Intelligence collective et données

Arthur Charpentier

Professeur, Université du Québec à Montréal

Ewen Gallic

Maître de conférences, Aix-Marseille Université

La psychologie des foules a longtemps insisté sur les dangers des comportements collectifs, à commencer par Charles Mackay, qui affirmait en 1841, « les hommes, a-t-on bien dit, pensent en troupeaux ; on verra qu'ils deviennent fous en troupeaux, alors qu'ils ne recouvrent leurs sens que lentement, et un par un ». Cinquante ans après, Gustave Le Bon reprenait cette vision en écrivant : « l'individu se trouve altéré par la foule, devient surtout soumis à l'inconscient, et régresse vers un stade primaire de l'humanité ». Mais on redécouvre depuis quelques années qu'il est possible, au contraire, de mettre à profit la « sagesse des foules » pour reprendre l'expression de James Surowiecki.

Des méthodes ensemblistes d'apprentissage

Lors de ses travaux sur l'application des probabilités sur les votes, Condorcet observait en 1785 que la probabilité qu'un jury se trompe, collectivement, était relativement faible, bien plus faible que la probabilité individuelle. Si p désigne la probabilité de se tromper au niveau individuel, la probabilité qu'un jury composé de n jurés se trompe (on supposera n impair pour éviter une indécision) s'écrit :

$$\sum_{k\geq(n+1)/2} \binom{n}{k} p^k(1-p)^{n-k}$$

si les votes sont indépendants. Avec $n = 11$ membres, si la probabilité individuelle de se tromper est de 40 %, la probabilité que le jury se trompe passe à 25 %. Cette agrégation, par la règle de la majorité, semblait permettre d'approcher la vérité, et ce d'autant mieux que les prédictions individuelles sont justes, et que les prédictions sont indépendantes (sinon l'agrégation n'apporte pas grand-chose). La notion d'*argumentum ad populum* semblerait d'ailleurs plus loin, comme le rappelle Charpentier [2018], en affirmant que la majorité a toujours raison.

En 1907, Sir Francis Galton a publié un article intitulé « Vox populi », racontant sa visite de la foire agricole du comté de Cornwall, et qui observe des résultats proches de ceux énoncés par Condorcet. Lors de la foire, un concours « devinez le poids du panier garni » était organisé (il fallait en réalité deviner

le poids d'une vache), et Galton a analysé l'ensemble des 787 participations. Si la moitié des propositions étaient entre 1162 et 1236 livres, la moyenne était à 1198 livres, soit moins d'une livre du poids réel de la vache. Là encore, il semble qu'agréger les prédictions individuelles des experts en prenant la moyenne donne de meilleurs résultats que suivre un expert pris au hasard. Et depuis le XIVe siècle, les paris organisés lors des élections papales au Vatican permettaient d'avoir des prédictions plus justes que de chercher l'expert le mieux informé, comme le rappelle Charpentier [2017].

Ces réflexions montrent qu'il est possible de tirer parti d'un ensemble de prévisions individuelles, en agrégeant de manière adéquate, soit en prenant la moyenne dans le cas où il faut prédire une valeur précise (on parlera d'un problème de régression), soit en considérant la classe majoritaire (dans un problème de classification). Cette interprétation dans un contexte de modélisation aura de nombreuses conséquences. L'idée de chercher « le meilleur modèle » n'est pas optimale selon Galton, et il serait préférable de considérer une moyenne des modèles, potentiellement une moyenne pondérée. Statistiquement, agréger cet ensemble de modèles est meilleur qu'en choisir un seul. Sous un angle computationnel, calculer une simple moyenne est bien plus simple que de chercher le meilleur modèle, qui est un problème d'optimisation. Enfin, si certains modèles sont contraints, prendre la moyenne permet d'obtenir des valeurs qu'aucun modèle individuel ne donnerait.

