ALLISTENE
l'alliance des sciences et technologies du numérique

Recommandations pour un Cloud pour l'IA

Jamal Atif (CNRS) et Frédéric Desprez (INRIA)

7 octobre 2019

Résumé

Ce rapport est issu d'une série de réunions organisées au CNRS entre Janvier et Mars 2019 et de discussions/visites avec les acteurs principaux du domaine.

Composition du groupe de Travail

- Jamal Atif (Université Paris-Dauphine, CNRS-INS2I, animateur)
- Frédéric Desprez (INRIA, animateur)
- Michel Dayde (CNRS, IRIT, Université de Toulouse, INP Toulouse)
- Christan Roux (IMT DRI, Paris)
- Laurent Simon (Bordeaux-INP, Talence)
- Stéphane Canu (Insa Rouen)
- Christophe Calvin (CEA)
- France Boillod-Cerneux (CEA)
- Alain Viari (INRIA)
- Agathe Guilloux (Université d'Évry Val d'Essonne, CNRS)
- Jean-Philippe Proux (GENCI)
- Laurent Crouzet (DGRI)

1 Introduction

L'alliance Allistene a été saisie par le MESRI, fin 2018, pour lancer une réflexion prospective, concernant un Cloud Recherche pour l'IA. Cette réflexion est à considérer comme une suite logique à celle menée fin 2017 dans le cadre d'une proposition de plate-forme de recherche pour l'IA (Dossier GENIAL [1]). Elle s'appuie sur plusieurs éléments:

- les recommandations faites sur ce thème dans le rapport Villani *Donner un sens à l'intelligence artificielle: pour une stratégie nationale et européenne* [8] où le besoin d'une telle infrastructure est mentionnée explicitement,
- l'inscription de cette problématique dans le cadre du programme national IA sur la dimension d'offres d'infrastructures de calcul,

- la nécessité de veiller à la complémentarité de cette éventuelle nouvelle offre par rapport aux infrastructures existantes et en particulier sur ce qui va être déployé au sein de GENCI/Idris pour la recherche en IA (plate-forme Jean Zay),
- l'importance d'associer à cette infrastructure des communautés scientifiques variées, usagers pour leur recherche propre de l'IA (Bio-Santé, SHS,...) et qui ne sont pas familières du domaine du Calcul Hautes Performances mais aussi les spécialistes de la recherche en IA qui ont besoin de d'attaquer les couches les plus basses des architectures et des systèmes,
- la possibilité d'élargir ensuite la proposition d'usage à d'autres champs thématiques de recherche que l'IA, en Sciences du Numérique et au-delà,
- prendre en compte les besoins en formation et en enseignement des futurs ingénieurs et chercheurs en IA,
- l'insertion de cette offre et sa compatibilité avec la doctrine générale, promue par l'Etat, en matière d'usage de l'informatique en nuage au sein des administrations publiques
- l'articulation et la cross-fertilisation de cette offre de Cloud avec les infrastructures Cloud déjà existante dans la communauté recherche (e.g. France GRILLES, GRID'5000/SILECS, ...)

Le document est structuré comme suit:

- une première section donne un état des lieux des plates-formes dédiées principalement à l'IA que ce soit pour les académiques ou les industriels.
- La Section 3 décrit les besoins par communautés et quelles sont les ressources et outils utilisés actuellement.
- Une Section compare les offres actuelles de GENCI et de Grid'5000/SILECS par rapport à ce qui pourrait être offert par un Cloud dédié à l'IA.
- Une dernière section liste nos recommandations et conclusions.

2 Etat des Lieux des Infrastructures de Calcul pour l'IA

2.1 En France

Jusqu'à présent, les chercheurs en IA avaient à leur disposition en France plusieurs moyens non-dédiés pour effectuer leur recherche comme les supercalculateurs de GENCI opérés par les grands centres nationaux (TGCC, IDRIS, CINES), l'infrastructure expérimentale Grid'5000/SILECS et parfois des clusters disponibles dans les laboratoires ou dans les centres de recherche Inria.

Nous détaillons maintenant plus particulièrement les offres de GENCI/mésocentres régionaux et de Grid'5000/SILECS.

