Utilisation des données et de l'intelligence artificielle aux Etats-Unis pour lutter contre la maladie

L'épidémie de COVID-19, causée par le virus SRAS-CoV-2, a été identifiée en décembre 2019 en Chine et déclarée pandémie mondiale par l'OMS le 11 mars 2020. La combinaison des données relatives au virus et à l'épidémie et de l'intelligence artificielle (IA) comporte un grand potentiel de synergie dans la lutte contre la pandémie de COVID-19 et les initiatives pour les utiliser se sont multipliées partout dans le monde. Nous parlons ici d'IA au « sens large », à savoir tous les outils permettant d'extraire du sens ou des motifs intéressants à partir des données relatives à l'épidémie ou au virus : l'utilisation du « machine learning », de la visualisation de données, de la vision par ordinateur, traitement du langage naturel, etc.

Dans ce rapport du *Service pour la Science et la Technologie de l'Ambassade de France aux Etats-Unis*, un premier bilan de ces initiatives aux Etats-Unis est dressé en examinant la contribution réelle et potentielle de l'IA à la lutte contre le COVID-19, ainsi que les limites actuelles sur ces contributions. Il vise à donner une lisibilité à un véritable foisonnement d'initiatives tous azimuts et à un corpus de travail dont la taille ne cesse de croître chaque jour. Au moment de la rédaction de ce rapport (16 avril 2020), une grande incertitude entoure encore les estimations quant à la gravité de la situation et le succès des réponses non-pharmaceutiques et pharmaceutiques.

Les principales conclusions sont les suivantes:

- *Il est indiscutable que l'écosystème américain du numérique s'est mobilisé à une échelle exceptionnelle en réorientant massivement ses activités de recherche et de développement vers la recherche de solutions. Si cette mobilisation se fait dans l'ordre dispersé caractéristique de cet écosystème, on observe aussi des initiatives spectaculaires permettant un meilleur partage des ressources.*

- *L'IA n'apporte pas de solution miracle en réponse à la crise. Des outils maintenant classiques sont mis au service de la gestion de la crise mais ceux qui reposent sur un apprentissage long ne parviennent pas à apporter des réponses opérationnelles dans une situation de crise.*
- *La crise fait bouger les lignes concernant les usages de l'IA et les questions éthiques. Les scientifiques s'efforcent de construire des outils qui permettent à la fois de prendre en compte les enjeux de la crise et de préserver la confidentialité des utilisateurs.*

Ainsi, de nombreuses applications sont développées pour faire face à la crise. Parmi ces applications, nous distinguons trois grandes familles : (A) les outils permettant de suivre et prévoir l'évolution d'une épidémie à une échelle macroscopique ou sur le plan scientifique, (B) ceux qui permettent de détecter, suivre et prévoir l'évolution de la maladie à l'échelle du patient et (C) les applications permettant de développer des traitements pour vaincre la maladie.

Collecte et analyse des données sur l'épidémie

Prévision de la crise

L'utilisation de grandes masses données pour prédire les épidémies existe depuis bien avant l'épidémie de COVID-19. Une des premières tentatives d'envergure a été faite par Google avec son outil, maintenant retiré du service, « Google Flue Trend » (GFT). Le principe de base est simple : avec une barre de recherche aussi utilisée que celle de Google, on peut s'attendre à ce que lorsqu'une épidémie survient beaucoup de gens vont entrer dans le navigateur des mots-clés particuliers (e.g. relatifs à la grippe, ou à des symptômes). En croisant ce pic de requête avec d'autres données captées sur internet, on peut ainsi espérer anticiper l'apparition de l'épidémie et gagner quelques précieux jours ou semaines pour s'y préparer.

Dans le même esprit, John Brownstein, de la Harvard Medical School, fait partie d'une équipe internationale qui utilise l'apprentissage automatique (*Machine Learning*) pour parcourir les médias sociaux et des bases de données publiques et

privées (hôpitaux, réseaux de transport et autres), afin de réaliser des analyses prévisionnelles en temps réel de l'épidémie. L'IA permet ici d'extraire des informations pertinentes d'un ensemble de données accessibles de plus en plus conséquent qu'il serait difficile de parcourir de manière non automatisée. D'autres initiatives de ce type existent et viennent d'universités (*e.g.* Carnegie Mellon University) comme d'entreprises (Stratyfid à Charlotte, Bluedots au Canada).

En ce qui concerne l'efficacité de ces outils, la capacité de la startup d'IA canadienne *BlueDot* à prévoir la crise du COVID-19 a eu un grand retentissement dans les médias grands publics. Elle a en effet envoyé une alerte à ses clients sur le coronavirus le 31 décembre 2019, soit bien avant que l'OMS n'émette une mise en garde, le 9 janvier 2020. Pourtant, *HealthMap, du Boston Children Hospital,* a émis un avertissement concernant de mystérieux cas de pneumonies à Wuhan un jour plus tôt que *BlueDot*, soit le 30 décembre 2019. Au-delà des articles admiratifs qui sont apparus à ce sujet, le succès de ces IA doit être relativisé. Notons en effet d'une part qu'un scientifique du *Program for Emerging Monitoring Diseases* a émis une alerte au sujet du COVID-19 seulement 30 minutes plus tard qu'HealthMap, sans s'appuyer sur un outil d'IA. De plus, l'alerte de HealthMap a beaucoup minimisé la sévérité de l'épidémie, puisque l'alerte n'était évaluée qu'avec une gravité de 3 sur 5, soit bien inférieure à la gravité réelle de l'épidémie de COVID-19. La crise du COVID-19 a donc montré que les IA n'étaient pour l'instant pas plus performantes que les humains pour la prédiction d'épidémies qui reste un problème complexe. Bien que ces outils puissent être d'une aide utile, force est de constater qu'il est plus que jamais nécessaire que l'homme reste dans la boucle pour pouvoir correctement contextualiser l'information. Une utilisation connexe et prometteuse de la science des données, est le suivi de l'épidémie.