Les données collaboratives et le partage de données

L'idée d'assemblage d'informations produites par une multitude d'individus pour construire un résultat plus global se rencontre dans de nombreux projets collaboratifs, comme l'encyclopédie en ligne Wikipédia ou le projet cartographique OpenStreetMap. Ces projets font appel à l'association des connaissances et des compétences individuelles pour cocréer une connaissance plus globale. Cette démarche peut être qualifiée d'intelligence collective, que le MIT Center for Collective Intelligence définit comme « des groupes d'individus effectuant des choses de manière collective qui paraissent intelligentes » [Malone *et al.*, 2009]. L'existence de ces projets est favorisée par Internet, qui permet leur coordination. Selon Lévy [2016], les outils informatiques favorisent l'intelligence collective, dans la mesure où ils permettent de « rendre les humains intelligents ensemble au moyen des ordinateurs ». La généalogie en est un bon exemple. L'agrégation des arbres produits par des millions d'utilisateurs de services en ligne de généalogie permet entre autres de documenter des changements démographiques [Blanc, 2020], des phénomènes de migration ou encore la mortalité de nos ancêtres [Charpentier et Gallic, 2020]. Ces études sont possibles grâce au travail minutieux réalisé par un très grand nombre de généalogistes mettant à disposition le fruit de leurs recherches sur Internet. A leur tour, les généalogistes voient la quantité d'effort à fournir considérablement réduite grâce à l'accès facilité aux données constituant la base de leur recherche (principalement, les registres paroissiaux et d'état civil qui sont de plus en plus disponibles gratuitement au format numérique).

La période récente, marquée par l'épidémie de Covid-19, illustre également l'importance du partage de données, tant pour alimenter la production scientifique, pour participer à l'information du grand public de la situation et de son évolution, ou encore pour aider les pouvoirs publics à prendre des décisions. La communauté scientifique a réagi à la récente crise sanitaire par une publication massive d'articles liés à la Covid-19. Près de 200 000 publications ont été recensées en 2020, dont plus de 30 000 documents en accès libre sur les plateformes de prépublication comme bioRxiv ou medRxiv [Else, 2020]. On peut noter au passage que durant la pandémie, de nombreux chercheurs ont partagé leurs travaux sous forme de prépublication en accès libre pour la première fois [Fraser *et al.*, 2021]. Beaucoup d'éditeurs de revues scientifiques ont de plus pris la décision de retirer les verrous d'accès payant (*paywalls*) des articles en lien

avec la Covid-19. Parallèlement, des données relatives à l'épidémie ont été rendues publiques afin de la documenter et de suivre son évolution. On pense notamment aux données agrégées et publiées par le Center for Systems Science and Engineering de l'Université Johns Hopkins ayant alimenté un tableau de bord permettant à tout le monde de suivre en temps réel la situation mondiale [Dong *et al.*, 2020], ou encore, à l'échelle de la France, aux données distribuées par Santé publique France et par l'Insee.

La prolifération de travaux scientifiques et un accès possible à des volumes non négligeables de données ont permis aux pouvoirs publics de prendre des décisions, notamment en matière de santé publique, peut-être davantage éclairées qu'à l'ordinaire. Les pouvoirs publics ont en effet pu bénéficier d'une importante variété d'indicateurs chiffrés pour prendre leurs décisions. Ils ont également pu tirer parti de la sagesse des foules. En effet, en l'absence de données, comme le notait Morgan [2019], les décideurs publics s'appuient sur les connaissances tirées des événements passés et sur les avis, parfois divergents, d'une poignée d'experts. Or ici, dès le début de la pandémie, l'appareil politique a pu rassembler les connaissances construites et partagées par des équipes de recherche de milieux variés, allant des épidémiologistes aux virologues, en passant par les économistes, les démographes ou les historiens.