TGIR GENCI Avec la TGIR GENCI l'écosystème HPC national a été structuré pour mettre à disposition des moyens de calcul diversifiés et complémentaires dans les 3 centres de calcul nationaux en lien avec 17 mésocentres en région par l'Equipex Equip@meso.

En 2018, il a été demandé à GENCI de mettre à disposition des moyens de calcul pour la recherche française en IA, ce qui se concrétisera par l'installation au printemps 2019 d'une machine convergée (HPC+IA) de 14 PFlops (avec notamment 1044 GPU Nvidia V100 répartis dans 261 serveurs de 4 GPU V100 chacun), nommée Jean Zay à l'IDRIS (CNRS) ainsi que la mise en place d'un

nouveau mode d'accès, dédié et adapté aux usages en IA (notamment en mode essai/erreur lors de la recherche d'hyperparamètres), dynamique et interactif, permettant rapidement et tout au long de l'année d'avoir accès à ces ressources de calcul. Cette machine convergée Jean Zay pour sa partie IA, vise tout d'abord à adresser les besoins de la communauté de recherche en IA, c'est-à-dire développant des nouveaux outils et méthodes, puis dans un second temps la communauté de recherche (climat, fusion, combustion, matériaux, biologie, ...) souhaitant utiliser l'IA. Sa configuration évoluera légèrement en 2019 (ajout de 150 à 200 GPU supplémentaires) mais surtout fin 2020 avec une extension forte de ses capacités (pour le périmètre HPC pour le moment).

Depuis 2 ans GENCI offre également un accès aux serveurs issus de sa cellule de veille technologique, notamment Ouessant composé de 12 serveurs de 4 GPU P100 représentant en ensemble de 48 GPU Nvidia P100 et qui est déjà très sollicité par des travaux en IA. D'autres matériels à base de GPU sont disponibles dans les 2 autres centres mais sont en priorité dédiés à des usages de post-traitement et de visualisation graphique des données générées par les calculateurs. Par ailleurs, d'autres évaluations sont en cours ou prévues par la cellule de veille technologique autour des technologies mémoire (mémoire Optane Intel), processeurs et accélérateurs de calcul (Intel Nervana, GPU AMD, processeur optique LightOn, ...) qui pourraient aussi bénéficier à la communauté IA française.

En région très peu sont les centres qui se sont dotés de matériels dédiés à un usage en IA à base de GPU. Comme pour les centres nationaux il s'agit plus d'un usage connexe au HPC. Cependant quelques expérimentations en IA de plus en plus importantes comme cela a été le cas sur la machine Ouessant viennent solliciter les équipements installés par les mésocentres. Elles permettent depuis peu aux centres régionaux d'appréhender les nouveaux services, équipements et socles logiciels indispensables au besoin de l'IA.

Le matériel disponible dans les mésocentres équipés de carte GPU au 02/2019 est déjà significatif :

- À Rouen, le Criann dispose de 48 K80 (12 serveurs x 4 K80) et de 19 P100 (8 serveurs x 2 P100 + 1 serveur x 3 P100), à Nice le Cicada dispose de 16 M2070Q (8 serveurs x 2 M2070Q), à Toulouse au Calmip c'est 48 V100 (12 serveurs x 4 V100) qui peuvent servir à faire de l'IA.
- L'Université Aix-Marseille met à disposition 28 K80 (7 serveurs x 2 x 2 K80), 4 P100 (2 serveurs x 2 P100, 4 K20Xm et 5 K40M pour tous les usages du HPC y compris de l'IA comme c'est le cas pour Grenoble à Gricad avec ses 9 K20 (9 serveurs x 1 K20) et ses 4 serveurs de 4 GPU V100 chacun.
- Quant à Reims dans le centre Roméo qui est de loin le mésocentre le plus avancé en matériel et en compétence, il dispose d'un ensemble de 280 P100 dans 70 serveurs composés de 4 P100 mais aussi d'un DGX1 de 8 V100 et de 260 K20X de ses 130 anciens serveurs de 2 K20X chacun.

Grid'5000/SILECS Grid'5000 est une infrastructure pour la recherche expérimentale dans tous les domaines de l'informatique, et en particulier le calcul à haute performance (HPC), le Cloud, le Big Data, et l'Intelligence Artificielle. L'infrastructure est composée de 8 sites, 31 clusters, 828 serveurs, 12328 coeurs, et 88 GPUs. Elle est utilisée chaque année par environ 600 utilisateurs, qui produisent 150 publications.

