x des entrées) à une autre (u, qui est une couche cachée au milieu du réseau) se fait par l'intermédiaire d'un ensemble de neurones artificiels.

Sur cette figure, c'est le premier neurone, celui qui calcule la première valeur u_1, qui compose la couche u. Ce neurone connecte un certain nombre d'éléments de la première couche (ici trois, à savoir x_1, x_2 et x_3, mais il peut y en avoir plus) à un seul élément de la deuxième, donc ici u_1. La formule calculée par le neurone est :

$$u_1 = \max(w_1 \times x_1 + w_2 \times x_2 + w_3 \times x_3 + w_4, 0).$$

Le neurone effectue ainsi une somme pondérée des trois entrées, avec trois poids (w_1, w_2, w_3), et on ajoute également w_4, qui est un biais. Puis le neurone calcule le maximum entre cette somme et zéro. On peut également utiliser une autre fonction que la fonction maximum, mais celle-ci est la plus populaire. Il s'agit d'une opération de seuillage. On peut la comparer aux neurones biologiques qui laissent ou non passer l'information suivant s'ils sont suffisamment excités ou pas. Ainsi, si la somme pondérée $w_1x_1 + w_2x_2 + w_3x_3 + w_4$ est plus petite que 0, alors le neurone renvoie la valeur $u_1 = 0$, sinon il renvoie la valeur de cette somme et la place dans u_1.



Neurone biologique
et
neurone artificiel
© G. Peyré

De tels réseaux de neurones ont été introduits par Frank Rosenblatt en 1957, qui les a appelés *perceptrons*. Les premiers perceptrons ne contenaient qu'une seule couche. De telles architectures avec une seule couche sont trop simples pour pouvoir effectuer des tâches complexes. C'est en rajoutant plusieurs couches que l'on peut calculer des fonctions plus élaborées. Les réseaux de neurones profonds utilisent ainsi un très grand nombre de couches. Depuis quelques années, ces architectures ont permis d'obtenir des résultats impressionnants pour faire de la reconnaissance d'images et de vidéos ainsi que pour la traduction automatique de textes. Ce sont ces recherches sur les réseaux

profonds qui ont permis au Français Yann Le Cun ainsi qu'aux Canadiens Geoffrey Hinton et Yoshua Bengio d'obtenir le prix Turing en 2018, considéré comme l'équivalent du prix Nobel en informatique. Pour se familiariser avec ces réseaux multi-couches, on peut utiliser l'application interactive **https://playground.tensorflow.org.**

Exemples d'images issues d'ImageNet, une base de données utilisées pour l'apprentissage.

© *Images issues de ImageNet (www.image-net.org)*

L'apprentissage supervisé d'un réseau de neurones

L'entraînement d'un réseau de neurones consiste à choisir les « meilleurs » poids possibles de l'ensemble des neurones qui compose un réseau (typiquement ici les poids w_1, w_2 et w_3). Il faut ainsi choisir les valeurs de ces poids afin de résoudre le mieux possible la tâche étudiée, et ceci sur un ensemble de données d'apprentissage. Pour la reconnaissance d'objets dans les images, il s'agit d'un problème d'*apprentissage supervisé* : on dispose à la fois des images x et des y (les probabilités de présence d'un chat ou d'un chien dans l'image). La figure ci-dessus montre quelques exemples d'images utilisées pour entraîner un réseau, pour lesquelles on sait ce qu'elles contiennent (la classe des chats et la classe des chiens). Il faut donc, en amont de la phase d'apprentissage, que des humains fassent un long et fastidieux travail d'étiquetage de milliers voire de millions d'images !

La procédure d'entraînement consiste ainsi à modifier les poids w de sorte que, pour chaque x, le réseau f_w prédise aussi précisément que possible le y associé, c'est-à-dire que l'on souhaite à la fin de l'entraînement que y soit « très proche » de $f_w(x)$. Un choix simple est de minimiser la somme $E(w)$

des carrés des erreurs, ce que l'on écrit mathématiquement

$$\min_{w} \mathrm{E}(w) = \sum_{(x,y)} \left(f_w(x) - y \right)^2.$$

Ceci correspond à un problème d'optimisation, car il faut trouver un jeu de paramètres qui optimise une certaine quantité d'intérêt. C'est un problème difficile, car il y a beaucoup de paramètres. Ces derniers, surtout ceux des couches cachées, influencent de façon très subtile le résultat.

Des merveilles d'ingéniosité et une révolution scientifique

Heureusement, il existe des méthodes mathématiques et algorithmiques performantes pour résoudre de façon satisfaisante ce type de problème d'optimisation. Elles ne sont pas encore totalement comprises sur le plan théorique ; c'est d'ailleurs un domaine de recherche en pleine explosion. Ces méthodes d'optimisation modifient les poids w du réseau pour l'améliorer et diminuer l'erreur d'entraînement $\mathrm{E}(w)$. La règle mathématique pour décider de la stratégie de mise à jour des poids s'appelle la *rétro-propagation* et est une merveille d'ingéniosité. C'est en fait un cas particulier d'une méthode mathématique et algorithmique qui s'appelle la *différentiation automatique à l'envers*.

Ces techniques d'apprentissage supervisé datent pour l'essentiel des années 1980. Mais c'est seulement en 2012 qu'un travail d'Alex Krizhevsky, Ilya Sutskever et Geoffrey Hinton crée un coup de tonnerre en montrant que les réseaux profonds permettent de résoudre efficacement des problèmes de reconnaissance d'images. Cette révolution a été possible grâce à la combinaison de trois ingrédients : des nouvelles bases de données beaucoup plus grandes qu'auparavant, des grosses puissances de calcul grâce aux processeurs graphiques (les « GPU », qui étaient auparavant cantonnés aux jeux vidéo), et l'introduction de plusieurs techniques d'optimisation qui stabilisent l'apprentissage.

Capter efficacement l'information présente dans les données

George Cybenko a démontré en 1989 qu'un réseau de neurones f_w avec deux couches peut approcher, aussi précisément que l'on veut, n'importe quelle fonction continue f^* (donc en quelque sorte résoudre n'importe quelle tâche, représentée par la fonction f^* inconnue, qui serait capable de reconnaître des objets dans n'importe quelle image), pour peu que la taille de la couche interne u (donc le nombre de neurones) soit arbitrairement grande. Ce n'est pas pour autant qu'un tel réseau f_w avec seulement deux couches fonctionne bien en pratique. Pour appliquer le théorème de Cybenko, il faut pouvoir

disposer d'un nombre de données d'apprentissage potentiellement infini, ce qui est très loin d'être le cas en pratique. Le but final de l'apprentissage n'est pas de minimiser l'erreur d'apprentissage $E(w)$, mais de pouvoir prédire aussi précisément que possible sur des nouvelles données. Si l'on dispose de peu de données, on risque de ne pas pouvoir apprendre assez précisément, et donc de faire des mauvaises prédictions : la fonction f_w sera en réalité « très loin » de la fonction f^* idéale que l'on voudrait apprendre si l'on disposait d'une infinité d'exemples.

Afin d'effectuer les meilleures prédictions possibles avec un nombre limité de données d'entraînement, on cherche donc les architectures de réseaux « les plus adaptées », qui peuvent capter efficacement l'information présente dans les données. Les réseaux de neurones profonds (avec de nombreuses couches) mais avec relativement peu de connexions entre les couches se sont avérés très efficaces sur les données très « structurées » comme les textes, les sons et les images. Par exemple, pour une image, les pixels ont des relations de voisinage, et on peut imposer des connexions spécifiques (une architecture) et ne pas connecter un neurone avec tous les autres mais seulement avec ses voisins (sinon il y aurait trop de connexions). De plus, on peut imposer que les poids associés à un neurone soient les mêmes que ceux associés à un autre neurone. On appelle ce type de réseaux les *réseaux convolutifs*. Pour l'instant, il n'y a pas d'analyse mathématique qui explique cette efficacité des réseaux convolutifs profonds. Il y a donc besoin de nouvelles avancées mathématiques pour en comprendre les comportements et les limitations !

G. P.

Pour en savoir (un peu) plus :

The perceptron, a perceiving and recognizing automaton Project Para. Frank Rosenblatt, Cornell Aeronautical Laboratory, 1957.

Learning representations by back-propagating errors. David Rumelhart, Geoffrey Hinton et Ronald Williams, *Nature* 323, 1986.

ImageNet classification with deep convolutional neural networks. Alex Krizhevsky, Ilya Sutskever et Geoffrey Hinton. *Advances in neural information processing systems*, 2012.

Deep learning. Yann LeCun, Yoshua Bengio et Geoffrey Hinton, *Nature* 521, 2015.

ImageNet: a largescale hierarchical image database. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li et Li Fei-Fei, 2009 IEEE conference on computer vision and pattern recognition, 2009.

Générer ou reconnaître des images : les réseaux de neurones à la rescousse

Gabriel Peyré

CNRS & DMA
PSL, École normale supérieure

On sait désormais comment entraîner de façon supervisée des réseaux de neurones (voir l'article précédent). Ceci permet de résoudre efficacement des problèmes de classification, par exemple de reconnaissance d'images. Ce qui est peut-être encore plus surprenant, c'est que ces mêmes réseaux de neurones sont également utilisés de façon non supervisée afin de générer automatiquement des textes ou des images « virtuels », ce que l'on appelle souvent des *deep fake*s. Savez-vous qu'il existe un lien entre l'apprentissage de réseaux de neurones génératifs et la théorie du transport optimal ? Cette dernière a été proposée par Gaspard Monge (1746 – 1818), conte de Péluse, au XVIII^{e} siècle et a été reformulée par Leonid Vitalievitch Kantorovitch (1912 – 1986) au milieu du XX^{e} siècle. Elle est maintenant devenue un outil de choix pour aborder l'explosion récente de la science des données.

Une utilisation « à l'envers » des réseaux de neurones

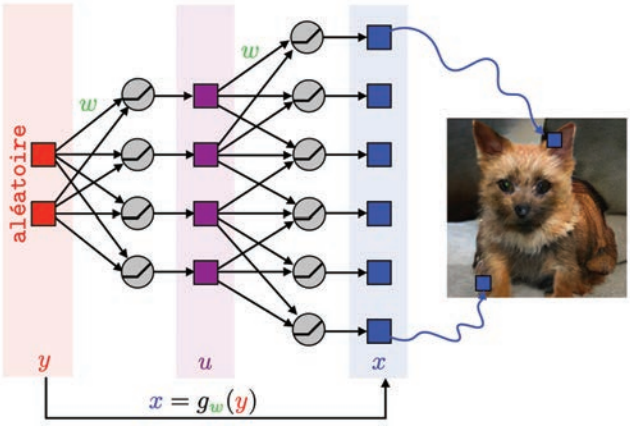
Au lieu d'utiliser des réseaux de neurones pour analyser des images, on peut les utiliser « à l'envers » afin de générer des images. Ces *réseaux de neurones génératifs* trouvent des applications pour les effets spéciaux, les jeux vidéo ou encore la création artistique. On retrouve des approches similaires dans l'apprentissage des voitures autonomes et la résolution de jeux de stratégie. La figure suivante montre la structure d'un tel réseau g_w, qui dépend de poids w. Les couches jouent en quelque sorte des rôles miroirs par rapport à l'architecture des réseaux de neurones discriminatifs de l'article précédent. À partir d'une entrée y composée d'un « petit » nombre de valeurs, qui sont typiquement tirées aléatoirement, on génère une image $x = g_w(y)$.

Le problème de l'apprentissage de tels réseaux est *non supervisé* : on dispose uniquement d'un grand nombre d'images d'apprentissage, sans indication sur ce qu'elles contiennent. Il n'y a plus besoin d'intervention humaine pour indiquer au réseau le contenu des images qu'il doit reconnaître !

* L'auteur remercie Gwenn Guichaoua pour sa relecture attentive, ainsi que Sébastien Racanière et Vincent Barra pour leurs corrections.

Un réseau de neurones génératif simplifié (un réseau permettant de générer des images aussi complexes possède en réalité plus de couches).

© G. Peyré



La collecte de données est plus facile que pour l'entraînement de réseaux discriminatifs. De plus, ce principe d'apprentissage non supervisé est « proche » de la façon dont les enfants apprennent, principalement en observant et manipulant le monde qui les entoure. Le but est alors de sélectionner les poids w des neurones du réseau g_w de sorte que les images aléatoires générées (les images fausses, *fakes* en anglais) ressemblent « le plus possible » aux images d'apprentissage.

L'apprentissage non supervisé : un problème de poids !