Suivi à grande échelle de l'épidémie

L'agrégation des données permise par les nouveaux outils numériques permet efficacement d'avoir une vision macroscopique de la situation. Cela permet de mieux comprendre la circulation du virus et de construire des scénarios pour mieux appréhender l'efficacité et les modalités de diverses mesures (distanciation sociale, restriction de mouvements, confinement, déconfinement).

Le tableau de bord développé par le centre d'épidémiologie de l'Université Johns Hopkins (image ci-dessus) est un des exemples qui s'est rapidement imposé pour faire un suivi statistique et géographique sommaire de l'épidémie. En dehors des exemples précédemment cités, à la frontière entre prévision et suivi, notons également divers outils de visualisation qui s'efforcent de mettre les données en perspective : *Our World in Data* (basé à Oxford, en collaboration avec Harvard, Stanford, Berkeley et le MIT), le projet *COVID Tracking* (lancé par le mensuel The Atlantic), le projet *EarthTime* (collaboration entre Carnegie Mellon et le Forum Economique Mondial) ou les cartes du CIDRAP (Université du Minnesota).

Comme pour les tentatives de prévision, il est crucial d'agréger les données de différentes sources en temps réel mais également de pouvoir le faire à l'échelle internationale pour pouvoir effectuer un suivi efficace de la propagation du virus. Au-delà des pures questions techniques, cela soulève notamment des problèmes de partage de l'information entre différentes juridictions : entre les pays, entre différents niveaux de gouvernance mais aussi entre différentes plateformes et équipes de recherche. Cet exercice pose également des questions de fond concernant la cohérence et l'interopérabilité des données (différences locales, changements de méthodologie, à dessein ou non, différence dans la définition des cas) qui complique les interprétations et prévisions que l'on peut en tirer. Par exemple, l'agrégation des cas décédés hors du milieu hospitalier au nombre de cas confirmés le dimanche 5 avril donne l'impression sur des tableaux de bord simplifiés que ce nombre était encore croissant en France alors qu'en réalité la situation commençait à se stabiliser. Dans une direction différente, mentionnons l'enjeu qui consiste à évaluer le nombre d'infections non répertoriées pour diverses raisons (travaux de Lucy Li du Chan Zuckerberg Biohub). Des solutions pour limiter les dégâts de ces incohérences sont envisagées par la communauté scientifique (voir notamment l'initiative CEDAR ci-dessous).

Certains outils complémentaires permettent de spatialiser les données et de les croiser avec des outils mesurant les déplacement (d'abord à une échelle macroscopique). C'est important au départ pour suivre l'évolution de l'épidémie et détecter les prochains "hot spots", mais cela joue un rôle aussi pour mesurer l'impact des mesures de confinement. Google a notamment publié sur son portail dédié au coronavirus des résultats préliminaires s'appuyant entre autres sur les

données de smartphones (COVID-19 Community Mobility Reports) qui mettent en évidence les changements de comportements dans les déplacements des populations avant et après la mise en place de mesures de confinement. Différents niveaux de détails sont disponibles en fonction des régions (par pays, par état).

Ces données spatialisées de l'infection sont d'une importance cruciale en particulier lorsqu'elles sont mises en regard d'informations sur les ressources médicales disponibles.

Columbia University (avec le centre hospitalier new-yorkais du Mount Sinai) et le MIT (COVID-19 Policy Alliance) ont également développé des cartes permettant d'identifier les zones dites "à risque", déterminées à partir de données démographiques (personnes âgées ou ayant des conditions de santé dégradées) et des capacités d'accueil (nombre de lits, d'équipements médicaux adaptés). A partir de ces informations, le même groupe du MIT produit également des suggestions de mesures pour le déconfinement préconisant l'utilisation de la télémédecine. Lien : Severe COVID-19 Risk Mapping

Des chercheurs de UC San Diego (UCSD) ont également travaillé à avoir une image macro de l'épidémie, en utilisant un objet connecté, l'anneau *Oura* d'une startup finlandaise. Cet anneau, initialement utilisé pour l'analyse du sommeil, permet de collecter différentes données corporelles (température, rythme cardiaque, etc.). En collaboration avec UC San Francisco (UCSF), ces données sont analysées pour alerter la personne qu'elle est susceptible d'être infectée par le coronavirus. Ces informations sont ensuite transmises à UCSD qui les anonymise et les utilise pour fournir une image macro de l'épidémie. Enfin, ces informations peuvent être croisées avec des questionnaires en ligne en libre-service et permet d'effectuer une déclaration sur son propre état de santé. A ce sujet on pourra voir l'article sur ce site et le site d'information de UC.

Information des décideurs et du grand public

La dissémination de l'information auprès des décideurs publics et de la population en général est importante et difficile. Les multiples sources d'information agrégées (le New York Times entretient aussi l'une d'elles) souvent accessibles au public

(averti) commentées par les experts des médias se sont souvent trouvées en contradiction avec le discours diffusé par la Maison Blanche. Il est frappant par exemple de s'apercevoir que les « modèles » présentés par l'exécutif pour soutenir les décisions de confinement sont eux-même une « synthèse » ad hoc de différents travaux scientifiques et qu'il n'est pas possible de remonter aux sources. Face à cette abondance de données et à la complexité de l'information, l'outil numérique est d'une aide précieuse pour structurer la pensée et arriver à distinguer les informations pertinentes et de source sûre des mauvaises informations (Communications plus spécifiques sur ces sujets à venir).