L'infrastructure fournit un environnement très flexible, fortement reconfigurable et observable, permettant des expériences à tous les niveaux de la pile logicielle. Pour cela, elle s'appuie sur du

déploiement bare-metal d'images systèmes (permettant à l'utilisateur d'utiliser l'environnement logiciel de son choix, et donc de lever toutes les contraintes pour installer des piles logicielles spécifiques), sur la possibilité d'avoir les droits "root" sur l'image par défaut (avec remise en état des ressources en fin de réservation). Grid'5000/SILECS fournit aussi plusieurs services pour la gestion de gros volumes de données.

Grid'5000/SILECS est déjà largement utilisé pour des travaux de recherche en IA, notamment dans les domaines suivants: traitement automatique des langues et de la parole, robotique, ordonnancement de systèmes complexes, vision, indexation de contenus multimédias, génie logiciel automatisé.

Teralab Teralab (`https://www.teralab-datascience.fr/`), faisant partie des plates-formes technologiques de l'institut Carnot Télécom et Société Numérique et conçu dans une logique de *Digital Innovation Hub* (DIH), est sélectionné comme tel par la commission européenne parmi les 30 DIH sur la thématique Intelligence Artificielle. Teralab propose une plateforme d'ingénierie et d'expertise en IA et Bigdata destinée aux acteurs tant du monde de la recherche et de la formation que des entreprises.

Réflexions générales Aujourd'hui, l'intelligence artificielle associée au calcul intensif peut constituer un outil majeur pour les scientifiques de toutes les disciplines, non seulement pour ce qui concerne l'apprentissage profond où le calcul intensif peut apporter une énorme plus-value lors de la phase d'apprentissage à large échelle, mais également pour apporter au calcul intensif des outils de post-traitement des données générées afin d'en permettre un filtrage précoce avant stockage ou un couplage inédit entre modèles de simulations et modèles appris pour accélérer la convergence des simulations numériques ou élargir les domaines explorés. Le calculateur Jean Zay de GENCI hébergé à l'IDRIS adressera ces deux objectifs en plus de celui d'être une machine de production en HPC. Il ne faut pas sous-estimer aussi de forts besoins de stockage de données, en préalable à l'utilisation de méthodes IA, qui peuvent être satisfait par cette offre Cloud. L'infrastructure Cloud offre de plus la possibilité « d'amener » les chaines de traitement (avec des machines virtuelles ou des conteneurs) vers les jeux de données lorsqu'ils sont de grande taille.

La convergence d'usages entre la simulation numérique et l'IA, couplant les codes de simulation avec des modèles appris, post-traitant à la volée les données générées par les simulations pour ne stocker que les données pertinentes et ainsi gagner temps et énergie, permettant un apprentissage massif et automatique (autoDL, autoML) de réseaux de neurones tout en développant une IA explicable (comme préconisé dans le rapport Villani) basée par exemple sur un couplage entre réseaux de neurones et IA symbolique. Elle permettra aussi l'avènement de nouveaux usages comme le couplage des supercalculateurs convergés avec les très grands instruments de recherche (TGIR) pour leur dimensionnement/calibration en amont mais aussi et surtout l'analyse et la valorisation de leurs résultats en aval, de par le volume des données générées par la nouvelle génération d'instruments. On notera que ce point est soulevé dans le Livre Blanc du CNRS [6] sur les Données au sein de l'INSU qui note l'utilisation des chaînes de traitement impliquant des moyens centralisés du type HPC et ressources plus distribuées ou périphériques.

Il existe avec les solutions cloud publics des différences significatives avec les supercalculateurs HPC. Elles sont souvent appropriées pour des applications très généralistes, très loin des applications intensives de simulation numérique qui nécessitent de mobiliser efficacement plu-

sieurs dizaines de milliers de processeurs en parallèle et/ou de stocker à haute performance et à très grande vitesse des données massives de résultats (par exemple modèles climatiques ou cosmologiques incorporant l'intelligence artificielle, modélisation haute-fidélité de la combustion, médecine personnalisée brassant données omiques hétérogènes massives, mise à disposition/retraitement de données instrumentales ou de simulation, recherches sur l'aide à la décision).