Tandis que certaines données relatives à l'épidémie de Covid-19 ont été rendues accessibles et ont conduit à une production scientifique substantielle, d'autres, comme le déplorent Cosgriff *et al.* [2020], n'ont pas été partagées. C'est le cas des données individuelles des patients. Mais ce sont loin d'être les seules. Paradoxalement, la plupart des données générées sont aujourd'hui collectées par le secteur privé. Comment utiliser pertinemment les possibilités d'information et de piste de ces données pour l'amélioration des politiques publiques ? On se souvient notamment des questionnements sur l'efficacité des confinements lors de la pandémie, et de la difficulté d'avoir des données pour permettre une modélisation fine des graphes sociaux des individus, afin d'estimer le nombre de cas contacts potentiels, en fonction de l'âge, de la profession, etc. Si le secteur public a depuis des années « libéré » des données administratives, gratuitement, le secteur privé essaie au contraire de monnayer ces données collectées souvent à l'insu de l'utilisateur. On peut néanmoins mentionner quelques initiatives médiatiques, comme les *Disaster Maps* de Facebook, qui cherchent à combler les lacunes des sources de données traditionnelles et à informer certains partenaires (Croix rouge, Croissant-rouge, Unicef) sur les efforts de secours plus ciblés lors de catastrophes naturelles, en partageant des données de localisation et de déplacement. S'il y a des efforts ponctuels, de nombreuses entreprises refusent de partager leurs données, même si de nombreuses études ont montré que c'est l'utilisation et la réutilisation des données qui leur donnent leur vraie valeur, et les transforment en bien public.

Les biais du détournement d'usage

Pour reprendre la classification de Rosenbaum [2018], les données collaboratives sont à mi-chemin entre les données d'expérience d'un côté, et les données d'observation (ou administratives) de l'autre. Dans le premier cas, les données sont collectées dans un but et une finalité bien précis, afin de répondre à une question. On peut penser à toutes les études réalisées pour tester un vaccin ou un médicament. De leur côté, les données d'observation sont des données massives, souvent collectées dans un but mais suffisamment exhaustives pour être réutilisées. On peut penser aux tickets de caisse des magasins (qui collectent toutes les transactions pour la comptabilité), aux données fiscales ou démographiques (qui sont un état des lieux précis de la richesse d'une nation, ou de l'état d'une population). Les données collaboratives sont, à l'instar des secondes, massives, mais entachées de biais comme les premières.

Pour comprendre le biais des données d'observation, on peut penser à l'exemple classique de l'analyse causale de l'analyse de la convalescence post-opératoire.

Supposons que l'on cherche à savoir si une personne opérée se remettra plus vite à l'hôpital, ou en rentrant chez elle. On peut récupérer des statistiques passées dans un hôpital, et regarder la probabilité de revenir dans les six mois consécutifs à une opération, en distinguant celles qui sont rentrées chez elles le lendemain de l'opération, et celles qui sont restées plus longtemps. Mais cette comparaison est biaisée, car le premier groupe est en réalité constitué des personnes qui ont été autorisées à rentrer, et étaient donc, a priori, en relativement bonne santé.

Dans le contexte des données collaboratives (et non pas administratives comme dans le cas des hôpitaux), on peut penser aux applications de déplacement en automobile, comme Waze, Here WeGo, ou Google Map. A titre personnel, je peux utiliser ces applications pour trouver un chemin rapide permettant d'aller d'un point A à un point B. Mais comme le montre Graaf [2018], il est possible d'utiliser ces applications dans un contexte de planification urbaine, pour voir les routes les plus empruntées, les carrefours qui sont des goulots d'étranglement, etc. Ce détournement d'usage, à des fins de maximisation du bien-être collectif peut être louable, mais il convient d'analyser ces données correctement. En effet, certaines applications anticipent les routes très encombrées, et peuvent proposer des itinéraires alternatifs à leurs utilisateurs. Comment tenir compte de ce biais comportemental dans les données que l'on va ensuite essayer d'analyser ?