Le but des réseaux de neurones génératifs n'est pas de résoudre une tâche telle que la reconnaissance d'objets dans les images. Afin d'entraîner les poids w du réseau, il faut formaliser mathématiquement le problème. Il s'agit de générer un ensemble d'images « virtuelles » (les fausses) qui « ressemblent » aux images réelles d'une base de données. Il ne s'agit pas simplement qu'une image générée « ressemble » à une image réelle, il faut mettre en correspondance les deux ensembles d'images. Par exemple, si dans la base de données il y a la moitié d'images de chiens et la moitié d'images de chats, il faut que le réseau génère également pour moitié des chiens et pour moitié des chats.

On va noter $\{z_1, z_2 \dots z_n\}$ l'ensemble des n images de la base de données. Le nombre n d'images est très grand, de l'ordre de plusieurs milliers ou millions. Étant donné un réseau de neurones génératif g_w, qui est paramétré par ses poids w, on note $\{x_1, x_2 \dots x_n\}$ un ensemble de n images « fausses » générées aléatoirement par le réseau. Pour générer la première image fausse x_1, ceci signifie que l'on tire aléatoirement les valeurs d'entrée y_1 et que l'on applique

le réseau à ces entrées, pour obtenir l'image virtuelle $x_1 = g_w(y_1)$. On a ensuite fait la même chose avec $x_2 = g_w(y_2)$, et ainsi de suite.

Le but de l'apprentissage non supervisé est donc de trouver des poids w de sorte que l'ensemble des fausses images $\{x_1, x_2 \dots x_n\}$ soit « le plus proche » de l'ensemble des images $\{z_1, z_2 \dots z_n\}$ de la base de données. Le problème d'optimisation s'écrit ainsi :

$$\min_w \text{Distance}\left(\{x_1, x_2 \dots x_n\}, \{z_1, z_2 \dots z_n\}\right).$$

Les images générées, $\{z_1, z_2 \dots z_n\}$, dépendent du réseau g_w et donc des poids w. On peut reformuler le problème précédent comme suit :

$$\min_w \text{Distance}\left(\{g_w(y_1), g_w(y_2) \dots g_w(y_n)\}, \{z_1, z_2 \dots z_n\}\right).$$

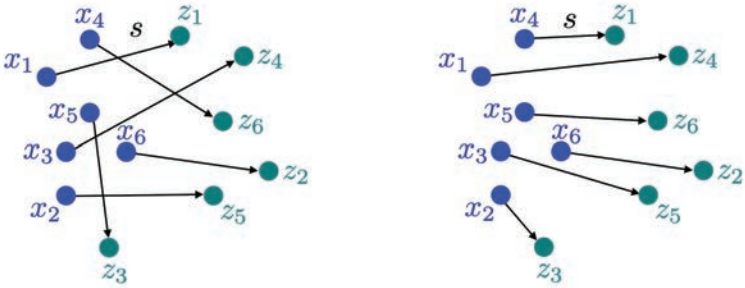
La question mathématique qui se pose est donc de définir une notion de distance entre deux ensembles de points. Il existe bien des façons de le faire ! L'une d'entre elles est particulièrement adaptée à ce problème d'apprentissage. Elle exploite la théorie du transport optimal.

Monge et le transport optimal, des tas de sable aux images

Le *problème du transport optimal* est formulé par Gaspard Monge en 1781, pour des applications militaires. Il s'agit de déterminer la façon la plus économe de transférer des objets depuis un ensemble de sources $\{x_1, x_2 \dots x_n\}$ vers un ensemble de destinations $\{z_1, z_2 \dots z_n\}$.

Pour Monge, il s'agit de transférer de la terre depuis des déblais pour créer des remblais. Mais cette question en fait trouve une multitude d'applications. Pour le problème d'entraînement de réseaux génératifs, les sources sont les images fausses générées par le réseau et les destinations sont les images de la base de données. Il s'agit alors de relier chaque source, par exemple x_1, vers un unique point de destination, que l'on va noter z_{s_1}, où s_1 est un entier entre 1 et n. De manière similaire, x_2 est relié à z_{s_2}, et ainsi de suite.

Sur la figure suivante, on a relié x_2 à z_5, ce qui signifie que $s_2 = 5$. Il faut également que chacune des n destinations soit approvisionnée par une source. Ceci signifie par exemple que x_1 et x_2 ne peuvent pas être reliés à la même destination : il faut relier toutes les sources à des destinations différentes. En d'autres termes, $\{s_1, s_2 \dots s_n\}$ doit être une permutation des n premiers entiers. Sur notre exemple simple (avec $n = 6$ éléments), on a choisi, sur la gauche, la permutation ($s_1 = 1$, $s_2 = 5$, $s_3 = 4$, $s_4 = 6$, $s_5 = 3$, $s_6 = 2$).



Exemple à gauche d'une permutation s non optimale
et à droite de la permutation optimale, dans le cas de six points en dimension 2.

© G. Peyré, 2020

Le problème de Monge consiste alors à trouver la permutation qui minimise la somme des coûts de transport. Monge a posé, pour ses besoins, que le coût de transport entre une source x et une destination z est égal à la distance euclidienne $\|x-z\|$ entre les deux points, mais on peut choisir un autre coût (temps de trajet, prix nécessaire en essence si on utilise des camions…). On doit ainsi résoudre le problème :

$$\min_{w} \text{Distance}\left(\{x_1, x_2 \dots x_n\}, \{z_1, z_2 \dots z_n\}\right).$$

Une fois que l'on a calculé une permutation $s^* = (s_1^*, s_2^* \dots s_n^*)$ optimale (donc solution du problème précédent), on définit la distance entre les ensembles de points comme la valeur du coût total de transport :

$$\text{Distance}\left(\{x_1, x_2 \dots x_n\}, \{z_1, z_2 \dots z_n\}\right) = \left\|x_1 - z_{s_1^*}\right\| + \left\|x_2 - z_{s_2^*}\right\| + \dots + \left\|x_n - z_{s_n^*}\right\|.$$

La difficulté pour calculer cette distance est que le nombre total de permutations à tester est très grand ! En effet, pour le choix de s_1 on a n possibilités, pour celui de s_2 il en reste $n-1$ (puisque la valeur de s_1 est prise), pour s_2 il y en a $n-2$… Donc le nombre total de permutations est égal à $n!$, la factorielle du nombre n, égale au produit $n(n-1)(n-2)\dots\times 2\times 1$.

Pour $n=6$, il existe donc $6!=720$ permutations possibles (faites le calcul !). Dans ce cas simple, on peut toutes les tester et choisir la meilleure, à savoir ($s_1=4$, $s_2=3$, $s_3=5$, $s_4=1$, $s_5=6$, $s_6=2$). Mais déjà pour $n=70$, on dénombre plus de 10^{100} possibilités, ce qui est à comparer aux 10^{79} atomes estimés dans l'univers connu… Et pour entraîner des réseaux de neurones, l'entier n est encore beaucoup plus grand ! Il a donc fallu attendre plusieurs révolutions mathématiques et algorithmiques pour pouvoir obtenir une méthode permettant de résoudre ce problème.

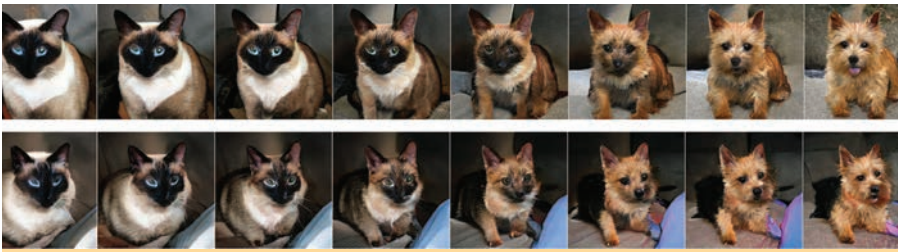
Théorie revisitée, division de la production et prix Nobel

Monge a remarqué que les solutions de son problème ont des structures très particulières. Par exemple, sur notre figure, à droite, les trajets optimaux ne se croisent pas, ce que le conte de Péluse avait prouvé dans son remarquable article. Mais cette remarque pertinente n'est pas suffisante pour résoudre le problème, car il existe encore énormément de trajectoires sans croisement.

Il a fallu plus de deux cents ans pour comprendre comment obtenir plus d'information sur les solutions afin de les calculer efficacement. C'est Leonid Kantorovitch qui a trouvé, en 1942, une nouvelle formulation du problème de transport optimal calculable rapidement. Il a autorisé chaque source à se « diviser » en plusieurs parties, par exemple deux parties égales avec une pondération de 1/2 chacune. Cette « division de la production » est intéressante car elle simplifie le problème d'optimisation. Elle est également naturelle pour les préoccupations de Kantorovitch, qui étaient de modéliser et de planifier la production en économie. Il a d'ailleurs obtenu en 1975 le prix Nobel pour cette idée.

Conjointement aux travaux pionniers de Kantorovitch, George Bernard Dantzig (1914 – 2005) a trouvé en 1947 *l'algorithme du simplexe*, qui permet de résoudre efficacement des problèmes de transport de grande taille. Sa complexité numérique pour résoudre un problème de transport optimal entre n points est de l'ordre de n^3, ce qui est beaucoup plus faible que $n!$. Cet algorithme est au cœur d'un nombre impressionnant de systèmes industriels qui doivent optimiser l'adéquation entre des moyens de production et de consommation.

On peut du coup l'utiliser également pour entraîner des réseaux de neurones génératifs ! Un court article de John Forbes Nash Jr (1929 – 2015), lui-même prix Nobel en 1994, introduit dès 1950 des algorithmes efficaces et des applications à la science des données. Depuis, l'exploration et la recherche continuent !



Deux exemples de *deep fakes* qui sont des images virtuelles interpolant entre chats et chiens.

© G. Peyré

Prendre en compte la géométrie des objets dans les images

Une difficulté pour appliquer le transport optimal pour entraîner des réseaux génératifs est qu'il faut choisir un « coût de transport entre deux images ». On pourrait calculer la distance euclidienne entre les pixels des images, mais ceci ne produit pas des résultats satisfaisants, car une telle distance ne peut pas prendre en compte la géométrie des objets présents dans les images.

Une idée très fructueuse a été introduite en 2014 par Ian Goodfellow et ses collaborateurs. Elle consiste à … utiliser un second réseau de neurones f pour déterminer ce coût de transport ! Ce nouveau réseau, nommé *réseau adversaire*, joue un rôle de discriminateur.

Alors que le but du générateur g est de générer des images fausses très ressemblantes, le but de f est au contraire de faire de son mieux pour reconnaître les vraies et les fausses images. Ces deux réseaux sont entraînés conjointement ; on parle de fait de *réseaux antagonistes*. L'entraînement de g et f correspond à un jeu à somme nulle, concept introduit en 1944 par John von Neumann (1903 – 1957) et généralisé ensuite par Nash en 1950.

Ces avancées récentes ont permis d'obtenir des résultats excellents pour la génération d'images. La figure en page précédente montre des résultats obtenus avec une méthode introduite par Andrew Brock, Jeff Donahue et Karen Simonyan, utilisée pour calculer des « chemins » d'images entre chiens et chats. Après le transport optimal et la géométrie, voici que la topologie se mobilise, ouvrant un nouveau champ à explorer !

G. P.

Pour en savoir (un peu) plus :

Mémoire sur la théorie des déblais et des remblais. Gaspard Monge, *Histoire de l'Académie royale des sciences*, 1781.

On the transfer of masses. Leonid Kantorovich, *Doklady Akademii Nauk* 37 (2), 1942.

Equilibrium points in n-person games. John Nash, *Proceedings of the National Academy of sciences* 36 (1), 1950.

Theory of games and economic behavior. Oskar Morgenstern et John von Neumann, Princeton University Press, 1953.

Origins of the simplex method. George Dantzig, dans *A history of scientific computing*, Addison-Wesley Publishing Company, 1990.

Generative adversarial nets. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et Yoshua Bengio, dans *Advances in neural information processing systems*, 2014.

Large scale GAN training for high fidelity natural image synthesis. Andrew Brock, Jeff Donahue et Karen Simonyan, *Proceedings of the 7th international conference on learning representations*, 2019.

L'intelligence artificielle sans les neurones

Enka Blanchard et Levi Gabasova

Loria, Université de Lorraine
Université Grenoble Alpes

Les réseaux de neurones, dont il a beaucoup été question dans les pages précédentes, ne sont qu'un type particulier d'intelligence artificielle (IA). Ils ne représentent qu'une famille de méthodes parmi d'autres.