Pourquoi ne pas mettre à disposition du monde scientifique une offre globale et permettre à tout utilisateur/développeur en IA, académique ou industriel, de pouvoir accéder le plus simplement en souplesse via un portail d'accès unique, et en fonction de son besoin (depuis l'expérimentation bas niveau, l'évangélisation et jusqu'aux activités lucratives) à des ressources de calcul variées et compétitives. Ce portail servirait à la fois à orchestrer de façon transparente les besoins sur l'infrastructure la plus appropriée (en termes de niveaux de service, confidentialité/accréditation, disponibilité) mais aussi à proposer des offres intégrées de type EaaS (*Expertise as a Service*), IAaaS/MLaaS (*IA et ML as a service*). Elle permettrait aussi de coller aux évolutions rapides des logiciels et du matériels. Ce continuum permettrait aussi d'avoir une interface mobile autorisant par exemple le débordement d'activités de recherche ouverte vers le Cloud (*burst mode*). Enfin elle attirerait les communautés de recherche déjà utilisatrices d'infrastructures comme GRID'5000/SILECS ou France GRILLES (dont le Cloud fédéré, certes de petite taille, est assez chargé) pour des usages qui relèvent de l'IA et sont déjà présents.

Repris en partie du Livre Banc sur les données au CNRS mais quelques éléments pertinents:

On observe une convergence entre l'analyse de données à grande échelle, le calcul haute performance et l'IA. Il semble clair que les systèmes centralisés (c.-à-d. centres HPC et systèmes Cloud) et décentralisés (plates-formes distribuées de services) doivent être intégrés/fédérés au sein d'un réseau de ressources et de services adressant la complexité et la diversité de la logistique des données tout au long des chaînes de production et d'utilisation des données. Cela requiert une nouvelle stratégie (méthodologique, technologique et culturelle) et une nouvelle architecture, avec de nouveaux enjeux logiciels, afin d'interfacer et d'interopérer une diversité de plates-formes technologiques incluant : Plates-formes périphériques de traitement et de réduction des données permettant le traitement, l'agrégation, et la réduction « intelligente » des flux (volumes, vitesses) au plus proche de leurs sources (grands instruments, systèmes d'observations, réseaux de capteurs…) dans des environnements souvent reculés, ainsi que le pilotage adaptatif de leurs systèmes d'acquisition.

Plates-formes de services de calcul et d'analyse de données, distribuées et fédérées mutualisant dans des environnements multi-utilisateur et multi-application, des services flexibles de communication, de logiciel, de stockage, de calcul (analyse de données, HPC), d'exécution adaptés au Big Data et aux nouvelles technologies de virtualisation ainsi qu'à l'utilisation croissante de méthodes de type statistique et apprentissage machine, avec des flux de traitement et d'analyse proches des vitesses d'accès aux données.

Plates-formes centralisées de type HPC et Cloud, c'est-à-dire à l'échelle des grands centres nationaux et régionaux. Elles concentrent des ressources de très haute performance dont l'utilisation doit être maximisée pour servir des communautés multiples, avec de nouveaux environnements permettant de supporter les technologies de virtualisation et adaptés à des configurations complexes de workflows couplant HPC et analyse de données, ainsi que l'utilisation croissante des méthodes de type Intelligence Artificielle (pour l'analyse de données, la représentation des connaissances et l'aide à la décision) : grands ensembles de simulations numériques couplées, ingestion et assimilation de grands jeux de données multi-source.

Plates-formes fédérées d'archivage, de curation et de distribution des données mutualisant ressources, services et expertises pour le stockage, la curation et la mise à disposition de données durant leur cycle de vie, et dont les volumes et la diversité impliquent aujourd'hui des capacités croissantes de stockage et de calcul.

2.2 Quelques exemples à l'étranger

Le rapport GENIAL [1] liste un certain nombre de plates-formes disponibles pour la recherche en IA dans le monde dans son annexe 2.

