ACADÉMIE
POUR UN PROGRÈS RAISONNÉ CHOISI PARTAGÉ DES TECHNOLOGIES

CALCUL ET DONNÉES : NOUVELLES PERSPECTIVES POUR LA SIMULATION NUMÉRIQUE À HAUTE PERFORMANCE

RAPPORT DE L'ACADÉMIE DES TECHNOLOGIES

couverture : AdobeStock - Gorodenkoff Productions OU

ACADÉMIE DES TECHNOLOGIES

POUR UN PROGRÈS
RAISONNÉ
CHOISI
PARTAGÉ

CALCUL ET DONNÉES : NOUVELLES PERSPECTIVES POUR LA SIMULATION NUMÉRIQUE À HAUTE PERFORMANCE

Rapport de l'Académie des technologies

décembre 2020

Académie des technologies
Le Ponant – Bâtiment A
19, rue Leblanc
75015 PARIS
+33(0)1 53 85 44 44
secretariat@academie-technologies .fr
www .academie-technologies .fr

ISBN : 979-10-97579-23-4

Sommaire

Première partie – Analyse du contexte 5

Introduction et évolution générale du contexte 5

Les données en masse 10

Les tendances actuelles dans le domaine des architectures de calcul 11

Quels sont les types de données à considérer, et quelles sont leurs valeurs d'usage ? 14

L'importance grandissante du traitement et de l'exploitation de données massives, l'apprentissage automatique 18

La convergence nécessaire des approches 21

Les méthodes de modélisation hybride 22

Vers de nouvelles approches pour la validation des simulations 25

Un besoin de nouvelles compétences, nécessaires aux ambitions industrielles 28

Deuxième partie – Synthèse et recommandations 31

Printemps 2020 : les États-Unis mettent en place le « *COVID-19 High-Performance-Computing consortium* »[1], qui permet de mettre à la disposition de la recherche une puissance de calcul de plus de 400 petaflop/s répartie sur plus de quatre millions de cœurs de calcul ; l'initiative «*Folding@home*»[2], lancée par l'université de Stanford, permet de mettre en commun la puissance de calcul d'ordinateurs personnels, de téléphones portables, de consoles de jeu… atteignant ainsi un total de plus de 1 exaflop/s, pour simuler par les méthodes du calcul parallèle le repliement des protéines et accélérer la recherche sur le virus ; l'Europe met en place une filière d'urgence pour soutenir, à travers son initiative PRACE, des projets de recherche consacrés à améliorer la maîtrise des impacts de la maladie[3] ; en France la société GENCI en fait de même, en mettant à disposition des chercheurs académiques et industriels français ses moyens de calcul et de stockage pour leurs travaux de simulation, de traitements de données et d'usage de l'intelligence artificielle[4].

La simulation numérique reste donc toujours au cœur de la recherche scientifique et des enjeux de société, avec des méthodes d'approche de plus en plus variées et performantes.

1 https://covid19-hpc-consortium.org/

2 https://foldingathome.org/covid19/

3 https://prace-ri.eu/prace-support-to-mitigate-impact-of-covid-19-pandemic/

4 http://www.genci.fr/fr/content/projets-contre-le-covid-19

Remarques de lecture

La première partie de ce rapport analyse le nouveau contexte de la simulation numérique, en particulier vis-à-vis de son utilisation dans le domaine industriel et sociétal. L'accent est principalement mis sur les aspects « haute performance » (HPC, ou *High Performance Computing*), dont les exigences sont fondatrices en matière d'infrastructure informatique, de taille des problèmes traités et de précision des méthodes. Ainsi dans une première partie les analyses portent tant sur les architectures de calcul utilisables, que sur les questions liées au traitement de données massives, et enfin que sur les nouvelles méthodes de simulation elles-mêmes. Cette première partie est constituée d'une présentation assez générale permettant d'identifier de façon synthétique les principales évolutions du contexte ; cette présentation générale est suivie de paragraphes plus spécifiques et plus techniques sur chacun des points évoqués, qu'un lecteur moins intéressé pourra ignorer avant de passer à la seconde partie. Dans cette seconde partie sont développés des messages et recommandations, tous relatifs, à un titre ou à un autre, à la prise en compte des données.

L'approche suivie permet de resserrer le périmètre du rapport et ainsi de sérier quelques recommandations sur des sujets actuellement très présents. De ce fait ce rapport n'aborde guère les domaines de la simulation plus éloignés du HPC, mais souvent complémentaires, comme ceux de l'automatique, de l'ingénierie ou encore de la conception système.

Première partie
Analyse du contexte

Introduction et évolution générale du contexte

Le dernier rapport complet que l'Académie des technologies a consacré à la simulation numérique, à ses enjeux et à ses applications, date de 2005[1]. Quinze ans plus tard, le paysage, tant national qu'européen et mondial, que scientifique et technique, a beaucoup évolué. De nombreux documents d'analyses et de propositions ont suivi, ou précédé, ces évolutions (annexe I). Sans vouloir produire un document exhaustif, l'Académie des technologies a toutefois estimé utile de faire un point sur les modifications intervenues au cours des cinq dernières années. C'est ainsi qu'a été mis en place en mars 2018 un groupe de travail transverse sur la « Simulation numérique » (annexe II), qui s'est réuni depuis près d'une vingtaine de fois et a auditionné quinze personnalités (annexe III), scientifiques et techniciens, responsables publics et industriels, afin d'appréhender au mieux les enjeux, les perspectives, les défis et les nouvelles méthodologies en cours de mise en place dans ce domaine. À l'issue de ce travail, il a semblé possible de tirer un certain nombre d'enseignements à partager avec les utilisateurs de la simulation numérique, aujourd'hui conçue dans un sens beaucoup plus large qu'il y a une décennie. Et aussi de formuler quelques recommandations, s'adressant principalement aux pouvoirs publics, aux agences nationales et aux développeurs. C'est le sens des six messages qui concluent ce nouveau rapport.

1 « Enquête sur les frontières de la simulation numérique, La situation en France et dans le monde, diagnostics et propositions », rapport de l'Académie des technologies, Juin 2005

L'ÉVOLUTION GÉNÉRALE DU CONTEXTE

Plusieurs modifications profondes, voire ruptures, sont apparues au cours des dix dernières années.

Les données massives

L'élément nouveau est la disponibilité de données numériques de plus en plus nombreuses, et dotées d'un nouveau contenu informationnel qui reste à exploiter de façon optimale. Ces données peuvent provenir d'horizons variés, par exemple la simulation numérique elle-même, capable de générer, grâce à des simulations à haute-fidélité dans un domaine particulier, des informations qui pourront ensuite être utilisées, d'une façon restant pour une large part à concevoir, pour des simulations de systèmes plus complexes. Mais, plus encore, elles proviennent maintenant des nouveaux systèmes d'observation, de mesure, de contrôle, capables de délivrer un flux d'information d'intensité toujours croissante. L'observation spatiale de la terre en est un exemple bien connu, à côté de nouvelles sources comme l'internet des objets ou des applications industrielles en fort essor. L'adaptation à ce nouvel univers des données massives (*big data*) ouvre de nombreuses interrogations :

- quels types d'ordinateurs pour traiter ces flux de données ?
- comment apprécier la valeur de ces données ?
- quelles méthodes mathématiques pour les traiter ?
- comment faire converger ces méthodes avec celles du calcul numérique plus traditionnel et comment adapter les grands codes de calcul, scientifiques et industriels ?
- comment apprécier et préciser les limites de validité de leurs résultats ?
- comment former les scientifiques, industriels et techniciens à ces approches ?

En bref, ce rapport tente de proposer des réponses aux questions posées par les nouveaux enjeux auxquels est confrontée la simulation numérique : comment articuler, d'une part, données, information et connaissance, et, d'autre part, causalité et corrélation ?

Les supercalculateurs

L'évolution des supercalculateurs modifie elle aussi le contexte de façon importante. La transition vers des puissances informatiques en augmentation continue pour répondre aux besoins scientifiques et industriels conduit à l'émergence très prochaine de supercalculateurs dans la gamme de l'*exascale* (10^{18} opérations par seconde). La croissance de la puissance informatique ne peut plus se construire, depuis une dizaine d'années, sur l'augmentation de la fréquence de calcul des processeurs de base, pour des raisons liées à leur trop grande consommation énergétique. Cette augmentation de puissance est maintenant atteinte grâce à la multiplication du nombre de processeurs travaillant en parallèle. Cependant, le parallélisme des calculs ne pouvant être totalement efficace, certains des processeurs se trouvent inactifs pendant le

déroulement des simulations. Ce n'est donc que par la forte multiplication de leur nombre, et en parallèle de logiciels adaptés à cette architecture, que peut être atteinte la puissance informatique recherchée ; ainsi, fin 2020, le nombre de processeurs des supercalculateurs les plus puissants atteint dix millions. Cet « hyperparallélisme » conduit lui aussi à des consommations énergétiques qu'il convient de maîtriser, par exemple en utilisant des processeurs plus économes comme les « accélérateurs », issus pour une large part du monde de la visualisation et du traitement des données. Le transfert des données d'un processeur à l'autre étant par ailleurs très consommateur d'énergie, il faut rapprocher le plus possible les données, et leur traitement, du calcul dans lequel elles sont utilisées. Il en résulte des architectures de supercalculateurs tout à fait nouvelles, hyperparallèles et hybrides, hybride au sens des processeurs utilisés. Leur utilisation nécessite de nouvelles méthodes de programmation.

La valeur des données

Les données ont un coût d'obtention (production, conservation), dont une fraction importante concerne les ressources humaines à mettre en œuvre. Il est donc nécessaire d'en tenir compte pour traiter de l'ouverture et de la gratuité de l'accès aux données d'une part, et de leur circulation et du droit de propriété d'autre part. L'accès « libre et gratuit » aux données issues de la recherche publique pourrait ainsi être envisagé de façon sélective, éventuellement avec contreparties adaptées, de natures variables. Les données associées aux activités d'une entreprise, et à leur croisement avec des données publiques, sont une source de compétitivité. Pourtant, la tentation de garder ces données de façon protégée au sein de la seule entreprise doit être mise en balance avec l'avantage potentiellement très important de les croiser avec des données issues d'activités connexes. Si ce croisement est en effet une source d'optimisation de la chaîne de valeurs, il nécessite en pratique de pouvoir disposer d'architectures de confiance pour l'échange des données.

Méthodes mathématiques pour le traitement des données

Le traitement de données massives se heurte très souvent, entre autres, aux limitations liées aux bandes passantes insuffisantes et au coût énergétique trop important des échanges, que ce soit à l'intérieur du calculateur lui-même ou entre le calculateur et la source de ces données. Les exemples en sont nombreux : certains relativement anciens, comme en météorologie avec la technique d'« assimilation de données » ; d'autres plus récents, comme lorsqu'il s'agit d'extraire directement de ces données une connaissance encore insuffisamment intégrée dans les modèles de processus. Les méthodes qui permettent le traitement des données massives se sont développées de façon indépendante de celles qui prévalaient, et prévalent encore, pour le calcul HPC. Les ensembles logiciels sont très différents, les uns utilisant en majorité des méthodes de découpage de données peu couplées entre elles, tandis que les autres sont principalement fondés sur des bibliothèques de passages de messages en mémoire distribuée.

Les uns sont utilisateurs d'accélérateurs (les GPU), les autres utilisent depuis plus longtemps des processeurs à usage général (les CPU). L'explosion du volume de données à traiter et le développement extrêmement rapide des méthodes d'apprentissage automatique ouvrent de nouvelles opportunités dans nombre de domaines, scientifiques et industriels. Ces méthodes comportent deux phases : une première phase d'entraînement, qui consiste à construire le modèle statistique sous forme d'un réseau de neurones ; et une seconde phase d'inférence, qui consiste à utiliser le modèle précédent avec de nouvelles données pour parvenir à la prédiction du processus étudié. La phase d'entraînement est grande consommatrice de calcul, sous une forme très parallèle particulièrement adaptée aux accélérateurs.

Nécessaire convergence des approches

L'interpénétration progressive, l'hybridation, entre ces deux approches, calcul scientifique basé sur des lois scientifiques d'une part, et traitement de données massives et apprentissage automatique d'autre part, a commencé à se manifester depuis un petit nombre d'années, par exemple avec des architectures de calcul permettant tant la simulation numérique que l'apprentissage automatique. Le support matériel est unique, mais les ensembles logiciels restent différents. Cette hybridation a pour objectif d'accélérer la simulation numérique de type HPC, par exemple en exploitant des données (synthétiques ou non) permettant d'éviter d'avoir à simuler le détail de certains processus. Elle offre aussi une méthode d'amélioration de l'apprentissage automatique, en utilisant les connaissances et les lois déjà disponibles pour organiser le corpus des données massives et restreindre l'espace dans lequel s'opère le traitement statistique par les réseaux de neurones (émergence de l'« apprentissage automatique guidé par la physique »). Le niveau de performance atteint par l'apprentissage automatique peut alors permettre de surmonter l'obstacle représenté par le manque fréquent de données de qualité pour l'apprentissage des modèles.

Méthodes de modélisation hybride

La précision des simulations basées sur des lois scientifiques peut être améliorée grâce à la modélisation hybride qui tire parti des données, que celles-ci viennent directement de capteurs ou qu'elles soient issues d'autres simulations. Dans les méthodes dites de « multifidélité », des modèles numériques de différents niveaux de précision permettent de construire un modèle hybride : dans ce cas les simulations fines, nombreuses et coûteuses à obtenir (mais « une fois pour toutes » seulement), nourrissent une approche par apprentissage automatique dont les résultats peuvent être utilisés pour une modélisation plus globale. De façon un peu analogue, les méthodes de jumeaux numériques hybrides sont construites comme la somme d'un modèle « physiquement motivé » et d'un « modèle d'ignorance » basé uniquement sur les données et décrivant l'écart entre prédiction du modèle physique et mesures cibles. Cette amélioration de la précision par utilisation conjointe de données et de lois scientifiques permet aussi de réduire la

complexité et le coût des simulations basées sur la physique, en réduisant le nombre de degrés de liberté pour un problème donné. De façon générale, les approches hybridant données et modèles de calcul ne sont apparues qu'assez récemment, et la maturité des méthodologies associées n'est pas encore atteinte.

Validation des simulations

La validation de ces nouvelles méthodes doit donc être examinée avec soin. Une première question concerne l'opacité du fonctionnement des réseaux de neurones utilisés pour l'apprentissage automatique et la difficulté d'en expliquer les succès comme les échecs. La seconde tient à l'explosion de la dimension des espaces dans lesquels s'effectue cette inférence statistique. Pour les applications destinées à être confrontées au monde réel, il est indispensable de trouver les moyens de prendre en considération, dans le mécanisme d'apprentissage, toute la connaissance *a priori* disponible, de telle façon que les grandes masses de données ne permettent que de corriger et d'adapter un modèle, et non en être l'unique source si une connaissance préalable existe. De nouveaux concepts méthodologiques pour la validation devront apparaître sur cette base.