Déjà, il semble impossible de mettre les chercheurs d'accord sur ce que l'on appelle « intelligence artificielle ». Une blague du domaine, connue sous le nom de *théorème de Tesler*, affirme précisément que *« l'intelligence artificielle est tout ce qui n'a pas encore été résolu »*. De nombreuses problématiques, comme reconnaître des caractères ou jouer aux dames, ne sont plus considérées comme des problèmes d'IA. Au vu des progrès sur l'apprentissage profond, on pourrait mettre à jour le théorème en le reformulant ainsi : *« L'intelligence artificielle est tout ce qui n'a pas encore été résolu, et ce qui a été résolu sans qu'on comprenne comment. »*

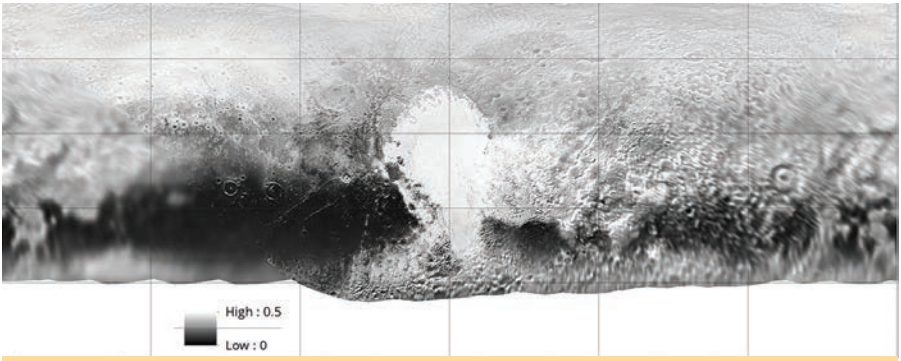
Des flamands roses, des iguanes et un classifieur linéaire

Les problèmes d'IA ont souvent une structure complexe : il s'agit souvent d'« apprendre », et de pouvoir prendre des décisions à partir d'informations incomplètes. Selon les problèmes concernés, on peut utiliser les réseaux de neurones, mais il existe généralementt des méthodes plus simples et moins coûteuses en tant de calcul. Par exemple, faire la différence entre des photos de flamands rose et d'iguanes est facile avec les méthodes vues dans les pages précédentes. Mais ce que l'on appelle un *classifieur linéaire* peut obtenir de bons résultats beaucoup plus simplement, en décidant entre les deux cas en fonction du nombre de pixels roses et de pixels verts.

Très fréquemment, on se trouve donc en présence de fonctions d'optimisation. Formellement, avec une fonction f, on a un objectif y, et on cherche à trouver un x tel que $f(x) = y$. Le plus souvent, on peut se contenter d'une valeur approchée, et on doit donc trouver x avec $f(x)$ « aussi proche que possible » de y.

Pour simplifier les notations, on utilise généralement une formulation équivalente : trouver x tel que $g(x) = 0$, avec $g(x) = |f(x) - y|$. On transforme ainsi un problème d'optimisation générale en problème de minimisation. De très nombreuses méthodes existent pour les résoudre, fruits de plus de soixante-dix ans de recherches. Concentrons-nous ici sur deux techniques : le recuit simulé et les algorithmes génétiques.

Voici un premier exemple pratique issu des sciences spatiales. Après avoir survolé Pluton en 2015, la sonde New Horizons de la NASA nous a transmis de très nombreuses photos de la surface (ce qui prit plusieurs années), couvrant un spectre allant de l'infrarouge aux ultraviolets en passant par le spectre visible. Les cartes résultantes ne sont pas uniformes : la précision dépend de la longueur d'onde ainsi que de l'endroit imagé. Chaque pixel en infrarouge n'est pas simplement une valeur d'intensité moyenne mais est associé à une fonction, correspondant à l'intensité de lumière perçue par la sonde, pour chaque longueur d'onde.



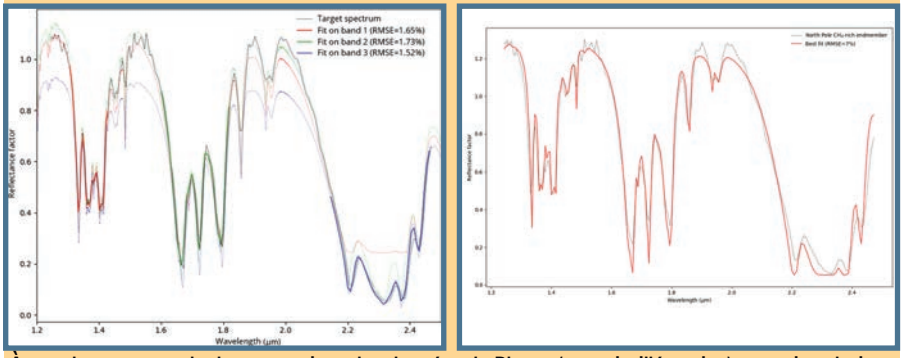
Carte de Pluton en réflectance visible (proportion de lumière réfléchie par la surface), compilée à partir des données de la sonde New Horizons. La résolution de l'image varie fortement selon les zones.

© *Levi Gabasova et New Horizons, NASA*

Les planétologues se posent alors une question difficile : à partir de ce spectre, est-il possible de retrouver la composition du sol de Pluton ? Ce n'est pas une question sans importance : une réponse précise permettrait de mieux comprendre l'évolution non seulement de cette planète naine, mais aussi des autres corps de notre système solaire.

Une variante du problème est relativement simple et peut se résoudre aujourd'hui : à partir du spectre venant du soleil, de la composition du sol, et des vitesses et positions relatives de Pluton et de la sonde, il est possible de calculer le spectre final. Dans la question qui intéresse les planétologues, cependant, on cherche à inverser l'équation : on veut retrouver la composition x à partir du spectre $f(x)$. C'est ce que l'on appelle le *problème de l'inversion du spectre*.

Si f correspondait à une fonction « simple » comme un polynôme, on pourrait donner l'antécédent x de $f(x)$, s'il existe, ou encore une liste des différents antécédents possibles. Le problème est que cette fonction n'est pas si « gentille ». Tout d'abord, le x qu'elle prend en entrée n'est pas une simple variable numérique, mais un ensemble de variables (correspondant aux proportions de différentes molécules et à leur agencement). Ensuite, $f(x)$ n'est pas non plus une valeur, mais une fonction (le spectre de la figure précédente). Notre fonction f va ainsi d'un espace à plus de trente dimensions vers un espace à deux cents dimensions.



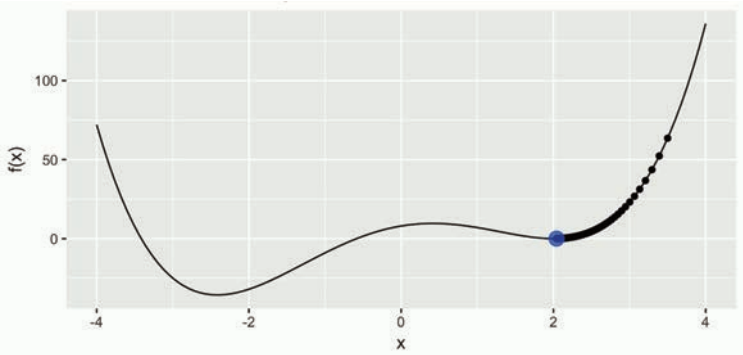
À gauche, un exemple de spectre issu des données de Pluton (en pointillés noirs), et trois solutions candidates, cherchant chacune à optimiser sur une partie du spectre et non pas l'ensemble du spectre. Le taux d'erreur rapporté (RMSE) ne correspond qu'à la partie du spectre optimisée. À droite, une solution candidate pour l'ensemble du spectre, avec un taux d'erreur moyen global plus bas, mais de moins bonnes performances locales.

© *L. Gabasova*

Si $f(x_1, x_2 \ldots x_{30})$ était simplement la somme des trente variables, on pourrait décrire mathématiquement les solutions (potentiellement en nombre infini). Il se trouve que f est une « boîte noire » et correspond à un calcul algorithmique compliqué. Les mathématiques disponibles ne sont pas encore suffisantes, loin s'en faut, pour trouver une équation propre permettant d'en calculer les antécédents. Les scientifiques se retrouvent donc avec une seule option : deviner une valeur potentielle pour x, calculer $f(x)$, et recommencer en ajustant x selon la proximité entre $f(x)$ et le spectre souhaité. La question devient ainsi : comment « deviner » des « bonnes » valeurs pour x ?

Une première idée serait de quadriller l'espace à l'aide d'un maillage fin, et de tester progressivement tous les nœuds de ce maillage. Hélas, même si toutes nos variables étaient comprises entre 0 et 1, et en ne testant que par paliers allant de 0,1 en 0,1, cela ne produirait rien. En effet, le temps de calcul nécessaire pour calculer $f(x)$ pour chaque nœud x serait déjà supérieur à l'âge estimé de l'univers, même en utilisant toute la puissance de calcul terrestre disponible !

Ces problèmes d'optimisation en apparence impossibles sont hélas courants, et il faut quand même les résoudre. On dispose heureusement de méthodes qui produisent en pratique de bons résultats, appelées des *méta-heuristiques*. On peut commencer par la plus simple : en partant d'un x_0 pris au hasard, on calcule $g(x_0)$, ainsi que $g(x_0+\varepsilon_i)$ pour des « petites » variations ε_i. Si l'une des valeurs calculées pour un $x_0+\varepsilon_j$ est inférieure à $g(x_0)$, on recommence à partir de $x_1=x_0+\varepsilon_j$. On s'arrête quand on ne peut plus trouver plus petit (ou alors on continue en diminuant les ε_i, jusqu'à atteindre la précision voulue). Si notre fonction possède un unique minimum, on aura convergé et atteint notre objectif. Par contre, s'il y a deux minimums, comme sur la figure suivante, on peut s'arrêter au mauvais endroit. Il faut alors utiliser des méthodes plus sophistiquées ; l'une des plus connues s'appelle le recuit simulé (tirant son nom d'une technique métallurgique).



Sur cet exemple où $f(x) = (x^2-4x+4)(x^2+4x+2)$, partant de $x_0 = 4$, la méthode de gradient va converger vers 2, qui est un minimum local, et y restera piégée, loin du minimum global, compris entre -3 et -2.

© *Imran Kocabiyik, 2019*

La méthode centrale du recuit simulé s'appelle l'*algorithme de Metropolis–Hastings*. On part, comme précédemment, d'un point aléatoire x_0, et on essaie de se déplacer autour afin d'améliorer successivement la solution. On applique donc à nouveau une petite modification ε, et on calcule $g(x_0+\varepsilon)$. Si $g(x_0+\varepsilon)<g(x_0)$, on a une amélioration, et on fixe donc ainsi $x_1=x_0+\varepsilon$.

Par contre, on n'arrête pas si $g(x_0+\varepsilon)>g(x_0)$. Au contraire, on calcule la quantité $p=\exp(((g(x_0)-g(x_0+\varepsilon))/\mathrm{T})$, où exp désigne la fonction exponentielle et où $\mathrm{T}>0$ est un paramètre, appelé *température* (voir ci-dessous). On va alors prendre une décision probabilistique. Par construction, p est bien un nombre compris entre 0 et 1. Avec probabilité p, on fixe $x_1=x_0+\varepsilon$, et avec probabilité $1-p$ on conserve $x_1=x_0$. On a donc tendance à se rapprocher du

minimum de la fonction, mais on peut aussi adopter temporairement des solutions « moins bonnes ». Plus la température T est élevée, plus $(g(x_0)-g(x_0+\varepsilon))/T$ est « proche » de 0, et donc plus p est élevé (proche de 1). Et plus p est élevé, plus on aura tendance à aller – temporairement – vers des solutions de moins bonne qualité. Cela nous permet en fait de nous échapper des pièges que sont les minimums locaux, comme sur la figure précédente. Par contre, cela n'aide pas à converger vers la solution optimale ! Il faut donc faire « baisser la température » progressivement, afin que l'on finisse par converger… tout en évitant les minimums locaux.

Des choix qui relèvent parfois plus de l'art que de la science

On se retrouve alors avec un problème assez similaire aux problèmes des réseaux de neurones : le choix des paramètres. Ici, on en a deux principaux, le premier concernant les variations ε considérées (par exemple, ε pourrait être choisi uniformément entre -1 et 1, selon une distribution normale ou selon une autre loi de probabilité). Le second, encore plus important, correspond à la fonction associant la température T_n au numéro n de l'itération. C'est naturellement une fonction décroissante, et il existe une multitude de telles fonctions (par exemple, l'inverse de n, ou l'inverse de son logarithme). On a bien des théorèmes démontrant que l'algorithme converge vers la « bonne » solution si l'on choisit les « bons » paramètres, mais ils n'indiquent hélas pas comment choisir ces paramètres en pratique. Souvent, on fait donc des tests, et le choix des paramètres reste plus souvent un art qu'une science.