Dans une optique plus pragmatique, pour répondre aux nombreuses questions concrètes que se posent les citoyens sur l'épidémie, l'entreprise IBM, en accord avec les agences gouvernementales, met à disposition sa plateforme *Watson Discovery* (renommée *Watson Assistant for Citizens* pour cette application spécifique) et ses capacités en IA permettant de calibrer des "assistants virtuels". Diverses institutions locales (Lancaster, CA, Ostego, NY, ou Austin TX) mais également des pays européens (Espagne, Grèce, Pologne, République Tchèque, Royaume-Uni) ont pu bénéficier de ce service pour mieux organiser leurs réponses au flux de questions sur le dépistage, les symptômes, la marche à suivre en cas d'infection, les consignes de nettoyage, le nombre de cas dans le voisinage. Il s'agit ici d'une utilisation "classique" de l'IA qui n'est liée à l'épidémie que par le contexte mais qui peut elle-même produire des informations d'utilisation intéressantes à reboucler avec le systèmes précédemment cités.

Données personnelles

La granularité des mesures considérées dans le cadre du suivi épidémiologique à une échelle macroscopique soulève peu de questions de confidentialité. Il s'agit d'agréger intelligemment des données plus ou moins publiques issues de sources différentes et pas toujours disponibles de manière cohérente. C'est encore le cas lorsqu'il s'agit de données concernant par exemple le code génétique des virus détectés, données qui partagées à l'échelle mondiale permettent de retracer l'histoire de la contagion (*Stainstream* ou *Department of Immunology and Microbiology* de Scripps Research, voir NDI-2020-0163634 et NDI-2020-0170673) sans identification d'un individu à proprement parler.

Néanmoins, ces questions entrent en jeu rapidement dès lors que les systèmes cherchent à exploiter des données plus fines, notamment des données cliniques détaillées concernant l'état de santé de patients. Certaines initiatives utilisent des données fournies volontairement par le patient lui-même. Mais la situation d'urgence engendrée par la crise bouscule les procédures habituelles. Les solides procédures de consolidation des données de terrain des CDC ne sont pas assez rapides pour cette situation exceptionnelle et cela amène certains acteurs à demander d'accélérer la diffusion d'informations personnelles pour pouvoir apporter des solutions. Darren Schulte, médecin et PDG d'Apixio, qui a conçu un algorithme IA pour extraire des informations des dossiers détaillés de patients, pense que ces informations devraient être ouvertes pour permettre une détection précoce des individus les plus sujets aux risques de complications. Il y voit à ce stade plus une difficulté technique, liée à la manière dont ces données sont stockées actuellement, qu'une réelle question de confidentialité. Il estime que ces données devraient d'ailleurs être partagées à l'échelle internationale en temps réel pour permettre d'alimenter les algorithmes.

Cet avis n'est pas unanimement partagé dans le pays. L'urgence de la situation force certains compromis : par exemple, le Département de la Santé a émis une directive assouplissant les règles de confidentialité en vue de simplifier la tâche des praticiens notamment pour leur permettre d'exercer à distance (Notification of Enforcement Discretion for Telehealth Remote). Les américains restent très attachés à la confidentialité de leurs données de santé et la recherche d'un compromis entre l'urgence d'efficacité de la lutte contre le virus et la nécessité de confidentialité et protection de la vie privée est âprement discutée, autant entre experts techniques que décideurs politiques (voir NDI-2020-0178754 au sujet du *contact tracing*). En effet, une réelle inquiétude existe sur la jurisprudence créée et la persistance de certaines pratiques vis-à-vis des données en dehors du cadre extraordinaire de la crise. (Coronavirus is forcing a trade-off between privacy and public health).

Pour autant, si d'un côté la technologie pose d'importantes questions sur l'utilisation des données, elle peut également participer à la conception de solutions pour protéger la vie privée. En effet, en parallèle de la crise sanitaire, les Etats-Unis sont en pleine campagne de recensement et le gouvernement fédéral en partenariat avec Apple et Facebook a développé la notion de *differential privacy*. Cette technique

mathématique permettant d'anonymiser les données offre en outre la possibilité de mesurer le niveau de protection obtenu suite à l'opération.

Notons, par ailleurs, diverses initiatives à caractère éducatif qui ont été lancées et qui tirent avantage du confinement pour sensibiliser la population américaine aux enjeux de la protection de la vie privée. Des cours en ligne et autres matériels pédagogiques sont mis à disposition par l'université de Berkeley (soutenue par la NSF), l'université d'Harvard, l'entreprise Intel ou l'association *Cyber Civics* afin que les jeunes parents puissent enseigner à leurs enfants les bonnes pratiques et réflexes à avoir dans l'utilisation d'outils numériques ou d'outils IA.