On peut aussi citer:

- La plate-forme Bridges-AI [3], aux Etats Unies, qui donne accès à 88 GPUs NVIDIA Volta via des containers Singularity [7] et des images disques pré-installées (dont Caffe, Caffe2, PyTorch, TensorFlow, ...),
- La plateforme Open Compass: `https://pscedu.github.io/AI-BD-website/`
- UIUC Deep Learning Instrument: `http://www.ncsa.illinois.edu/enabling/data/deep_learning`
- La plateforme ICE `https://ice.sics.se`, en Suède un datacenter orienté recherche qui met à disposition des environnements dédiés à l'IA
- Toujours en Suède, le centre HPC (Swedish National Infrastructure for Computing) est en passe d'acquérir une machine dédiée à l'IA/ML `http::www.snic.se`

2.3 Clouds publics

La plupart des Clouds publics ont des offres ciblant l'IA comme par exemple Amazon, Google, Microsoft Azure ou OVH.

Amazon L'offre AWS AI d'Amazon met à disposition des entreprises ou particuliers un environnement de développement et un ensemble d'outils de Machine Learning, en particulier de deep learning pour le déploiement rapide d'algorithmes sur des cas d'usages. Pour ce faire des Jupyter notebooks pré-construits sont mis à disposition ainsi que des outils d'omtimisation des hyper-paramètres (AutoML). L'offre AWS ML propose une facturation selon la tâche (e.g. Amazon Transcribe pour la reconnaissance vocale, Amazon Rekognition pour l'analyse d'images et de vidéo, etc.). La tarification par exemple pour Amazon Transcribe est de 0,0004 USD PAR SECONDE. Source `https://aws.amazon.com/fr/transcribe/pricing/`

Google L'offre Cloud de Google propose le Cloud ML Engine `https://cloud.google.com/ml-engine/`. Ce service (MLaaS) intègre les frameworks Scikit-learn, XGBoost, Keras, tensorflow, permet l'utilisation de conteneurs personnalisés, l'entraînement distribué, le transfert de modèles via la SDK TensorFlow, des outils de hypermarameters tuning et d'AutOML, etc. La tarification pour la phase d'entraînement d'un modèle, en Europe, et en mode basique est de BASIC 0,3212 USD PAR SECONDE. Plus d'informations sur les tarifications se trouvent à cette adresse: `https://cloud.google.com/ml-engine/pricing?hl=fr` Il convient de noter que Google met désormais à disposition, gratuitement, l'environnement Google Colab `https://colab.research.google.com` un environnement online sous forme de Jupyter Notebook avec accès à des ressources GPU ou TPU.

Microsoft Azure Microsoft via Azure propose Microsoft Azure Machine Learning Studio qui offre un environnement de développement et de déploiement d'algorithmes de Machine Learning. Le service propose des abonnements pour l'utilisation du Studio et une tarficiation spécifique à l'usage de GPU. A titre d'exemple, la réservation d'un GPU V100 (semblable aux ressources Jean Zay) est de 3,781 Euros/heure en incluant des fonctionnalités spécifiques au ML (source : `https://azure.microsoft.com/fr-fr/pricing/details/machine-learning-service/`). A l'instart de Google Colab, Microsoft met à disposition un notebook gratuit : `https://notebooks.azure.com/`

Une étude comparative des offres précédentes en termes de fonctionnalités, est disponible sur un site spécialisé [2]

OVH OVH met à disposition des solutions d'hébergement Cloud. L'offre OVH pour un cloud de recherche en IA est disponible en annexe de ce document.

Par ailleurs, OVH travaille sur une offre de *Machine Learning as a Service* appelée Pré-science. Celle-ci, basée sur des outils open-source, est utilisée en interne pour l'instant.

Autres Clouds Les offres Cloud fleurissent actuellement en France et dans le monde et certains fournisseurs ont, ou auront, des offres ciblant spécifiquement l'IA (MLaaS). On peut citer par exemple IBM Watson, ATOS/Bull Codex AI, vast.ai, paperspace, etc.

3 Analyse des besoins par communautés

L'intelligence artificielle trouve des champs d'application dans l'ensemble des disciplines, en particulier celles fondant leur démarche scientifique sur la mesure, les études de terrain, et de façon générale sur des données d'observation. En particulier, nous observons un intérêt croissant des différentes communautés scientifiques pour l'utilisation des techniques d'apprentissage automatique; intérêt suscité par les succès récents de l'apprentissage profond (*deep learning*). Ci-après nous dressons une cartographie partielle des grands champs scientifiques utilisant ou trouvant un intérêt dans l'utilisation des techniques d'intelligence artificielle moderne.

