SURVIVRE À LA PROCHAINE AVALANCHE : DÉVELOPPEMENT DES COMPÉTENCES ET NOUVEL UNIVERS DES DONNÉES ADMINISTRATIVES

MICHAEL HAAN

Michael Haan, Ph.D., est professeur agrégé et titulaire d'une chaire de recherche du Canada en migration et relations ethniques à la Western University. Il s'intéresse à la migration, à l'immigration et au marché du travail canadien. Il est également directeur de la Collaborative Graduate Specialization in Migration and Ethnic Relations à Western et directeur universitaire du Centre de données de recherche de Statistique Canada à Western.

Dans un article de 1982, le philosophe canadien Ian Hacking a documenté une augmentation stupéfiante du volume de données collectées au début du XIXe siècle, qu'il qualifiait d'avalanche de chiffres imprimés. Il a montré qu'entre environ 1800 et 1830, la quantité de données statistiques augmentait de manière exponentielle, à une époque où le mot écrit n'augmentait que de manière linéaire. Sans aucun doute, il s'agissait d'une révolution scientifique de la même nature et de la même ampleur que ce que Thomas Kuhn avait décrit dans *La structure des révolutions scientifiques* (1962). Les façons de voir et de penser le monde furent transformées ; non seulement devint-il possible de collecter et de stocker des quantités de données, mais il devint également souhaitable, voire essentiel, que les États-nations modernes et en voie de modernisation comptent et mesurent là où ils étaient en mesure de le faire. Considérons que les États-Unis (1790), le Royaume-Uni (1801) et la France (1801) ont tous organisé leur premier recensement national à peu près à la même époque.

Contrairement à d'autres révolutions scientifiques, celle-ci n'a pas permis d'obtenir initialement des produits matériels. Certes, les résultats du recensement ont été publiés sous forme imprimée, mais les résultats ont été minimes par rapport aux données collectées. Rien de tel que la création d'une presse à imprimer ou d'une machine à vapeur. De cette époque, nous ne disposons que des piles de livres devenus poussiéreux, avec d'innombrables tableaux croisés. En effet, bon nombre des outils que nous utilisons aujourd'hui pour comprendre les statistiques n'existaient pas encore. Bien que des moyennes existaient, les médianes, les écarts types, les variances, les corrélations et les régressions ne l'étaient pas ; même le mot « statistiques » ne faisait pas partie de la langue anglaise jusqu'à la fin du XVIIIe siècle, lorsqu'il a été adapté du mot allemand « statistik ».

Le volume considérable de données collectées, ainsi que l'absence relative de techniques permettant de travailler avec ces données, ont créé les conditions

propices à l'épanouissement des pionniers de la statistique. De brillants chercheurs, comme les Anglais Francis Galton et Karl Pearson, ont mis au point une infrastructure statistique dont une grande partie reste intacte aujourd'hui.

Ce qui a contribué à faire avancer la révolution des données statistiques au XIXe siècle, c'est qu'elle nous a donné une nouvelle façon de penser le monde, une nouvelle façon d'appréhender le monde. L'époque de l'indéterminisme était révolue, alors que nous entrions dans un monde de probabilités et de vraisemblances. La statistique est devenue le moteur ontologique du jour.

Je crois que nous sommes actuellement à un moment similaire dans l'histoire. Nous avons maintenant plus de données que jamais auparavant, et les outils nécessaires pour analyser ces données sont soit totalement absents, soit pas utilisés couramment dans de nombreuses disciplines. Pour les spécialistes des sciences sociales, cela signifie que nous devrons soit mettre à jour nos trousses à outils, soit regarder les autres disciplines nous dépasser.