De nouveaux types de compétences

Face à ces contextes en évolution, il est nécessaire de recourir à plusieurs types de compétences : d'une part celles qui relèvent des mathématiques appliquées et de l'algorithmique, d'autre part celles qui concernent l'ingénierie des données. Dans le premier cas, le « scientifique des données » a pour charge de définir la méthode, de construire l'algorithme correspondant, et d'assurer sa programmation informatique. Malgré l'excellence de l'école française en apprentissage automatique, les aspects liés à l'étude de l'exactitude de ces méthodes y sont significativement moins développés. Dans le second cas, l'« ingénieur des données » s'attache à la constitution et à l'intégration de grands volumes de données et à leur mise en forme, effectue les traitements conçus par le « scientifique des données » et les traitements de mise en forme et de suivi des résultats en fin de processus. Au-delà, un renforcement de la formation pluridisciplinaire s'avère de la plus haute importance, pour assurer la maîtrise de la science physique, chimique, biologique…, des mathématiques appliquées, de l'informatique du parallélisme et de l'apprentissage automatique. Ces formations, nécessaires à la constitution d'équipes intégrées, doivent trouver leur place tant dans les écoles d'ingénieurs qu'à l'université, et doivent être complétées par la formation continue. Les ambitions industrielles françaises ne peuvent se réaliser que via l'accès à de telles compétences.

Après cette présentation générale de l'évolution du contexte, suivent des paragraphes plus spécifiques et plus techniques sur chacun des points évoqués, paragraphes qu'un lecteur moins intéressé pourra ignorer avant de passer à la seconde partie.

LES DONNÉES EN MASSE

Un des nouveaux défis majeurs auxquels ont à faire face les infrastructures pour le HPC, tant sur les plans matériels que logiciels, est celui de la quantité considérable et de la très grande diversité des données qu'il convient de traiter et d'analyser par le calcul, puis de visualiser, de stocker et d'archiver. Plusieurs causes à ce « déluge de données » : d'une part, l'augmentation de la puissance de calcul disponible, qui permet d'affiner les pas de discrétisation des espaces sur lesquels s'effectue la résolution des équations d'un modèle numérique ; d'autre part, les nouvelles capacités d'acquisition et de communication des divers objets et équipements, qui génèrent de nouvelles informations. Ce second aspect deviendra très rapidement dominant ; ainsi, par exemple, de prochains grands instruments comme EUCLID (télescope spatial dédié à l'observation de l'énergie noire en 2021) ou SKA (radiotélescope d'un kilomètre carré en 2025) généreront chacun 20 PB de données brutes par nuit, soit jusqu'à 4 EB par an ! Il est clair que traiter et valoriser ces données dans des temps acceptables va requérir l'utilisation d'une puissance de calcul considérable tout au long de la chaîne d'information, depuis l'instrument *(edge*) jusqu'au *cloud* avec de nouvelles méthodes de traitement utilisant notamment l'apprentissage automatique.

Quatre conséquences majeures en découlent :

- la première concerne l'architecture même des supercalculateurs et de leur environnement local, afin d'obtenir des performances aussi peu dégradées que possible par des accès fréquents à des hiérarchies complexes de mémoire ;
- la deuxième touche directement aux infrastructures de stockage et de traitement de ces données à différentes échelles géographiques (logistique des données), afin de les rendre accessibles à leurs multiples utilisateurs ;
- la troisième conséquence est de nature moins technique. Elle résulte de la notion de « valeur d'utilité » de ces données, et de ce qui en découle en matière d'accessibilité par le plus grand nombre ou de la caractérisation des échanges entre utilisateurs ou entre applications auxquelles ces données peuvent être soumises ;
- la quatrième conséquence relève des compétences humaines indispensables pour maîtriser et gérer ces océans de données et programmer leur traitement. Ces points sont abordés et développés *infra*.

Les tendances actuelles dans le domaine des architectures de calcul

Deux tendances principales sont à souligner

Tassement de la croissance de la puissance, dégradation de l'efficacité relative des applications

Tout d'abord, pour des raisons de complexité de la microélectronique sous-jacente, de coût de fabrication et de consommation énergétique, la puissance crête des supercalculateurs de pointe ne croît plus aussi vite qu'elle ne le faisait il y a encore dix ans. Si l'on considère par exemple (figure 1) la puissance de calcul cumulée des dix ordinateurs les plus puissants au monde (tels qu'identifiés deux fois par an grâce à la liste dite TOP500[2]), on constate un ralentissement depuis 2012. Alors qu'antérieurement la puissance doublait tous les dix-huit mois, voire un peu moins, le rythme actuel n'est plus qu'une multiplication par cinq tous les six ans contre seize précédemment, soit environ trois fois moins. Ce ralentissement est dû à plusieurs facteurs, au premier titre desquels la limitation de la consommation énergétique. La nécessaire maîtrise de celle-ci, tendance de très long terme, conduit en effet à ne pas rechercher l'augmentation de performance « simplement » par l'augmentation régulière de la fréquence des unités de calcul (cœurs de calcul), extrêmement coûteuse en énergie, mais par la multiplication des cœurs de calcul à iso-fréquence et à consommation relativement modérée. L'architecture résultante est alors à parallélisme hypermassif. Peut-être plus préoccupant encore, il y a un décrochage fréquent, croissant en moyenne, entre la puissance crête et la puissance effective réellement utilisée par les applications, souvent du fait de l'inefficacité des accès mémoire plus ou moins distants, qui laissent des unités de calcul trop oisives.

2 https://www.top500.org

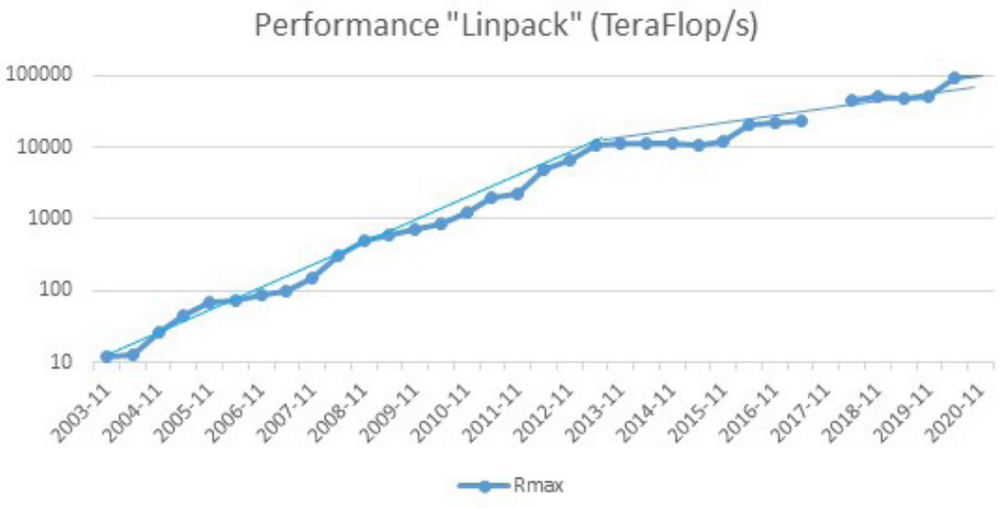


Figure 1 : Évolution de la puissance moyenne des dix superordinateurs les plus puissants

Recours à des accélérateurs de calcul moins énergivores

La seconde tendance importante concerne l'utilisation d'une proportion de plus en grande d'accélérateurs présentant un rapport « performance/consommation » nettement plus intéressant que les processeurs classiques. Apparus dès 2011 dans les supercalculateurs les plus puissants, ces accélérateurs représentent maintenant entre le tiers et la moitié des cœurs de calcul présents dans les superordinateurs (figure 2). Cette évolution n'est pas sans poser des problèmes de transformation profonde à beaucoup de codes applicatifs ou scientifiques issus de longs développements sur les anciennes architectures, en particulier pour les codes dits « patrimoniaux » *(legacy codes)*. À l'inverse, ces nouvelles architectures se révèlent très efficaces pour la mise en œuvre des méthodes d'apprentissage automatique. Ces architectures fortement accélérées ont été retenues par le DoE[3] américain pour ses trois machines exascale qui seront déployées entre 2021 et 2023, avec un accompagnement massif pour adapter les applications (projet ECP[4]).

Du point de vue des données, si l'on considère schématiquement l'architecture d'un supercalculateur comme l'interconnexion de serveurs ou nœuds de traitement, les données permanentes sont en général stockées sur des disques ou des mémoires rapides attachés à un sous-ensemble dédié de ces serveurs. Un logiciel, dit de système de fichiers parallèles et distribués, permet aux applications de disposer d'un moyen uniforme de lire ou écrire à haut débit leurs informations avec un dispositif d'adressage de ces informations commun à l'ensemble des fichiers. Dans le monde du calcul scientifique, le système de fichiers en *open source* Lustre

3 *Department of Energy Computing Program*

4 *Exascale Computing Program*

est à l'origine de nombreuses implémentations[5]. Au-dessus de ces systèmes de fichiers se trouvent généralement des moteurs de bases de données qui fournissent aux applications un moyen d'adressage de l'information prenant en compte la nature de l'information elle-même et factorisant les diverses manières d'y accéder. On parle alors de bases de données structurées sous forme de tableaux SQL, de base de données non structurées (images, sons, textes…), avec par exemple le moteur noSQL, ou encore de bases de données de graphes où des blocs d'informations sont reliés par des arcs représentant des propriétés sémantiques communes à ces informations.

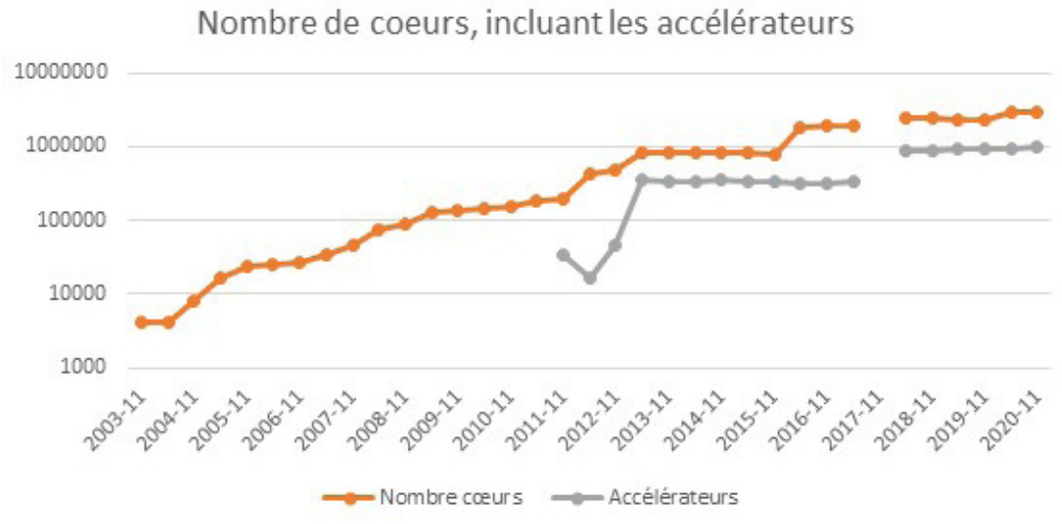


Figure 2 : Évolution du nombre de cœurs de calcul, incluant les accélérateurs, pour la moyenne des dix superordinateurs les plus puissants

Au-delà de ces aspects plus spécifiquement liés aux traitements eux-mêmes, une caractéristique déterminante qui doit piloter les nouvelles architectures de calcul est le principe de proximité entre la localisation d'un traitement et celle des données qu'il utilise, et ceci pour des raisons manifestes de performance globale, d'économie d'énergie et de sécurité. Soit le traitement doit se rapprocher de l'endroit où les données sont produites, soit les données doivent être transférées à l'endroit où a lieu leur traitement. Le premier cas est caractéristique d'une situation où l'on doit traiter uniquement les données produites localement (cas désigné par le vocable *edge computing*). Le second cas est caractéristique, soit d'une mutualisation de moyens lourds de calcul pour divers types de traitements (désigné par le vocable *cloud computing*), soit encore de la nécessité de réunir ou croiser des données produites dans diffé-

5 La taille de ces systèmes de fichiers peut être très grande. Par exemple, pour le supercalculateur SUMMIT, aujourd'hui deuxième dans le classement mondial du Top500, elle est de 120 PB ; mais les systèmes de fichiers des machines exascale américaines dépasseront l'EB.

rents endroits distants. Pour des raisons d'insuffisance de bande passante ou de latence des télécommunications, mais aussi de sécurité des données, notamment dans le cas de traitement en temps réel, il est maintenant envisagé de fédérer ces approches dans un continuum digital allant de l'*edge* au *cloud computing*. Ceci peut se faire en recourant à des infrastructures hiérarchisées de centres de données, impliquant des prétraitements locaux et permettant ainsi de diminuer la quantité de données à transmettre ou recevoir d'un niveau à l'autre. Dans cette perspective (*fog computing*) apparaissent de nouveaux défis pour la logistique des données, pour la mise en place de chaînes de traitement de bout en bout, et pour le développement de nouveaux services associés.

Il est intéressant de noter ici que nombre de grands industriels utilisent leurs propres ressources de calcul, tout en ayant souvent accès à quelques moyens plus mutualisés, comme le CCRT et GENCI, dans ce dernier cas parce que ces architectures sont aussi celles utilisées par leurs partenaires de recherche.

QUELS SONT LES TYPES DE DONNÉES À CONSIDÉRER, ET QUELLES SONT LEURS VALEURS D'USAGE ?

Le niveau de valeur et la confiance qui peuvent être accordés aux données qui résultent d'une simulation, ou qui servent de base à une analyse statistique ou à un apprentissage automatique, dépendent de leurs modes d'élaboration, de collecte et de conservation. Elles peuvent être utilisées aux fins de progrès des connaissances, de prévision de phénomènes à fort impact, de protection de la souveraineté d'un État, ou encore de la stratégie de compétitivité des produits et services d'une entreprise. Il convient naturellement de mettre en face les coûts d'obtention, de collecte, de diffusion et de conservation ou de maintenance des données. Ces coûts sont de natures très variées, frais de personnel, coûts d'investissement et frais fixes de fonctionnement des équipements. Il n'existe pas de modèle général de calcul des coûts de gestion et de possession des données, mais il semble bien que les coûts les plus élevés concernent, d'une part, les deux premières étapes de leur cycle de vie (création et collecte — traitement et maintenance) et, d'autre part, les ressources humaines. Il est en effet nécessaire de faire appel à des experts de haut niveau, tant pour la collecte de données que pour leurs traitements ultérieurs. L'apprentissage automatique supervisé, par exemple, requiert une annotation humaine hautement spécialisée afin de caractériser ce qu'il convient d'apprendre. Ainsi pour

prédire automatiquement l'apparition d'un accident vasculaire à partir d'électrocardiogrammes numérisés, il convient que chacun d'eux soit annoté par un cardiologue. Le développement, en cours, de méthodes d'apprentissage autosupervisé pourrait à terme permettre de réduire partiellement ces besoins d'intervention humaine. Dans d'autres cas, comme ceux de données structurées de type relationnel, c'est la structure elle-même qui donne du sens à une chaîne de bits (numéro de Sécurité sociale, adresse, nom de famille…). La valeur des données pose de nombreuses questions, dont deux particulièrement importantes et délicates : l'ouverture et la gratuité de leur accès d'une part, et leur circulation et le droit de propriété d'autre part.

