L'OSTP met en place deux programmes public-privé de grande envergure pour accélérer la recherche scientifique

Dès le début de la propagation du SARS-CoV-2 aux Etats-Unis, plusieurs **agences gouvernementales** ont été **ouvertement critiquées**, notamment : (i) la **FDA** (*Food and Drug Administration*) pour avoir complexifié les procédures de validation des tests malgré l'urgence de la situation ; et (ii) les **CDC** (*Centers For Disease Control and Prevention*), l'une des principales composantes opérationnelles du ministère de la santé et des services sociaux, pour avoir pris des décisions qui ont considérablement ralenti la mise en place des tests RT-PCR pour diagnostiquer le COVID-19[1][2][3]. D'autre part, sur le **plan épidémiologique** les Etats-Unis étaient en train de devenir le pays ayant le plus grand nombre de cas de COVID-19 et de mortalité associée au monde. Dans ce contexte difficile, l'**OSTP** (*Office for Science and Technology Policy*) a initié plusieurs initiatives dont 2 partenariats public-privé de grande envergure pour accélérer la recherche scientifique :

« Appel à l'action » pour l'analyse de la littérature scientifique

Via l'OSTP, la maison blanche a constitué un consortium regroupant les acteurs suivants :

- la *National Library of Medicine* (NLM), la plus grande bibliothèque médicale du monde (https://www.nlm.nih.gov),
- le *Allen Institute for AI*, un institut de recherche indépendant à but non lucratif (https://allenai.org/),
- la *Chan Zuckerberg Initiative* (CZI), qui se définit comme « entreprise philanthropique » visant à apporter des solutions technologique à la société

(https://chanzuckerberg.com/),
- le *Center for Security and Emerging Technology* (CSET) de *Georgetown University* (https://cset.georgetown.edu),
- *Microsoft*

Un « appel à l'action » (*call to action*) a été lancé le 16 mars 2020 auprès de la communauté technologique (ex : en sciences des données, intelligence artificielle) afin que de **nouvelles techniques** d'exploration de **textes** (ex: *Natural Language Processing* ou NLP) et de **données** soient développées pour aider la communauté scientifique à accéder à la littérature pertinente et donc combattre le COVID-19[4].

L'OSTP a dans ce contexte identifié dix questions prioritaires émanant de l'OMS et du comité sur les maladies infectieuses émergentes et les menaces sanitaires du 21ième siècle de l'académie des sciences (*National Academies of Sciences, Engineering, and Medicine)* (voir Annexe 1).

Le consortium met à la disposition des analystes de données un corpus traitant du COVID-19 et des coronavirus (dont SARS, MERS, etc.) publiés dans les bases de données PubMed, bioRxiv, ou medRxiv notamment. Initialement constitué de 29.000 articles (et 44.000 méta-données), le corpus est actuellement composé de plus de 69.000 publications en texte intégral[5] (avec plusieurs sites miroirs dont au MIT) et le nombre d'articles ne cesse de s'accroître. Le **COVID-19 Open Research Dataset** (CORD-19) a été téléchargé plus de 75.000 fois depuis sa mise-en-ligne[6], et plus de 30 mises-à-jour[7] ont été effectuées[8].

Ce corpus est disponible sur la plateforme web organisant des compétitions en science des données : *Kaggle* (Alphabet/Google, https://www.kaggle.com/) dans un format (JSON) déjà accessible pour les algorithmes. Pour chaque question, *Kaggle* offre une récompense de 1 000 $ à l'équipe dont les performances sont les meilleures selon 3 critères principaux : justesse et précision des résultats ; documentation de la méthodologie et possibilités de réutilisation du code ; efficacité et qualité de la présentation des résultats (dont visualisation des données).

Mise à disposition du calcul haute performance

L'OSTP a également fédéré une initiative centrée sur le calcul haute performance (https://covid19-hpc-consortium.org/). La gouvernance est répartie entre 2 co-présidents -Paul Dabbar (DoE) et Dario Gil (IBM)- et un directeur exécutif Barb Helland (DoE)[9]. Cette initiative réunit des agences fédérales (NSF et NASA), les laboratoires nationaux (dont Los Alamos et Oak Ridge), des industriels (*Amazon Web Services*, *Google*, *Hewlett Packard*, *IBM*, *Microsoft*, etc.) et des universités (ex. : MIT, UT Austin, UCSD) qui offrent du temps et des ressources de calcul (ex : capacité de stockage) sur leurs supercalculateurs pour des recherches computationnelles complexes (voir la liste des partenaires en annexe 2). Ceci permet l'accès à une partie des ressources des 40 supercalculateurs du réseau constitué, représentant plus de 480 pétaflops (soit 10e15 FLOPS, *FLoating-point Operations Per Second* ou nombre d'opérations en virgule flottante par seconde) ; 130 000 nœuds ; 5 millions de cœurs de CPU (*Central Processing Unit*) et 50 000 GPU (*Graphics Processing Unit*)[10].

