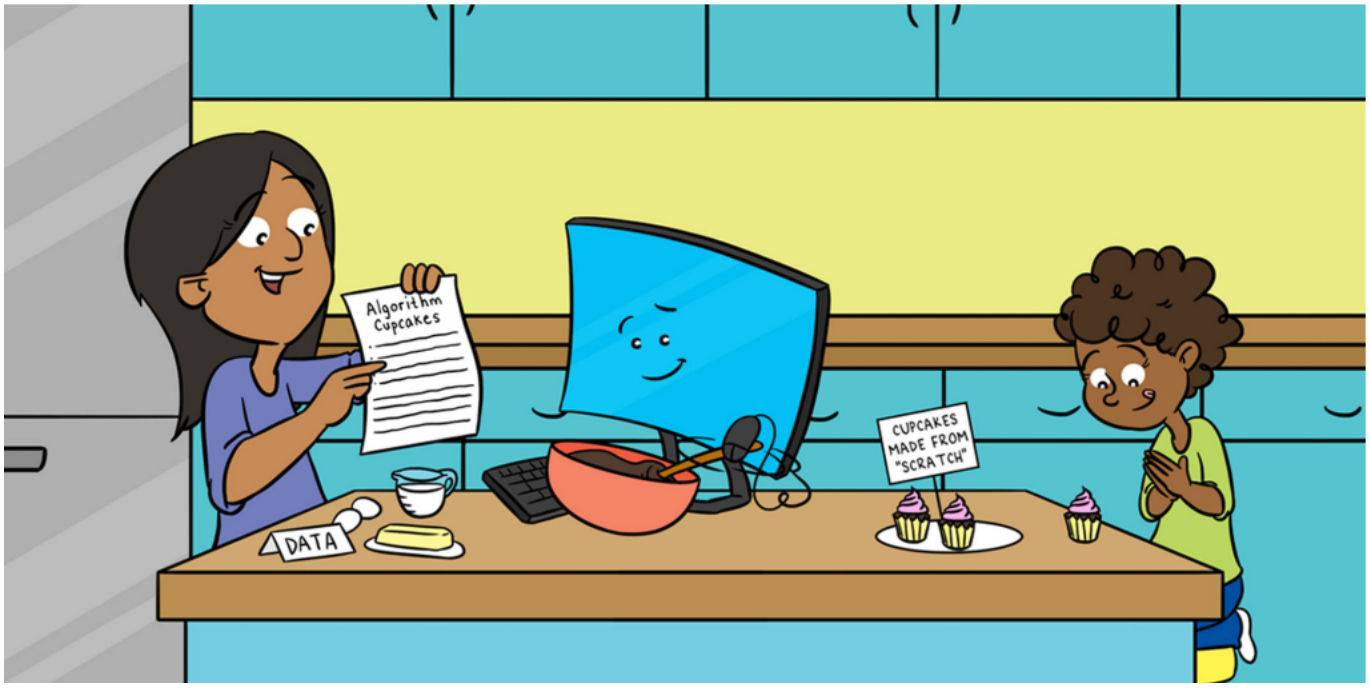
Article original : doi 10.3389/frym.2021.632923



LA MEILLEURE RECETTE POUR COMPRENDRE LES MÉGADONNÉES : LA SCIENCE DES DONNÉES !

*Dr. Andrea Rau**

Université Paris-Saclay, INRAE, AgroParisTech, GABI, Jouy-en-Josas, France

Les mégadonnées (données massives ; en anglais, les « *Big Data* ») sont des informations collectées en très grandes quantités. Les mégadonnées changent notre façon de concevoir et de répondre aux questions dans des domaines nombreux et variés, comme les prévisions météorologiques ou la biologie. Pour nous aider à stocker, traiter, analyser, et comprendre une telle masse d'informations, nous avons besoin d'ordinateurs. La science des données est un vaste champ multidisciplinaire qui s'appuie sur des approches issues de plusieurs domaines, comme la statistique, les mathématiques et l'informatique, pour donner du sens aux données brutes. Les scientifiques des données écrivent des algorithmes, c'est-à-dire des instructions étape par étape, pour apprendre aux ordinateurs à identifier une logique au sein de ces données. Pour aider les ordinateurs à comprendre ces instructions, les algorithmes doivent être traduits en un langage de programmation à partir de la question initiale posée par un·e scientifique des données. Les résultats doivent ensuite être retraduits afin que les humains puissent les comprendre. Cela signifie que les scientifiques des données sont simultanément des détectives, des programmeurs, et des traducteurs !