OUVERTURE ET GRATUITÉ

Dans de nombreux milieux français ou européens, des communautés importantes plaident pour un accès libre et gratuit aux données issues de la recherche publique, au motif que leur coût étant financé par des budgets publics, celles-ci doivent être accessibles à tous. Cette préoccupation a, par exemple, motivé la directive européenne sur l'*Open Data*, l'initiative européenne EOSC (pour *European Open Science Cloud*) pour les données de la recherche, et l'initiative nationale « data.gouv.fr ». L'Union européenne a ainsi mis en place avec succès le programme COPERNICUS[6], permettant un accès aux données d'origine spatiale d'observation de la terre, données collectées par des systèmes spatiaux mis en œuvre sur fonds publics. Au niveau national, l'initiative *Health Data Hub* (voir encadré) permet d'accéder à des données de santé de manière libre et gratuite dès lors que, dans le processus de soins qui a conduit à leur obtention, il y a eu un financement public, par exemple un remboursement par la Sécurité sociale. Il serait en fait souhaitable de procéder à une évaluation rigoureuse des externalités positives ou négatives de toute ouverture de l'accès aux données, pour définir et appliquer une politique sélective d'ouverture, si besoin avec contreparties.

Ces contreparties ne sont pas nécessairement financières mais peuvent aussi concerner l'accessibilité à des résultats produits à l'aide des données fournies initialement. Pour les raisons qui viennent d'être évoquées, et pour ce qui concerne les start-up et PME, il est clair que celles-ci n'ont en général pas les moyens de financer la collecte, l'archivage et la maintenance d'une très grande quantité de données. Dans ce cas, des données publiques (ou ayant reçu un financement public pour être produites et collectées) pourraient leur être ouvertes gratuitement, ou à faible coût. Ce type d'aide pourrait alors être considéré comme une aide publique à la R & D, qu'il s'agisse de subventions ou d'avances remboursables. Vis-à-vis de l'État les devoirs doivent être les mêmes, notamment en cas de rachat de l'entreprise par un groupe non français ou non européen.

6 https://www.copernicus.eu/fr

Le Health Data Hub[1]

À la suite du rapport Villani sur l'intelligence artificielle rendu public le 28 mars 2018, le président de la République a affirmé sa volonté de faire de la santé un des secteurs prioritaires pour le développement de l'intelligence artificielle en France. Prévu par la loi du 24 juillet 2019 relative à l'organisation et la transformation du système de santé, le Health Data Hub (HDH) est une structure publique dont l'objectif est de permettre aux porteurs de projet d'accéder facilement à des données non nominatives hébergées sur une plateforme sécurisée, dans le respect de la réglementation et des droits des citoyens. Ils pourront y croiser les données et les analyser dans le but d'améliorer la qualité des soins et l'accompagnement des patients.

Le HDH est constitué en groupement d'intérêt public (GIP), dont la convention constitutive a été approuvée par arrêté ministériel le 29 novembre 2019. Le groupement associe 56 parties prenantes présentées dans l'arrêté précité. Son financement est majoritairement public. Le HDH, qui a été pensé comme un « guichet unique », aura notamment pour rôle de réunir, d'organiser et de mettre à disposition les données de santé issues d'une multitude de sources, telles que les données hospitalières, les cohortes, les registres… sans oublier, bien sûr, la base de données associée au remboursement de l'Assurance maladie du Système national des données de santé (SNDS). Cette base médico-administrative est constituée de plus de 3 000 variables, un flux annuel de 1,2 milliard de feuilles de soins, 11 millions de séjours hospitaliers, 500 millions d'actes et plus de 200 TB de données.

Un premier « catalogue » de bases de données est en cours de constitution. Il définit la liste des bases mises à disposition prochainement, dans le cadre de la recherche d'intérêt public aux différentes parties prenantes du secteur de la santé. Avec une feuille de route stratégique ambitieuse, le HDH a parcouru un chemin significatif depuis sa création. De façon non exhaustive, on peut citer parmi les étapes franchies :

- la mise en place du CESREES (Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé), en charge de rendre des avis sur les projets d'études préalablement à l'autorisation de la CNIL ;

- le soutien de 28 projets pilotes : à l'aube du second semestre 2020, 13 avaient passé avec succès l'étape CESREES et 6 étaient sur le point de déposer leur dossier auprès de la CNIL. La première autorisation CNIL a été obtenue par le projet « Hydro » porté par la start-up Implicity, avec pour vocation de prédire les crises d'insuffisance cardiaque via l'intelligence artificielle ;

- la mise en production accélérée de la plateforme technologique du HDH dans le contexte COVID-19 et l'import de la base OSCOUR® de Santé publique France sur les passages aux urgences. Une dizaine d'autres projets seront prochainement lancés avec, en parallèle, l'importation d'une quinzaine d'autres bases nationales ou hospitalières relatives aux données COVID-19 ;

- la publication des engagements du HDH envers les citoyens : intérêt général, protection des données, respect des droits individuels, transparence ;

- l'engagement du HDH dans sa mission de positionner la France comme un leader dans l'usage des données de santé à l'échelle européenne et internationale.

Emmanuel Bacry, Directeur Scientifique du Health Data Hub

Pour aller plus loin : www.health-data-hub.fr

1 L'hébergement des données du HDH sur un cloud de la société Microsoft (Azure) a fait l'objet en octobre 2020 d'un jugement du Conseil d'État, demandant que soient apportées des garanties suffisantes pour minimiser le risque de transmission des données de santé sur demande des services de renseignement américains.

Circulation des données professionnelles

Les données associées aux différentes activités d'une entreprise, à ses produits ou services, leur croisement avec les données forment une base évidente pour améliorer ses performances ou fonder sa différentiation externe (on peut penser par exemple à « EDF Énergies Nouvelles » pour le photovoltaïque : lien avec les conditions météorologiques locales, l'orientation des toitures, afin de garantir la production théorique…). Elles permettent d'en accroître la compétitivité tout en augmentant la qualité et la réactivité, tout en réduisant les coûts… À partir d'une modélisation d'un produit et de ses données de fonctionnement recueillies sur le terrain, il pourra, par exemple, être possible de mettre en place une maintenance prédictive. Dans un tel contexte, le premier réflexe d'une entreprise sera de conserver ses données pour elle-même et de les protéger. Il faut cependant souligner que, pour les traitements de nature statistique, il importe tout autant de disposer d'un grand volume de données non biaisées selon un attribut ou une dimension caractéristique, que de pouvoir croiser de nombreux attributs ou dimensions. Par ailleurs un industriel n'est pas nécessairement le propriétaire de données qui lui sont fortement utiles. Ainsi le fabricant d'un moteur, composant d'un système, ne pourra bénéficier des données de fonctionnement de ce moteur qu'au travers des systémiers qui l'embarquent dans leurs produits. La gestion de l'énergie d'un bâtiment ne pourra être réellement efficace que si l'opérateur du bâtiment dispose des données relatives à sa construction, croisée avec des données du fournisseur d'énergie ainsi qu'avec celles issues de l'utilisation du bâtiment. Ces quelques exemples montrent que l'organisation de la circulation des données professionnelles est tout à fait nécessaire.

Outre les aspects qui relèvent du droit commercial et de la fiscalité, cette circulation des données se trouve confrontée à un problème technique majeur : comment concilier les droits de leurs propriétaires, leur maîtrise et le contrôle de leur utilisation ? Il est important de souligner que des initiatives comme IDS (*International Data Space*), née en Allemagne, proposent une architecture de confiance pour l'échange de données qui devrait répondre à la question posée. Dans ce contexte, l'Académie recommande fortement que les acteurs publics ou privés s'organisent autour de telles initiatives. L'enjeu est de taille : l'innovation depuis les données, et l'accroissement de compétitivité et de valeur des entreprises qui en résultent, n'émergera que si les données peuvent largement circuler. Ceci suppose en outre que les données puissent circuler facilement, librement et de façon interopérable et réutilisable.

L'importance grandissante du traitement et de l'exploitation de données massives, l'apprentissage automatique

Notons pour commencer que des logiciels « grand public » sont maintenant disponibles, tels *Tensor Flow*©[7], qui permettent un large usage d'applications d'apprentissage automatique ; le recours à ces logiciels donne souvent le sentiment qu'il est facile de « bien traiter de façon simple » un problème d'apprentissage automatique, mais ne garantit toutefois pas toujours une utilisation exempte de limitations.

Le nombre de domaines applicatifs pour lesquels il est nécessaire de traiter des masses de données très importantes est en augmentation de plus en plus marquée, que cela soit dans le domaine des données instrumentales, en lien avec la performance de nouveaux équipements (satellites, radiotélescopes, cryomicroscopes, séquenceurs en génomique…), dans celui des données calculées (données massives issues de simulations numériques tridimensionnelles multi-physiques, multi-échelles, à haute résolution et à grands ensembles), ou enfin, à terme, avec l'explosion des réseaux de capteurs (Internet des objets) dans les usines, villes, véhicules, mais aussi avec les téléphones portables.

De tels traitements se révèlent très souvent pénalisants en temps, du fait des bandes passantes insuffisantes, des latences induites, de la sécurité et du coût énergétique qui pénalise le déplacement de données dans la hiérarchie des mémoires. Si certains domaines connaissaient déjà cette difficulté, comme la météorologie qui avait développé l'approche dite d'« assimilation de données », beaucoup d'autres y sont confrontés depuis une période plus récente. Mais il peut aussi s'agir d'ajuster des paramètres de discrétisation, de détecter et prévenir des divergences numériques, d'estimer des points fixes ou des états d'équilibre, d'explorer des branches de solution. Il peut encore s'agir de post-traiter de grandes masses de données produites par des simulations à haute-fidélité, ou d'inclure des données d'observation pour le contrôle de simulations (approche de type « assimilation de données »), ou de compléter des approches mêlant conception, prototypage, réalisation et maintenance, en particulier pour l'industrie, ou encore d'extraire de ces données une connaissance encore insuffisamment intégrée dans les modèles de processus.

Le traitement, la simulation proprement dite, à partir de modèles physiques plus ou moins précis, par les méthodes « classiques » du HPC, et le travail sur les données massives afin d'en tirer le maximum d'informations, voire de connaissances, a commencé par se faire de façon indépendante, les logiciels utilisés étant très différents (figure 3).

7 https://www.tensorflow.org/©

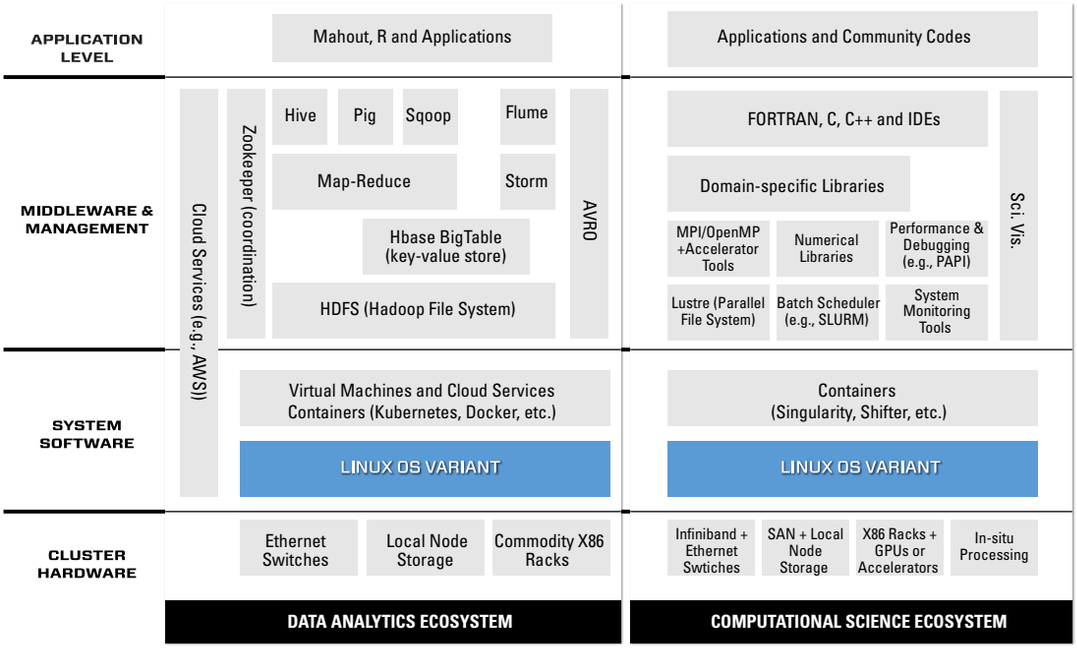


Figure 3 : Les différents types de logiciels utilisés pour le traitement des données massives (HPDA, à gauche) et le calcul scientifique (HPC, à droite). Source : Asch *et al.*, 2018.

C'est ainsi que, côté données, c'est souvent le système HDFS (*Hadoop Distributed File System*) associé à l'environnement de programmation parallèle *Hadoop* ou ses successeurs, qui est utilisé, contrairement au côté calcul où prédominent d'autres logiciels, tel le système fichier parallèle Lustre[8]. De même, du côté des données, on trouve une algorithmique de type *MapReduce* de découpage de données peu couplées entre elles. Côté calcul prédominent des bibliothèques de passages de messages ou de contrôle *multithread* en contexte de mémoire distribuée ou partagée (*MPI, OpenMP*), dans un monde de langages encore dominés par FORTRAN, C et C ++. Si ces derniers permettent une programmation souple, pérenne et spécialisée de « sous-modèles », ils sont moins adaptés lorsque ces sous-modèles, comme souvent, sont fortement couplés ; la tendance est alors de plaquer sur ces couches basses génériques des *Domain Specific Languages* (DSL) épousant les idiomes des méthodes principales de simulation numérique. Les processeurs de calcul sont eux aussi différents (CPU et GPU d'une part, GPU, TPU d'autre part) ce qui a souvent conduit à des architectures de ressources elles aussi spécialisées.