David Hand [2020] désigne par *dark data* ces données invisibles, manquantes, qui sont inévitables quand on se contente d'observer. Il propose une classification en une quinzaine de types, entre les données que l'on sait manquantes, celles qu'on ne sait pas manquantes, le biais de sélection, les volontaires qui participent aux enquêtes ou au contraire les participants qui sortent des enquêtes, les données qui changent dans le temps, les données agrégées ou bruitées (pour éviter des soucis d'anonymat), etc. Avec un peu de chance, il est possible de comprendre ce biais, de le modéliser, et de le corriger. Mais dans certains cas, ces biais rendent toute analyse caduque.

En 2008, Google a lancé un service Web appelé Google Flu Trends destiné à prédire les épidémies de grippe. La méthodologie décrite par Ginsberg *et al.* [2009] consistait essentiellement à rechercher des mots-clés associés aux syndromes grippaux, car « certains termes de recherche sont de bons indicateurs de l'activité grippale » comme l'indiquait la foire aux questions (FAQ) du site. Cette information recueillie par Google pouvait alors être exploitée pour suivre l'évolution de la grippe saisonnière, au jour le jour. La fréquence d'observation journalière était bien plus élevée que celle des bulletins hebdomadaires que les Centres pour le contrôle et la prévention des maladies publient avec un retard d'en général une à deux semaines.

Néanmoins, le service de la multinationale a cessé de fournir ses prédictions en 2015, à la suite d'une série d'échecs. En effet, l'algorithme n'a pas été en mesure de prévoir la pandémie non saisonnière de grippe A (H1N1) de 2009 et a par la suite surestimé la prévalence de la grippe pendant 100 semaines sur 108 entre 2011 et 2013, comme l'ont noté Lazer *et al.* [2014]. Les auteurs mettent en avant deux raisons de l'échec de Google Flu Trends. Premièrement, l'orgueil du big data. L'exercice de prédiction confié à l'algorithme s'accompagne d'un risque de surapprentissage. Il consiste à modéliser un peu plus d'un millier de points, à l'aide de plusieurs millions de termes de recherche. La question de l'utilité du recours à un tel volume de données se pose d'autant plus que la précision des prédictions ne semble guère meilleure que celle fournie par un modèle faisant intervenir les observations des semaines précédentes et corrigeant des effets saisonniers. Deuxièmement, l'échec de Google Flu Trends peut être attribué aux modifications apportées en permanence au service. Les données utilisées dans le modèle proviennent de l'utilisation du moteur de recherche de Google. Elles sont donc liées aux changements permanents (au moins un par jour) opérés sur ce dernier non seulement par l'entreprise, mais également par ses utilisateurs. Un événement notable est l'introduction de l'auto-complétion par le moteur de recherche ayant rendu complètement inefficace l'algorithme de prévision de grippe (l'auto-

complétion allant jusqu'à proposer aux utilisateurs des diagnostics pour des recherches incluant des symptômes physiques).

Mais la dérive la plus grande est probablement celle de l'évaluation collective, racontée récemment par Coquaz et Halissat [2020]. Nées d'une défiance croissante vis-à-vis des experts, certaines plateformes ont proposé aux utilisateurs de noter leurs restaurants, leurs professeurs, leurs séjours à l'hôpital... Partager mes expériences culinaires me permet, en échange, d'avoir des avis d'autres gourmets, qui m'aideront à dénicher un bon restaurant si je vais dans une ville que je ne connais pas. Mais là encore, de nombreux exemples ont montré que ces données étaient aussi détournées, par exemple pour licencier un serveur qui m'a fait mettre une note « moyenne » au restaurant. Comme le notait Pasquier [2014], « sous couvert de donner à tout un chacun le pouvoir de s'exprimer (...) les algorithmes de l'industrie du Web participatif permettent de construire à moindres frais des modèles commerciaux rentables ». Et cette utilisation pose de nombreux problèmes sur lesquels nous reviendrons.

Bibliographie

Blanc G., "Demographic Change and Development from Crowdsourced Genealogies in Early Modern Europe", HAL, 2020. *https://hal.archives-ouvertes.fr/hal-02922398.*

Charpentier A., « *Fake news, post-truth*, Wikipedia et *blockchain* : vérité et consensus », *Risques*, n° 115, 2018, pp. 133-138.