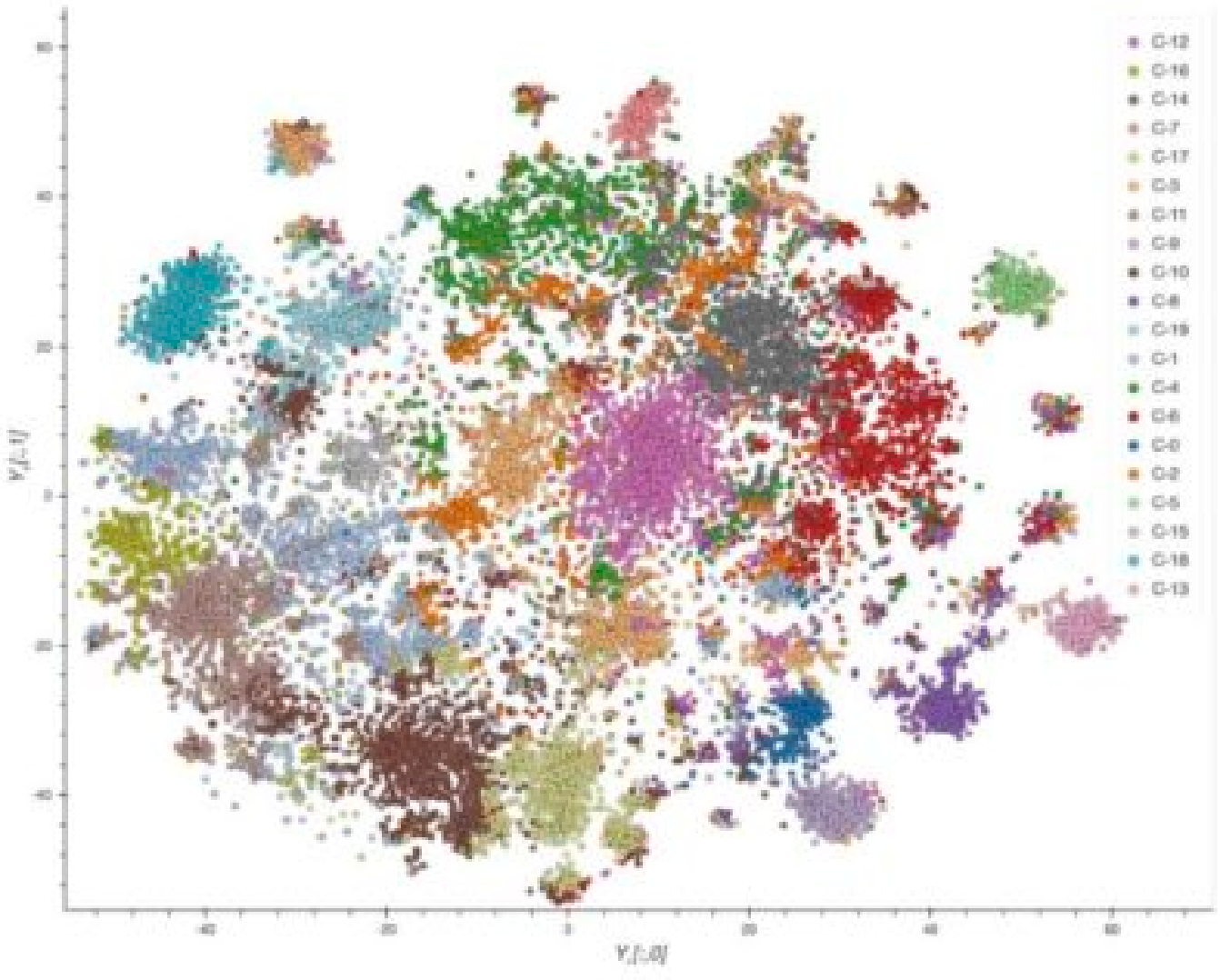
Des données de tous types, y compris des descriptions de cas cliniques, sont aussi accessibles sur la plateforme Kaggle (une entreprise du groupe Alphabet) ; fournies par diverses sources (en principe respectueuses des normes de confidentialité en vigueur dans les pays d'où elles viennent), elles permettent à des informaticiens ou à des équipes de contribuer directement à leur exploitation en proposant des algorithmes permettant l'extraction d'information structurée. Par exemple, on trouve des modèles de prévision à partir de données spatialisées, des analyses des effets du confinement, des évaluations des risques de complications associées à certaines pathologies ou des estimations de la durée d'incubation. Le format "concours" associé à chaque jeu de donnée (parfois doté par la structure fournissant les données en soulevant des questions spécifiques) assure la popularité de la plateforme et suscite de nombreuses contributions ; les plus significatives sont évaluées par des votes.

Information scientifique

Outre la diffusion d'information vers le grand public et les décideurs publics, un autre aspect déjà partiellement mentionné est l'échange d'informations au sein de la communauté scientifique pour faire avancer la recherche et les difficultés qui peuvent subvenir. L'effort d'ouverture doit encore s'accélérer et prend de nouvelles formes grâce à différentes initiatives.

Déjà mentionnée dans la NDI-2020-0166556 et la NDI-2020-0177610, la base de données CORD-19 publiée par l'OSTP sur Kaggle et qui rassemble les « publications

» scientifiques sur le coronavirus a suscité un engouement exceptionnel. Elle a permis notamment de faire émerger des outils de navigation et traitement de l'information très utiles. A ce jour, la contribution la plus populaire a été produite par une équipe du *Malware Research Group* de l'Université de Maryland à Baltimore County. L'algorithme proposé permet de grouper et visualiser les articles scientifiques pertinents en fonction de mots-clés fournis par l'utilisateur (voir illustration des groupements). Cela peut être utile pour effectuer des analyses bibliographiques mais c'est également un gain de temps pour les scientifiques à la recherche d'un résultat ou d'une étude donnée.



Regroupement d'articles scientifiques sur le coronavirus par mots-clés – illustration tirée de Kaggle. Chaque point sur le graphique représente un article de recherche.

En cliquant sur un point, on obtient les informations (titre, auteurs etc.) et un lien vers l'article. A chaque groupe d'articles trouvé par l'algorithme, une couleur est associée.

Le consortium de recherche public-privé C3.ai DTI (*Digital Transformation Institute*) qui regroupe plusieurs universités prestigieuses américaines et Microsoft a pour vocation d'accélérer le développement de l'intelligence artificielle à destination des entreprises et des gouvernements. Il a annoncé focaliser ses efforts de recherche sur la lutte contre le COVID-19. Doté d'un budget total de 367 millions de dollars sur 5 ans, l'Institut compte concentrer ses recherches sur l'IA, le machine learning, la cyber sécurité, le big data, l'éthique et les politiques publiques associées à ces sujets. Une enveloppe de 5,8 M$ a été prévue pour financer l'appel à projets initial pour le COVID-19. La liste complète des sujets de recherche liés à la crise sanitaire est disponible sur le site de l'Institut.

Le consortium a également mis à disposition des chercheurs le *COVID-19 Data Lake* qui est une plateforme incorporant diverses bases de données de recherche sur le COVID-19 (entre autres celle de l'OMS, du Centre européen de prévention et de contrôle des maladies et des centres homologues indien, italien et indonésien) couplée à la suite logicielle développée par C3.ai. La plateforme est annoncée accessible dès le 20 avril de manière complètement gratuite.