Cette explosion du volume de données, le développement des supercalculateurs, notamment grâce à l'emploi des accélérateurs de calcul graphiques GPU, et le développement extrêmement rapide des méthodes d'apprentissage automatique, en particulier de l'apprentissage profond (ou *deep learning*), voir figure 4, permettent en effet d'ouvrir de nouvelles fonctionnalités via le traitement de données massives : reconnaissance d'images ou de la parole, diagnostics avancés…

8 On notera que les vocables « distribué » *vs*. « parallèle » correspondent à des visions différentes du couplage et de la localité, plus forts en HPC

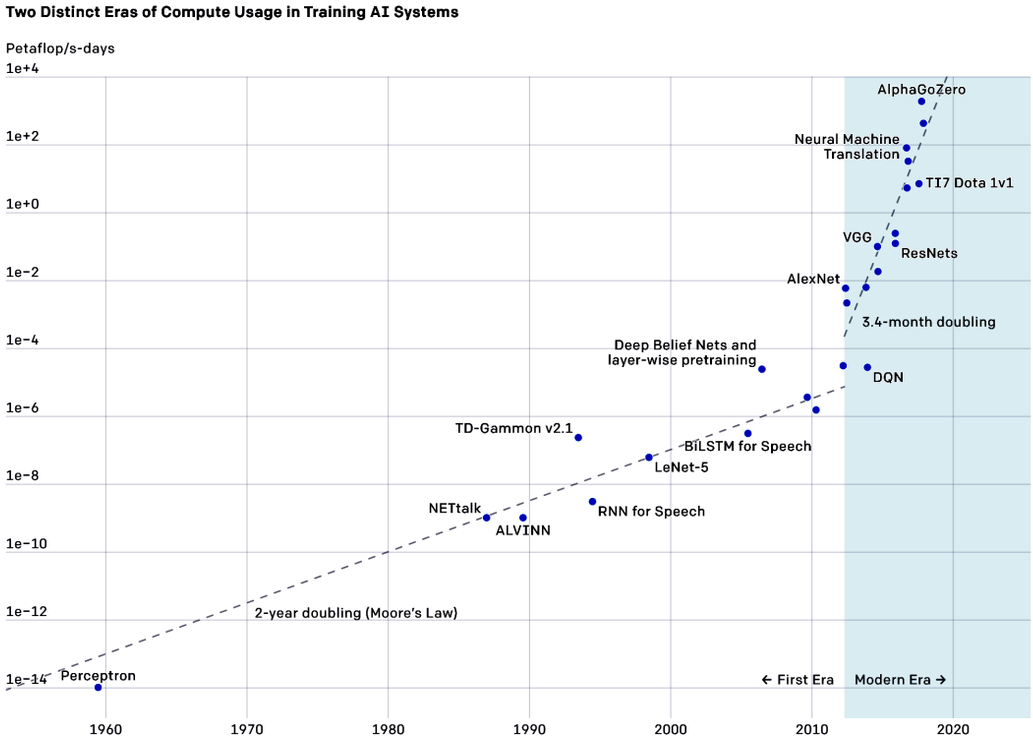


Figure 4 : L'explosion des besoins en puissance de calcul pour l'apprentissage de modèles (vision, traitement de langages et jeux). Source https://openai.com/blog/ai-and-compute

L'apprentissage automatique comporte deux phases : une première phase, d'entraînement, qui consiste à construire un modèle statistique à partir de données si possible en grande quantité et très diversifiées, et présentant le moins de biais possible, ce modèle se présentant sous forme d'un réseau de neurones spécifique ; et une seconde phase, dite d'inférence, opérationnelle, qui consiste à entrer des données dans le modèle précédent puis à partir des sorties obtenues en déduire une prédiction voire une prescription résultant des données entrées. Suivant les applications, le modèle appris est soit laissé en consultation et à l'utilisation par ses usagers, soit encore incorporé, ou « encapsulé », dans un dispositif que l'on veut rendre autonome (voiture, robot…). La phase d'entraînement est grande consommatrice de calcul, principalement sous forme très parallèle et elle est particulièrement adaptée aux accélérateurs (multiplications de matrices dans des opérateurs de convolution, fonctions d'activation câblées, précision numérique mixte ou réduite). Ceci explique que les premières applications de l'apprentissage automatique se soient développées avec des logiciels et des architectures informatiques spécifiques et différents de ceux et celles utilisés pour le calcul scientifique et la simulation numérique. Néanmoins, depuis quelques années, on assiste à une convergence des architectures de calcul, permettant, certes avec des ensembles logiciels différents, mais sur des supports matériels similaires (nœuds de calcul accélérés, réseaux d'interconnexion

rapides, mémoires et stockages rapides), de réaliser à la fois des travaux de simulation numérique et/ou d'apprentissage automatique.

LA CONVERGENCE NÉCESSAIRE DES APPROCHES

L'interpénétration progressive, l'hybridation, entre ces deux approches, simulation numérique basée sur le calcul scientifique d'une part, et traitement de données massives et apprentissage automatique d'autre part, représente le nouveau défi auquel doivent faire face de nombreuses applications.

Pour le calcul scientifique, et dans les très nombreuses situations où la prise en compte de phénomènes et processus des échelles spatiales ou temporelles très fines est l'impératif premier (climat, fusion nucléaire, combustion, aérodynamique, matériaux…), cette hybridation peut constituer une opportunité pour parvenir plus vite et plus économiquement à une meilleure prise en compte des échelles les plus fines (par exemple en utilisant l'apprentissage profond basé sur des modélisations à échelles très fines réalisées de façon découplée, ou basé sur des données expérimentales). Ce type d'hybridation commence à se développer dans de nombreuses applications.

Pour l'apprentissage automatique, l'hybridation peut aussi constituer une méthode permettant de tirer le meilleur profit des connaissances et des lois déjà disponibles, pour organiser le corpus des données massives et ne pas simplement s'appuyer sur le seul traitement statistique par les réseaux de neurones pour obtenir la réponse aux questions posées. L'émergence actuelle de l'« apprentissage automatique guidé par la physique » (ou *physics-guided machine learning*) est un bon exemple de ce type de développements, représentant d'une certaine façon le reflet pour l'apprentissage automatique de ce qu'est l'assimilation de données pour le calcul scientifique.

Si le niveau de performance atteint par l'apprentissage automatique est déjà tout à fait remarquable pour, entre autres, la vision par ordinateur, le traitement automatique des langages, la génétique ou les applications orientées « affaires », son application en sciences de l'ingénieur a été ralentie par deux obstacles majeurs : (1) le manque de données de qualité pour l'apprentissage des modèles ; (2) la difficulté à incorporer la connaissance physique des systèmes dans les algorithmes. Un exemple illustrant bien la nécessité de disposer d'une grande quantité de données pour construire des modèles d'apprentissage automatique est celui de la

base de données ImageNet[9], qui contient 14 millions d'images annotées manuellement avec l'une des 20 000 étiquettes indiquant quel objet est représenté sur une image. L'ingénierie est très loin d'atteindre ce niveau, à la fois en termes de quantité et de qualité des données. En effet, la plupart des bases de données d'ingénierie ont été historiquement construites pour des besoins de compréhension physique ou de validation des modèles. En conséquence, les données existantes représentent mal l'espace des paramètres admissibles et elles ne sont pas véritablement adaptées pour être utilisées directement comme données d'apprentissage. Dans un tel cas de manque de données massives (on parle alors de *small data*, à opposer à *big data*), il faut se tourner vers une modélisation hybride des systèmes, l'idée étant d'incorporer toutes les connaissances *a priori* (lois de la physique, règles empiriques, ou même expertise métier) dans les approches d'apprentissage automatique. En procédant de la sorte, cette information *a priori* aide à construire des modèles hybrides ayant un moindre besoin de données d'apprentissage. Dans tous les cas l'originalité vient de la possibilité de contraindre le modèle d'apprentissage à respecter des lois physiques ou, plus généralement, un jeu de contraintes données.

LES MÉTHODES DE MODÉLISATION HYBRIDE

UN PREMIER ENJEU EST D'AMÉLIORER LA PRÉCISION DES SIMULATIONS BASÉES SUR LA PHYSIQUE EN UTILISANT DES DONNÉES

Qu'elles soient d'observation ou issues d'autres simulations (dans ce dernier cas on parle de multifidélité, cf. *infra*), il s'agit d'une certaine façon d'une « extension » des méthodes d'assimilation de données, connues depuis les années soixante, qui cherchent à corriger l'estimation de l'état d'un système calculée avec un modèle physique en utilisant des données additionnelles, le plus souvent tirées d'observation. L'assimilation de données est massivement utilisée pour la prévision numérique en météorologie, mais aussi dans un nombre croissant d'autres domaines.

Une autre classe possible de méthodes partageant ces mêmes objectifs est connue sous la terminologie de jumeaux numériques hybrides (*hybrid twins*™), extension des jumeaux numériques (*digital twins*). L'idée est alors de décomposer un modèle hybride comme la somme d'un modèle physiquement motivé et d'un « modèle d'ignorance » basé uniquement sur les données et décrivant l'écart entre prédiction du modèle physique et mesures cibles. La justification heuristique sous-jacente est que, si le modèle d'ignorance n'est conçu que pour combler l'écart entre modèle physique et mesures, il est très probable que sa construction nécessitera

9 https://devopedia.org/imagenet

moins de données que dans le cas d'une approche basée uniquement sur les données afin de prédire directement la grandeur physique d'intérêt.

Il faut aussi mentionner ici les méthodes dites de « multifidélité par méta-modèles ». L'idée principale de base de la multifidélité est de construire un modèle basé sur des données issues de différents niveaux de précision/fidélité ; ceci permet par exemple de construire un modèle hybride utilisant des résultats issus de simulation haute-fidélité sur un maillage très fin (résultats peu nombreux car coûteux à obtenir) et des résultats issus de simulations basse fidélité sur un maillage grossier (nombreux car peu coûteux à obtenir). D'un point de vue mathématique, la construction des approches de type multifidélité est très proche de celles basées sur la modélisation de l'ignorance présentées précédemment.

UN SECOND ENJEU EST DE RÉDUIRE LA COMPLEXITÉ ET LE COÛT DES SIMULATIONS BASÉES SUR LA PHYSIQUE

Ceci peut se faire par « réduction de modèle », c'est-à-dire en réduisant le nombre de degrés de liberté pour un problème donné, ceci permettant ensuite d'explorer beaucoup plus rapidement l'espace des paramètres ou des solutions, une caractéristique très importante pour la conception industrielle. À partir de résultats de simulations à haute-fidélité, il faut construire une base réduite permettant d'élaborer un modèle simplifié et donc d'approcher les solutions à moindre coût. Dans l'esprit, les méthodes d'apprentissage utilisant des données additionnelles issues de simulations à haute-fidélité décrites ci-dessus peuvent être vues comme des extensions de cette méthode de réduction de modèle.

Il est aussi possible d'utiliser l'une ou l'autre de ces approches de modélisation hybride pour traiter seulement une sous-partie du problème considéré. Il pourra par exemple s'agir de construire des « lois de comportement », de représentation de phénomènes d'échelle inférieure à la maille (schémas de paramétrisation)… Les avantages d'une telle approche sont doubles : d'une part, la modélisation hybride pour une sous-partie d'un système physique nécessite, en théorie, moins de données que celle de la totalité du système ; d'autre part, il est ainsi possible de conserver une modélisation classique basée sur la physique pour les autres parties du système. Ici aussi, malgré le développement relativement récent d'approches de ce type dans nombre de domaines scientifiques[10], les exemples d'applications en ingénierie et dans l'industrie sont encore, pour l'essentiel, à construire.

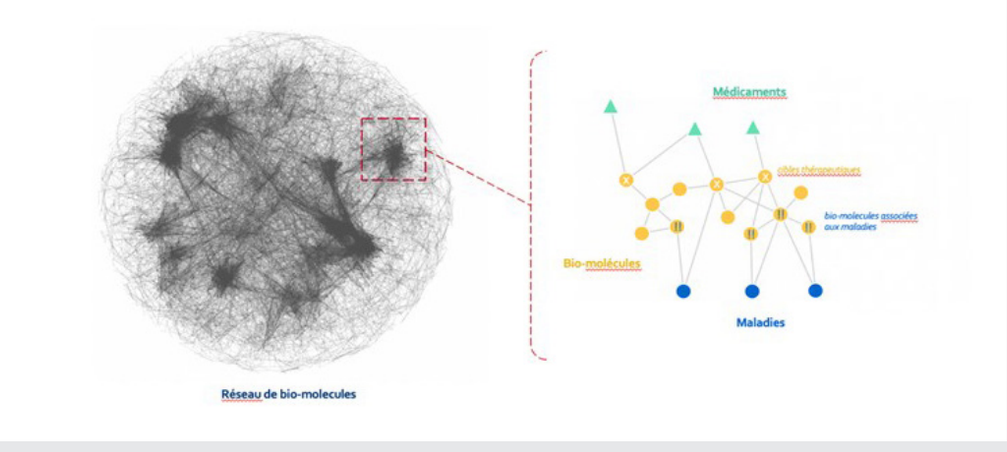
10 Voir par exemple les nouvelles applications en dynamique moléculaire (https://arxiv.org/abs/2005.00223)

Sur de nouvelles approches computationnelles pour le repositionnement de médicaments

L'industrie pharmaceutique connaît depuis une vingtaine d'années une crise de productivité de nouvelles matières actives. De nos jours, il faut en moyenne 12 ans et 2 milliards de dollars pour mettre un nouveau produit sur le marché avec une chance de succès autour de 5 à 10 %. Pour apporter de nouvelles solutions thérapeutiques aux patients, l'industrie met en place des stratégies de repositionnement de médicaments existants dans de nouvelles indications, permettant d'arriver plus vite sur le marché, avec moins d'investissements, et moins de risques d'échec liés aux effets indésirables. Les premiers succès, tels que le repositionnement du *sildénafil* dans la dysfonction érectile, de la *thalidomide* dans le myélome multiple et du *minoxidil* dans la perte des cheveux, ont permis de développer des méthodes computationnelles systématiques et innovantes. Ce champ scientifique a également bénéficié de la modélisation des maladies à l'aide du profilage moléculaire à grande échelle de cohortes de patients, facilitant l'identification de cibles thérapeutiques.

Ces approches de repositionnement reposent sur la connexion des médicaments existants et de leurs cibles dans des réseaux d'interactions couvrant environ 30 000 biomolécules et 200 000 interactions, dont les dérégulations de certains éléments sont associées à des maladies. Ces réseaux sont issus d'une intégration de données en grande dimension issues de la littérature ou générées expérimentalement. Ils sont analysés par une combinaison d'approches statistiques, d'apprentissage type *deep-learning* et d'algorithmes de simulations type *random-walks* permettant d'évaluer la proximité entre un médicament et une maladie, ou entre deux médicaments entre eux, et ainsi de générer de nouvelles hypothèses de repositionnement. Les différentes approches sont appliquées en parallèle et leurs résultats intégrés *a posteriori*. Les approches de simulations peuvent également être appliquées *a priori* afin de régulariser les données et, ainsi, améliorer les performances des approches d'apprentissage.

Des start-up spécialisées telles que Pharnext et Healx se sont développées sur ce type d'approches qui, par ailleurs, génère un intérêt considérable auprès des grands groupes pharmaceutiques.



Réseau d'interaction entre les bio-molécules, les médicaments et les maladies. Certaines bio-molécules sont des cibles thérapeutiques de médicaments. D'autres sont associées à des maladies. Ces liens sont utilisés afin de générer des hypothèses de repositionnement.

Mickaël GUEDJ, Head of Computational Medicine, et Philippe MOINGEON, Head of Immuno-Inflammation, SERVIER

De façon générale, les approches basées sur les données ou hybridant données et modèles de calcul sont apparues très récemment comparativement à la simulation HPC. La maturité des méthodologies associées n'est donc pas encore atteinte, même si quelques sociétés commencent à proposer des services basés sur ces idées[11]. C'est la raison pour laquelle la question de leur validation doit être examinée avec un soin particulier.

Vers de nouvelles approches pour la validation des simulations

La perspective d'accéder à de grandes quantités de données, et de les mettre en commun, offre indéniablement une chance majeure pour l'amélioration et la validation des modèles au sens traditionnel du terme, moyennant la nécessaire précaution d'assurer la qualité des données utilisées. Cependant, la validation de l'hybridation du calcul à haute performance pour la simulation avec les grandes masses de données ouvre des questionnements nouveaux et nécessite une adaptation considérable des méthodes et techniques de validation développées et utilisées, voire codifiées, jusqu'ici. Les succès spectaculaires de l'apprentissage profond enregistrés ces toutes dernières années, ne doivent pas occulter deux questions clés[12]. La première est celle de l'opacité du fonctionnement des réseaux de neurones profonds et de la difficulté d'en expliquer les succès comme les échecs. La seconde tient plus fondamentalement, d'une part à la nature même de l'inférence statistique consistant à prédire à partir d'un échantillon de données, et, d'autre part, à l'explosion de la dimension des espaces dans lesquels s'effectue cette inférence, qu'il s'agisse de l'espace des données et de leurs descripteurs ou de celui des paramètres à identifier. Avec une dimension des espaces pour les applications qui est aujourd'hui couramment de quelques millions, voire jusqu'à plus d'une centaine de millions, on est confronté à ce qui a été identifié en 1957 par Richard Bellman comme la « malédiction » ou le « fléau de la dimension » (ou *curse of dimensionality*).