La procédure du recuit simulé est en fait proche d'un processus biologique très commun : l'évolution des organismes asexués. À chaque génération, l'organisme crée un ou plusieurs clones de lui-même (avec des erreurs correspondant à notre ε), et chacun survit avec une certaine probabilité, dépendant de son adaptation au milieu (qui correspond ainsi en quelque sorte à la qualité de la solution). Ce parallèle a inspiré un autre type d'algorithme : les algorithmes génétiques. Contrairement au recuit simulé, on n'a plus alors une évolution asexuée. À la place, on garde un ensemble de solutions candidates. À chaque itération, ces solutions se « reproduisent », c'est-à-dire qu'on fait un mélange de leurs propriétés.

Prenons un exemple concret, où chaque solution correspond à trois variables (x, y, z). On commence avec trois solutions, (1, 0, 0), (0, 2, 0) et (0, 0, 3). Après une itération, on prend certaines des possibilités de mélange (où l'on va garder une partie d'une solution, et une partie d'une autre). Par exemple, on pourrait obtenir (1, 0, 3) et (0, 0, 0) en mélangeant respectivement la première et la troisième solution, ou la troisième et la deuxième solution. On

évalue la qualité de ces nouvelles solutions, et on garde les meilleures, avant de recommencer. Souvent, on introduit aussi des « petites mutations aléatoires » (imitant ainsi le recuit simulé).

Alors même que ce mécanisme est beaucoup plus proche des méthodes génétiques que l'on peut observer dans la nature, on se retrouve avec une énigme. Dans la nature, la reproduction sexuée est beaucoup plus présente, ce qui semble indiquer qu'elle offre un solide avantage évolutif. Cependant, les algorithmes génétiques simulés ne sont presque jamais plus efficaces que d'autres méthodes d'optimisation comme le recuit simulé, et donnent souvent de moins bons résultats. Pourquoi ? Une réponse partielle à cette énigme fut enfin donnée en 2008, par une équipe internationale de quatre chercheurs. Sous des hypothèses raisonnables, ces derniers ont prouvé que les solutions d'un algorithme génétique ne convergent pas vers les solutions idéales. Au lieu de cela, elles convergent vers des solutions *stables*, c'est-à-dire des solutions qui restent de « bonne qualité » quand elles se mélangent entre elles, ce qui assure un certain niveau de diversité, et permet donc d'éviter un anéantissement total de la population quand l'environnement change (ou que la fonction à optimiser est un peu différente). Mais on n'a pas toujours convergence vers l'optimum désiré.

La recherche de nouveaux algorithmes et de résultats de convergence est actuellement en pleine explosion et mobilise des centaines de mathématiciens de par le monde, aussi bien dans le secteur public que dans le secteur privé. Les applications sont en effet innombrables et nécessitent d'introduire beaucoup de belles mathématiques venant de toutes les spécialités !

E. B. & L. G

Pour en savoir (un peu) plus :

Pluto surface composition from spectral model inversion with metaheuristics. L. Gabasova, N. Blanchard, B. Schmitt, W. Grundy, C. Olkin, J. Spencer, L. Young, K. Ennico-Smith, H. Weaver, A. Stern et The New Horizons COMP Team, EPSC-DPS joint meeting, 2019, disponible en ligne.

Global compositional cartography of Pluto from intensity-based registration of LEISA data. L. Gabasova, B. Schmitt, W. Grundy, T. Bertrand, C. Olkin, J. Spencer, L. Young, K. Ennico-Smith, H. Weaver, A. Stern et The New Horizons Composition Team, Icarus, 2020.

A mixability theory for the role of sex in evolution. Adi Livnat, Christos Papadimitriou, Jonathan Dushoff et Marcus Feldman, *Proceedings of the National Academy of Sciences of the United States of America* 105 (50), 2008, disponible en ligne.

Des maths partout, même dans les jeux vidéo !

Nicolas Nguyen*

Professeur de mathématiques en classes préparatoires

Les mathématiques fournissent aux entreprises des outils incomparables pour l'analyse et la gestion de systèmes complexes. On retrouve souvent en bonne place le *data mining* (fouille et analyse de données) et l'intelligence artificielle, mais également le traitement du signal et de l'image, la modélisation, la simulation, l'optimisation, la recherche opérationnelle (en logistique, typiquement), le calcul haute performance, les statistiques, le calcul stochastique, la sécurité de l'information, la cryptographie, la bio-informatique… De fait, bien qu'invisibles, les mathématiques sont omniprésentes !

On parie ?

Vous êtes sans doute nombreux à pratiquer les jeux vidéo et à apprécier leur fluidité, sans savoir que derrière l'animation se cachent des mathématiques avancées (voir encadré). Elles interviennent aussi bien pour analyser les données de jeu qu'en amont, lors de leur conception, où l'on a recours aux statistiques et aux probabilités pour créer de nouveaux univers et les animer.

Pour illustrer ce propos, Samuel Sergeant, analyste de données (*data analyst* dans le jargon), a accepté de nous parler de son métier. Après un master « Ingénierie statistique et Numérique », il a rejoint le leader national de l'industrie du jeu vidéo, Ubisoft, il y a trois ans. Une évidence pour lui : *« Au sein de nos équipes, nous ne partageons pas uniquement une activité professionnelle, mais aussi une passion : les jeux vidéo ! »* Aujourd'hui, ils ne sont qu'une poignée d'analystes de données, un domaine en forte croissance ; le secteur est donc *« à la recherche de nouveaux talents formés aux sciences de la donnée »*.

*L'auteur remercie Samuel Sergeant pour sa participation.

Des indicateurs de performance en support de la production

Un analyste de données travaille pour la production, avec un objectif simple : contribuer à améliorer l'expérience des joueurs. Les données analysées sont recueillies par télémétrie : le joueur joue en ligne, ses données de jeux sont envoyées aux serveurs où elles sont formatées afin d'être analysées. Une part du travail consiste à définir des indicateurs de performance clés (ou KPI pour *key performance indicator*) et à les mesurer. Par exemple, Just Dance 2020 (Ubisoft, 2019) consiste à reproduire les mouvements du danseur présent à l'écran. Plusieurs modes de jeu y sont disponibles : « Kids » (« enfants »), « Just Dance » (« danse ») et « All Stars » (« vedette »). Pour analyser la performance d'un mode de jeu, on peut considérer les KPI suivants : proportion de joueurs qui essaient ce mode ; proportion de joueurs qui retournent sur ce même mode ; temps de jeu moyen passé dans ce mode. Il s'agit ensuite d'expliquer les résultats observés lors de sessions dédiées (les *playtest*s, où les joueurs sont invités à essayer le jeu) ou grâce à des sondages. Toutes ces données sont indispensables à la production : elles donnent une vision d'ensemble d'un jeu et vont permettre d'orienter des choix stratégiques (augmenter le temps de jeu, améliorer l'utilisation d'un mode de jeu…).

Améliorer par l'expérimentation : les statistiques et l'*AB testing*

Les équipes de production ont également la possibilité d'expérimenter des modifications durant la vie d'un jeu. L'idée est simple : afin de décider quelle version du jeu est jugée « la meilleure », on fait tester deux versions aux joueurs simultanément. La version de contrôle A (non modifiée) ou la version B (modifiée). On évalue ensuite les indicateurs de performance des différents modèles proposés aux joueurs pour enfin retenir l'option jugée la plus performante. Afin d'éviter tout biais, les joueurs ignorent s'ils s'exercent sur la version A ou B. Ces *méthodes d'AB testing* ont initialement été utilisées dans le *e*-marketing afin d'améliorer la performance des sites Internet (Google a par exemple réalisé le choix de couleurs de son logo *via* cette méthode). Il s'agit donc, typiquement, de proposer à un échantillon aléatoire de joueurs une version alternative du jeu (nouvelles façons d'afficher le contenu, différentes façons de le recommander…). Les jeux développés dans les studios intègrent déjà en amont des versions alternatives, facilement paramétrables en vue de faire l'objet d'AB tests.

Sur Just Dance 2020, deux interfaces différentes ont été proposées et testées.

En haut, la version de contrôle.

En bas, la version modifiée, avec ajout de la ligne « Recommandé pour vous ».

© *Ubisoft, 2020*

Les algorithmes et le *machine learning* entrent en piste

Finalement, les mathématiques interviennent également en temps réel pour améliorer l'expérience du joueur, grâce à l'apprentissage automatique (*machine learning*). Le principe consiste à analyser les données du jeu pour en tirer des prédictions.

Un exemple d'algorithme basé sur l'historique de l'ensemble des joueurs pour proposer un contenu personnalisé à chacun.

© *Ubisoft, 2020*

Jeux vidéo et mathématiques : un duo gagnant !

Un *quaternion* q s'écrit à l'aide de quatre réels sous la forme $q = a + x\,i + y\,j + z\,k$ où i, j et k vérifient les relations $i^2 = j^2 = k^2 = ijk = -1$. Parmi ces nombres, on retrouve donc les nombres complexes $\mathbb{C}$ ($y = z = 0$) et les nombres réels $\mathbb{R}$ ($x = y = z = 0$). L'ensemble $\mathbb{H}$ des quaternions (nommé ainsi en l'honneur de Sir William Rowan Hamilton, qui les découvrit en 1843) est muni d'une addition et d'une multiplication (comme $\mathbb{R}$ et $\mathbb{C}$), qui font de $\mathbb{H}$ un corps non commutatif. Tout comme les nombres complexes constituent un outil puissant pour l'étude de la géométrie plane, les quaternions jouent un rôle majeur pour l'analyse dans l'espace ; ils sont notamment abondamment utilisés en aérospatiale et dans l'animation 3D des jeux vidéo.

Un quaternion *imaginaire pur*, de la forme $v = x\,i + y\,j + z\,k$, s'identifie au vecteur de l'espace v de composantes (x, y, z). Si u est un quaternion *unitaire* (de norme 1), la transformation qui à q associe uqu^{-1} induit une rotation sur les quaternions imaginaires purs, c'est-à-dire une rotation sur les vecteurs de l'espace. En fait, toute rotation de l'espace est représentée par un quaternion unitaire.

D'autres façons de représenter une rotation existent à l'aide des trois angles d'Euler ou d'une matrice 3×3. Cependant, la représentation par un quaternion unitaire a des avantages notables. Les calculs sont de fait bien moins gourmands que les calculs matriciels. Il suffit de quatre réels pour représenter un quaternion quand il en faut neuf pour une matrice ! Ainsi, le produit de deux quaternions coûte vingt-huit opérations élémentaires seulement, contre cinquante-quatre pour le produit de deux matrices 3×3.

L'utilisation des quaternions en animation prend tout son sens lorsque la caméra d'un joueur évolue au sein d'une scène : ce que l'on voit à l'écran, c'est l'environnement transformé par une suite de rotations spatiales. Mathématiquement, les transitions pour passer d'un quaternion unitaire q_0 à un autre q_1 sont définies par $q_s = q_0\,(q_0^{-1} q_1)^s$ pour s dans $[0, 1]$. À chaque instant s, q_s est une rotation, qui va varier de la façon la plus naturelle qui soit entre q_0 et q_1. C'est ainsi que lorsque votre personnage de jeu vidéo préféré se déplace et balaie du regard une scène 3D, la fluidité est parfaite !

L'apprentissage automatique est une branche de l'intelligence artificielle (IA) qui consiste à utiliser des algorithmes pour apprendre automatiquement à prédire des phénomènes ou des comportements (voir les articles dans cette brochure). Ici, ces algorithmes peuvent servir à proposer du contenu personnalisé,

apprendre à une IA à reproduire des comportements humains, faire de la reconnaissance d'image… Pour Just Dance 2020, un algorithme de recommandation a été ainsi développé. Afin d'aider le joueur dans sa sélection de musiques, cet algorithme essaye d'anticiper ses choix en lui proposant en priorité les titres qui seront « le plus susceptibles de lui plaire ». Le but est de réduire le temps de recherche de la prochaine chanson dans le catalogue, qui contient plus de cinq cents titres.

Les entreprises toutes friandes de mathématiques !

Aujourd'hui, plus encore qu'hier, les grands groupes n'hésitent plus à faire appel aux mathématiques pour doper leur innovation. Du développement de nouveaux produits jusqu'à leur commercialisation, en passant par les procédés de fabrication, la reine des sciences contribue à tous les niveaux de la chaîne de valeur d'un produit.

Les mathématiques se distinguent par leur caractère universel. Galilée, déjà, ne disait-il pas au XVII^{e} siècle que le livre de la nature est écrit en langage mathématique ? Elles se nourrissent de nombreux autres champs et irriguent moult disciplines, et c'est probablement ce qui explique l'importance de leur contribution au développement de technologies clés, leviers de l'économie française. Pour toutes celles liées au numérique, aux transitions climatique et énergétique, à la santé comme à la finance, les mathématiques apportent un véritable avantage concurrentiel. En fait, selon le ministère de l'Industrie, elles sont déterminantes pour le développement de 49 % des technologies clés pour 2020, en fournissant des outils pour la modélisation, la simulation et l'ingénierie numérique. Retrouvez nombre de ces aspects dans les pages de cette brochure !