3.1 Sciences humaines et sociales (SHS)

3.1.1 Grandes thématiques utilisant l'IA

Actuellement, les thématiques concernées sont celles qui utilisent les corpus qui peuvent être constitués de texte, de parole, de vidéos ou de documents. Cela concerne plus précisément les thématiques suivantes:

- Archéologie: fouille de carnets d'archéologues
- Droit: justice prédictive, analyse de contrats (cf. *Robot Lawyer*)
- Sciences politiques: analyse de discours, gestion d'archives, comparaisons, etc.
- Histoire: domaine en pointe des humanités numériques. Analyse de documents anciens (inscriptions sur des monuments, manuscrits, archives de presse) ou comparaison de la façon dont des évènements ont été relatés dans différents pays.
- Géographie: analyse de cartes, de relevés, de photographies aériennes.

- Sociologie: analyse de réseaux sociaux, étude de l'émergence de normes
- Psychologie et psychologie cognitive: analyse d'écrits ou de conversations et plus généralement analyse de données comportementales.
- Economie : l'IA est déjà très présente dans la finance au travers des techniques de Machine Learning. De nouvelles problématiques émergent avec la disponibilité de grandes masses de données: économie des plateformes, crypto-monnaies, fintech, évolution du capital, changements des modes de travail, etc.

Techniques et outils utilisés

- Classification automatique, apprentissage automatique
- Extraction d'information (automates ou modèles probabilistes)
- TAL (essentiellement pour l'enrichissement)
- Indexation et recherche d'information (logiciel ElasticSearch ou Solr), y compris sur documents structurés XML
- Analyse d'images et OCR
- Traitement du signal et parole
- Analyse de graphe (par ex logiciel Gephi)
- API de moissonnage de grosses bases de données

Dans la plupart des cas il s'agit d'utiliser des méthodes de data mining ou de text mining assez classiques. Les techniques avancées du Machine Learning, en particulier du deep learning ne sont pas encore pleinement exploités par la communauté SHS (à part en économie). A noter que pour le TAL ou l'analyse vidéo, ces outils sont aujourd'hui la norme.

Ressources mobilisées actuellement: cloud, mésocentres, etc. La principale ressource nationale vient de la TGIR HumaNum[1] qui fournit logiciels de traitement et espaces de stockage. Cette dernière utilise les ressources du centre de calcul de l'IN2P3.

3.2 Biologie-Santé

Grandes thématiques utilisant l'IA

- Imagerie biologique/biomédicale : neuro-imagerie pour l'analyse et l'interprétation des images (et leur croisement avec d'autres données), cardiologie, radiologie, histologie et anatomopathologie, biologie cellulaire, par exemple, pour la reconnaissance de structures cellulaires en microscopie(s) optique(s)
- Biologie moléculaire et "omiques" : génomique, pour la détection de variants, protéomique et métabolomique, pour l'interprétation des données de spectrométrie de masse, autres omiques (notamment transcripto-, methylo-, métalobolo- miques, single cell), principalement pour croiser des données de natures diverses.
- Biologie structurale : relation structure-fonction, simulation moléculaire
- Agriculture: monitoring de culture et utilisation des sols et de l'eau

1. https://www.huma-num.fr/

- Epidémiologie et santé publique : pharmaco-vigilance, évaluation de politique publiques. Les données utilisées sont celles des bases SNIIRAM (Système national d'information inter-régimes de l'Assurance maladie), PMSI (Programme de médicalisation des systèmes d'information), CepiDC (Centre d'épidémiologie sur les causes médicales de décès), et les cohortes, par exemple Constance.
- Recherche clinique : exploitation des données administrativo-hospitalières, données textuelles, d'imagerie, génétique
- Médecine personnalisée: aide au diagnostic et à la thérapeutique
- Technologies pour la santé: capteurs biomédicaux, robotique médicale

Techniques et outils utilisés Toutes les techniques du machine learning et de l'intelligence artificielle sont d'intérêt. En imagerie et radiologie, les techniques de deep learning propres à l'image sont massivement utilisées (segmentation, segmentation sémantique, reconnaissance, génération). Les techniques de deep learning sont également utilisées en biologie cellulaire, génomique (prédiction de phénotypes). Les techniques d'apprentissage par transfert sont utilisées notamment pour les données omiques. Enfin les données de cohorte et administrativo-hospitalières contiennent des textes (comptes-rendus) qui nécessitent l'utilisation de techniques de deep learning pour le texte. Ces techniques (LSTM, etc) sont également utilisées pour les données de santé publique et de cohortes pour prendre en compte leur aspect temporel.