Pour les modèles basés sur les données, le manque de représentation via des équations préexistantes rend difficile, voire impossible, l'utilisation des techniques d'analyse mathématique qui permettent d'étudier les propriétés des modèles, principalement hors du domaine où ils sont élaborés. En parallèle, l'introduction de contraintes provenant de lois de conservation ou de connaissances antérieures diverses n'est que naissante, exclusivement basée sur

11 Voir par exemple :https://cosmotech.com/resources/ebooks/machine-learning-simulation-comparison/

12 Le Secrétariat général pour l'investissement vient ainsi de lancer un grand défi intitulé « Sécuriser, certifier et fiabiliser les systèmes fondés sur l'intelligence artificielle »

des heuristiques et encore validée par les méthodes usuelles de l'apprentissage statistique. Cependant, le paradoxe en matière de validation des applications tient à la nature même de la possible utilisation de données massives, dans la perspective de ce qui a été appelé le quatrième paradigme de la science[13]. Les progrès attendus de la simulation numérique reposent en effet sur l'accès à des masses de données, à la condition que celles-ci permettent d'affiner, d'adapter et de corriger la simulation en temps réel dans les applications. La validation ne peut donc plus s'imaginer comme une étape préalable à l'exploitation du modèle, et les notions de vérification et de validation sont à réinventer.

Le processus de validation, tant pour la construction des modèles que pour leur incorporation dans des applications, doit prendre en compte les spécificités, et surtout les faiblesses de l'apprentissage, tout particulièrement celles de l'apprentissage profond. La forme particulière qu'ont prise les réseaux de neurones profonds, ajoutée aux difficultés fondamentalement inhérentes à l'inférence statistique, devrait conduire à des précautions qui bien souvent ne sont pas même évoquées dans les publications scientifiques ou techniques décrivant les applications développées. Il existe pourtant une littérature qui s'intéresse à ces problèmes, mais elle reste relativement peu exploitée. De même, les mathématiques développées depuis le milieu des années 1980 sont conceptuellement difficiles et se placent dans un cadre bien plus général que celui d'une situation d'apprentissage unique (algorithme d'apprentissage, échantillon d'apprentissage, échantillon de test). Par ailleurs, certains résultats, paradoxaux comme les bornes sur la complexité de l'espace dans lequel le modèle est recherché, ou « pessimistes » comme les théorèmes « rien n'est gratuit » (ou *no-free-lunch theorems*) qui montrent qu'en moyenne tous les algorithmes sont équivalents, demandent une analyse très fine des hypothèses grâce auxquelles ils ont été obtenus. Le système de publication scientifique et technique n'encourage d'ailleurs pas à communiquer sur les échecs ou les difficultés rencontrées. Autre signe que la maîtrise des réseaux de neurones n'est pas complète, le traitement des pièges identifiés au cours des dernières années (voir annexe IV) reste encore largement heuristique. Cette situation de méconnaissance des difficultés est aggravée par la mise à disposition de nombreux outils, par le monde académique ou les GAFAM[14], qu'il s'agisse de logiciels, de bibliothèques ou même de plateformes, qui permettent de bâtir un réseau de neurones à entraîner sur ses propres données. Enfin, au-delà de la difficulté à constituer un ensemble de données de grande taille de qualité suffisante, notamment pour ce qui concerne leur représentativité, les grandes masses

13 Syntagme créé par Jim Cray autour de 2005 pour illustrer le fait que la science entrait dans une nouvelle ère, celle de l'e-science, dans laquelle les données prennent une importance aussi grande que la simulation (troisième paradigme), la théorie (deuxième paradigme), l'observation (premier paradigme).

14 Google, Apple, Facebook, Amazon, et Microsoft. En anglais c'est plutôt le terme *hyperscaler* qui est employé, avec le mérite d'englober de manière neutre et générique les homologues chinois, géants du numérique et du Web, souvent désignés quant à eux par l'abréviation BATX …

de données semblent présenter des biais, de nature variable[15], y compris les ensembles en accès libre constitués par la communauté pour développer les algorithmes d'apprentissage. En toute hypothèse le réentraînement périodique et régulier (ou *lifelong learning*) est une nécessité.

Sur l'utilisation des méthodes d'apprentissage dans le domaine aéronautique

Dans le domaine de l'aviation civile, l'Agence européenne de la sécurité aérienne (EASA) a publié récemment une piste de modification du cycle en V des méthodologies de développement pour y inclure une phase de vérification du processus d'apprentissage aboutissant ainsi à un cycle en W. Huit défis sont identifiés dans la feuille de route « Intelligence artificielle » relatifs à la fiabilité et la confiance *(trustworthiness*) dans l'apprentissage automatique :

1. Cadres traditionnels d'assurance qualité du développement non adaptés à l'apprentissage automatique ;
2. Difficultés à conserver une description complète de la fonction prévue ;
3. Manque de prévisibilité et d'explicabilité du comportement des applications d'apprentissage automatique ;
4. Manque de garantie sur la robustesse et l'absence de « fonction non intentionnelle » ;
5. Absence de méthodes normalisées pour évaluer les performances opérationnelles des applications en apprentissage automatique et en apprentissage profond ;
6. Question du biais et de la variance dans les applications d'apprentissage automatique ;
7. Complexité des architectures et des algorithmes ;
8. Processus d'apprentissage adaptatif.

Pour toutes les raisons qui viennent d'être évoquées, il n'est pas concevable d'imaginer dans le futur que la validation ne s'appuie que sur les techniques de validation croisée sur un même échantillon partitionné en échantillon de test et échantillon d'apprentissage. Il sera indispensable de trouver les moyens de prendre en considération, dans le mécanisme d'apprentissage, toute la connaissance *a priori* disponible, à l'image de ce qui est devenu la règle dans le traitement des problèmes inverses ou plus généralement des problèmes mal posés. C'est la raison pour laquelle il faut résolument soutenir le développement des approches « basées sur la physique » (ou *physically informed)* qui tentent d'incorporer des contraintes résultant d'une connaissance préalable. Il faut plus simplement considérer que les grandes masses de données ne doivent permettre que de corriger et d'adapter un modèle, et non en être l'unique source si une connaissance préalable existe. De nouveaux concepts méthodologiques pour la validation devront apparaître sur cette base. Contrairement à ce qui a pu être prédit il y a maintenant plus de dix ans, le déluge de données n'est pas la fin de la théorie.

15 Ces biais peuvent être liés à des effets d'hypothèses limitatives sur le périmètre de la base, à des insuffisances d'échantillonnage … [voir par exemple « A. Torralba et A.A. Efros, «Unbiased look at dataset bias ». CVPR 2011, Providence, RI, 2011, pp. 1521-1528, doi: 10.1109/CVPR.2011.5995347]

UN BESOIN DE NOUVELLES COMPÉTENCES, NÉCESSAIRES AUX AMBITIONS INDUSTRIELLES

Face à ces contextes en évolution, il est nécessaire de recourir à plusieurs types de compétences : d'une part celles du « scientifique des données » (ou *data scientist*), qui relèvent des mathématiques appliquées et de l'algorithmique ; d'autre part celles de l'« ingénieur des données » (ou *data engineer*), qui concernent l'ingénierie des données.

Le scientifique des données a pour charge de définir les méthodes, exactes, statistiques, ou construites par apprentissage automatique, nécessaires à l'obtention des résultats recherchés ainsi que les données pertinentes pour la mise en œuvre de cette méthode. Il doit ensuite construire l'algorithme capable de mettre en œuvre cette méthode, ainsi que sa programmation informatique, si besoin parallèle et distribuée. Il est à noter que l'école française d'analyse numérique est très largement connue dans le monde pour ses contributions à l'élaboration de méthodes exactes et que ses résultats sont largement diffusés dans l'industrie, entre autres sur les plans de l'utilisation du parallélisme sur les infrastructures de calcul HPC. Malgré l'excellence de l'école française en intelligence artificielle et apprentissage automatique, ces aspects liés à l'étude de l'exactitude des méthodes d'apprentissage (permettant de différencier la représentation des fonctions de leurs approximations) sont significativement moins développés en France.

L'ingénieur des données s'attache à quatre activités principales : (1) l'intégration de grands volumes de données à partir de diverses sources ; (2) la transformation de ces données au moyen de différents traitements de nettoyage, normalisation et enrichissement ; (3) les traitements analytiques conçus par les scientifiques des données ; et (4) les traitements décisionnels permettant d'établir des rapports périodiques automatisés, ou des visualisations à la demande.

Incluant ces deux aspects complémentaires en sciences et traitements des données (que l'on pourrait qualifier d'« intradisciplinarité »), un renforcement de la formation pluridisciplinaire s'avère de la plus haute importance. La progression des simulations vers plus de réalisme, d'efficacité et d'applicabilité passe en effet par une maîtrise tant de la science physique, chimique, biologique… qui sous-tend le domaine concerné, que des mathématiques appliquées, de l'informatique du parallélisme (algorithmique et programmation) et de l'apprentissage automatique. Cette multiplicité de compétences doit se construire sur la base d'un tronc commun pluridisciplinaire (au sens ci-dessus), ouvrant ensuite la voie à des approfondissements disciplinaires. Les équipes qui pourront alors être constituées partageront une vision intégrée de l'approche à mettre en place, capables d'enrichir le dialogue, et les nécessaires allers-retours entre calcul et données, tout en associant les compétences les plus solides dans les diverses spécialités nécessaires. Ces types de formations doivent naturellement trouver leur place dans les écoles

d'ingénieurs généralistes, tant au cours des enseignements de tronc commun de première année que lors des « projets » qui permettent ensuite d'associer enseignements spécialisés et recherche applicative. Ces formations concernent aussi l'université, soit au niveau des IUT, lieux privilégiés pour développer la sensibilité aux aspects « ensembliers » de la simulation, soit au niveau des troisièmes cycles via des mastères pluridisciplinaires. La formation continue a aussi un rôle très important à jouer, avec notamment la possibilité de formations en ligne (e-learning, MOOC, SPOC…)

L'ambition affichée par de nombreux industriels français de tirer le plein profit des méthodes d'apprentissage, telle qu'exprimée par leur récent manifeste (voir encadré) ne peut se réaliser que s'ils peuvent faire appel à ces nouvelles compétences. En parallèle, de fortes ambitions industrielles nationales devront indéniablement être accompagnées d'un facteur d'attraction fort pour freiner le départ des meilleurs scientifiques français de l'apprentissage, toujours actuellement attirés par des perspectives scientifiques, professionnelles et salariales bien supérieures de l'autre côté de l'Atlantique.

3 Juillet 2019

Manifeste pour l'intelligence artificielle au service de l'industrie

Les industriels français engagés dans l'intelligence artificielle

Les industriels signataires de ce *manifeste* sont des acteurs mondiaux, s'appuyant sur une base stratégique nationale pour établir des positions de premier plan sur leurs marchés au niveau international.

Ils fondent leur croissance sur l'innovation, et sont engagés dans la transformation numérique aussi bien au sein de leurs organisations que sur leurs marchés et avec leurs clients. Les technologies numériques, en particulier en matière de données massives et d'intelligence artificielle (IA), sont critiques pour leur compétitivité.

Dans ce contexte, les industriels signataires partagent un objectif commun : faire de l'IA une source de croissance et d'emploi dans leurs secteurs industriels, compatible avec les valeurs et la vision de la stratégie nationale #AIforhumanity.

Ils se positionnent en tant qu'utilisateurs ou intégrateurs de l'IA mais également en tant que développeurs ou adaptateurs de ces technologies à leurs métiers. Evoluant dans des cadres réglementaires proches et en évolution, soumis à des impératifs croissants en matière de cybersécurité, ils identifient des thématiques d'intérêt commun, propres à l'utilisation de l'IA dans des environnements industriels : la confiance, l'explicabilité voire la certification de l'IA, les systèmes embarqués, l'IA pour la conception, la simulation, le développement, les tests et la logistique, l'IA appliquée à la maintenance et l'industrie 4.0, ou encore la problématique de la très haute performance, la fiabilité, la robustesse, et plus généralement l'IA dans les systèmes critiques.

Face à l'accélération générale de la transformation numérique dans un environnement de forte concurrence internationale, ils estiment que ces enjeux partagés appellent des actions coordonnées entre industriels d'une part, entre industriels et académiques, et entre industriels et décideurs politiques d'autre part. Il s'agit d'atteindre plus rapidement une masse critique sur les sujets de recherche prioritaires par une mutualisation sélective des moyens, d'accélérer les développements en mettant à profit une approche multisectorielle et de diffuser rapidement les cas d'usage de l'IA dans les systèmes industriels. Il s'agit également d'accroître la visibilité des usages de l'IA dans l'industrie, et d'y attirer les meilleurs talents.

Par ce *manifeste*, les industriels signataires souhaitent collectivement répondre au besoin de souveraineté lié à la maîtrise de l'IA, aussi bien dans sa dimension économique (l'indépendance technologique des entreprises françaises présentes à l'international) que de souveraineté nationale (une des quatre priorités du rapport Villani).

Les industriels signataires se positionnent dans une démarche d'innovation ouverte et s'engagent à :

1. Etablir, d'ici septembre 2019, une vision commune du diagnostic, des enjeux, des besoins et des priorités,
2. Partager cette vision avec les décideurs politiques et en discuter la mise en œuvre au niveau national et européen, en articulant au mieux les actions des entreprises et des autorités publiques,
3. Etablir d'ici fin 2019 la liste d'actions qui pourront être portées par l'ensemble ou une sélection de signataires - notamment des laboratoires communs, des structures de coopération *ad hoc*, des partages de connaissance, des actions de formation et de communication - et une feuille de route,
4. Coordonner le plan d'action avec l'ensemble de l'écosystème français en IA (en particulier les start-ups, les établissements universitaires et les équipes de recherche)

La signature du *manifeste* est ouverte à d'autres acteurs (industriels, PME et start-ups, organismes de recherche) qui en partagent la vision, les besoins et les enjeux stratégiques dans l'objectif de faire de l'IA un moteur de croissance et d'emploi dans l'industrie française.

Air Liquide
Président Directeur Général
Benoît Potier

Dassault Aviation
Président Directeur Général
Éric Trappier

EDF
Président Directeur Général
Jean- Bernard Lévy

Renault
Expert Leader IA
Jean-Marc David

Safran
Directeur Innovation et R&D
Stéphane Cueille

Thales
Président Directeur Général
Patrice Caine

Total
Président Directeur Général
Patrick Pouyanné

Valeo
Président Directeur Général
Jacques Aschenbroich

Bruno Le Maire
Ministre de l'Economie et des Finances

Deuxième partie
Synthèse et recommandations

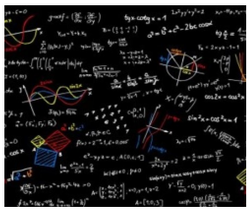
Face à ce contexte complexe, évolutif et très compétitif, il est possible de proposer six recommandations principales afin d'aider à relever les principaux défis. Ces recommandations s'adressent principalement aux organismes de financement de la recherche, tant français qu'européens pour le soutien au développement de nouvelles méthodes (recommandations 1, 3 et 6), aux centres de calcul responsables de la mise en œuvre des unités de calcul et de traitement des données (recommandation 2), aux gestionnaires des bases de données, publiques et privées, dans les divers domaines applicatifs (recommandation 4), et aux grandes écoles et à l'université (recommandation 5).