LES DONNÉES SONT PARTOUT !

Les données sont un ensemble d'informations – nombres, mesures, mots ou descriptions – qui ont été rassemblées et stockées pour une

raison spécifique. Récemment, de nouveaux outils ont été développés pour faciliter la collecte de très grandes quantités de données. Lorsque les données sont disponibles en très grandes quantités, elles sont souvent appelées des mégadonnées. Ces mégadonnées changent notre façon de concevoir et de répondre à des questions issues de nombreux domaines, comme par exemple prévoir la météo, trouver des itinéraires afin de ne pas rester coincer dans un embouteillage ou suggérer une nouvelle série télévisée qui te plaira en fonction des émissions que tu as précédemment regardées.

MÉGADONNÉES. Ensembles de données extrêmement volumineux et complexes, difficiles à stocker, traiter, analyser et interpréter. Les scientifiques doivent souvent utiliser des outils et des méthodes spécialisés pour travailler avec les mégadonnées.

LES MÉGADONNÉES : UN DÉFI D'IMPORTANCE EN BIOLOGIE !

Les mégadonnées ont également contribué à faire avancer la recherche en biologie, une science qui s'intéresse à l'étude des êtres vivants comme les humains, les animaux, les plantes et les microbes. Des outils très spécialisés permettent aujourd'hui la collecte de grandes données biologiques dans les laboratoires de recherche, les hôpitaux, la nature, et même à domicile ! Par exemple, les appareils connectés tels que les montres intelligentes peuvent avoir des capteurs permettant à un médecin de surveiller en temps réel la qualité de ton sommeil. Les drones, en survolant les fermes et en prenant des photos des champs, peuvent donner une vue d'ensemble de l'état des cultures. De nouvelles techniques de laboratoire permettent désormais de lire facilement la totalité de l'information génétique d'une personne, longue d'un peu plus de 3 milliards de lettres* (pour te donner une idée de l'échelle, 3 milliards de secondes représentent environ 90 ans !). Avec toutes les informations disponibles dans les mégadonnées, c'est un grand défi de les stocker, les traiter, les analyser et les interpréter, et nous avons besoin d'ordinateurs pour nous aider.

* : L'ADN est composé de deux chaînes de lettres (nucléotides), chaque chaîne est composée d'un peu plus de 3 milliards de nucléotides.

MATHÉMATIQUES+STATISTIQUES+INFORMATIQUE +MÉGADONNÉES = LA SCIENCE DES DONNÉES

Les mégadonnées sont d'une telle ampleur qu'un nouveau domaine passionnant a dû être développé pour y faire face : la science des données. La science des données rassemble des outils de nombreux autres domaines, notamment la statistique, les mathématiques, et l'informatique, afin de donner du sens aux données complexes. Les scientifiques des données passent beaucoup de temps à mettre de l'ordre dans leurs données avant de poursuivre leur travail. Afin de répondre à une question spécifique, un·e scientifique des données doit trouver ou créer un jeu de données, ou bien un ensemble de jeux de données. Certains jeux de données sont accessibles au public et peuvent être retrouvés avec une simple recherche par mot-clé en utilisant un moteur de recherche comme la Recherche d'ensembles de données de Google[1]. D'autres jeux de données, comme ceux qui contiennent des informations médicales personnelles, ne sont

SCIENCE DES DONNÉES. Domaine multidisciplinaire qui combine des outils de plusieurs domaines (statistiques, mathématiques, informatique…) pour tirer profit des jeux de données complexes, y compris les mégadonnées.