Ressources mobilisées actuellement: cloud, mésocentres, etc. Il existe de nombreuses plateformes de bio-informatique (par exemple à l'institut Curie, Imagine – Hopital Necker, ...) dans lesquelles sont traitées les données omiques et cliniques. Le Health Data Hub devrait permettre de se doter d'outils pour les données du SNDS.

3.3 Sciences de l'Univers

Les sciences de l'univers regroupent les domaines d'astronomie et d'astrophysique, de l'océan, de l'atmosphère, des sciences de la Terre, des surfaces et interfaces continentales, dont le but scientifique est comprendre et prévoir le fonctionnement et l'évolution de ces systèmes dans leurs environnements. Ces domaines partagent une culture scientifique et des pratiques de recherche fondées sur l'observation (sol, air, mer, spatial) à long terme (¿30 ans) des systèmes naturels, intégrant un large spectre d'échelles spatiales et temporelles et incluant une vaste palette d'outils d'analyse in situ d'échantillons issus des milieux naturels terrestres ou extraterrestres. Ces disciplines intègrent en plus de la gestion (au sens large) des données, le traitement et l'analyse de grands jeux de données multi-sources pour en extraire de nouvelles informations et les distiller sous des formes réutilisables. Ce sont des domaines où l'utilisation des techniques de machine learning est traditionnelle.

Grandes thématiques utilisant l'IA Les approches de type inférence/inversion (p. ex. cosmologie et astrophysique, sismologie, géodésie, gravimétrie) permettent, dans un cadre probabiliste, d'améliorer la description des modèles de systèmes naturels, ainsi que la reconstruction de leurs états. Les approches d'assimilation de données (p. ex. climat et météorologie, champs magnétiques terrestres et planétaires), dans un cadre variationnel ou statistique, permettent d'améliorer les

prévisions de l'évolution de ces modèles en reconstruisant un état initial aussi consistant que possible avec leur dynamique.

Inférence et assimilation de données sont des workflows complexes orchestrant de multiples phases HPC (simulation numérique) et HDA (traitement et analyse de données). Ils exploitent aujourd'hui des approches probabilistes au travers d'ensembles de simulations numériques, permettant d'explorer des espaces de modèles complexes, et leurs conditions initiales, ainsi que de quantifier les incertitudes directes et inverses. Ils combinent souvent « apprentissage machine », optimisation stochastique, théorie de l'information et transport optimal.

Techniques et outils utilisés Tous les domaines de machine learning, de gestion de masses de données, de traitement de signal. L'utilisation du deep learning est de plus en plus croissante notamment pour l'analyse de signaux.

Ressources mobilisées actuellement: cloud, mésocentres, etc. Forte utilisation des moyens nationaux mais aussi infrastructures – souvent communautaires – distribuées faisant appel à de la virtualisation (conteneurs ou autres) pour les services.

Cependant des besoins sont exprimés en termes de plates-formes de services, de calcul et d'analyse de données qui fédèrent et mutualisent des services flexibles de stockage, de calcul (HTC, HPC), de communication et de logiciel permettant des flux de traitement proches des vitesses d'accès aux données. Ces infrastructures de supportent des utilisateurs et des applications multiples (p. ex. le service labellisé Terapix à l'IAP pour l'exploitation des données MegaCAM du CFHT ou encore l'ARC node ALMA à l'IRAM pour la réduction des données de l'interféromètre millimétrique ALMA de l'ESO) dans un environnement collaboratif, réactif et résilient qui intègre des éléments de HPC et HDA avec des services Cloud de traitement en flux, des technologies de virtualisation (conteneurs) et d'exécution adaptés au Big Data (p. ex. Spark, Storm) qui pourraient donc bénéficier d'une offre Cloud plus musclée en France.