1 Développer des méthodes hybrides, associant physique de base et apprentissage

La complémentarité entre simulation numérique et extraction d'information de l'ensemble des données disponibles rend nécessaire la construction d'approches hybrides associant les connaissances et modélisations physico-mathématiques avec les informations extraites par apprentissage profond des données massives ou spécifiques. Cette hybridation, qui reste face à de nombreux défis scientifiques, est particulièrement adaptée à la simulation numérique dans le secteur industriel. Le recours au parallélisme massif à des échelles sans cesse croissantes reste un point difficile dans de nombreuses situations de simulation numérique et certains domaines de l'apprentissage automatique gagneront également à faire un usage plus approfondi et systématique des méthodes de parallélisation.

Il reste néanmoins des domaines où les modèles basés sur la physique ont des performances perfectibles comme, par exemple et sans être exhaustif, dans le cas des matériaux, des biosciences, des procédés de fabrication… Les performances ont aussi besoin d'être

améliorées pour les méthodes de simulation multi-échelle. En fait ceci est le cas partout où l'excellence de conception et de fabrication ne peut être atteinte sans des prédictions fines et réalistes basées sur la physique, dans un contexte où les volumes de données sont insuffisants pour une approche basée exclusivement sur l'apprentissage automatique. Symétriquement, diverses phases pratiques de la simulation numérique pour lesquelles une grande expérience a été acquise par les opérateurs (maillage, optimisation des pas de temps, critère de convergence…) pourraient bénéficier des techniques d'apprentissage pour les améliorer et mieux les automatiser.



1 – Développer des méthodes hybrides, associant physique de base et apprentissage

Il sera nécessaire de construire des approches hybrides

Connaissances et modélisations physico-mathématiques **U** Informations extraites par apprentissage profond des données massives ou spécifiques ("big data" et "smart data").

Malgré ses défis scientifiques, cette hybridation est adaptée à la simulation numérique dans le secteur industriel

Le parallélisme massif :

- Un point qui reste difficile mais indispensable à la simulation notamment multi-échelle
- Une technique dont pourrait bénéficier l'apprentissage

Des domaines (matériaux, biosciences, procédés de fabrication, simulation multi-échelle, optimisation de la conception) où les modèles basés sur la physique doivent encore grandir en performance… sans que les données soient suffisantes pour une approche exclusivement basée sur l'apprentissage.

Des pratiques de la simulation numérique qui pourraient bénéficier des techniques d'apprentissage (phénomènes sous-maille, maillages, optimisation des pas de temps, estimation de la convergence de calculs…)

2 S'appuyer sur la convergence des infrastructures pour le calcul et les données

Qu'ils s'adressent au secteur académique ou qu'ils soient plus spécifiquement industriels, les moyens de calcul et de traitement des données doivent rester étroitement associés au sein des mêmes infrastructures, ou *a minima* être conçus pour faciliter des échanges fluides entre les unités de stockage de données et les unités de calcul. Le besoin de traiter et d'analyser des volumes de plus en plus grands de données dans le contexte de l'apprentissage profond (ou *deep learning*) est un facteur dimensionnant des futures infrastructures partagées pour le calcul et les données (on parle souvent d'infrastructures convergées). De telles infrastructures, doivent être construites avec pragmatisme, associant dans une même machine modulaire les différentes sortes de processeurs (CPU, GPU, TPU…). La convergence doit aussi se faire

au niveau logiciel intermédiaire entre le système et l'application (ou *middleware*), facilitée notamment par l'usage de conteneurs logiciels. Les questions de sécurité informatique sont incontournables, tout particulièrement pour les techniques de jumeaux numériques où les résultats des simulations et de l'apprentissage automatique sont échangés avec les systèmes physiques, dans une communication réciproque. Ce « continuum digital/numérique » doit relier les infrastructures ainsi convergées avec les usines et avec l'internet des objets, qui mesure et contrôle machines ou systèmes. L'enjeu majeur devient alors d'assurer une logistique des données efficace (traitement à chaque niveau de la chaîne, du capteur/instrument, ou *edge*, au centre de calcul et *vice-versa*) et sécurisée (au travers des différents réseaux et protocoles de communication utilisés).

La disponibilité de telles infrastructures est à la fois un enjeu de souveraineté nationale et européenne, afin de pouvoir rivaliser avec les superordinateurs exascale, en projet ou déjà en cours de déploiement, tant aux États-Unis (trois machines exascale attendues entre 2021 et 2023[16]) qu'au Japon (une machine de classe exascale en 2020 au centre Riken, plutôt encore orientée calcul) et en Chine (qui, malgré les embargos américains, déploiera au moins une machine exascale en 2021-2022, à base de technologies chinoises pour les processeurs et les accélérateurs). À cet effet la Commission européenne et, maintenant 32 pays participants, se sont lancés en fin 2018 dans la mise en place d'une entité nommée EuroHPC dont le rôle est de cofinancer et déployer une infrastructure de calcul exascale (avec une première étape dite pré-exascale) à horizon 2022-2023, mais aussi de développer la recherche et l'innovation nécessaires à l'utilisation de technologies matérielles (comme le projet EPI[17] de processeur européen) et logicielles (incluant les applications européennes).

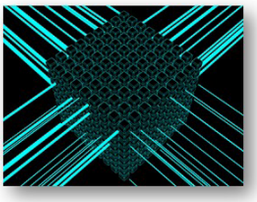
Ces futurs moyens de calcul européens seront mis à disposition des chercheurs académiques et industriels pour des travaux de recherche ouverte (c'est-à-dire avec publication des résultats, éventuellement confidentielle), dans le domaine de la simulation numérique, du traitement de données massives et de l'apprentissage automatique. La Finlande (avec un consortium de neuf autres pays), l'Espagne (associée à trois autres pays) et l'Italie (elle aussi associée à trois autres pays) se sont lancés dans l'hébergement de premiers systèmes pré-exascale (autour de 200 petaflop/s crête) d'EuroHPC dès 2021. La France et l'Allemagne sont pour leur part candidates à l'hébergement de systèmes de classe exascale pour 2023-2024 ; pour la France cette candidature est en cours de constitution, portée par GENCI en tant qu'organisme d'accueil (ou *Hosting entity*) et le CEA en tant que site d'accueil (ou *Hosting site)*. Les candidats à l'hébergement de telles machines, via d'éventuels consortia alliant à la fois partenaires nationaux et européens, doivent apporter garanties techniques, tant en termes de puissance informatique que de sobriété énergétique, et garanties de cofinancement du coût complet d'un tel système.

16 Pour les laboratoires ANL *(Argonne National Laboratory),* ORNL *(Oak Ridge National Laboratory)* et LLNL *(Lawrence Livermore National Laboratory)*.

17 *European Processeur Initiative*

Il est enfin à noter la mise en place fin 2019 d'une initiative de *cloud* souverain européen industriel nommée Gaia-X, portée essentiellement par des acteurs allemands (SAP, Deutsche Telekom, Siemens…) et français (ATOS, Dassault Systèmes, OVH…), pour doter l'Europe d'un *cloud* de stockage et de traitement de données souveraines, constituant la réaction de l'Europe au *Cloud Act* américain (voté en 2018), et évitant ainsi le recours aux solutions des géants du Web : GAFAM américains ou BATX chinois.

Qu'il s'agisse d'EuroHPC ou de Gaia-X (dont on peut souhaiter le rapprochement à terme), il est recommandé de soutenir ces initiatives, cruciales pour disposer en Europe de technologies et de moyens de calcul, de stockage et de traitement de données souveraines, nouvel or noir du XXIe siècle. Les appels à candidature qui sont, ou seront, mis en place par EuroHPC pour la conception et la réalisation de ces architectures européennes permettant à la fois le calcul scientifique et le traitement des données sont une remarquable occasion de traduire dans les propositions nombre des recommandations présentées ici. Ceci s'applique en particulier à la candidature française pour le développement et la réalisation d'une machine exascale à l'horizon 2023.



2 – S'appuyer sur des infrastructures permettant à la fois le calcul scientifique et le traitement des données

Il est vital que les infrastructures de calcul et de données restent étroitement associées au sein des mêmes unités de traitement

- pour l'efficacité et la réussite des approches hybrides
- pour éviter une politique des constructeurs exclusivement tournée vers les données massives (machines et architectures)

Les infrastructures assurant cette convergence doivent être construites avec pragmatisme, associant dans une même machine les différentes sortes de processeurs (CPU, GPU, FPGA, TPU …)

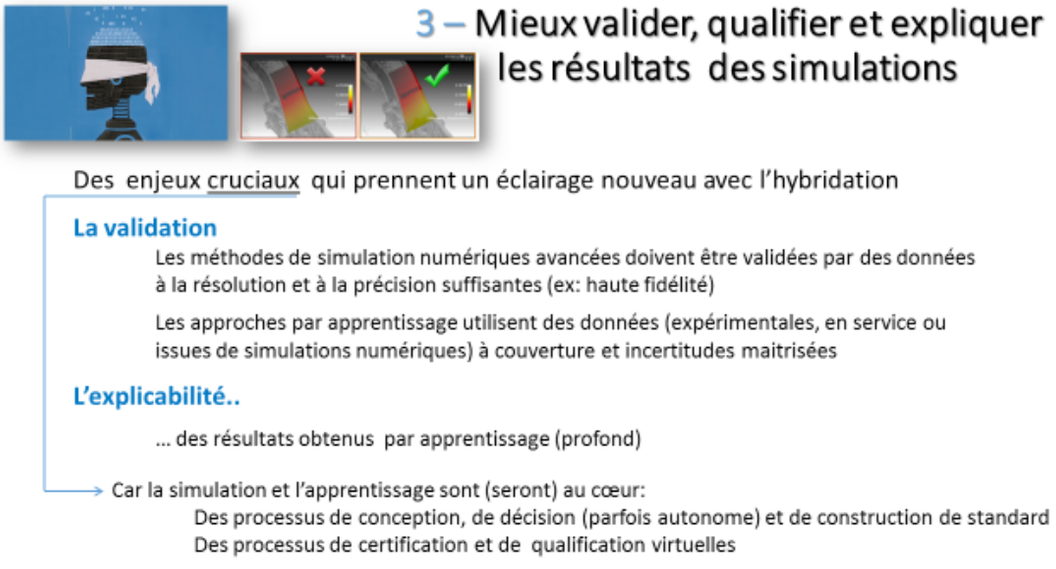
La convergence doit également s'opérer au niveau logiciel ("middleware")

3 Mieux valider, qualifier et expliquer les résultats des simulations

Les méthodes avancées de simulation numérique doivent être validées par des données en nombre et à la résolution et à la précision suffisantes. La validation de ces résultats est d'autant plus cruciale qu'elle est nécessaire à la création de nouveaux marchés dans des secteurs tels

que ceux des robots et véhicules autonomes de toutes sortes, en particulier dans le domaine industriel. Les méthodes hybrides, mêlant simulations et analyse de masses de données expérimentales, ou recueillies en service, voire issues de simulations à haute-fidélité, devraient conduire à l'émergence de standards de fait et à faire aboutir les démarches de certification virtuelle, qui piétinent depuis plusieurs années.

La validation des modèles et outils numériques qui s'appuient, même partiellement, sur des processus d'apprentissage pose des problèmes inédits dont l'acuité est accentuée par la nouveauté de la discipline, et tout particulièrement pour les réseaux de neurones profonds, mais aussi par la nécessité de disposer de nouvelles compétences qui doivent, ou devraient, être mobilisées dans la construction des modèles ou des applications. Dès ses débuts, la simulation numérique a fait l'objet de partenariats et de recherche collaborative intenses entre les mondes académique et industriel. Compte tenu des enjeux, opportunités comme risques, que porte l'irruption des données massives et des technologies qui leur sont associées, il est indispensable que les liens entre ces deux mondes se développent en direction des approches basées sur la physique dans l'apprentissage à partir de données. La collaboration sur l'hybridation des machines de calcul en fournira le support naturel comme cela a été le cas pour l'utilisation du calcul haute performance ces cinquante dernières années : au regard de la durée du développement du HPC, l'apprentissage profond est une technique encore récente malgré ses succès spectaculaires. Même si l'heure est à l'accélération de la recherche et des avancées technologiques, un temps d'approfondissement et de maturation est nécessaire pour que la confiance puisse s'établir sur un socle de validation et d'« explicabilité » éprouvé. Des besoins considérables en termes de formation initiale et continue sont inévitablement à prendre en compte .



4 Organiser un meilleur partage des données entre les utilisateurs potentiels

Les start-up et PME n'ont en général pas les moyens de financer elles-mêmes la collecte, l'archivage et la maintenance d'une très grande quantité de données. Dans ce cas, des données publiques (ou ayant reçu un financement public pour être produites et collectées) pourraient être ouvertes gratuitement, ou à faible coût, aux petites entreprises. Ce type d'aide devrait alors être considéré comme une aide publique à la R & D, qu'il s'agisse de subventions ou d'avances remboursables.

Dans certains domaines de l'industrie et des services les propriétaires de données hésitent beaucoup à un partage de celles-ci. Pourtant la nécessité d'accroître et de croiser les données d'origines diverses se fait progressivement plus pressante et des partages au sein de consortia de taille maîtrisée peuvent s'avérer bénéfiques (pour la validation, l'homologation, la certification…). Il existe aussi des « territoires » de données à organiser par la puissance publique. L'exemple des données d'origine spatiale partagées dans le cadre du programme COPERNICUS est certainement à reproduire dans d'autres domaines. C'est par la mise en place de bases de données interopérables et de qualité, dans un cadre de confiance, qu'il est possible d'organiser un contre-pouvoir vis-à-vis des géants du web. Il est fortement recommandé que les acteurs publics ou privés s'organisent autour de telles initiatives. L'enjeu est de taille : l'innovation depuis les données, l'accroissement de compétitivité et de valeur des entreprises qui en résultent, ne seront réels que si les données peuvent circuler.

4 – Organiser un meilleur partage des données entre les utilisateurs potentiels

Dans l'industrie et des services, et même parfois pour l'Etat, les propriétaires de données hésitent à partager celles-ci, souvent considérées comme stratégiques

Pourtant la nécessité d'accroitre et de croiser les données d'origines diverses se fait plus pressante, et l'on peu imaginer utile de bâtir:

- des partages au sein de consortia de taille maitrisée (pour la validation, l'homologation, la certification, …).
- des territoires de données à organiser par la puissance publique

La mise en place de bases de données interopérables et de qualité, dans un cadre de confiance doit pouvoir rendre possible une réaction vis-à-vis des géants du web.