JEU DE DONNÉES. Collection structurée d'informations (nombres, mesures, mots ou descriptions) rassemblées et stockées ensemble pour une raison spécifique.

[1]https://datasetsearch.research.google.com

accessibles qu'à un nombre restreint de personnes. De plus, un·e scientifique des données peut même être amené·e à collecter de nouvelles données pour répondre à sa question. Par exemple, si tu souhaites connaître la couleur préférée de tes camarades de classe, tu pourrais rédiger un sondage pour recueillir leurs réponses.

DES DONNÉES DÉSORDONNÉES, DES DONNÉES BIEN RANGÉES

Une grande partie du travail d'un·e scientifique des données consiste à réorganiser ses données pour qu'elles soient dans un format utilisable. Par exemple, imagine que tous tes LEGOs® sont éparpillés chez toi : un vrai fouillis ! Avant de trier les blocs pour commencer ta construction, il faut d'abord faire un peu de rangement et tous les rassembler dans la même pièce. La plupart des données réelles sont aussi très « désordonnées », ce qui signifie qu'elles peuvent contenir des erreurs, des fautes de frappe, ou même des valeurs manquantes. Par exemple, certaines réponses de ton enquête sur la couleur préférée de tes camarades de classe peuvent inclure « bleu », « Bleu », « BLUE », et « bleue ». Pourtant toutes ces différentes réponses correspondent à la même couleur ! Pour rendre ces données plus faciles à utiliser, tu auras donc besoin de les ranger en changeant toutes ces variations en une unique valeur, comme « bleu ».

LES ALGORITHMES, UNE RECETTE POUR LA SCIENCE DES DONNÉES

Une fois que tous tes LEGOs® sont rassemblés au même endroit, ton prochain objectif pourrait prendre plusieurs formes : par exemple, grouper tes blocs en ensembles, ou aider tes parents à deviner le prochain ensemble qui te plairait. Si tu n'as qu'un petit nombre de blocs LEGOs®, il serait peut-être facile de le faire à la main, mais pour un nombre massif de blocs, il faudra des outils spécifiques pour aller plus vite. De la même manière, l'apprentissage automatique est un outil qui permet de faire face aux mégadonnées. L'objectif est d'apprendre à un ordinateur comment explorer et extraire lui-même de l'information utile issue des données, sans être guidé par un humain. Pour ce faire, un·e scientifique des données doit fournir à l'ordinateur un algorithme (Figure 1), c'est-à-dire un ensemble d'instructions détaillées étape par étape. Ces instructions doivent être écrites pour que l'ordinateur puisse les comprendre : c'est ce que l'on appelle le codage. Un algorithme, c'est un peu comme une recette pour faire un gâteau. La recette commence par un ensemble d'ingrédients (tes données) et t'indique exactement comment les mélanger pour obtenir la pâte et la cuire au four (ton algorithme) afin d'obtenir un dessert savoureux (tes résultats). Il y a cependant une grande différence entre une recette et un algorithme : les instructions d'un algorithme doivent être *extrêmement* précises pour que l'ordinateur sache *exactement* quoi faire. Dans une recette, au lieu de

APPRENTISSAGE AUTOMATIQUE. Utilisation d'algorithmes pour indiquer à un ordinateur comment apprendre automatiquement à partir des données et s'améliorer au fur et à mesure, sans l'aide d'un humain.

ALGORITHME. Ensemble d'instructions détaillées à suivre, étape par étape, par un ordinateur.

CODAGE. Utilisation d'un langage de programmation pour communiquer avec un ordinateur et lui fournir des instructions, appelées un algorithme.

dire « ajouter une pincée de sel à la pâte », cela reviendrait à dire « ajouter 1g de sel à la pâte et remuer 3 fois avec une cuillère en bois ».