Charpentier A., « Les marchés prédictifs comme technique de prévision », *Risques*, n° 111, 2017, pp. 117-121.

Charpentier A. ; Gallic E., « La démographie historique peut-elle tirer profit des données collaboratives des sites de généalogie ? », *Population*, vol. 75, n° 2, Cairn.info, 2020, pp. 391-421. *https://doi.org/10.3917/popu.2002.0391*

Coquaz V. ; Halissat I., *La nouvelle guerre des étoiles. Enquête : nous sommes tous notés*, Kero, 2020.

Cosgriff C. V. ; Ebner D. K. ; Celi L. A., "Data Sharing in the Era of Covid-19", *The Lancet Digital Health*, vol. 2, n° 5, Elsevier BV, 2020, p. 224. *https://doi.org/10.1016/s2589-7500(20)30082-0*

Dong E. ; Du H. ; Gardner L., "An Interactive Web-Based Dashboard to Track Covid-19 in Real Time", *The Lancet Infectious Diseases*, vol. 20, n° 5, Elsevier BV, 2020, pp. 533-534. *https://doi.org/10.1016/s14733099(20)30120-1*

Else H., "How a Torrent of Covid Science Changed Research Publishing-in Seven Charts", *Nature*, vol. 588, 2020, p. 553. *https://doi.org/10.1038/d41586-020-035 64-y*

Fraser N. ; Brierley L. ; Dey G. ; Polka J. K. ; Pálfy M. ; Nanni F. ; Coates J. A., "The Evolving Role of Preprints in the Dissemination of Covid-19 Research and Their Impact on the Science Communication Landscape", edited by Ulrich Dirnagl, *PLOS Biology*, vol. 19, n° 4, Public Library of Science (PLoS), e3000959, 2021. *https://doi.org/10.1371/journal.pbio.3000959*

Ginsberg J. ; Mohebbi M. H. ; Patel R. S. ; Brammer L. ; Smolinski M. S. ; Brilliant L., "Detecting Influenza Epidemics Using Search Engine Query Data", *Nature*, vol. 457, n° 7232, Springer Science+Business Media LLC, 2009, pp. 1012-1014. *https://doi.org/10.1038/nature07634*

Graaf S. (van der), "In Waze We Trust: Algorithmic Governance of the Public Sphere", *Media and Communication*, vol. 6, n° 4, 2018, Cogitatio, pp. 153-162. *https://doi.org/10.17645/mac.v6i4.1710*

Hand D. J., *Dark Data: Why What You Don't Know Matters*, Princeton University Press, 2020.

Lazer D. ; Kennedy R. ; King G. ; Vespignani A., "The Parable of Google Flu: Traps in Big Data Analysis", *Science*, vol. 343, n° 6176, American Association for the Advancement of Science (AAAS), 2014, pp. 1203-1205. *https://doi.org/10.1126/science.1248506*

Lévy P., « L'intelligence collective, en quelques mots... », Pierre Levy's blog, 2016. *https://pierrelevyblog.com/2016/03/03/lintelligence-collective-en-quelques-mots/*

MALONE TH. ; LAUBACHER W. R. ; DELLAROCAS CH., "Harnessing Crowds: Mapping the Genome of Collective Intelligence", MIT Sloan Research Paper, n° 4732-09, SSRN, 2009. *https://doi.org/10.2139/ssrn.1381502*

MORGAN O., "How Decision Makers Can Use Quantitative Approaches to Guide Outbreak Responses", *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 374, n° 1776, The Royal Society, 20180365, 2019. *https://doi.org/10.1098/rstb.2018.0365*

PASQUIER D., « Les jugements profanes en ligne sous le regard des sciences sociales », *Réseaux*, vol. 1, n° 183, La Découverte, 2014, pp. 9-25.

ROSENBAUM P., *Observation and Experiment. An Introduction to Causal Inference*, Harvard University Press, 2018.