On le voit, au-delà de l'ouverture des données, un enjeu tout aussi, voire encore plus important, est la qualité des données échangées. L'acronyme FAIR (*findable, accessible, interoperable, reusable*) témoigne des efforts importants à consacrer afin de s'assurer que les données soient trouvables, accessibles, puissent fonctionner avec différents systèmes et être réutilisables pour différentes applications.

L'une des propositions faites pour remédier à ce défi est celle du CEDAR (*Center for Expanded Data Annotation and Retrieval*) qui prône la normalisation des "meta-données", ces informations associées à chaque jeu de données qui permettent de déterminer ce qu'elles représentent. Mark Musen, professeur au *Stanford Center For Biomedical Informatics Research* de l'Université Stanford explique que cette normalisation permet un partage efficace et un traitement automatisé des informations collectées dans le monde, notamment dans la situation de crise

actuelle. On trouvera plus de détails sur ces aspects en lien avec la crise actuelle sur le site VODAN.

Diagnostic, pronostic, suivi des patients et gestion des données individuelles.

Pour accélérer la prise en charge des patients, l'outil numérique s'avère également précieux bien que toujours perfectible. Dans les paragraphes suivants, nous analysons les initiatives américaines en télédiagnostic, les travaux utilisant l'IA pour l'analyse d'images médicales, pour la pronostic, et le suivi individuel des patients.

Télédiagnostic

Des solutions de télédiagnostic ont été mises à disposition du grand public. Elles n'ont pas la prétention de fournir un réel diagnostic médical, mais d'apporter des recommandations, dans le but que seuls des gens qui en ont vraiment besoin ne viennent se présenter aux urgences, dans un contexte de ressources médicales limitées.

Le *Center for Disease Control (*CDC), avec la coopération de *Microsoft Azure Bot service*, a mis en ligne un chatbot afin de guider les gens sur la conduite à tenir s'ils ont l'impression d'avoir des symptômes du coronavirus (Testing for COVID-19). Un produit relativement similaire a été proposé par une association entre les deux entreprises de santé Cigna et Express Scripts et l'entreprise de technologie Buoy (partenaire de *Boston Children's Hospital* et *Harvard Medical School*). Il permet d'interagir gratuitement en ligne avec un chatbot qui pose une série de questions permettant d'évaluer le niveau de risque d'être infecté par le coronavirus et d'avoir besoin rapidement d'une prise en charge. Ce chatbot suit les directives du CDC. Cependant, cet outil n'apporte pas à proprement parler un diagnostic.

Ces produits semblent donner satisfaction mais contrairement aux outils modernes d'IA qui sont pilotés par les données, ces outils fonctionnent comme des systèmes experts, à savoir que les connaissances d'experts ont été injectés à défaut de données. Ils ne découlent donc pas d'une utilisation réelle des données relatives au

COVID-19.

Le CDC travaille aussi avec divers partenaires industriels comme la firme *Kinsa Health* qui exploite les données provenant de thermomètres connectés (*AI enabled!*) répartis dans tout le pays et contribue à une cartographie prévisionnelle.

Verily, société appartenant à Alphabet propose ses services pour l'aide à la détection des malades potentiels à San Francisco en participant au projet baptisé *ProjectBaseline*, à destination de l'état de Californie, qui utilise les données des utilisateurs (et prétend les traiter avec de l'IA) afin de développer des technologies de télédiagnostics.

On relève de multiples autres initiatives plus ou moins ciblées pour répondre à des demandes spécifiques à l'aide d'outils numériques à grande échelle dans lequel l'IA joue un rôle plus ou moins "marketing" ; mentionnons par exemple le travail de Feifei LI, co-directeur du HAI (*Human AI*) de l'Université Stanford pour une approche IA permettant de conseiller les personnes âgées à domicile, mais en respectant des normes éthiques et de sécurité, ou l'outil CURAI prétendant proposer des diagnostics personnalisés. Pour plus de pistes sur les développement d'outils numériques ciblés, le MIT a organisé un Hackathon.

Utilisation de l'IA pour l'aide au diagnostic par analyse automatique de *CT scans*.

Le diagnostic le plus fiable et le plus utilisé pour diagnostiquer un patient au SARS-CoV-2 est un test génétique dit « RT-PCR ». Dans beaucoup d'endroits du monde, ceux-ci se sont malheureusement rapidement révélés être disponibles en trop faible quantité. Par contre, les machines radiologiques sont présentes dans tous les hôpitaux, et des médecins ont donc voulu y substituer des diagnostics se basant sur des images à rayons X des poumons, dits « *CT scans* ». Il s'agit pour le radiologue d'essayer de déceler dans ces radiographies des poumons des anomalies caractéristiques des pneumonies.

Dans le cas du SARS-CoV-2, des initiatives sont rapidement apparues pour entraîner des algorithmes à reconnaître dans les images de *CT scans* les patterns

caractéristiques de l'affection afin d'assister les praticiens. Les premiers cas ayant été diagnostiqués en Chine, c'est dans ce pays où les premières données ont été générées, et donc là où les premiers algorithmes d'analyse automatique de *CT scans* sont donc apparus en Chine : Alibaba, Baidu, inferVISION (société sino-américaine), PingAn healthcare, Deepwise Technology, Iflytek, ont publié ainsi des performances impressionnantes pour cette tâche de détection de l'affection au SARS-CoV-2 dans des images de *CT scans.* De tels systèmes ont été déployés dans plusieurs hôpitaux en Chine. Des bases de données ont ensuite commencé apparaître sur le site de challenge Kaggle, qui comporte des images de *CT scans* relatives au coronavirus. Puis, ces applications ont suivi en occident, l'université de Waterloo a ainsi mis en ligne un réseau de neurones ouvert pour l'aide au diagnostic, et beaucoup d'entreprises partout dans le monde, comme Qure.ai, proposent des solutions.