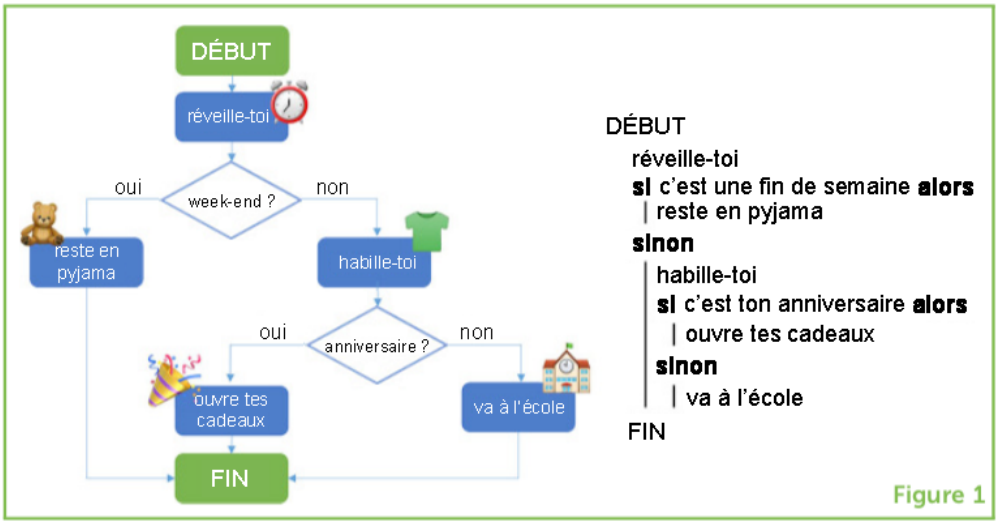


Figure 1. Un algorithme est un ensemble d'instructions à effectuer pas à pas pour un ordinateur. Un moyen utile de visualiser et de construire un algorithme consiste à dessiner un organigramme pour relier chaque étape à une autre. Dans les organigrammes, les rectangles peuvent représenter des actions et les losanges une décision. Le matin, tu pourrais utiliser un organigramme comme celui de gauche pour décider si tu peux rester en pyjama ou t'habiller, ouvrir des cadeaux d'anniversaire ou aller à l'école. Après avoir dessiné un organigramme, tu pourrais ensuite traduire les étapes de ton algorithme en une description plus détaillée, comme indiqué à droite.

PARLES-TU LA MÊME LANGUE QUE TON ORDINATEUR ?

Le codage est un moyen de traduire une question scientifique dans un langage compris par un ordinateur. Il existe de nombreuses langues parlées partout dans le monde (français, anglais, italien, allemand etc.) et de même, il existe de nombreux langages de programmation qui peuvent être utilisés pour écrire un algorithme (Figure 2).

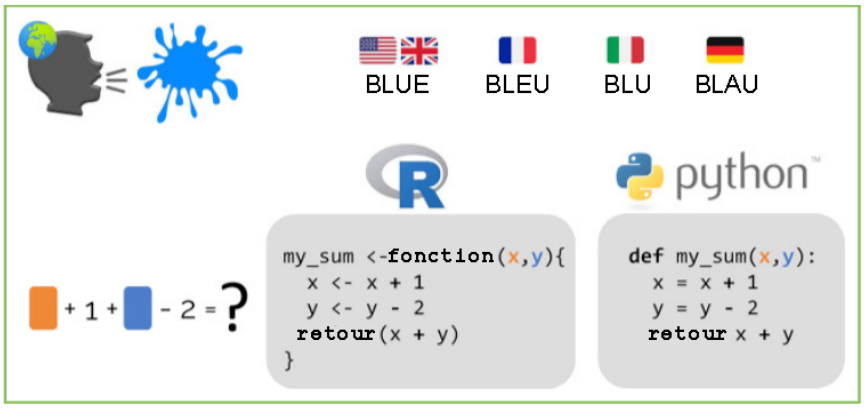


Figure 2. Les algorithmes peuvent être écrits en différents langages de programmation, tout comme les idées peuvent être exprimées en différentes langues. Disons que nous voulons écrire un algorithme qui prendrait deux nombres quelconques, ajouterait 1 au premier et soustrairait 2 au second, puis les additionnerait. Si nous commençons avec 2 et 4, nous voulons apprendre à l'ordinateur à nous fournir comme réponse (2 + 1) + (4 – 2) = 5. Notre algorithme, que nous avons appelé « my_sum », semble similaire dans les langages de programmation R et Python. Mais si tu regardes attentivement, tu verras quelques différences…