LA GRANDE RÉVOLUTION (SOUVENT ADMINISTRATIVE) DES DONNÉES

Comme je le dis plus haut, le rythme auquel les données sont à nouveau collectées a de nouveau créé une situation dans laquelle les techniques sont en retard par rapport au développement des données. Le nouveau régime de données est très prometteur, car il nous permet de poser et de répondre à des questions auparavant impossibles. Cela dit, ces nouvelles opportunités ne sont pas libres ; les données volumineuses et/ou administratives ne sont généralement pas collectées principalement à des fins de recherche ; elles sont donc souvent confuses, disparates et mal documentées.

Les fichiers de données volumineuses sont non seulement énormes, mais extrêmement complexes. La base de données sur l'immigration longitudinale (BDIM), par exemple, représente plus de 30 gigaoctets une fois les différentes années d'imposition liées aux dossiers de résidence temporaires et aux registres de débarquement des résidents permanents. La plupart des progiciels sont tout simplement incapables de travailler avec des fichiers aussi volumineux : impossible de tout simplement lancer SPSS[1] et commencer à exécuter des tableaux croisés sur la BDIM. Lorsque mes étudiants commencent à travailler avec ces fichiers, ils réalisent souvent assez rapidement qu'ils ne possèdent pas les compétences nécessaires pour travailler avec ces données, ce qui nécessite des mois de travail de correction. Ce n'est pas une réflexion négative sur eux, mais cela montre bien qu'ils ne reçoivent pas les compétences nécessaires dans leurs cours.

Les utilisateurs ont maintenant besoin de connaître les formes complexes de données manquantes, les liens probabilistes par opposition aux déterministes, souvent avec une documentation incomplète. Les étudiants, du moins dans mon université, ne reçoivent pas souvent cette formation et sont au contraire forcés d'apprendre ces choses au travail ou pas du tout. Dans les 100 prochaines années, il est difficile de savoir avec certitude ce que l'avenir nous

1 SPSS - *Statistical Package for the Social Sciences*, un progiciel de pointe utilisé pour l'analyse de l'information statistique.

apportera, mais voici quelques tendances actuelles qui, selon moi, vont perturber les pratiques habituelles pour beaucoup d'entre nous.

RÉPONDRE À DES QUESTIONS DE RECHERCHE SANS CADRE THÉORIQUE EXPLICITE

Les mégadonnées nous permettent souvent de poser des questions pour lesquelles nous n'avons aucune orientation théorique. Par exemple, les recommandations (étrangement précises) de films par Netflix, alors que la seule chose qu'ils savent à votre sujet est ce que vous avez regardé dans le passé. Auparavant, quand nous ne disposions que des techniques développées pendant l'Avalanche, nous aurions utilisé l'âge, le sexe, le lieu de résidence, le revenu et à peu près tout ce que nous pourrions obtenir du spectateur pour développer un cadre prédictif de suggestions d'émissions.

Maintenant, la plupart de ces informations sont soit utilisées de manière tangentielle, soit pas du tout. Si les algorithmes de Netflix ont plus de succès avec leurs suggestions qu'un cadre empirique bien-fondé (et, ceci semble être le cas dans mon expérience anecdotique), de nombreuses disciplines doivent repenser fondamentalement leur programme d'enseignement. Quel rôle joue un cours théorique dans une discipline des sciences sociales lorsqu'il n'existe pas de lien évident entre un cadre théorique et les méthodes que les spécialistes des sciences sociales utiliseront pour obtenir leurs réponses ?

Bien que cela semble aller à l'encontre des méthodes de recherche en sciences sociales, cela peut aussi être perçu comme une occasion de réfléchir différemment à la nature de la recherche.

OÙ VA LA STATISTIQUE INFÉRENTIELLE ?

Il est de plus en plus probable que les futurs jeux de données représentent des populations entières, et pas seulement des échantillons d'intérêt. Cela signifie que les résultats que nous obtiendrons seront nécessairement une représentation exacte d'une population, éliminant ainsi la nécessité d'utiliser des statistiques inférentielles. Par définition, toute tendance dérivée de données démographiques est statistiquement significative, car elle inclut toutes les personnes à partir desquelles nous essayons de généraliser.