Si les résultats annoncés sont impressionnants, l'impact réel de ces outils dans la lutte contre le COVID-19 semble faible. En effet, pour être réellement exploitables et apporter un gain de temps au radiologue, ces outils doivent bien s'insérer dans le workflow de celui-ci. Pour cela, les outils d'apprentissage profond pour l'imagerie médicale nécessitent de très grandes quantités de données de bonne qualité. Dans le cas présent, ces données consistent en un jeu de *CT scans* annotés. Ces annotations doivent être faites par un expert radiologue, et demandent beaucoup de temps, d'autant que ce travail doit être fait conjointement par le radiologue, et les ingénieurs qui vont désigner l'outil et entraîner le réseau de neurones. Ainsi, dans un article (https://arxiv.org/abs/2003.11336) cosigné par l'OMS, l'*United Nations Global Pulse*, et le *Montreal's Mila institute for AI*, les auteurs concluent que peu de ces systèmes ont atteint ce niveau de maturité à cause de données trop peu nombreuses, et trop dispersées. Une collaboration renforcée permettant des bases de données plus ouvertes permettrait de combler ce manque. Ainsi, beaucoup d'efforts ont été effectués qui permettront probablement à l'avenir d'améliorer de tels systèmes, mais l'aide apportée par l'exploitation des données pour l'estimation de modèles d'apprentissage profond sera probablement limité. Nous n'avons pas trouvé de trace d'utilisation de l'IA pour le diagnostic en cours en situation réelle aux Etats-Unis. Le journaliste Devan Coldeway a ainsi dit que, ces technologies n'étant pas assez mûres, '"aucun patient n'aura de diagnostic du COVID-19 par une IA" aux Etats-Unis. (AI and big data won't work miracles in the fight against

coronavirus)

Prédiction de risque de troubles respiratoires par IA

Outre l'aide au médecin pour l'analyse d'images médicales, une grande classe d'application de l'IA est la détermination de patients à risques sur la base d'un ensemble de critères plus ou moins vastes : âge, sexe, taux détectés dans des analyses sanguines, etc. Des travaux dans ce sens ont été menés par des chercheurs de l'université de New-York en collaboration avec des hôpitaux chinois, sur la base de données chinoises, et ont abouti à des modèles permettant de prédire avec une précision significative la survenue de complications graves chez les patients atteints du COVID-19. Contrairement à ce qu'on pouvait imaginer au vu des premières observations, l'âge et le genre ne semblent pas être les critères les plus importants d'après les modèles, mais plutôt des indicateurs moins intuitifs comme les niveaux de certaines enzymes du foie ou le type d'hémoglobine. Ce travail ne peut directement donner naissance à un outil réellement fiable à court terme, mais il montre que les méthodes d'IA peuvent permettre de détecter des motifs (*patterns*) statistiques inattendus et par là faire avancer notre compréhension de la maladie.

En Chine, des chercheurs ont proposé avec ce type de modèles un outil permettant d'aider les médecins à déterminer les chances de survie respectives de patients en détresse respiratoire à partir d'un prélèvement sanguin (Should AI help make life-or-death decisions in the coronavirus fight?). Cet algorithme, détectant les patients les plus à risques, a fourni de bons résultats à l'hôpital de Tongji à Wuhan. Mais là encore, l'utilisation d'un algorithme mal "expliqué" et reposant sur un jeu de données relativement faible reste délicate quand il s'agit de l'utiliser pour prendre des décisions engageant le pronostic vital d'un patient, même en situation d'urgence.

Aux Etats-Unis, nous n'avons cependant identifié aucune utilisation de ce type de l'IA en situation réelle. Pour aborder ce type de questions, le Colorado a créé des directives afin de déterminer quel patient traiter en priorité dans le cas où le nombre de patients en détresse respiratoire était trop élevé par rapport aux capacités d'accueil de l'hopital (Colorado Is Creating Guidelines To Help Make

Excruciating Coronavirus Care Decisions), mais ces directives ne reposent pas sur des algorithmes d'apprentissage.

Suivi (*contact tracing*) des patients infectés et gestion des données individuelles

Afin de renforcer la lutte contre l'épidémie de COVID-19, le suivi individuel, ou « *contact tracing* », est à présent ouvertement envisagé par l'exécutif français. Il permet de suivre les personnes infectées, avec qui elles sont ou ont été en contact, ou de surveiller l'application des mesures de confinement et de distanciation sociale.

L'idée la plus simple consiste à collecter les informations sur les déplacements individuels à l'aide des GPS des smartphones. De telles technologies ont été adoptées très tôt dans la lutte contre le coronavirus par les autorités asiatiques et elles sont présentées comme relativement efficaces. Aux Etats-Unis un tel suivi individuel n'a pas été imposé de manière unilatérale par le gouvernement mais l'urgence, l'importance des enjeux et le flou actuel sur des réglementations suscite d'âpres discussions le compromis entre la santé publique la protection de la vie privée. On pourra voir à ce sujet cet article sur CNBC ou ce compte-rendu du think-tank Carnegie. Le World Economic Forum a publié, par l'intermédiaire de Kay Firth-Butterfield, qui en dirige le groupe IA et machine learning, une déclaration exhortant les entreprises à ne pas s'éloigner d'une utilisation saine des données pour agir plus rapidement: "*We need to keep in mind that the big ethical challenges around privacy, accountability, bias, and transparency of artificial intelligence remain*".

Dans cette optique, les chercheurs, notamment Ramesh Raskar au MIT, ont proposé des protocoles qui permettraient de crypter suffisamment les informations stockées pour préserver l'anonymisation. Face à cette difficulté s'est développée aussi l'idée, portée notamment par Cristina White de l'Université de Stanford, d'exploiter non pas la position absolue des individus mais leurs positions relatives : le réseau Bluetooth de chaque appareil est capable de détecter la proximité d'autres appareils. Si deux personnes se sont croisées, les deux appareils peuvent garder la trace de l'événement sans même collecter les positions GPS. Si l'un des deux se

révèle infecté et le signale, l'autre peut-être prévenu (modulo validation d'un médecin). Un algorithme astucieux de stockage des informations collectées permet de rendre le processus complètement anonyme : seules les personnes concernées reçoivent finalement l'information les concernant. Cela semble être la solution la plus prometteuse dans la mesure où elle pourrait satisfaire les exigences de toutes les parties.

