https://doi.org/10.14745/ccdr.v41i09a03f

Commentaire

Les données massives font évoluer la lutte contre les maladies infectieuses

Links MG[1, 2*]

[1]Centre de recherches de Saskatoon, Agriculture et Agroalimentaire Canada, Saskatoon (Saskatchewan)
[2]Département des sciences informatiques, Université de la Saskatchewan, Saskatoon (Saskatchewan)

*Correspondance : Matthew.Links@usask.ca

Résumé

Les données massives étaient traditionnellement associées aux passionnés d'informatique et aux entreprises commerciales, mais elles sont devenues bien ancrées dans de nombreuses disciplines scientifiques, notamment la prévention et le contrôle des maladies infectieuses. L'utilisation des données massives a permis de déterminer les tendances des maladies et de retracer, voire de prédire, les origines des éclosions. Le volume des données massives ne diminue pas. Nous sommes confrontés à des difficultés qui nous poussent à perfectionner notre capacité d'analyse pour faire face à l'énorme « rapport signal/bruit » au moyen d'une puissance de calcul adéquate et d'équipes pluridisciplinaires capables de gérer des quantités de données en constante augmentation. Grâce aux données massives, il sera également possible de créer de futures applications de traitement *sur mesure* (ou personnalisé).

Introduction

Les données massives semblent être une évolution récente qui, pour beaucoup, est liée à des expressions telles que l'« informatique en nuage ». Or, le terme est né avant le changement de millénaire, dans les années 1990, lorsque John Mashey était informaticien en chef chez Silicon Graphics (SGI). À l'époque, SGI était à l'avant-garde de l'infographie et luttait pour faire face à des besoins de calcul en considérable expansion qui dépassaient les capacités du matériel disponible. Mashey a rédigé à la fin des années 1990 un exposé qui annonçait la collision imminente entre les données massives et la performance de calcul (1).

Aujourd'hui, l'expression « données massives » est couramment utilisée pour décrire des situations dans lesquelles les volumes de données sont caractérisés par des propriétés incluant, entre autres, la taille, la vitesse d'évolution dans le temps et la nature hétérogène des données elles-mêmes (2). Le terme « données massives » fait généralement référence à d'importants volumes de données qui peuvent être structurés (p. ex. bases de données relationnelles) ou non (p. ex. flux d'actualités Twitter) et qui sont analysés pour en extraire de l'information. Bien que, par le passé, les données massives soient nées dans les domaines des sciences informatiques, de la statistique et de l'économie (3), elles ont depuis été plus largement adoptées dans toutes les disciplines scientifiques.

On dit généralement qu'un problème implique des données massives lorsque le volume est si important qu'il devient impossible de convertir les données en connaissances. Dans le cas de la recherche sur les maladies infectieuses, les données massives ont des répercussions considérables. La capacité à effectuer un suivi des maladies en temps réel et à prévoir les éclosions s'est fondée sur des données non structurées pour faire évoluer la manière dont les maladies infectieuses sont prises en charge. Par exemple, grâce à l'utilisation de diverses sources d'actualité, le RMISP a été utilisé pour signaler de façon précoce de nouvelles infections (telles que le SRAS et le CoV-SRMO) qui a servi de base à une intervention de santé publique en réponse aux éclosions qui ont suivi (4).

Les données structurées sont particulièrement utiles lorsqu'il s'agit de compiler des renseignements provenant de multiples sources en fonction d'une structure prévisible des données. Dans le cas de la surveillance de santé publique, il est désormais possible de chercher des données structurées pouvant

servir de source d'information en remplacement d'une confirmation en laboratoire ou d'une déclaration de cas émise par un médecin. Muchaal et ses collègues ont démontré que l'utilisation de médicaments constitue une source possible d'information précoce de substitution (5).

Quelle est l'ampleur des données massives?

Il est difficile de se rendre compte d'à quel point le volume de données massives est réellement important. Dans un récent article, Stephens et ses collègues ont replacé les données massives de deux disciplines dans un contexte relatif (6). Bien que la palme du volume faramineux de données revienne actuellement aux études d'astronomie, Stephens et ses collègues ont indiqué que la génomique devrait atteindre le volume des données en astronomie d'ici à 2025. Pour avoir une idée de l'échelle, les données de la génomique en 2025 pourraient être équivalentes à 8 milliards de fois la capacité des plus gros iPhones disponibles aujourd'hui (128 Go de mémoire en 2015), soit l'équivalent des données contenues dans un iPhone par personne sur Terre, chaque année.

Les données massives et les origines d'une éclosion

L'application de la génomique aux maladies infectieuses peut contribuer à déterminer d'où proviennent réellement les éclosions de maladies infectieuses. Le travail d'enquête réalisé pour faire face à une éclosion de rougeole pendant les Jeux olympiques de 2010 en constitue un bon exemple (7). Au moyen du séquençage complet du génome, Gardy et ses collègues ont réussi à identifier exactement bon nombre des cas signalés (30 sur 82). Un constat important a été que l'éclosion était causée par plusieurs types de virus de la rougeole. Alors que le génotypage traditionnel du virus de la rougeole se concentre sur les séquences de la phosphoprotéine et de l'hémagglutinine, deux gènes spécifiques utilisés pour distinguer les isolats, Gardy et ses collègues ont démontré que d'autres gènes de la rougeole contenaient des variations supplémentaires qui pouvaient être utilisées pour définir plus précisément les lignées virales.