Moteur de l'innovation, les mathématiques sont également créatrices de valeurs. Elles ont permis l'émergence de nouveaux marchés et de nouveaux métiers, comme ceux d'*analyste de données* et de *scientifique des données* (*data scientist*). Couplées à l'informatique, et en interaction avec les disciplines du secteur d'activité concerné, elles permettent de construire et de manipuler des modèles complexes, de proposer des simulations numériques à la base de création de valeur, dans l'industrie et les services. La gestion de risque et l'aide à la décision sont indispensables à la stratégie et à la prospective.

En matière de conception de produit, les mathématiques permettent de réduire les coûts et d'accélérer la recherche. Du côté de la production, en modélisant pièces et assemblages, elles optimisent les procédés. En contrôle qualité, grâce notamment aux statistiques et aux technologies d'analyse d'image, le contrôle de la production à distance améliore son efficacité. La maîtrise des coûts d'énergie et le contrôle financier ont un impact direct sur la gestion de l'entreprise. La modélisation de la relation client, l'analyse du comportement client et la gestion des prix (*pricing management*) intéressent le secteur commercialisation-vente…

Les petites et moyennes entreprises (PME) ne sont pas en reste. Elles évoluent elles aussi au sein de marchés de plus en plus complexes. Les outils d'analyse, de simulation et de prédiction que leur apportent les mathématiques constituent des avantages concurrentiels indéniables !

N. N.

Maths et entreprise : un partenariat gagnant

Les entreprises n'ont pas toujours la capacité financière d'intégrer des mathématiciens au sein de leurs équipes, ou la visibilité nécessaire pour le faire. Il faut donc trouver des solutions adaptées à plusieurs échelles de temps. De nombreux laboratoires universitaires de mathématiques proposent aujourd'hui un large panel de solutions, allant d'une étude exploratoire d'une semaine à une thèse de trois ans afin que chaque entreprise puisse trouver une collaboration adaptée à ses objectifs et à ses moyens. Les difficultés majeures sont de sensibiliser ces entreprises aux atouts qu'elles peuvent tirer des mathématiques et de leur faire connaître l'offre universitaire.

Dans ce contexte, l'association Le Temps des Sciences, en partenariat avec l'Agence Lebesgue de Mathématiques pour l'innovation, coorganisent une journée de rencontres entre laboratoires de recherche mathématique et entreprises de Bretagne et des Pays de la Loire, intitulée Journée innovation et mathématiques. Trois cent cinquante entreprises sont ainsi attendues pour la prochaine édition, le 17 mai 2021, au couvent des jacobins à Rennes (Ille-et-Vilaine).

Renseignements : www.journee-innovation-et-mathematiques.com

Et si on savait tous compter ?

Jean-Marie De Koninck

Université Laval, Québec, Canada

Comment comprendre les différents enjeux liés à la dure réalité de la Covid-19 si on ne sait pas ce que signifie une « croissance exponentielle » ? Les mathématiques font de plus en plus partie de notre quotidien, et ceux et celles qui en sont détachés sont souvent laissés pour compte, voire vulnérables.
Avant 1454, soit avant l'invention de l'imprimerie, le citoyen pouvait se débrouiller pour vivre en société sans savoir lire. Après 1454, ceux qui savaient lire avaient un net avantage sur ceux qui ne jouissaient pas de cette compétence. Jusqu'à 1950, il était encore possible de fonctionner en société et de « réussir dans la vie » sans avoir des connaissances de base en mathématiques. Vers la fin des années 1970, soit au moment où les ordinateurs personnels sont fabriqués à grande échelle, l'humanité connaît une véritable révolution numérique. Or, comme le squelette de cette révolution numérique repose sur les mathématiques, il est devenu difficile, pour celui ou celle qui ne détient pas un bagage minimal en mathématiques, d'être performant au travail et même de répondre adéquatement aux exigences de la vie quotidienne.

Alors qu'on peut définir la *littératie* comme l'ensemble des connaissances en lecture et en écriture permettant à un individu d'être fonctionnel en société, la *numératie* correspond quant à elle à la capacité d'une personne de comprendre et d'interpréter de l'information numérique dans une variété de contextes. Parmi ces contextes, on retrouve la gestion d'un budget familial, le calcul du montant d'intérêt à payer pour une offre de prêt et la compréhension des textes schématiques, entre autres choses.

Une étude du Conseil canadien de l'apprentissage démontrait, en 2014, que 55% des adultes canadiens présentent un niveau de connaissances en mathématiques insuffisant pour répondre aux exigences de la vie quotidienne, et que 49% des adultes canadiens sont incapables d'interpréter adéquatement des tableaux, des graphiques ou des formulaires. Il y a tout lieu de croire que le même genre de constat s'applique en France.

Savoir compter, c'est d'abord et avant tout avoir le sens du nombre

Aller à la source nous permet de résoudre les problématiques de façon durable. C'est pourquoi nous devons nous concentrer sur la formation et l'éducation. Nous avons effectivement tous un rôle important à jouer auprès des plus jeunes quant au maintien de leur intérêt envers les mathématiques, que l'on soit enseignant, parent ou personne d'influence. En effet, chaque enfant possède une aptitude naturelle pour tout ce qui est numérique (voir l'article *Un cerveau naturellement conçu pour les mathématiques*). De fait, il est prouvé que les enfants d'âge préscolaire ont un potentiel énorme en mathématiques. D'ailleurs, ces dernières s'avèrent l'un des moyens les plus efficaces par lesquels ils peuvent apprécier et comprendre le monde qui les entoure. C'est pourquoi, afin que les enfants développent le sens du nombre et acquièrent au fil du temps différentes compétences mathématiques, nous partageons la responsabilité de leur montrer la richesse de cette discipline et, surtout, d'entretenir leur instinct et leur talent naturel pour celle-ci. Trop souvent, nous sous-estimons leur capacité d'émerveillement, et finissons par les éloigner des mathématiques en leur présentant la matière comme ardue et rébarbative au lieu d'en montrer le côté ludique et réjouissant.

Cette approche finit par nous coûter cher collectivement quand ces jeunes deviennent des citoyens présentant un niveau de numératie inadéquat ! Ils sont ainsi moins à même de participer efficacement à la vie en société. D'où l'importance de développer, mais surtout de maintenir l'intérêt pour les mathématiques de façon continue.

Et les décideurs dans tout ça ?

Plusieurs de nos décideurs rentrent dans la catégorie des jeunes s'étant éloignés des mathématiques. Ils manquent souvent d'outils ou de connaissances de base pour bien gouverner et prendre des décisions éclairées. Un jour, on arrivera peut-être à concevoir un cours de mise à niveau en mathématiques à l'intention de tout individu se préparant à un rôle de gestion. Ce cours comprendrait, entre autres, un peu d'arithmétique (tables de multiplication, règle de 3…), les calculs d'intérêts composés et d'hypothèques, les notions de croissance de populations, de croissance exponentielle et de logarithme, les notions de base en statistiques, *etc*. Non seulement un tel cours d'appoint intéresserait sans doute un grand nombre de citoyens et les inciterait peut-être eux-mêmes à améliorer leurs connaissances en mathématiques, mais cette remise à niveau rassurerait certainement les électeurs et employés quant au bien-fondé des orientations prises et celles à venir.

Comprendre la pandémie de Covid-19 et agir en conséquence

Voyons pourquoi pour bien saisir les enjeux de la pandémie de Covid-19 qui sévit, il serait souhaitable d'être familier avec la notion de croissance exponentielle. Une personne infectée par la grippe traditionnelle contaminera en moyenne 1,4 personne, alors qu'une personne porteuse du coronavirus Covid-19 en contaminera en moyenne 3. Ainsi, au bout de dix relais de transmission du virus, la première personne grippée aura en bout de ligne été la source de l'infection d'environ $(1{,}4)^{10} = 28{,}9\ldots$ personnes, alors que la première personne porteuse du coronavirus Covid-19 aura en bout de ligne été la source de l'infection d'environ $3^{10} = 59\,049$ personnes. Il y a donc lieu de croire que ceux qui comprennent ces chiffres et saisissent donc bien le potentiel de propagation du nouveau coronavirus adopteront un comportement responsable, en particulier en ce qui a trait au confinement.

Apprécier avec modestie l'immensité de l'Univers

En marge des aspects utilitaires d'un excellent bagage de connaissances en mathématiques, il y a aussi le potentiel d'appréciation ludique des phénomènes que l'on peut percevoir. Ainsi, quiconque sait compter pourra davantage apprécier l'immensité de l'Univers dans lequel on vit, en prenant conscience qu'on y occupe une place bien modeste. Un exemple classique illustrant bien cette immensité est qu'il existe autant d'étoiles dans l'Univers connu que de grains de sable sur Terre. En effet, les astronomes savent qu'il existe dans l'Univers connu environ deux cent cinquante milliards de galaxies et que chacune de ces galaxies contient en moyenne entre deux cents et cinq cents milliards d'étoiles. Cela veut dire qu'il y a approximativement 10^{23} étoiles dans l'Univers observable.

On évalue par ailleurs qu'il existe environ 10^{23} grains de sable sur l'ensemble des plages de notre planète. C'est pourquoi on peut conclure que le nombre d'étoiles dans l'Univers et le nombre de grains de sable sur Terre sont du même ordre de grandeur. Ce constat ne peut faire autrement que nous inspirer une certaine humilité.

Parallèlement, on peut aussi apprécier l'immensité de l'Univers en prenant connaissance d'un des plus beaux héritages du télescope Hubble, soit une photo d'un coin sombre du ciel où les astronomes ont pu identifier pas moins de dix mille galaxies.

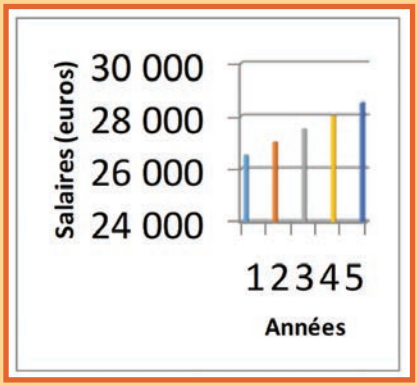
Les galaxies abondent dans l'Univers,
comme le prouve cette exceptionnelle image d'une portion de ciel de la taille de la Lune,
formée de centaines de clichés pris depuis seize ans par le télescope Hubble
et qui contient plus de deux cent mille galaxies.

© *Nasa, ESA, mai 2019*

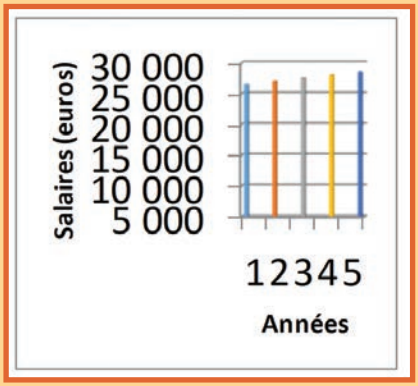
Le premier coup d'œil est souvent trompeur

Placé devant des données numériques présentées sous forme de graphiques, il est important de savoir les saisir dans leur globalité pour en retenir l'essentiel, ce qui permettra de porter un jugement éclairé. C'est ainsi que l'on peut facilement imaginer deux graphiques à première vue différents, mais qui en définitive représentent exactement les mêmes données. Par exemple, examinons les deux diagrammes ci-dessous, représentant tous deux les salaires annuels offerts aux employés d'une entreprise sur cinq années consécutives (de 26 500 € à 28 500 €), l'un selon le point de vue patronal et l'autre selon le point de vue syndical.

Le premier semble indiquer une progression fulgurante du salaire annuel, l'autre une progression très modeste. Il ne faut donc jamais se laisser berner par les premiers coups d'œil.



Le point de vue patronal.



Le point de vue syndical.

© *J.-M. D. K.*

L'évaluation des risques est aussi affaire de mathématiques

Très souvent, nous avons à évaluer les risques avant de prendre une décision. Certes, il y a les risques financiers (choix d'une hypothèque, comparaison des prix de différents articles de consommation…), mais il y a aussi les décisions quotidiennes dans plusieurs domaines d'activité, tels que les risques associés aux différents modes de transport. Fort heureusement, plusieurs études se sont penchées sur la question. Par exemple, l'Agence ferroviaire européenne a confirmé qu'entre 2008 et 2010, sur l'ensemble des territoires des vingt-sept pays de l'Union européenne, le nombre de morts dans des accidents, par milliard de voyageurs-kilomètres, est de 48,94 pour les motos, 3,14 pour la voiture individuelle, 0,20 pour les passagers de car, 0,13 pour le train et enfin 0,06 pour l'avion. Donc, l'avion est de loin le mode de transport le plus sécuritaire. Par contre, le coût du billet d'avion est peut-être prohibitif pour certains. Alors, quoi choisir pour se déplacer entre Paris et Lille (Nord), disons ? Or, il s'avère que les chiffres donnés ci-dessus nous confirment que la voiture est vingt-quatre fois plus risquée que le train. Voilà un bel exemple montrant que la capacité à interpréter des données mathématiques peut nous mener à faire de meilleurs choix, quel que soit le domaine.