Tout comme une recette écrite en français ou en anglais qui communique la même idée mais de deux manières, les différents langages de programmation assemblent de façons diverses les instructions pour un ordinateur. De nouveaux langages de programmation sont inventés chaque année ! Il existe même un langage de programmation, appelé Scratch[2] [1], créé spécialement pour les enfants de 8 à 16 ans. Aujourd'hui, deux langages de programmation populaires sont souvent utilisés par les scientifiques des données pour écrire leurs algorithmes : R et Python. Ces deux langages sont *open-source*, ce qui signifie que les scientifiques des données qui écrivent leurs algorithmes avec ces langages, peuvent les partager avec tout le monde. Cela permet aux scientifiques des données de travailler facilement ensemble et de s'entraider pour améliorer les algorithmes de chacun !

[2] https://scratch.mit.edu

OPEN-SOURCE. Langage de programmation développé et pris en charge par la communauté. Les codes et progiciels *open-source* sont généralement gratuits : ils peuvent être utilisés ou partagés par tout un chacun.

DES RECETTES INFORMATIQUES À UN LIVRE DE CUISINE EN SCIENCE DES DONNÉES

Pour répondre à la question posée, un·e scientifique des données doit parfois écrire plusieurs algorithmes et les mettre ensemble. Tout comme un chef cuisinier pourrait vouloir rassembler plusieurs recettes dans un livre de cuisine, un·e scientifique des données peut créer ou utiliser des ensembles d'algorithmes, appelés progiciels. Lorsque ces progiciels sont écrits dans un langage de programmation *open-source*, comme R ou Python, cela peut favoriser un travail transparent et reproductible. La reproductibilité implique que d'autres personnes peuvent facilement répéter et réutiliser le travail d'un·e scientifique des données. La reproductibilité permet d'avoir confiance en la justesse d'un algorithme. Cela aide chacun à travailler plus efficacement et à partager facilement ses résultats. De la même manière, tu peux donner ton livre de cuisine préféré à un ami pour qu'il puisse réaliser son propre gâteau !

PROGICIEL. Collection organisée d'algorithmes liées qui fonctionnent ensemble pour une tâche particulière ou qui ont une fonction similaire.

CONCLUSIONS

Les mégadonnées ne cessent de grossir, que ce soit en biologie, en finance, ou en commerce, et elles continueront donc d'avoir un impact énorme sur nos vies. Cependant, les conséquences de la collecte de mégadonnées sur la vie privée représentent également une préoccupation croissante.

Lorsque tu t'inscris à un service en ligne ou à une application gratuite (comme les réseaux sociaux, les mails, la diffusion vidéo, ou les services de géolocalisation), tu dois généralement accepter en échange de laisser une entreprise privée collecter des données te concernant. Ces données peuvent inclure les mots-clés que tu recherches, les sites internet que tu parcours, les vidéos que tu aimes, ou les endroits de ton quartier que tu as visités. Les entreprises peuvent se servir de ces données pour créer des publicités ciblées spécifiquement pour toi,

dans le but d'augmenter leurs ventes ! Tu peux prendre des mesures pour savoir quels types de données sont collectés à ton sujet, par exemple en regardant dans les paramètres des applications. Cela peut t'aider à éviter la collecte de certains types de données, et également à identifier les applications et les services auxquels tu fais confiance ainsi que ceux que tu pourrais éventuellement désinstaller.