APPRENTISSAGE PAR MACHINE

En termes simples, l'apprentissage automatique va probablement mettre certains d'entre nous, ou nos étudiants, au chômage. Les définitions de l'apprentissage automatique varient, mais elles semblent toutes avoir pour principal objectif l'apprentissage autonome par ordinateur. Cela signifie que les humains n'auront plus besoin de dire aux ordinateurs quoi faire une fois tous les algorithmes requis téléchargés. Des quantités de données seront analysées sans grande intervention d'un analyste humain. Même si ça ressemble à de la science-fiction, les ordinateurs commencent à s'enseigner eux-mêmes, ce qui aura une incidence sur la manière dont les données seront collectées et analysées à l'avenir.

L'apprentissage automatique/intelligence artificielle suscite un enthousiasme considérable, avec des utilisations qui pourraient aller de l'utilisation d'un algorithme d'apprentissage automatique à la formulation de recommandations de lois et de politiques, à la recherche du meilleur endroit pour vivre. Les

juges pourraient prendre des décisions plus justes s'ils sont informés des résultats. Le taux de rétention des immigrants pourrait augmenter s'ils pouvaient calculer dans quel pays ils seraient le plus susceptibles de réussir et d'être heureux.

CONCURRENCE HORS SCIENCES SOCIALES

En tant que spécialistes des sciences sociales, la plupart d'entre nous avons vécu une vie assez paisible dans nos disciplines. Chacun de nous a un sujet, et tout le monde nous laisse plus ou moins tranquille. Les individus ? Tout le monde sait que l'étude des individus appartient à la psychologie. Nations ? La science politique se charge d'étudier celles-ci. Société ? Tout le monde est au courant que c'est le domaine de la sociologie. Mais que se passerait-il si la sociologie pouvait en faire plus en prenant l'avant-scène pour une fois ?

Si vous pensez que seulement 0,5 % de toutes les données sont en cours d'analyse (mars 2015) et que Google gère et enregistre à lui seul les paramètres de 40 000 recherches à la seconde, il semblerait que les analystes en sciences sociales seront plus que jamais surpassés par les données à l'époque des données massives. Il existe deux voies possibles : nous pouvons laisser passer cette opportunité et rester paisibles dans nos mondes individuels ou nous pouvons franchir cette nouvelle frontière dans le but de démontrer l'utilité des sciences sociales.

CONCLUSION

Statistique Canada a récemment célébré son 100e anniversaire et mérite d'être félicitée pour ses nombreuses contributions dans les domaines de la collecte, de la gestion et de l'analyse des données. Au Canada, nous avons non seulement un organisme statistique centenaire, mais un organisme internationalement reconnu. En même temps, ce n'est pas le moment de faire une pause : les agences privées de collecte de données sont extrêmement actives et ce n'est qu'une question de temps avant que nous ne nous perdions. Nous avons également un accord assez répandu sur le fait que les statistiques sont un élément essentiel de la démocratie moderne (le contrecoup de la décision du gouvernement Harper d'annuler le recensement détaillé de 2011 en est la preuve). Ce dont nous avons maintenant besoin, c'est d'un accroissement collectif des compétences nécessaires pour profiter au maximum de ce que les données ont à nous offrir, afin d'assurer notre survie et celle de notre environnement lors de l'arrivée de la prochaine avalanche.

RÉFÉRENCES

HACKING, IAN. 1982. "Biopower and the avalanche of printed numbers," *Humanities in Society* 5 :279–295.

KUHN, THOMAS. 1962. *The Structure of Scientific Revolutions.* Chicago : University of Chicago Press.

MARR, B. (2015). "Big Data : 20 mind-boggling facts everyone must read," *Forbes* magazine. www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#39c4da3a17b1

STATISTICS CANADA. "History of the Census." www150.statcan.gc.ca/n1/edu/power-pouvoir/ch2/history-histoire/5214912-eng.htm. Retrieved May 7, 2019.