Quantifier la chance, infime, de gagner aux jeux de hasard

Au Québec, la loterie la plus populaire se nomme la 6/49. Elle consiste en un tirage au sort de six boules numérotées de 1 à 49. C'est le même principe que le Loto en France. Ainsi le nombre de tirages distincts à la 6/49 est égal

à $\frac{49!}{6!43!}$ où, pour tout entier naturel n, $n!$ (la factorielle de n) désigne le produit $n\times(n-1)\times(n-2)\times\ldots\times2\times1$. Le nombre de tirages différents vaut donc, après simplification, $\frac{49\times48\times47\times46\times45\times44}{6\times5\times4\times3\times2}$, soit 13 983 816.

La probabilité de « deviner » la combinaison gagnante est donc d'environ une chance sur quatorze millions. Comment peut-on visualiser une telle probabilité ? Imaginons que nous nous amusions à aligner des pièces de monnaie de vingt-cinq cents (ou une pièce d'un euro) le long de l'autoroute reliant Québec et Montréal et qu'une seule d'entre elles soit identifiée comme la pièce gagnante. On peut se demander si on aurait davantage de chances de choisir au hasard cette pièce que de gagner le gros lot de la 6/49. En effet, sachant qu'une pièce de vingt-cinq *cents* mesure environ 2,4 cm de diamètre et que la distance entre les centres-villes de Québec et Montréal est de 270 km, cela veut dire que nous pourrions aligner onze millions deux cent cinquante mille pièces le long de la route entre les deux villes.
Or, comme $1/11\,250\,000 > 1/13\,983\,816$, on peut conclure que gagner le gros lot de la 6/49 serait encore moins probable que de tirer au hasard la pièce de vingt-cinq *cents* désignée gagnante ! Il faut bien comprendre que les jeux de hasard constituent un impôt volontaire, dont le principal objectif est de remplir les coffres de l'État. Devant ce constat, le citoyen doit être conscient qu'en s'adonnant aux jeux de hasard, il le fait pour le plaisir et non dans l'espoir de faire un gain qui viendrait grossir son enveloppe de revenus.

Éviter les arnaques et vivre plus heureux !

En tentant d'améliorer la qualité de son français, on arrivera à développer des habiletés de communication. En étudiant l'histoire, on apprendra à comprendre le comportement des humains en tant que collectivité. En étudiant la biologie, on développera un respect pour la vie et pour l'environnement. En se familiarisant avec les concepts de base en mathématiques, on développera des compétences telles que la rigueur intellectuelle, le raisonnement critique et la résolution de problèmes.

En définitive, si nous savions tous compter, nous nous sentirions plus compétents dans nos milieux de travail respectifs, nous serions mieux éclairés lorsque vient le temps de faire des choix (par exemple d'ordre financier), nous serions davantage à l'abri des arnaques et nous pourrions mieux accompagner nos enfants dans leurs études. Somme toute, osons le dire, nous serions plus heureux !

J.–M. D.K.

Les maths, ça sert… à être heureux !

Emmanuël Houdart

Mathématicien et comédien

« Les maths, à quoi ça sert ? » Ma bibliothèque est remplie d'une pléthore d'ouvrages qui justifient l'apprentissage de ma discipline adorée. Je me souviens d'ailleurs que je m'en servais allégrement pour répondre à ces doigts audacieux qui se levaient chaque mois de septembre : *« M'sieur ! Mais à quoi ça sert, vos maths ? »* J'attendais presque ces instants. Fier de moi, j'énumérais alors les différentes raisons qui allaient forcément les convaincre des bienfaits de cette matière injustement honnie par trop d'étudiants. Je terminais ma plaidoirie en faveur des mathématiques en décrochant ma flèche fatale : *« Parce que, enfin, vous allez pouvoir comprendre le monde ! »* Fermant les yeux, je savourais l'écho du silence qui devait ainsi mettre un point final au plus résistant des esprits farouches et je m'apprêtais à poursuivre face à un auditoire forcément convaincu de l'utilité d'une solide formation mathématique.

« Pourquoi infliger ce supplice à toute la population ? »

Jusqu'au jour où, lorsque j'ouvris les yeux, ce ne fut pas pour découvrir des êtres enclins à l'apprentissage mais bien plutôt des moues désabusées loin d'être satisfaites de la réponse entendue. J'acceptais de relever le défi : il allait s'agir cette fois d'être plus pragmatique et percutant. Je leur ouvris les yeux sur les modèles mathématiques développés pour répondre aux grandes questions climatiques, aux enjeux énergétiques, aux progrès médicaux ; je leur démontrai comment le paradoxe de l'amitié permettait de neutraliser une pandémie en détectant des individus dits « centraux ». Plus aucun doute, il était bien impérieux d'étudier les mathématiques. CQFD.

Mais une goutte de sueur me perla le front lorsqu'une étudiante m'assaillit d'un : *« Mais pourquoi diable fallait-il infliger ce supplice à toute la population ? »* Oh, ils avaient écouté attentivement ma diatribe, et tous voulaient bien admettre que le monde ne pouvait pas tourner rond sans les mathématiques. Mais du moment que quelques adeptes s'en chargeaient, cela

n'était-il pas suffisant ? Après tout, tout le monde convient de l'utilité d'un téléphone portable, est-ce pour autant que notre société impose à chacun d'en connaître le fonctionnement ?

Diable ! Que répondre ? Des bribes argumentaires me vinrent bien à l'esprit mais cette étudiante avait fait mouche : après tout, pourquoi imposer à tout le monde l'apprentissage des mathématiques ? Vingt-cinq ans plus tard, mon expérience me permettrait de répondre différemment à cette question. C'est assez curieux mais mon désir insatiable de défendre les intérêts des mathématiques m'a tout d'abord amené à… quitter mon métier d'enseignant. Pourquoi ? Parce qu'au fur et à mesure de quinze années enrichissantes et pleinement satisfaisantes, j'avais bien compris que le réel frein à l'apprentissage des mathématiques n'était pas un manque de capacités des élèves (qui en douterait ?), mais un manque de motivation. Il était donc nécessaire de donner goût aux mathématiques.

C'est sur la base de cette idée (et d'une solide conviction) que j'ai fondé, à Quaregnon, en Belgique, la Maison des maths. Un endroit unique dont le seul objectif était de faire découvrir le plaisir des mathématiques. Et le pari audacieux s'avéra largement gagnant. Durant trois années, ce sont des milliers de visages que j'ai vus irradiés de sourires. Quel que soit l'âge de nos visiteurs, ils semblaient heureux de déjouer les tours facétieux d'une énigme coriace, enthousiastes face à l'ingéniosité de notre système de numération, admiratifs de la créativité des mathématiciens. Et c'est durant ces trois années que j'ai compris que, finalement, les mathématiques ont l'incroyable pouvoir de rendre heureux.

Les mathématiques ont l'incroyable pouvoir de rendre heureux

Si vous en doutez, expliquez-moi alors l'incroyable succès de Martin Gardner (1914–2010), qui tint en haleine – durant plus de vingt ans – des millions de lecteurs grâce à une chronique intitulée « Jeux mathématiques ». Auteur prolifique, Martin Gardner a été l'un des premiers vulgarisateurs à comprendre la puissance de l'effet « ha-ha » des mathématiques.

Beaucoup plus récemment, dans son livre *Alex et la magie des nombres* (Robert Laffont, 2015), l'auteur britannique Alex Bellos comparaît les mathématiques à une blague. Et il ne dit pas ça pour rire. Loin de là. En guise d'introduction, Alex nous explique qu'une plaisanterie est un récit constitué d'un développement et d'une chute. On l'écoute attentivement jusqu'au bouquet final qui provoque le rire. Il en va de même pour un raisonnement mathématique.

Bien entendu, c'est une narration d'un autre genre, où les protagonistes sont des nombres, des symboles, des formes et des schémas. Mais si l'on suit la démonstration jusqu'au moment de la récompense, alors bingo, les neurones s'affolent et une vague de satisfaction intellectuelle balaye le sentiment initial de confusion… ce qui provoque la joie de la compréhension.

Si les mathématiques ne rendaient pas heureux, comment pourrait-on expliquer cette déferlante mondiale qu'ont engendrée les grilles de Sudoku ? Combien de fois m'est-il arrivé d'observer, dans le métro, la joie d'un navetteur noircissant la quatre-vingt-unième et dernière case ? Si les mathématiques ne rendaient pas heureux, comment expliquer autrement les casse-tête et énigmes mathématiques qui fleurissent sur la Toile durant cette difficile période de confinement ?

Manu Houdart
© Stephane-Kerrad-KBStudios

Si je peux vous en parler avec autant de conviction, c'est parce j'ai l'occasion d'expérimenter régulièrement cet effet « Waooh ! » des mathématiques. J'en ai d'ailleurs fait le *gimmick* de mon spectacle « Very MATH Trip ».

Présenté pour la première fois au festival d'Avignon en juillet 2019, je n'ai pas manqué de surprendre un public très étonné d'entendre le mot « mathématiques » au cœur de la Cité des papes. Pour les non-initiés, sachez qu'il est coutume lors de ce festival de tracter durant la journée, car c'est tout de même mille cinq cents (!) spectacles différents qui sont joués quotidiennement. Me voici donc sous un soleil de plomb – invitant tout un chacun à venir découvrir un spectacle unique en son genre :

un one-*math*-show !

Je me régalais d'observer les attitudes des festivaliers. Une minorité angoissée (irrécupérable ?) détalait à toute allure, de peur sans doute d'être rattrapée par d'effroyables souvenirs trau-math-iques ; d'autres toujours aussi minoritaires me désarmaient avec violence de mes précieux dépliants, tout heureux de retrouver leur discipline préférée.

Mais la grande majorité s'approchait de moi, l'intérêt piqué au vif par mon audace de porter les mathématiques sur le devant de la scène et curieuse d'en découvrir le contenu. La plupart du temps, j'arrivais à argumenter suffisamment bien pour que le festivalier en quête de spectacle décide qu'il avait enfin trouvé son bonheur. Ne me restait plus alors qu'à confirmer l'essai lors de la représentation.

« Ha-ha » et effet « waooh ! » : la joie de la compréhension

C'est un ressenti particulier d'avoir chaque soir, face à soi, une salle remplie. J'adore, dissimulé derrière le rideau, en écouter le bruissement. Un public, alléché par mes promesses enrobeuses, attend d'être séduit par les mathématiques. Certains avec enthousiasme, d'autres par défiance. Ça y est, le rideau se lève. Silence complet. Les premières minutes du spectacle me permettent de prendre la température de la salle quand soudainement, presque par surprise, un premier effet « Waooh ! » surgit. Les spectateurs se regardent, étonnés. Ils viennent d'être touchés en plein cœur par le « ha-ha » si cher à Martin Gardner. À partir de cet instant, je sais que je n'ai plus rien à craindre car les effets « Waooh ! » vont s'enchaîner, à toute allure, plongeant le spectateur dans un univers mathémagique. Le pétillement de leurs yeux trahit ce qu'Alex Bellos, et d'autres avant lui, ont si bien décrit : la joie de la compréhension.

Certains penseront qu'il était téméraire de monter sur scène avec les mathématiques comme toile de fond. Je leur répondrais que c'est plutôt facile de s'abriter derrière elles. Il serait bien impertinent de ma part de croire que le succès du show provient de ma performance scénique ! Certes, j'espère y contribuer, bien entendu, mais je ne suis pas dupe.

Ce sont bien les mathématiques qui provoquent cette joyeuse atmosphère régnant dans la salle. Et ce sont encore et toujours elles qui offriront ainsi aux spectacteurs un moment intense… de bonheur !

E. H.

Pour en savoir (un peu) plus :

Very Math Trip. Emmanuël Houdart, Flammarion, 2019.

One Zero Show Du point à la ligne. Denis Guedj, Le Seuil, 2001.

À l'endroit de l'inversion, petit essai en clownologie mathématique. Cédric Aubouy, L'Île Logique, 2017.

Ils vous font vivre les maths autrement !

Nous vous présentons ci-dessous une liste non exhaustive de mathématiciens qui suivent le chemin tracé par Denis Guedj (1940 – 2010) avec ses pièces « *One zéro show* » et «*Du point à la ligne* » (Le Seuil, 2001).

Julie-Anne Leblanc, du projet québécois Sciences et mathématiques en action piloté par Jean-Marie De Koninck, propose des conférences-spectacles « *Show Math* » et des pièces de théâtre durant lesquelles acteurs et public s'amusent avec les mathématiques.