À venir

L'une des dernières applications des données massives consiste en ce que Jennifer Gardy a appelé le traitement *sur mesure* (ou personnalisé) (8). Par exemple, les approches examinant le génome complet appliquées à des milliers d'isolats peuvent permettre de repérer une variation génomique associée à des phénotypes antimicrobiens de *Mycobacterium tuberculosis* (9). À l'avenir, cette méthode pourrait être appliquée de manière généralisée pour déterminer les meilleurs traitements à adopter contre les bactéries présentant une résistance aux antimicrobiens.

Comment allons-nous interpréter toutes ces données?

En dépit de tout ce potentiel de progrès, des problèmes clés restent à surmonter pour l'utilisation des données massives dans toutes les disciplines. À mesure que les données deviennent de plus en plus massives, il devient plus difficile de les interpréter : soit l'intégration des données est tellement complexe qu'elle est difficile à suivre sans de sérieuses ressources de calcul, soit leur simple échelle dépasse l'entendement.

Avec l'utilisation du texte non structuré provenant des flux d'actualité analysés aux fins de connaissance et de surveillance des maladies, les données massives servent à trouver du sens dans un déluge de bruit. L'échelle de la quantité de texte analysée, entre les articles scientifiques publiés dans les revues, les nouvelles décrivant de nouveaux enjeux et les gazouillis de 140 caractères, est vraiment intimidante, si l'on considère qu'un demi-milliard de micromessages sont publiés sur Twitter par jour (6). Pour compliquer encore davantage les choses, la surveillance des maladies ne se limite pas simplement à compiler des flux d'actualité. Au contraire, la clé d'une stratégie significative de surveillance des maladies axée sur les données massives est la détermination du risque potentiel. Ainsi, les approches telles que le RMISP (4) sont cruciales pour que les pays et la communauté internationale soient prêts à intervenir en cas d'éclosion de maladie. Comme les données massives continuent à croître à une vitesse exponentielle, tenter d'améliorer notre capacité à les analyser devient une quête du Saint-Graal en constante évolution.

Trop souvent, les données massives sont acquises dans le cadre d'une étude pluridisciplinaire et remises à une seule personne (p. ex. étudiant diplômé, étudiant postdoctoral ou boursier) en espérant que celle-ci pourra, à elle seule, réussir à comprendre ce que tout cela signifie. Confier à une seule personne la responsabilité de relations de plus en plus complexes émanant de volumes écrasants de données ne constitue tout simplement pas une stratégie viable. C'est pourquoi la tendance est à l'élaboration d'approches pluridisciplinaires pour l'interprétation des données massives (4).

Conclusion

À moins d'imaginer un âge des ténèbres de l'informatique, il est difficile de croire que le volume des données massives va diminuer. Par conséquent, les disciplines scientifiques ont développé leur capacité à exploiter des volumes de données en constante augmentation. La prudence est de rigueur pour ne pas se faire submerger par les données massives. De fait, c'est l'évolution vers une analyse pluridisciplinaire des données massives qui permet aux équipes de suivre les tendances des maladies et même de prédire les éclosions avant qu'elles n'aient lieu. Les données massives sont en position de se déplacer progressivement de la santé publique au milieu clinique; le traitement *sur mesure* (ou personnalisé) des maladies infectieuses pourrait bientôt être à notre porte.

Financement

Le programme de recherche du D^{r} Links a été ou est actuellement financé par Agriculture et Agroalimentaire Canada (AAC), l'Initiative de recherche et développement en génomique (IRDG) du gouvernement du Canada, les Instituts de recherche en santé du Canada (IRSC), le Conseil national de recherches du Canada (CNRC), l'Agriculture Development Fund (ADF) de la Saskatchewan et le Programme canadien pour la sûreté et la sécurité (PCSS) du gouvernement du Canada (auparavant appelé Initiative de recherche et de technologie chimique, biologique, radiologique et nucléaire [IRTC]).

Conflit d'intérêts

Aucun.

Références

(1) Mashey J, ed. Big Data and the next wave of infrastress. USENIX Annual Technical Conference; 1998; Monterey, California, USA: Usenix.
(2) Asokan GV, Asokan V. Leveraging "big data" to enhance the effectiveness of "one health" in an era of health informatics. J Epidemiol Glob Health. 2015 Mar 5.
(3) Diebold FX. A personal perspective on the origin(s) and development of 'Big Data': The phenomenon, the term, and and the discipline. Department of Economics, University of Pennsylvania. PIER Working Paper No. 13-003. November 2012.
(4) Dion M, AbdelMalik P, Mawudeku A. Big Data and the Global Public Health Intelligence Network (GPHIN). Can Commun Dis Rep. 2015;41:211-216.
(5) Muchaal PK, Parker S, Meganath K, Landry L, Aramini J**.** Evaluation of a national pharmacy-based syndromic surveillance system Can Commun Dis Rep. 2015;41:204-211.
(6) Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or genomical? PLoS Biol. 2015 Jul;13(7):e1002195.
(7) Gardy JL, Naus M, Amlani A, Chung W, Kim H, Tan M, et al. Whole-genome sequencing of measles virus genotypes H1 and D8 during outbreaks of infection following the 2010 Olympic Winter Games reveals viral transmission routes. J Infect Dis. 2015 Jul 6.
(8) Gardy JL. Towards genomic prediction of drug resistance in tuberculosis. Lancet Infect Dis. 2015 Jun 23.
(9) Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: A retrospective cohort study. Lancet Infect Dis. 2015 Jun 23.