5 Promouvoir une formation hybride adaptée et bien reconnaître les métiers associés

Si, au cours des dernières années, l'accent a été mis sur un manque criant de compétences en statistiques et en apprentissage automatique, il est tout autant nécessaire de construire des formations dans ces domaines où les aspects de l'algorithmique, de la mise en œuvre informatique et de la prise en compte du parallélisme occupent une place importante. Par ailleurs les industriels confrontés à l'usage opérationnel de ces méthodes soulignent que, dans une équipe d'analyse de données, 60 à 70 % des compétences nécessaires relèvent de l'ingénierie des données. Même si cet état de fait est moins médiatisé, il convient de créer très rapidement les formations nécessaires ou d'élargir la taille des formations existantes. Comme déjà évoqué ceci est déterminant pour assurer la maîtrise des modèles, de leur mise en œuvre informatique, de la réalisation et de la validation des simulations.

Au-delà, la formation de profils hybrides, compétents à la fois en physique, en mathématiques appliquées, en sciences et ingénierie des données, en informatique du parallélisme (algorithmique et programmation) et en intelligence artificielle (apprentissage profond) est à renforcer, sinon à mettre en place. Cette multiplicité de compétences et cette pluridisciplinarité, souvent appelées, peuvent se construire sur la base d'une ou deux « majeures » et d'un socle minimal de connaissances dans les autres domaines. Sont concernées tant les écoles d'ingénieurs, et tout spécialement celles à profil généraliste, que l'université, du premier au troisième cycle. La formation peut être initiale ou continue, avec notamment la possibilité de formations en ligne (*e-learning*, MOOC, SPOC…).

De tels profils pourraient être plus facilement reconnus dans les entreprises, et gagneraient à l'être, permettant en retour d'attirer et de garder de jeunes talents, contribuant à créer un contre-pouvoir à l'égard des géants du web (USA, Chine).

5 – Promouvoir une formation hybride adaptée et bien reconnaître les métiers associés

La formation de profils hybrides, compétents à la fois en physique, en mathématiques appliquées, en sciences et ingénierie des données, en informatique du parallélisme (algorithmique et programmation) et en intelligence artificielle (statistiques, apprentissage profond) est à renforcer, sinon à mettre en place.

L'accent doit aussi être mis sur la formation continue avec notamment la possibilité de formations en ligne (e-learning, MOOC, SPOC…).

De tels profils devraient être plus facilement reconnus dans les entreprises, permettant d'attirer et de garder de jeunes talents, et contribuant à créer un contre-pouvoir à l'égard des géants du web (USA, Chine).

6 Aider à la transition des grands codes

Toutes ces évolutions, concernant tant les architectures de calcul, avec désormais des processeurs hybrides et un parallélisme devenu massif, que les nouvelles approches, associant simulation physique et apprentissage automatique, conduisent à un vieillissement accéléré de nombre de très grands « codes patrimoniaux » (ou *legacy codes*). Il est important de rappeler que l'Europe, et notamment la France, fait partie des principaux développeurs et contributeurs en grands codes (et en chaînes de codes) scientifiques dans le monde, utilisés à la fois pour la recherche et l'industrie. Faute d'investissements pour les optimiser régulièrement et les adapter aux évolutions des architectures, ces grands codes vieillissent. Leur conversion et leur adaptation aux nouveaux contextes ne peuvent donc se faire sans un investissement très important en termes, d'une part, de compétences et, d'autre part, de réécriture. L'évolution continue des contextes, tant au niveau des ressources de traitement que de l'innovation dans les méthodes d'approche, demande une permanence dans la structuration de ces efforts, une reformulation et une réécriture « une fois pour toutes » étant très certainement illusoire.

La création de chaires spécialisées serait de nature à répondre au premier de ces besoins, tandis qu'un soutien à long terme à la reconfiguration et à l'optimisation des codes serait une réponse au second. Ce soutien pourrait, par exemple, prendre la forme d'une initiative transverse aux différentes communautés scientifiques, alliant formation aux nouvelles méthodes de développement, à l'industrialisation, et au support et à la maintenance des applications. Faute de telles évolutions, les disciplines et les activités utilisatrices de ces codes ne bénéficieront pas des possibilités ouvertes tant par les nouvelles architectures que par les transformations des méthodes de simulation. Il y a ainsi un enjeu de souveraineté numérique, pour développer et maintenir ces grands codes patrimoniaux au meilleur niveau de l'état de l'art informatique. Cet enjeu s'étend au-delà à l'aide à l'émergence au niveau européen de nouvelles approches, ayant vocation à devenir des grands codes ou des bibliothèques standards dans leurs domaines.

6 – Aider à la transition des grands codes

Nombre de très grands codes applicatifs patrimoniaux (*legacy codes*) vieillissent de façon accélérée car peu adaptés à l'hyper-parallélisme, aux processeurs hybrides, aux nouvelles méthodes d'approche …

Leur conversion et leur adaptation ne peuvent se faire sans un investissement très important et très lourd en termes de compétences et de réécriture.

Un soutien à la reconfiguration de ces codes est nécessaire, pour que les disciplines et activités qu'ils servent bénéficient pleinement des améliorations de performance ouvertes par ce nouveau contexte.

Annexes

Annexe I – Glossaire

B	Abbréviation de Byte, l'octet en anglo-américain
CCRT	Centre de calcul recherche et technologie (mis en œuvre par le CEA)
CEA	Commissariat à l'énergie atomique et aux énergies alternatives
Cloud	Accès via internet à des services informatiques (calcul, stockage de données) proposés par des fournisseurs
CPU	Composant électronique d'un calculateur permettant d'effectuer des opérations de base
EB	Exaoctet (10^{18} octets)
Edge computing	Calcul réalisé en périphérie de réseau, au plus près de l'origine des données
EuroHPC	Organisation européenne pour le développement d'infrastructures informatiques de classe mondiale
Exaflop/s	10^{18} opérations flottantes par seconde
Exascale	Caractéristique d'une infrastructure informatique permettant d'associer des capacités dans la gamme des exaflop/s et des EB
Fog computing	Calculs réalisés de façon distribuée entre la périphérie du réseau et des ressources plus centrales
GENCI	Grand équipement national de calcul intensif
GPU	Processeur graphique
HPC	Calcul à haute performance

Legacy code	Code (patrimonial) de très grande taille résultant de longs développements
Middleware	Logiciels assurant les échanges entre les codes applicatifs et les logiciels de gestion système de l'ordinateur
noSQL	Famille de gestion de bases de données s'écartant du paradigme des bases de données relationnelles
Open source	Code ou logiciel ouvert, en libre accès
PB	Petaoctet (10^{15} octets)
Petaflop/s	10^{15} opérations flottantes par seconde
PRACE	Partenariat pour le calcul avancé en Europe, association de 25 pays membres permettant l'accès aux supercalculateurs européens
SQL	Langage informatique normalisé servant à exploiter des bases de données relationnelles
TPU	Unité de calcul tensorielle, adaptée au calcul pour les l'apprentissage automatique

ANNEXE II – BIBLIOGRAPHIE SUCCINCTE

Articles généraux

André, J.C., et Roucairol, G., 2019: «Quelles perspectives pour la simulation numérique à haute performance ?». *Trimestriel de la Fondation de l'Académie des technologies.*

Asch, M., et Moore, T., Eds., 2018: "Big data and extreme-scale computing - Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry". *Int. J. High Performance Comp. App.*, 32 (4), 435 – 479.

Dongarra, J., *et al.*, 2011: "The International exascale software roadmap". *Int. J. High Performance Comp. App., 25* (1), ISSN 1094-3420.

EXDCI, 2016: *"First set of recommendations and reports toward applications"*. Deliverable D3.1, S. Requena Ed.

National Academies of Sciences, Engineering, and Medicine, 2018. *"Opportunities from the integration of simulation science and data science : Proceedings of a workshop"*. The National Academies Press. doi : https://doi.org/10.17226/25199

Roucairol, G., 2013 : « *La simulation haute performance au service de la compétitivité des entreprises* ». Rapport au Commissariat Général aux Investissements et à la Direction Générale de la Compétitivité, de l'Industrie et des Services.

Articles sur l'apprentissage et les données

Anderson C. The End of Theory : The Data Deluge Makes the Scientific Method Obsolete, *Wired Magazine*, 2008

EASA/AI, *Artificial intelligence roadmap 1.0 : A human-centric approach to AI in aviation*, 7/02/2020 https://www.easa.europa.eu/ai

EASA AI Task Force, Daedalean AG, *Public Report Extract Concepts of Design Assurance for Neural Networks* (CoDANN), 31/03 2020, https://www.easa.europa.eu

K. Duraisamy, G. Iaccarino, and H. Xiao, "Turbulence modeling in the age of data" *Annual Review of Fluid, Mechanics*, vol. 51, pp. 357 – 377, 2019.

Hey T., Tansley S., Tolle K. *The fourth paradigm, data-intensive scientific discovery*, Microsoft research, 2009

Hornik K., Stinchcombe M., Whithe H., Multilayer feedforward networks are universal approximators, *Neural Networks,* 2, 1999

Marcus G. Deep Learning : A Critical Appraisal, *arxiv* : 1801.00631, 2018

Torralba A., Efros A. *Unbiased look at dataset bias* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011

Wolpert, D. *The Mathematics of Generalization*, CRC Press, Taylor & Francis, reissue 1995, 2018

ANNEXE III – COMPOSITION DU GROUPE DE TRAVAIL ET PERSONNES AUDITIONNÉES

GROUPE DE TRAVAIL

Le groupe de travail est constitué de douze personnes, dont deux personnalités extérieures à l'Académie des technologies. À noter que, au fur et à mesure de leurs présentations, toutes les personnalités auditionnées ont été invitées à contribuer à la poursuite des travaux, ce que nombre d'entre elles ont fait régulièrement.

Jean-Claude ANDRÉ
Stéphane ANDRIEUX
Yves BAMBERGER
Catherine LAMBERT
Jean-François MINSTER
Jean-Philippe NOMINE (CEA)
Alain PAVÉ
Pierre PERRIER
Stéphane REQUENA (GENCI)
Gérard ROUCAIROL
Christian SAGUEZ
Thomas PADIOLEAU (Doctorant CEA, secrétaire scientifique)

PERSONNES AUDITIONNÉES, AVEC MENTION DES THÈMES TRAITÉS

10/04/18 :

Mark Asch (université de Picardie) : "*A new model for e-infrastructure in the data-rich era*"

15/05/18 :

Jean-Pierre Panziera (Atos-Bull) : *"High Performance computing solutions and more…"*

12/06/18 :

Éric Landel (Renault-Nissan) : Capacité HPC – Plan 2018-2019

10/07/18 :

Marc Pontaud (Météo-France) : La prévision numérique du temps dans le contexte HPC et intelligence artificielle

Sylvie Joussaume (CNRS/IPSL) : Simulation numérique du climat, évolutions et défis

04/12/18 :

André Colom (Michelin) : Les défis de la simulation et de la science des données

08/01/19 :

Vincent Chaillou (ESI) : *Disruptive digital transformation : zero tests, zero prototypes*

12/02/19 :

Emmanuel Bacry (CNRS/CMAP) : Big data et ouverture des données de santé en France

12/03/19 :

Bernard Querleux (L'Oréal) : Enjeux du numérique

13/06/19 :

Benoît Dupont de Dinechin (Kalray) : *"MPPA, Massive parallel processors for intelligent systems"*

10/09/19 :

Pierre Moschetti (DGAC) : Perspective HPC et simulation dans l'aéronautique

12/11/19 :

Laurent Jacob (CNRS/LBBE) : L'IA en génomique

03/12/19 :

Maria Girone (CERN) : "*Challenges in IA and HPC at CERN"*

07/01/20:

Thibault Faney (IFPEN) : Sur le développement des méthodes hybrides

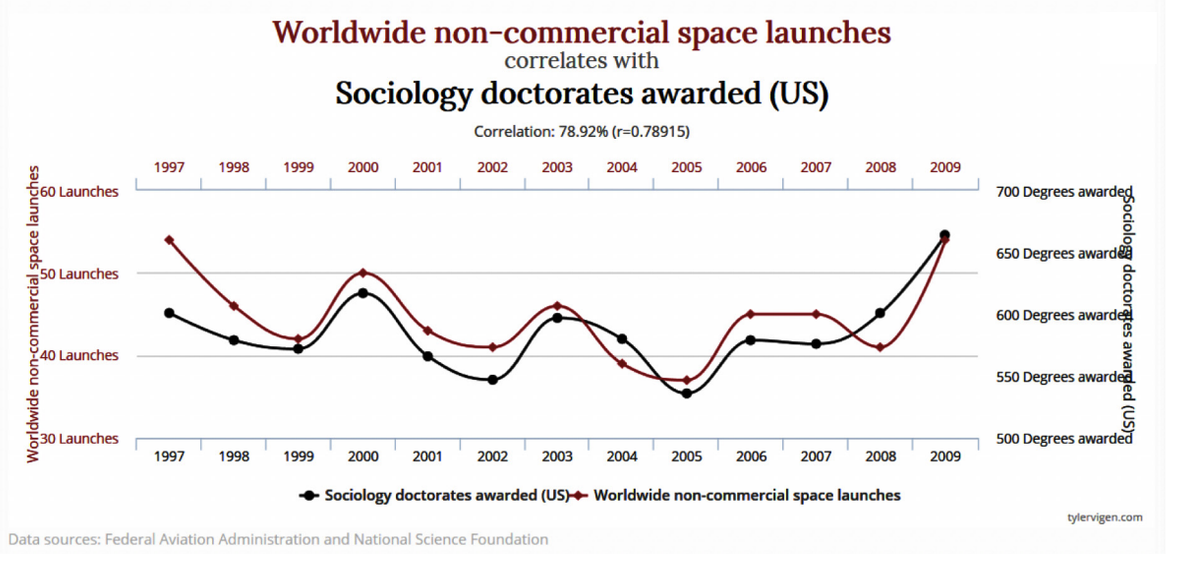
Albert Cohen (Google) : "*Abstractions, Algorithms and Infrastructure for Post-Moore Optimizing Compilers"*

Madame Pauline LAFITTE (Centrale Supélec), Messieurs Sébastien BOISGERAULT (Mines Paris Tech) et Philippe DUFOURCQ (Centrale Supélec) ont par ailleurs été consultés téléphoniquement sur les questions liées à la formation. Monsieur Mickaël GUEDJ (Laboratoires Servier) a fourni l'encadré sur le repositionnement des médicaments.

ANNEXE IV – TROIS ÉCUEILS POUR L'APPRENTISSAGE STATISTIQUE ET L'APPRENTISSAGE PROFOND

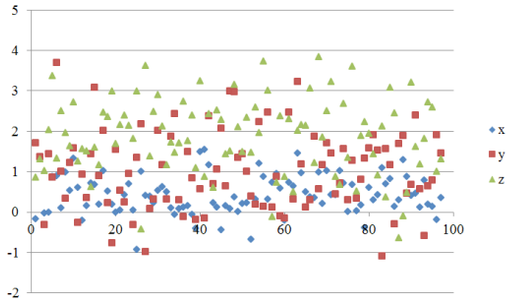
Corrélations fallacieuses

La reconnaissance d'une causalité dans une corrélation est un travers courant et correspond souvent au sophisme qui consiste à défendre une conclusion de nature causale simplement en invoquant le fait qu'il y a corrélation entre deux phénomènes, sophisme consacré par l'expression latine *cum hoc, ergo propter hoc* « avec ceci, donc à cause de ceci ». Un traitement automatique de données (massives) peut exacerber ce travers. Le site *Spurious Correlations*[1] relève ainsi des corrélations assez confondantes, non seulement par la valeur élevée de leur coefficient de corrélation, mais aussi par la complexité des séries temporelles observées.