Dominique Souder est passé maître dans l'art de la « mathémagie ». Cet auteur prolifique (*Maths & Magiques*, SOS Éducation, 2015, 2016, 2018) émerveille petits et grands avec ses tours de magie qui reposent sur des principes mathématiques.

Daniel Justens a traqué les mathématiques dans l'œuvre de Philippe Geluck (*la Mathématique du Chat*, Casterman, 2008). Il a collaboré avec Henri Vernes au tome 78 des aventures de Bob Morane, *les Gardiens d'Ishango* (Les Éditions Bob Morane Inc., 2019).

Jean-Christophe Deledicq, organisateur du festival « La nuit des maths » dans la région Centre-Val de Loire… et au-delà, n'hésite pas à se mettre en scène dans des spectacles étonnants, comme « Le vin et les maths : vers l'ivresse de l'infini ».

Aussi à l'aise avec le grand public qu'avec les plus jeunes, Laure Cornu, médiatrice en mathématiques, est la nouvelle pépite du Palais de la découverte. Son imagination, sa bonne humeur et ses prestations font mouche à tous les coups !

Mickaël Launay (*le Grand Roman des maths*, Flammarion, 2016) enchante plus de quatre cent cinquante mille abonnés avec sa chaîne Youtube, « Micmaths, bric-à-brac mathématique et ludique » (**https://www.youtube.com/micmaths**).

Ils vous font vivre les maths autrement !

Reportage photographique : É. Thomas

Benoît Rosemont (*Memento de la mémoire*, Fantaisium, 2015) est mnémotechnicien. Ses spectacles de magie et de mentalisme « *J'ai oublié un truc… mais ça va revenir !* » et « *Mnémosys, une vraie mémoire de fou* » allient mathématiques et humour déjanté.

La compagnie L'Île Logique (Cédric Aubouy, auteur de *À l'endroit de l'inversion*, 2017, et David Latini), accompagnés d'Anissa Benchelah. Ou comment l'art du clown peut aider à appréhender la logique et les mathématiques.

Le « mathémusicien » Moreno Andreatta, ici avec son comparse Laurent Mandeix, fait littéralement chanter les maths. Ses concerts-conférences « Math'n'Pop » sont toujours l'occasion de découvrir des liens inattendus entre musique et mathématiques !

Robin Jamet (*Vous avez dit maths ?*, Dunod, 2019) est médiateur scientifique au Palais de la découverte. Il n'hésite pas à bricoler des dispositifs mécaniques ou des appareillages pour illustrer un point de mathématiques. Il collabore à *Science et Vie Junior*.

Avec son acolyte le mathématicien Florent Hivert, le jongleur Vincent de Lavenère, de la compagnie Chant de balles, explore la modélisation de la jonglerie musicale lors d'interventions spectaculaires.

François Perrin, de la compagnie Terraquée, et la comédienne Caroline Benassy. Par le jeu théâtral (ateliers « Mathéâtre », spectacles « Pi, le nombre à deux lettres », « Il est rond mon ballon »), la troupe apporte un nouveau regard sur les mathématiques.

Complément d'enquête

Les mathématiques de l'apprentissage

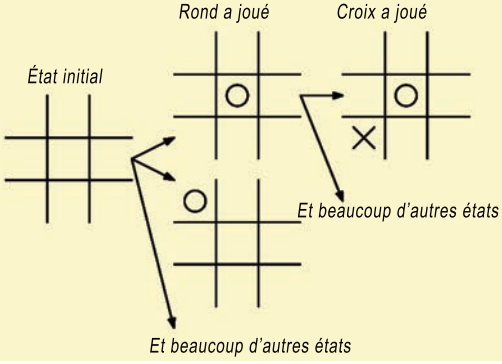
Quentin Labernia

Chercheur en apprentissage automatique
et intelligence artificielle à Corpy&Co

Les années 2000 ont apporté leur lot de nouveautés, et la plus en vogue en ce moment est certainement les réseaux de neurones profonds, dont il a été question dans plusieurs articles précédents. Mais l'IA, ce ne sont pas que des réseaux de neurones, et si vous en avez assez de n'entendre parler que d'eux, suivez-nous ! Découvrons quatre exemples de techniques d'intelligence artificielle (IA) qui ne requièrent pas de réseaux de neurones : rechercher une aiguille dans une botte de foin, gérer le raisonnement avec Prolog, apprendre et interpréter avec des arbres de décision, et laisser un agent renforcer son comportement par lui-même.

Morpion, recherche dans les grands espaces et heuristiques

Savez-vous comment gagner à coup sûr au morpion ? En prenant un papier et un crayon, et un peu de temps, il est possible d'énumérer toutes les combinaisons possiblesqui partent d'une grille vide. Prenez ensuite le chemin qui vous mène à la victoire ! Mais n'essayez pas avec le jeu de go, vous n'aurez pas assez d'atomes dans l'univers pour écrire l'ensemble des combinaisons… Il convient donc de faire preuve de stratégie pour gagner au go ; ces stratégies s'appellent des heuristiques.



L'état initial (à gauche) correspond à une grille vide. Lorsque c'est le tour du joueur Rond, on peut dessiner l'ensemble des états possibles qui partent de l'état initial après avoir dessiné un rond. Chaque grille est un nœud de notre graphe G, et les flèches sont les arcs qui relient les états entre eux.

© Q.L.

Considérons un jeu à deux opposants, comme le morpion (de taille 3×3). Les règles de ce jeu lui confèrent une propriété déterministe : avant de jouer, il est possible de dessiner un graphe orienté $G = (V, E)$ dont les nœuds V représentent l'ensemble des états possibles du plateau, et les arcs, une action entreprise par l'un ou l'autre des joueurs : $E = \{\text{croix, cercle}\} \times \{\text{position1, position2… position9}\}$. Un élément v de V est un vecteur à neuf entrées représentant la grille. Lorsque le jeu débute, le système est dans l'état V_0 où la grille est totalement vide. De cet état, il existe exactement neuf arcs sortants (en considérant avoir déjà choisi le premier joueur). Comme il n'est pas sportif d'effectuer l'action de l'opposant lorsqu'il a le dos tourné, nous allons utiliser le graphe G pour obtenir une description précise du « meilleur coup à jouer » à chaque tour et ainsi gagner la partie. L'objectif est de trouver un nœud *terminal* (qui ne possède aucun arc sortant) gagnant. Le problème du morpion se résume donc à un problème d'exploration de graphe ! Le graphe du morpion est relativement « petit » (il possède exactement cinq mille quatre cent soixante-dix-huit nœuds). L'histoire est tout autre pour un jeu comme le go : pour évaluer de manière exhaustive les états possibles avec quatre coups, il faut explorer plus de 10^{11} états…

L'exploration de vastes espaces mathématiques (« rechercher une aiguille dans une botte de foin ») est à la base de nombreuses techniques d'intelligence artificielle. Les réseaux de neurones explorent un espace vectoriel dont le volume croît exponentiellement avec le nombre de couches considérées. Cet espace est continu, à la différence de notre graphe G, qui lui possède une structure discrète. Il est naturellement possible de se déplacer dans un espace continu et différentiable *presque partout* (c'est-à-dire sauf en un nombre fini de points, comme c'est le cas pour les réseaux de neurones) en utilisant la *rétropropagation du gradient*.

Revenons à notre graphe G. La plus simple des techniques est l'énumération exhaustive de l'ensemble des éléments v de V, ce qui n'est réalisable en pratique que lorsque le cardinal de V est « petit ». On utilise alors des *heuristiques* : ce sont des techniques qui réduisent l'espérance du nombre d'éléments rencontrés avant le point d'intérêt (un nœud terminal gagnant). L'*algorithme minmax* ou la *recherche arborescente Monte-Carlo* travaillent dans des espaces discrets et sont de célèbres exemples d'heuristiques. Le premier explore des branches de manière optimisée selon les différentes actions possibles du joueur adverse, alors que la seconde explore aléatoirement des possibilités, joue de manière aléatoire jusqu'à une position terminale et utilise cet état final pour décrire la qualité de l'action prise et ainsi éviter les branches dont l'espérance de gain est faible.

Un moteur d'inférence : l'intelligence en des termes logiques

Explorer, c'est bien. Raisonner, c'est bien aussi. Grâce à un langage de programmation comme Prolog (abréviation de « *programmation logique* ») mis au point en 1972 par Alain Colmerauer et Philippe Roussel, il est possible d'exprimer des raisonnements logiques de manière très naturelle, en les décomposant en faits et en règles : les *faits* sont des concepts de base, figés, alors que les *règles* permettent de mettre en relation les faits. Les nombres entiers (1, 2…) et le signe d'addition, ce sont des faits. Une règle bien connue est la suivante : $1+1=2$. Voilà le cœur de Prolog : comment encoder cette logique.

Considérons la création d'un plan de table à trois personnes (a, b, c) pour illustrer Prolog (les trois sièges forment une ligne droite). Parlons des faits. Dans le cas de notre plan de table, il s'avère que la personne c veut être à la gauche de b : c'est une définition que l'on impose *a priori*. Déclarons donc le fait comme suit : ordre(c, b). Le fait est une fonction *n*-aire (elle peut prendre un nombre quelconque d'arguments) : elle a donc un nom, « ordre », et dans le cas présent deux arguments ($n=2$), c et b. Petit point de rigueur : en Prolog, toutes les lignes se terminent par un point final.

Passons aux règles. Voilà déjà quelques règles de syntaxe : les majuscules en argument correspondent à des variables, alors que les faits sont constitués de minuscules. La virgule dénote le ET logique alors que le point-virgule correspond au OU. Une règle se construit comme suit :

```
nomregle(arguments) :– regle.
```

Définissons la règle R1 :

```
ordre (A, B, C) :– ordre(A, B); ordre(B, C).
```

Ajoutons une autre règle, R2 :

```
plandetable(A, B, C) :– permutation([a, b, c], [A, B, C]), ordre(A, B, C).
```

Nous venons de définir deux règles R1 et R2. *Inférer* une règle, c'est trouver la valeur des variables qui se trouvent dans les arguments à gauche de la règle. Le mécanisme est simple : le moteur d'inférence de Prolog va chercher les faits qui valident les règles. Si l'on demande à Prolog la requête suivante : « plandetable(A, B, C) », il nous renvoie le résultat : A = *a*, B = *c* et C = *b*. La règle R1 peut alors se lire :

« Chercher les valeurs de A, B et C telles que ordre(A, B) existe
ou que ordre(B, C) existe. »

Quant à la règle R2, elle peut s'énoncer :

« Chercher les valeurs de A, B et C telles que nous avons une permutation de [A, B, C]
qui respecte également la règle ordre R1. »

Il est aussi possible de demander à Prolog plandetable(b, B, C), c'est-à-dire de fixer *b* au bout à gauche de notre table. Dans ce cas, Prolog nous retourne B = *b* et C = *a*, ce qui satisfait effectivement notre règle. À l'aide d'un fait et de deux règles, nous avons résolu notre problème de plan de table ! En utilisant une base conséquente de faits et de règles, la logique encodée peut devenir très complexe.

L'intelligence en des termes purement logiques consiste à stocker des faits et établir des règles pour combiner ces faits ensemble. Prolog est un langage naturel pour exprimer cette logique. Pour les fous de programmation, il existe une multitude de bibliothèques qui permettent d'utiliser la logique Prolog intégrée dans d'autres langages de programmation (Pylog pour Python, ...).

Interprétabilité, arbres, forêts aléatoires et industrie musicale

Les réseaux de neurones sont des êtres compliqués : il est terriblement difficile de comprendre pourquoi ils ont pris telle ou telle décision. C'est là un de leurs grands défauts. Les arbres de décisions, sont, eux, beaucoup plus terre à terre, puisqu'un humain peut déchiffrer sans souci chaque étape qui a mené à leur décision en suivant un *diagramme de décision*. On peut également combiner ces arbres de décisions en organisant un vote, telles les présidentielles (mais entre sorties d'arbres de décisions) : il s'agit de *forêts d'arbres décisionnels*. Elles offrent un avantage statistique indéniable. Mais quel est le rapport avec l'industrie musicale ?

Les programmes logiques créés avec le langage Prolog sont de remarquables exemples d'algorithmes de décision dits « interprétables ». Des secteurs comme les assurances, les banques ou le domaine médical ont besoin d'interprétabilité : étant donné le vecteur de paramètres d'entrée *x* d'un algorithme A, on dit que A est *interprétable* lorsqu'il est décomposable en une suite de décisions, que l'on peut comprendre simplement. Les réseaux de neurones sont difficilement interprétables : à chaque couche, l'espace est déformé sur la base d'une combinaison linéaire des descripteurs en entrée. Il n'est pas rare qu'à cette étape, le réseau de neurones mélange des choux et des carottes, sans se soucier de leur sémantique. Il est ensuite très difficile d'apporter une quelconque interprétation au résultat obtenu !