Dans les années à venir, nous aurons besoin de nombreux scientifiques des données capables de fournir un sens aux mégadonnées grâce à des méthodes d'apprentissage automatique. Il sera particulièrement important que des personnes de tous horizons contribuent à ce que chacun puisse bénéficier de manière égale aux résultats de ces analyses. La science des données représente un métier passionnant : nous sommes des détectives de données, des concepteurs graphiques, des programmeurs informatiques et des traducteurs, le tout réunis en un seul métier !

RÉFÉRENCE

[1] Maloney, J., Resnick, M., Rusk, N., Silverman, B., and Eastmond, E. (2010). The Scratch Programming Language and Environment. *ACM Transactions on Computing Education*. 10 (4): 1–15. doi:10.1145/1868358.1868363.

VERSION FRANÇAISE

Cet article d'accès libre est une traduction avec modifications d'un article publié par Frontiers for Young Minds (doi : 10.3389/frym.2021.632923 ; Rau A (2021) Cooking Up Knowledge From Big Data Using Data Science. Front. Young Minds 9:632923).

TRADUCTION : Andrea Rau, INRAE, Jouy-en-Josas, France

ÉDITION : Catherine Braun-Breton, Association Jeunes Francophones et la Science.

MENTOR SCIENTIFIQUE : Océane Paris, Association Jeunes Francophones et la Science.

JEUNES ÉDITEURS :

MATTHYS, AYLEEN, LOLA, KHELAN, SAMY, JULIEN, MORGAN, LÉA, ZOÉ, PAULINE, RAYAN, NOÉ, ALEXANDRE, 14-15 ANS

Ils sont élèves au collège d'Arbaud à Salon de Provence dans le sud de la France ; ils sont vivants, sportifs, imaginatifs et se sont impliqués avec enthousiasme et sérieux dans leur mission de jeunes éditeurs. Ils ont trouvé intéressant que les scientifiques doivent ranger leurs données avant de pouvoir les utiliser. Cet article leur a montré que la science ce n'est pas toujours que des expériences, mais aussi de l'organisation et du codage, que c'est un métier qui mélange plein de trucs, et ça donne envie d'en savoir plus.

REMERCIEMENTS : Merci à Benjamin Vierne pour son accueil et son implication dans l'édition de cet article par ses élèves.

ARTICLE ORIGINAL (VERSION ANGLAISE)

SOUMIS le 24 novembre 2020 ; ACCEPTÉ le 19 mars 2021 ;
PUBLIÉ EN LIGNE le 15 avril 2021.

ÉDITEUR : Norma Ortiz-Robinson, Grand Valley State University, United States

CITATION : Rau A (2021) Cooking Up Knowledge From Big Data Using Data Science. Front. Young Minds 9:632923. doi: 10.3389/frym.2021.632923

DÉCLARATION DE CONFLIT D'INTÉRÊT.

Les auteurs déclarent que les travaux de recherche ont été menés en l'absence de toute relation commerciale ou financière pouvant être interprétée comme un conflit d'intérêt potentiel.

DROITS D'AUTEURS

JEUNES EXAMINATEURS

JASMINE, 11 ANS

Je m'appelle Jasmine. J'aime les jeux stratégiques coopératifs. J'aime aussi lire les romans d'Agatha Christie et d'autres romans policiers classiques. Je m'amuse beaucoup en skiant, nageant, pratiquant le kayak et les arts martiaux. Je suis ceinture noire de karaté, ma plus belle réussite car cela m'a demandé beaucoup d'efforts et que j'ai mis 6 ans pour y arriver.

AUTEUR

ANDREA RAU

Je suis scientifique des données, biostatisticienne, et chercheuse à l'Institut National de Recherche pour l'Agriculture, l'Alimentation, et l'Environnement (INRAE) à Jouy-en-Josas en France. Je développe des modèles statistiques et fais du codage pour aider des biologistes à donner du sens à leurs données génomiques complexes. En plus d'écrire des algorithmes dans le langage de programmation R, je parle anglais et français au travail. Pendant mon temps libre, j'adore cuisiner

de nouvelles recettes et jouer avec ma fille Elise et mon chien Bella.
*andrea.rau@inrae.fr