A partir de ces considérations, plusieurs solutions techniques se dessinent et risquent de se trouver en concurrence. La solution du MIT, *Private Kit: Safe Paths* mêle information GPS cryptée et données Bluetooth, le consortium *Covid-Watch* s'est formé autour de la solution Bluetooth développée par Stanford mais la startup californienne *Nodle* développe de son côté l'application *Coalition* sur le même principe alors que diverses solutions open-source sont regroupées sur la plateforme *github*. Les choix techniques des opérateurs Google et Apple sur le sujet ne sont pas explicites et le paysage est tel qu'on ne voit pas comment une réponse unique pourra s'imposer. L'accord récent entre les deux géants pour assurer une interopérabilité de leurs systèmes Bluetooth (à l'aide d'une application à télécharger dès le mois de mai, avant de l'insérer dans les systèmes d'exploitation dans les mois qui viennent) suggère qu'une couche supplémentaire autorisant une interaction de plusieurs solutions permettrait de tirer le meilleur parti de cette concurrence et offrirait la perspective d'une efficacité à l'échelle internationale.

Une startup américaine propose aussi une surveillance du respect des mesures de distanciation sociales s'appuyant sur le réseau existant de caméras de surveillance: Social distancing detection for COVID-19.

Produire des traitements thérapeutiques / un vaccin.

Un autre pan d'applications de l'analyse de données est son utilisation pour le développement et la production de traitements thérapeutiques ou de vaccins. En effet, l'IA est depuis plusieurs années envisagée comme un outil essentiel pour aider à accélérer la découverte et la production de nouvelles molécules thérapeutiques,

repositionner des thérapies existantes, mais aussi aider à optimiser l'efficacité des essais cliniques. Rappelons qu'en moyenne, moins de 10% des médicaments proposés passent avec succès les tests cliniques, et l'IA est un candidat naturel pour essayer d'augmenter ce taux. Dans les circonstances de crise actuelles, ces initiatives d'utilisation de l'IA pour la pharmacologie se focalisent actuellement sur le coronavirus.

Décodage de la structure du SARS-CoV-2

Une première étape dans la recherche de traitements contre le coronavirus est de comprendre sa structure. Des laboratoires de recherche, mais aussi des grandes et petites entreprises américaines ont travaillé pour identifier les protéines codées par les gènes du SARS-CoV-2, voir avec quelles protéines humaines elles interagissent au cours de l'infection, ce qui permet d'envisager des traitements qui permettent de perturber le cycle de réplication du virus.

DeepMind a ainsi annoncé qu'elle rendait publique le décodage des structures de plusieurs protéines liées au COVID-19 afin faciliter les travaux de recherche de tous les laboratoires qui travaillent sur l'analyse du SARS-CoV-2. Ces structures ont été « prédites » avec la dernière version du système *AlphaFold*, un algorithme qui permet d'estimer rapidement la structure de nouvelles protéines. La startup *Insilico*, qui en temps normal utilise l'IA pour travailler davantage sur les problématiques de vieillissement, a également publié six structures de protéines du coronavirus.

De même, notons que grâce à la plateforme de prédiction de structures protéiques Robetta qui utilise des algorithmes à réseau de neurones, l'Université de Washington et son *Institute for Protein Design* ont réussi à créer une carte à l'échelle atomique de la protéine S (protéine de spicule ou *spike protein*) partie du virus qui lui permet "d'agripper" et infecter les cellules humaines. Suite à ces résultats, les scientifiques cherchent à développer des protéines pouvant neutraliser le coronavirus en s'attachant aux protéines S avant infection des cellules saines.

Dans l'optique de mieux comprendre le virus, l'entreprise *Adaptive Biotechnologies* basée à Seattle utilise les capacités de Machine Learning offertes par la plateforme cloud de Microsoft Azure pour passer en revue les données existantes sur les

récepteurs des cellules en T humaines et analyse la réponse de ces récepteurs au virus. Cette technique éprouvée en 2018 pour la maladie de Lyme devrait permettre de cartographier les réponses immunitaires au SARS-CoV-2 à l'échelle de la population américaine : pour plus de détails, voir l'article de Pureai.

En dépit des différentes avancées permises par la science des données et l'IA, notons que la discussion reste ouverte quant à leur valeur ajoutée pour le développement d'un vaccin.

D'un côté, Kate Broderick, Vice-Présidente pour la R&D de l'entreprise biotech de San Diego Inovio Pharmaceuticals (interview) ou Stephen Hoge, président de l'entreprise Moderna (article) pensent que la force des nouvelles méthodes numériques qui permettent de produire le design complet d'un vaccin à partir d'un algorithme d'apprentissage automatique en quelques heures à partir du code génétique du virus ouvre des perspectives assez larges.

Au contraire, David Baker, directeur de l'*Institute for Protein Design*, estime que les chances sont faibles pour que l'IA soit utilisée efficacement pour les recherches de vaccin (How AI is helping scientists in the fight against COVID-19, from robots to predicting the future). Il considère que son usage sera plus fructueux pour les recherches liées aux traitements médicamenteux.

Création de nouvelles molécules, repositionnement thérapeutique

Face au virus, une autre approche thérapeutique qui bénéficie de l'usage d'outils numériques est la recherche de molécules pouvant interagir avec celui-ci. Beaucoup d'entreprises essaient, par exemple, de créer de nouvelles molécules susceptibles de ralentir la réplication du virus.