Les chercheurs sont invités à soumettre des projets de recherche sur le COVID-19 via le portail en ligne XSEDE dédié (https://www.xsede.org/covid19-hpc-consortium). Les critères d'évaluation sont : (i) avantages potentiels pour la réponse à COVID-19 ; (ii) faisabilité de l'approche technique ; (iii) nécessité du calcul haute performance ; (iv) connaissances et expérience en matière de calcul haute performance de l'équipe projet ; (v) estimation des besoins en ressources informatiques.

Priorisant les projets, un groupe d'experts composé de scientifiques de divers horizons et de chercheurs en informatique, travaille ensuite avec l'équipe projet pour : évaluer les avantages des travaux pour la santé publique, identifier l'institution partenaire ayant les ressources informatiques adéquates.

Les projets priorisés sont ceux pouvant garantir des résultats rapides dans des domaines comme la bioinformatique, l'épidémiologie et la modélisation moléculaire pour « comprendre la menace à laquelle nous sommes confrontés et élaborer des stratégies pour y faire face ».

La liste actuelle, de plus de 60 projets, concerne principalement les aspects de conception de médicaments (*drug design*) et de découverte de médicaments (*drug discovery*, dont repositionnement de molécules)[11].

Auteur : Renaud SEIGNEURIC (SST Houston)

Notes :

[1] https://www.nature.com/articles/d41586-020-00823-w

[2] http://perkinslab.weebly.com/uploads/2/5/6/2/25629832/perkins_etal_sarscov2.pdf

[3] https://www.nature.com/articles/d41586-020-00772-4

[4] https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/

[5][5] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[6] https://www.semanticscholar.org/paper/CORD-19%3A-The-Covid-19-Open-Research-Dataset-Wang-Lo/bc411487f305e451d7485e53202ec241fcc97d3b?utm_campaign=CORD-19&utm_medium=email&_hsmi=88563295&_hsenc=p2ANqtz-_T6eMa6pyj-n72tihUsHe363PzeJ1H1MWemT6ic5f3Qnnf7AlMwWBTW8YqhEqq09b1Ic-tXyFoPRXRH1m5SiR5Z1dWgVv8MujlCpotprsPlgxiUZc&utm_content=88563295&utm_source=hs_email

[7] https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases.html

[8] Au 16 juin 2020

[9] https://covid19-hpc-consortium.org/who-we-are

[10] https://covid19-hpc-consortium.org/ Dernière mise-à-jour au 16 juin 2020.

[11] https://covid19-hpc-consortium.org/projects Dernière mise-à-jour au 16 juin 2020.

Annexe 1

Liste des 10 questions prioritaires ou *Tasks* (https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks) (au 14 avril 2020) :

Questions prioritaires
Que sait-on sur la transmission, l'incubation et la stabilité environnementale ?
Que savons-nous des facteurs de risque du COVID-19 ?
Que savons-nous sur la génétique, l'origine et l'évolution des virus ?
Que savons-nous des vaccins et des approches thérapeutiques ?
Qu'est-ce qui a été publié sur les soins médicaux ?
Que savons-nous des interventions non pharmaceutiques ?
Que savons-nous sur le diagnostic et la surveillance ?
Aidez-nous à comprendre comment la géographie affecte la propagation du virus

Qu’est-ce qui a été publié sur les considérations éthiques et de sciences sociales ?
Qu’est-ce qui a été publié sur le partage de l’information et la collaboration intersectorielle ?

Annexe 2

Liste des partenaires du consortium public-privé utilisant les supercalculateurs (https://covid19-hpc-consortium.org/), dernière mise-à-jour au 16 juin 2020 :

Industry

- IBM
- Amazon Web Services
- AMD
- BP
- D. E. Shaw Research
- Dell Technologies
- Google Cloud
- Hewlett Packard Enterprise
- Microsoft
- NVIDIA
- Intel

Academia

- Massachusetts Institute of Technology
- Rensselaer Polytechnic Institute
- University of Illinois
- University of Texas at Austin
- University of California – San Diego
- Carnegie Mellon University
- University of Pittsburgh
- Indiana University
- Massachusetts Green High Performance Computing Center (MGHPCC)
- University of Wisconsin-Madison
- Ohio Supercomputer Center
- UK Digital Research Infrastructure
- CSCS – Swiss National Supercomputing Centre

Department of Energy National Laboratories

- Argonne National Laboratory
- Lawrence Livermore National Laboratory
- Los Alamos National Laboratory
- Oak Ridge National Laboratory
- Lawrence Berkeley National Laboratory
- Sandia National Laboratories
- Idaho National Laboratory

Federal Agencies

- National Science Foundation
    - XSEDE
    - Pittsburgh Supercomputing Center (PSC)
    - Texas Advanced Computing Center (TACC)
    - San Diego Supercomputer Center (SDSC)
    - National Center for Supercomputing Applications (NCSA)
    - Indiana University Pervasive Technology Institute (IUPTI)
    - Open Science Grid (OSG)

        - National Center for Atmospheric Research (NCAR)
- NASA