Les arbres de décision sont un modèle d'apprentissage automatique supervisé qui, au contraire, font preuve d'une interprétabilité naturelle. Un *arbre de décision* D = (V, E) est un graphe acyclique orienté, où l'on associe à chaque nœud *v* de l'ensemble V une fonction de décision. Une fonction de décision prend en entrée une donnée $x = (x_1, x_2 \dots x_N)$, où chaque élément est un *descripteur* (variable contenant une information), et décide ensuite du nœud suivant où aller *via* un arc *e* de l'ensemble E. Typiquement, les inéquations de la forme $f(x) > t$ sont souvent utilisées dans les arbres de décision *binaires*

(pour lesquels il existe deux arcs sortants de chaque nœud non terminal). La valeur *t* est également mise à jour au cours de l'apprentissage. Chaque donnée d'entrée *x* va alors cheminer dans le graphe D jusqu'à rencontrer un nœud terminal : la valeur de ce nœud est la réponse souhaitée ! Un exemple d'algorithme de classification supervisé très fameux est C4.5 (Ross Quilan, 1993), autrement appelé J48 et très utilisé dans l'industrie du disque.

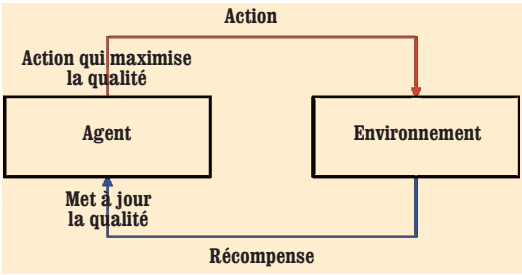
Comme l'union fait la force, des chercheurs ont eu l'idée de considérer un ensemble d'arbres de décision, et donc de créer *une forêt d'arbres décisionnels*. Chaque arbre de décision est entraîné sur un sous-ensemble des données (X, Y) qui sont tirées aléatoirement avec remise, et un sous-ensemble de descripteurs, tirés aléatoirement sans remise. Lorsqu'une donnée *x* arrive en entrée de la forêt, elle est évaluée par l'ensemble des arbres de décision selon les modalités de chaque arbre. Le nombre de réponses y_i est égal au nombre d'arbres. Un vote est organisé et la réponse majoritaire est retournée. Cette technique est appelée *apprentissage ensembliste*. Elle possède un effet de masse qui évite souvent les réponses farfelues grâce aux tirages aléatoires qui génèrent chaque arbre.

L'apprentissage par renforcement pour les agents intelligents

Les arbres ne vous plaisent pas ? Imaginez-vous alors sur une île déserte… Vous ne vous étiez pas préparés à une telle aventure et il est maintenant nécessaire de survivre. Vous êtes un *agent* : un système autonome qui évolue dans son environnement. Après avoir mangé des crabes délicieux et des serpents venimeux, vous vous rendez compte que vous n'êtes pas malade après avoir mangé les crabes, *a contrario* des serpents. Vous apprenez de vos erreurs et essayez de maximiser la qualité de votre état de santé : c'est tout le principe de l'*apprentissage par renforcement* ! Cette méthode est très utilisée pour la mise au point de systèmes robotiques ou d'agents « intelligents » : un robot devant se déplacer de manière autonome dans un labyrinthe, un capteur pivotable qui suit le soleil pour maximiser la quantité d'énergie reçue…

L'apprentissage par renforcement est une technique supervisée où l'agent « apprend » selon les informations provenant de l'environnement. En résumé, l'agent est autonome et la supervision intervient par la prise en compte des retours de l'environnement et de leur corrélation avec les actions entreprises. Cet environnement peut exister dans la vie réelle ou bien être simulé.

Explorons le *Q-learning*. Dans la même veine que les algorithmes génétiques ou les réseaux de neurones, il existe une forte inspiration provenant de la biologie. En l'occurrence, le principe du *Q-learning* repose sur la notion de *qualité* Q. Cette qualité est définie à chaque instant pour l'ensemble des actions A réalisables par l'agent dans des conditions S. Chaque action octroie une certaine récompense R à l'agent. La qualité mesure la récompense espérée



L'agent évolue dans son environnement. Il choisit de réaliser une action qui maximise la qualité (que représente son acquis) qui va avoir un impact sur l'environnement. Il reçoit en retour une récompense, qu'il utilise pour mettre à jour la qualité de l'action choisie (était-ce une action adaptée à la situation ?).

© *Q.L.*

en effectuant une action spécifique. La fonction Q prend en argument l'état du système et une potentielle action, puis calcule un nombre réel qui mesure la qualité. L'agent a alors intérêt à effectuer l'action qui maximise la qualité (par exemple, manger un crabe délicieux plutôt qu'un serpent venimeux).

Prenons l'exemple d'un robot à quatre pattes, l'agent, qui se trouve sur une table, l'environnement. On définit les actions possibles pour l'agent : $A = \{\text{bouger patte 1}\ldots\ \text{bouger patte 4}\}$. Les états du système sont $S = \{\text{avance, recule, sur-place, à gauche, à droite}\}$. La fonction de récompense est fixée arbitrairement, constante, égale à une valeur Q_0.

Au temps t_0, on choisit une action parmi celles qui maximisent la récompense, ici, la vitesse du robot (ou encore, votre état de santé). Au temps t, après avoir choisi une action $a(t)$ dans A, l'agent observe l'état du système $s(t)$ dans S et met à jour la qualité en se basant sur la formule suivante :

$$Q_{new}(s(t), a(t)) = Q(s(t), a(t)) + \text{Facteur}_{\text{apprentissage}} \times (R + \text{Estimation}_{\text{qualité}}(t+1) - Q(s(t), a(t))).$$

La qualité Q est mise à jour avec la récompense effectivement reçue et une estimation de la future valeur de qualité. En pratique, la fonction Q est une simple table d'association dont les lignes sont les états possibles S du système, et les colonnes sont les différentes actions A que l'agent peut effectuer. L'algorithme est indépendant du modèle qui décrit la récompense Q. À la place d'une table d'association, on pourrait par exemple utiliser… un réseau de neurones ! Dans ce cas, la fonction, définie sur un domaine continu, peut généraliser à des paires (a, s) non précédemment rencontrées.

Ces quelques exemples illustrent que le paysage de l'IA ne devrait plus être uniquement constitué de réseaux de neurones. De nombreux autres modèles d'apprentissage supervisé sont utilisés, comme les machines à vecteurs de support. Des méthodes d'apprentissage non supervisées permettent également de compresser l'information, de découvrir des tendances ou des groupes de données similaires, d'encoder la sémantique des mots dans des phrases… En multipliant les approches et les outils, on augmente le champ des possibles !

Q. L.

Cette brochure a été réalisée par le
Comité International des Jeux Mathématiques
sous la direction de
Marie José Pestel,
et d'Édouard Thomas qui en a assuré la relecture.

Imprimée grâce à
Animath,
Inria et le Crédit Mutuel Enseignant

Elle réunit les signatures de

Jean-Luc Berthier
Jean-Paul Delahaye
Jean-Pierre Demailly
Étienne Pardoux
Lionel Roques
Éric Blayo, Laurent Debreu et Christine Kazantsev,
Jocelyne Erhel
Hervé Lehning
Clément Cartier
Gabriel Peyré
Enka Blanchard et Levi Gabasova
Nicolas Nguyen
Jean-Marie De Koninck
Emmanuël Houdart
Quentin Labernia

Les Maths, Oui ça sert ! Express, née en plein confinement, vous offre plusieurs pistes de réflexion sur la gestion de notre avenir et des exemples concrets de l'importance des mathématiques dans des domaines qui touchent de près notre vie sur Terre.

Nous remercions particulièrement nos auteurs de nous avoir fourni dans ces circonstances exceptionnelles, et dans des délais records, les articles passionnants de cette édition dont nous sommes particulièrement fiers.

Merci à Loïc Michel pour sa vigilance tout au long de cette aventure
Maquette de couverture : Elsa Godet – www.sciencegraphique.com
Illustration chapitres : « *Impromtu* » – Métamorphose des accords – © Valentin Afanassiev : www.afanasieff.ru
Illustration fin de chapitres : © apmm
Réalisation : Patrick Arrivetz
Impression : Presses de CIA GRAPHIC – 03 86 90 96 10

Mai 2020

NOUVEAU !

l'abonnement numérique

INTÉGRAL MATHS

pour seulement

3,99 €

par mois

28

nos de Tangente

sur

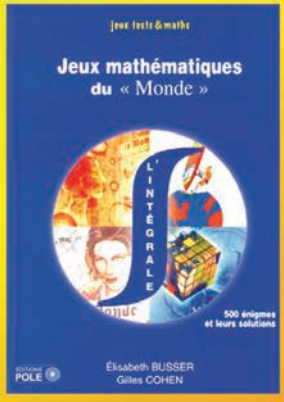
tangente-mag.com

17 hors séries

Et aussi...

sur

affairedelogique.com



400 problèmes du Monde et leurs solutions

www.infinimath.com/librairie

LA BANQUE DU MONDE DE L'ÉDUCATION

Crédit photos : Gettyimages – Fotolia.

MA BANQUE EST DIFFÉRENTE, CEUX QUI LA GÈRENT SONT COMME MOI.

UNE BANQUE CRÉÉE PAR DES COLLÈGUES, ÇA CHANGE TOUT.

Le Crédit Mutuel Enseignant est la bancassurance dédiée au monde de l'éducation dans sa définition la plus large. A ce titre, elle est ouverte à tous les personnels de l'éducation nationale, de la recherche et de la culture. Elle vous propose une offre adaptée à vos besoins, à chaque étape de votre vie.

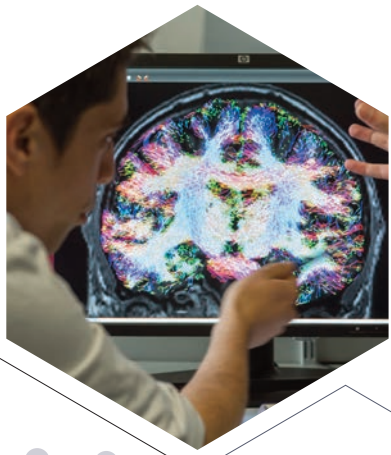
Crédit Mutuel
Enseignant

Crédit Mutuel Enseignant Île-de-France

Antony • Aubergenville • Bobigny • Cergy • Créteil
Evry-Courcouronnes • Melun • Paris Quartier Latin • Paris Haussmann
Paris La Défense • Serris • Versailles

Caisse Fédérale de Crédit Mutuel et Caisses affiliées, société coopérative à forme de société anonyme au capital de 5 458 531 008 euros, 34 rue du Wacken, 67913 Strasbourg Cedex 9, RCS Strasbourg B 588 505 354 - N° ORIAS : 07 003 758. Banques régies par les articles L.511-1 et suivants du code monétaire et financier.

Centre de recherche Inria de Paris en bref



600
personnes dont 520 scientifiques
36 équipes de recherche

- **22** bourses ERC depuis 2009
- **1** à **2** nouvelles start-up par an

En partenariat avec :

le CNRS, l'EHESS, l'ENPC, l'ENS, Inserm, Mines ParisTech, Paris Dauphine, Sorbonne Université, Université de Paris, Université Paris-Est Marne-la-Vallée.

Membre de :

PSL RESEARCH UNIVERSITY PARIS

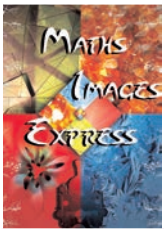
ALLIANCE SORBONNE UNIVERSITÉ

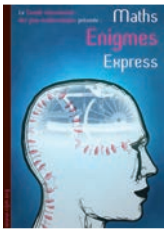
Inria.fr/Centre/Paris

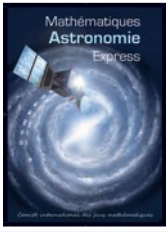
@inria_paris

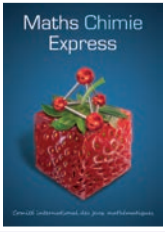
Maths Express
Une collection CIJM

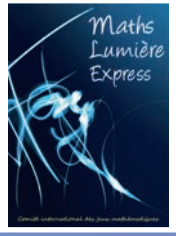
cijm.org/accueil/productions-cijm/90-maths-express

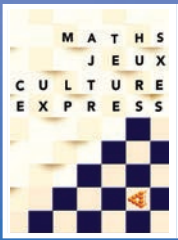












CIJM
Institut Henri Poincaré
11 rue Pierre et Marie Curie
75231 Paris Cedex 05
www.cijm.org