Le site spécialisé BenchSci répertorie 152 *startup* américaines (parmi 230) qui utilisent l'IA pour découvrir de nouveaux médicaments. Cette utilisation de l'IA peut être effectuée de beaucoup de façons différentes : pour modéliser le comportement d'une cellule malade et voir comment des médicaments peuvent la guérir, pour explorer automatiquement un grand nombre de molécules et recouper certaines de

leurs propriétés, pour analyser automatiquement de larges bases de données de brevets, d'articles, et de données cliniques pour en tirer des relations de causalité entre des gènes, des maladies, des protéines, etc.

De façon non exhaustive, citons deux exemples de telles initiatives :

IBM Recherche a utilisé de l'IA pour identifier 3000 nouvelles molécules candidates pour contrer la réplication du SARS-CoV-2. Ces molécules ont été publiées en open source et les chercheurs peuvent les étudier dans un outil de visualisation ad hoc (IBM Releases Novel AI-Powered Technologies to Help Health and Research Community Accelerate the Discovery of Medical Insights and Treatments for COVID-19).

Une autre initiative entre dans le cadre d'une coopération entre Iktos, une startup française lauréate 2018 de NETVA, l'accélérateur du Service pour la Science et la Technologie de l'Ambassade de France aux Etats-Unis. Iktos est une société d'IA pour la conception de médicaments, elle est basée sur des réseaux de neurones profonds et permet de concevoir automatiquement de nouvelles molécules virtuelles candidates à être un médicament contre le SARS-CoV-2. La coopération avec SRI International permet de coupler cette technologie avec le système de synthèse chimique de SRI, accélérant ainsi le développement de produits chimiques, et apportant de nouveaux médicaments à la clinique plus rapidement et à moindre coût. Cette plate-forme, appelée SynFini, automatise la conception, le criblage et l'optimisation des réactions, et la production de molécules cibles. (Iktos and SRI International Announce Collaboration to Combine Artificial Intelligence and Novel Automated Discovery Platform for Accelerated Development of New Anti-Viral Therapies – SRI)

Cependant, toute nouvelle molécule, même conçue à l'aide de l'IA, devra passer par des tests cliniques ce qui implique des délais supplémentaires avant d'obtenir un traitement autorisé. Par conséquent, une approche alternative poursuivie par d'autres organismes partout dans le monde est de faire ce que l'on appelle du *repositionnement thérapeutique* (médicaments dont le développement a été arrêté, réutilisé pour un autre traitement) ou de la *réutilisation de médicaments* (réutilisation de médicaments autorisés pour un autre traitement) pour accélérer le

déploiement d'un traitement. Dans ces deux cas, l'utilisation de l'IA pour analyser les données existantes sur les médicaments est utile afin d'identifier des médicaments déjà autorisés et les interactions médicamenteuses déjà connues.

Des chercheurs américains et sud-coréens ont ainsi utilisé des algorithmes de réseaux de neurones profonds pour modéliser l'interaction entre le médicament et sa cible, appelé « Molecule Transformer-Drug Target Interaction (MT-DTI) ». Cela a permis d'identifier l'atazanavir, un antiviral existant, qui pourrait être repositionné contre le SARS-CoV-2 (Predicting commercially available antiviral drugs that may act on the novel coronavirus (2019-nCoV), Wuhan, China through a drug-target interaction deep learning model).

Des travaux qui nécessitent de la puissance de calcul

Afin de pouvoir faire tourner les algorithmes et obtenir les différents résultats mentionnés, il est important de noter qu'une grande puissance de calcul est nécessaire en plus des données. Dans cette optique, le consortium *COVID-19 High Performance Computing*, mis en avant par le Président Trump le 23 mars dernier, mobilise la puissance informatique du DoE, la NSF, la NASA, de plusieurs universités (UT Austin, MIT, UC San Diego) et de grands groupes (IBM, Microsoft, Google, Amazon) pour soutenir des algorithmes de recherche contre le COVID-19. Sur la page du consortium, de nombreux projets en cours sont présentés dont la plupart concernent du repositionnement thérapeutique. L'utilisation de cette puissance de calcul est ouverte à des projets de recherche venant de l'étranger. Pour plus de détails on pourra se référer à la NDI-2020-0177610. Une initiative similaire quoique de moindre ampleur mais impliquant un chercheur français, Jérôme Baudry, est en cours en Alabama où l'University of Alabama à Huntsville (UAH) met à disposition ses supercalculateurs pour les chercheurs et les entreprises de biotechnologie d'Alabama.

Comme dans le cas du diagnostic, nous voyons en conclusion que l'IA est mobilisée dans le cadre de la pharmacologie pour tenter de répondre à la crise du COVID-19, mais compte tenu des délais incompressibles de tests des molécules qui pourraient être ainsi conçues, il est très probable que celles-ci ne seront disponibles qu'une fois

le pic de la crise passée.

Ce tour d'horizon donne un aperçu du foisonnement des initiatives de l'écosystème américain du numérique en réponse à la crise. On perçoit aussi comment ces initiatives se structurent petit à petit, même si, à ce stade, aucune action unifiée n'apporte de solution globale. Le dynamisme de l'écosystème et la violence de la crise contribue à faire avancer les débats sur l'usage qui peut être fait des données, en terme de partage ou d'utilisation par des algorithmes.

Comme cela a été souligné à plusieurs reprises, les avancées de l'IA de ces dernières années ne contribuent qu'à la marge à résoudre les problèmes urgents qui se posent, en partie justement parce qu'on dispose de peu de données pour la phase d'apprentissage quand on aborde un problème nouveau. Pourtant les expérimentations entreprises pendant cette période particulière, certaines de celles mentionnées ici lorsqu'elles auront produit leurs effets, comme d'autres moins directement tournées vers la lutte contre la pandémie mais induites par celle-ci, notamment par la vie en confinement, apporteront vraisemblablement un lot d'innovations qui auront un impact au delà de la crise.

Auteurs : Jean-Baptiste Bordes (SST San Francisco), Xavier Bressaud et Kevin Kok Heang (SST Washington).