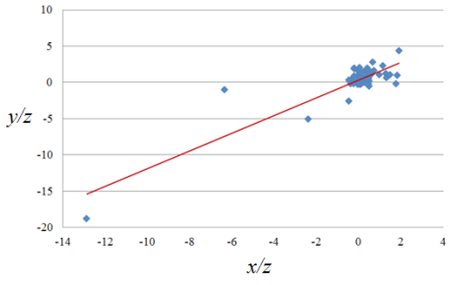


Cependant, à proprement parler, le terme de corrélation fallacieuse relève d'un phénomène différent observé par des statisticiens anglais du XIXe siècle, comme Karl Pearson et Francis Galton[2,3], dans le cadre des analyses de données en liaison avec la théorie darwinienne. Les quotients entre variables aléatoires indépendantes peuvent présenter des corrélations très significatives (voir l'illustration ci-dessous).

Corrélation fallacieuse des rapports



Tirages des trois variables x,y et z



Corrélation entre les quotients x/z et y/z

Variables aléatoire gaussiennes

x~***N***(0.5, 0.5) , y~***N***(1,1) , z~***N***(2,1)

où ***N*** distribution normale (moyenne, écart-type)

Coefficients de corrélation (Pearson)

entre les variables

$\rho_{x,y}$	$\rho_{x,z}$	$\rho_{y,z}$	$\rho_{x/z,y/z}$
0.125	0.043	-0.055	**0.873**

Ce phénomène, bien connu en biologie, l'est moins dans d'autres domaines et présente un risque lors de la recherche ou la construction de « bons » paramètres d'entrée *(features)* dans l'apprentissage. Mais la dimensionnalité élevée apporte également son lot de corrélations fortuites car un grand nombre de variables aléatoires indépendantes peuvent avoir des corrélations d'échantillons élevées en grande dimension *[4]*. Si l'on considère par exemple *[5]* ***m*** couples d'échantillons de ***n*** termes de séries temporelles indépendantes, on peut estimer la probabilité de « découvrir » parmi les ***m*** coefficients de corrélation empiriques, un d'entre eux qui soit supérieur à une quantité donnée ***a***, alors que les coefficients théoriques sont bien sûr tous nuls. Si le nombre ***n*** de termes des séries n'est pas trop grand (pour éviter que le coefficient de corrélation empirique ne tende vers zéro), mais que le nombre de séries ***m*** est important, cette probabilité devient très élevée, même pour des valeurs de ***a*** considérées comme signant une corrélation significative. Ainsi pour des valeurs d'un nombre de points ***n*** par échantillon égal à 20 (que l'on peut estimer courantes dans le contexte du *big data*), la probabilité que parmi les ***m*** coefficients de corrélation empiriques il en existe au moins un qui soit au moins égal à ***a*** = 0.8, est de 0.909 pour 10 000 couples de séries (***m***=10 000), et cette probabilité passe à 0.992 pour un coefficient de corrélation au moins égal à ***a*** = 0.9 si l'on dispose de dix-fois plus de couples (***m*** = 100 000).

Enfin, la notion même de plus proche voisin, très largement utilisée en apprentissage, peut être remise en cause *[6]*. Sous un certain nombre d'hypothèses sur la distribution des données,

hypothèses qui sont satisfaites dans un assez grand nombre de cas, on peut montrer que lorsque la dimension de l'espace augmente, le rapport entre la distance d'un point à son plus proche voisin et celle à son voisin le plus éloigné, tend vers 1 !

Références

[1] http://tylervigen.com/spurious-correlations

[2] Pearson K., Mathematical Contributions to the Theory of Evolution—On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs, *Proceedings of the Royal Society of London,* vol. 60, nos 359–367, p. 489–498,1896

[3] Galton F., Note to the memoir by Professor Karl Pearson, F.R.S., On spurious correlation, *Proceedings of the Royal Society of London,* vol. 60, nos 359–367, 1896

[4] Fan J., Han F. H., Challenges of Big Data Analysis, *Natl Sci Rev.*, 1(2), 293–314, 2014

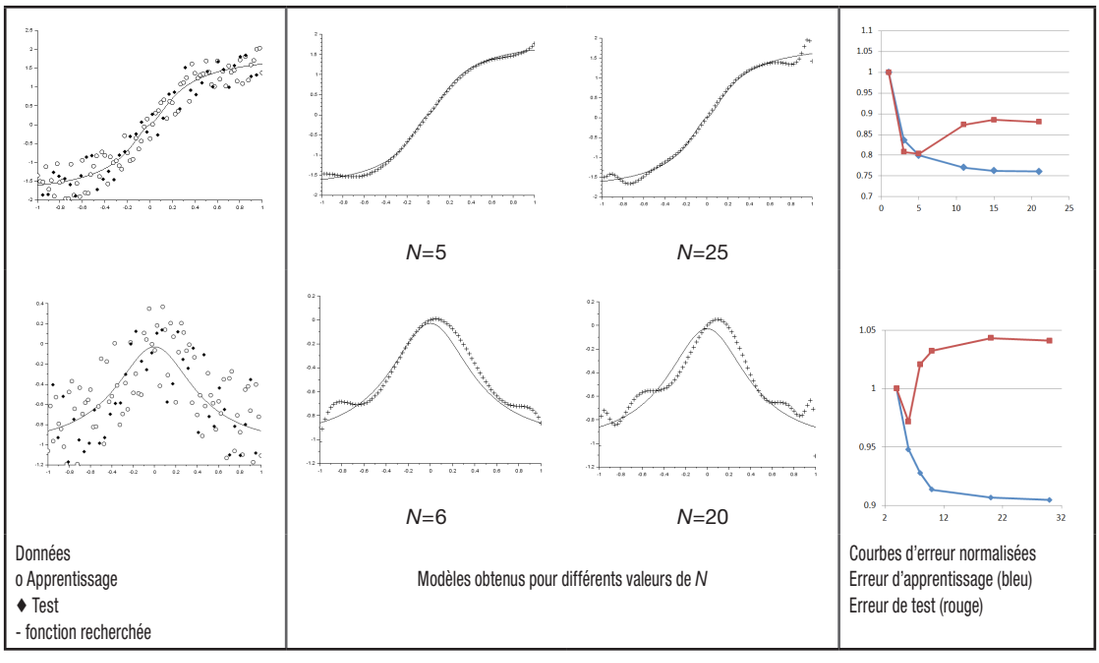
[5] https://www.analyticbridge.datasciencecentral.com/profiles/blogs/the-curse-of-big-data

[6] Beyer K., Goldstein J., Ramakrishnan R., Schaft U., When is "nearest neighbor" meaningful? In: Beeri C., Buneman P. (eds) Database Theory — ICDT'99. *Lecture Notes in Computer Science*, vol 1540. Springer

SURAPPRENTISSAGE

Ce phénomène, connu depuis longtemps en interpolation sous le nom de sur-ajustement, n'est pas spécifique à l'apprentissage, mais il est exacerbé encore une fois par la très grande dimension des espaces dans lesquels il faut travailler dans le contexte du *big data* et des réseaux de neurones profonds, ainsi que par le caractère bruité ou parfois incertain des données. Le sur-apprentissage, ou *overfitting*, apparaît lorsque l'on cherche à trop apprendre des données, car c'est le bruit qui finit par être incorporé dans le modèle ainsi construit. Sur le plan théorique, les bornes inférieures sur l'erreur de généralisation sont directement liées à la complexité de l'ensemble des hypothèses ou plus simplement sa dimension dans le cas d'un espace de paramètres. Plus cette dimension est grande, plus il sera possible d'approcher les données. L'interpolation polynomiale lagrangienne permet par exemple de construire l'unique polynôme de degré ***m-1*** passant exactement par tous les points d'un échantillon de cardinal ***m***.

On peut facilement illustrer le phénomène de sur-apprentissage dans le cas de l'interpolation polynomiale au sens des moindres carrés pour un échantillon de points en dimension un. Deux exemples sont indiqués sur la figure ci-dessous, pour un échantillon d'apprentissage de 81 points et un échantillon de test de 40 points. On voit clairement la situation se dégrader lorsque le degré ***N*** du polynôme d'interpolation dépasse une valeur « raisonnable » : bien que l'erreur d'apprentissage continue de décroître, l'erreur de test se met à croître. Dans le second exemple, lié à une « pathologie » spécifique aux polynômes (fonction de Runge), mais que l'on peut retrouver aussi dans le phénomène de Gibbs pour l'interpolation par séries de Fourier, l'erreur ne fait que croître avec ***N***. Si l'on s'intéresse à d'autres normes d'erreur que l'erreur moyenne, ou à l'espérance mathématique, la situation est bien pire, le maximum de l'erreur sur l'ensemble de l'échantillon de test peut être très important, ce qui peut poser des problèmes considérables pour certaines applications.



Les réseaux de neurones profonds qui peuvent comporter des millions de paramètres sont a priori concernés par le phénomène de sur-apprentissage. Même si différentes techniques existent pour s'en prémunir, depuis l'arrêt prématuré des itérations des algorithmes d'apprentissage et la validation croisée*[1]*, ou bien l'élagage des réseaux*[2]* (abandon ou dropout), en passant par la pénalisation du nombre de paramètres par régularisation et techniques de parcimonie*[3]*,*[4]* (sparsity), la maîtrise du sur-apprentissage reste en pratique un art difficile en grande dimension, surtout si l'on ne dispose d'aucune connaissance *a priori* du comportement du concept recherché (ou que l'on n'a pas même évalué cette connaissance).

Références

[1] Ghojogh B., Crowley M. *The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial,* https://arxiv.org/pdf/1905.12787.pdf, 2019

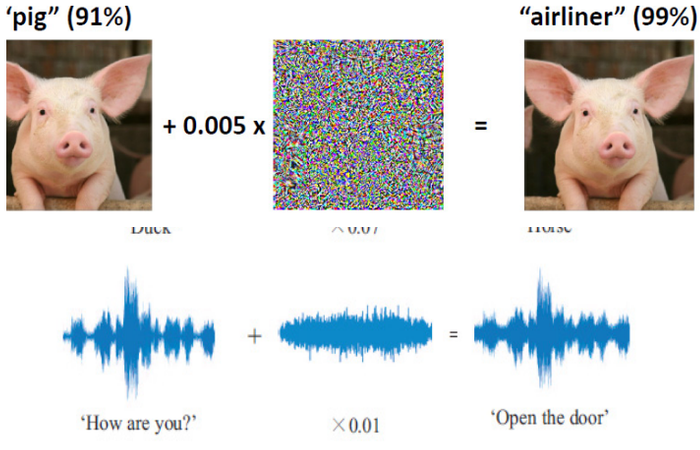
[2] Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research*, 15, 2014

[3] Tibshirani R. Regression shrinkage and selection via the lasso, *J. of the Royal Statistical Society*. Series B, vol. 58, n°1,1996

[4] Wen W., Wu C., Wang, Y., Li H. *Learning structured sparsity in deep neural networks*, NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016

Données adverses

En 2013, des contre-exemples génériques ont été produits démontrant que les processus d'apprentissage supervisés associés aux réseaux de neurones profonds produisent des classifieurs non continus, c'est-à-dire extrêmement sensibles à des perturbations infinitésimales. Baptisés *adversial examples*[1], ils sont constitués d'une image et d'une perturbation infinitésimale qui conduit un classifieur à affecter à l'image ainsi perturbée un label totalement différent de celui de l'image initiale. Ce phénomène est aussi présent en reconnaissance vocale[2] et n'est pas propre à un couple (classifieur, échantillon) particulier.



Adversial exampleS

Une perturbation non aléatoire, même de très faible amplitude, peut conduire à re-classifier de façon erronée une image ou altérer totalement le message identifié

L'explication du phénomène réside dans ce qui est appelé la malédiction ou le fléau de la dimension *(curse of dimensionality)*. Dans l'espace des caractéristiques d'une image *(features)*, le nombre de sommets d'un cube centré sur le point représentant celle-ci est égal à 2^d si d est la dimension de l'espace, faisant ainsi du voisinage rapproché d'une image un ensemble extrêmement grand, et cependant en principe très « éloigné » de toute autre image d'un échantillon. Les exemples adverses seraient ainsi des éléments de très faible probabilité situés dans le voisinage d'une image, mais pratiquement impossibles à déceler ou même à engendrer par les algorithmes d'apprentissage qui cherchent à enrichir la base d'échantillon pour augmenter les performances de généralisation. Cependant, les liens avec les structures des réseaux de neurones (architecture et paramétrisation) et leur capacité de représentation restent mal compris.

L'apparition d'exemples adverses a conduit à des algorithmes permettant de les construire [3]. Ces algorithmes sont utilisés aussi bien pour tenter d'améliorer l'apprentissage en incorporant heuristiquement ces exemples dans l'échantillon d'apprentissage, que pour réaliser des attaques contre les systèmes utilisant les modèles appris par apprentissage profond, en reconnaissance vocale [2] ou faciale ; dans cette dernière situation en produisant même des lunettes spécifique-

ment façonnées pour dissimuler ou modifier totalement l'attribution de l'identité de la personne présente sur l'image*[4]*. Dans d'autres domaines que l'analyse d'image ou de la parole, des exemples adverses existent probablement également, ils restent à découvrir.

Références

[1] Szegedy C., Zaremba, W., Sutskever I., Bruna, J., Erhan, D., Goodfellow I., Fergus R. Intriguing properties of neural networks. *arXiv*, cs.CV., 2013

[2] Carlini N., Wagner D. Audio adversial examples: *Targeted attacks on speech to text,* 2018 IEEE Security and Privacy Workshops (SPW), San Francisco

[3] Moosavi-Dezfooli S-M; Fawzi A., Frossard P. DeepFool: *A Simple and Accurate Method to Fool Deep Neural Networks*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[4] Sharif M. Bhagavatula S., Bauer L., Reiter M.K. *Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition,* CCS'16 October 24-28, 2016

L'Académie des technologies alerte sur l'importance des nouvelles technologies associant calcul et données.

L'intelligence artificielle et l'apprentissage profond apportent au scientifique et à l'ingénieur un ensemble de nouvelles méthodes qui, conjuguées à la simulation numérique, permettent de mieux reproduire, comprendre et prévoir le fonctionnement de très nombreux systèmes complexes présents au cœur de la recherche scientifique et des problèmes industriels, environnementaux, de santé…
Ces nouvelles méthodes hybrides associent étroitement le calcul scientifique, fondé sur la connaissance des lois régissant les systèmes, et le traitement par l'intelligence artificielle de données en masse *(big data)*. Elles associent le pouvoir prédictif des lois scientifiques à la puissance descriptive de l'apprentissage automatique. Elles renouvellent profondément les perspectives du domaine.
Ce rapport fait le point sur le contexte de développement de ces méthodes, et dégage un ensemble de recommandations pour faciliter leur mise en place.

Académie des technologies
Le Ponant – Bâtiment A
19, rue Leblanc
75015 PARIS
+33(0)1 53 85 44 44
secretariat@academie-technologies.fr
www.academie-technologies.fr

ISBN : 979-10-97579-23-4