3.4 Physique

Plusieurs domaines de la physique partagent les mêmes fondements théoriques avec l'apprentissage statistique. Par ailleurs, une sous-classe de réseaux de neurones trouvent leurs fondements en physique statistique (réseaux de Hopfield, machines de Boltzmann par exemple), et de grands espoirs reposent sur l'exploitation des concepts en physique statistique pour expliquer le fonctionnement des réseaux de neurones profonds. Parmi les domaines de recherche prospectif en apprentissage automatique, on trouve le quantum machine learning. Nous dressons ci-après quelques liens entre apprentissage automatique et sous-domaine de la physique. Ce qui suit repose grandement sur l'article "Machine learning and the physical sciences" [5].

Grandes thématiques utilisant l'IA

Physique statistique Il s'agit ici non d'appliquer les concepts d'apprentissage machine, mais d'utiliser la physique statistique et de ses outils pour comprendre les avancées récentes en apprentissage profond. A titre d'exemple, la théorie de Spin Glass a été exploitée pour comprendre la paysage de la fonction de cout en deep learning.

Physique des particules et cosmologie Les techniques d'apprentissage automatique sont utilisées pour l'identification de particules et la détection et sélection précoces d'évènements

(sélection d'un sous-ensemble de collisions pour une tâche donnée: recherche du Boson de Higgs, supersymètrie ou matière noire). Le challenge Kaggle HiggsML pour la détection du boson de Higgs à partir de données du LHC a été parmi les challenges les plus populaires de la plateforme Kaggle. D'autres applications concernent la classification de jet (gerbes hadroniques composées de quarks et de gluons), la détection de neutrinos, triggering, ondes gravitationnelles, trous noirs, etc.

Many-Body Quantum Matter Des architectures neuronales adaptées à l'encodage d'états quantiques ont été introduites. Notons que les succès récents de l'apprentissage profond, et en particulier des réseaux génératifs adversariaux est dû à leur capacité à échantillonner dans un régime de très grande dimension, ce qui est d'un intérêt majeurs pour des applications en physique. Un autre champ d'application est la simulation efficace de multi-corps (many-body simulations), grâce à l'utilisation des techniques d'apprentissage pour accélérer l'échantillonnage Monte Carlo.

Calcul quantique Des travaux émergent sur le quantum machine learning, mais aussi pour l'utilisation des techniques d'apprentissage automatique pour le contrôle et la préparation des qbits et le design de codes correcteurs.

Techniques et outils utilisés Toutes les techniques du Machine Learning. On constate un intérêt croissant pour les techniques d'apprentissage profond à la fois dans un cadre supervisé, non supervisé et par renforcement.

Ressources mobilisées actuellement: cloud, mésocentres, etc. La communauté de la physique disposent de très grands instruments et dinfrastructures synchrotron, neutron et laser à électrons libres, certaines ayant une dimension européenne ou même mondiale. Ces infrastructures portent collectivement le nom de PaN pour "photon and neutron" et incluent les synchrotrons SOLEIL (http://www.synchrotron-SOLEIL.fr) et ESRF (http://www.esrf.eu), les infrastructures de diffusion neutronique ILL (http://www.ill.eu) et Orphée (http://www-centre-saclay.cea.fr/fr/Reacteur-Orphee) et dans le futur l'ESS (https://europeanspallationsource.se), ainsi ue depuis peu, le laser à électrons libres, X-FEL (https://www.xfel.eu). Pour le volet traitement via des outils de deep learning, la communauté française ne dispose pas à ce jour d'outils dédiés. Le calculateur Jean-Zay peut combler ce manque.

3.5 Chimie et sciences des matériaux

L'intérêt pour l'utilisation de l'apprentissage automatique dans l'ensemble des domaines de la chimie et des sciences des matériaux, connaît un vrai engouement, corroboré notamment par des résultats récents montrant par exemple comme les outils d'apprentissage ont guidé à la découverte de matériaux depuis les calculs chemico-quantique à la synthèse. Le récent article publié dans Nature [4] dresse un panorama de l'apport de l'apprentissage automatique aux différents domaine de la chimie : chimie organique et supramoléculaire, chimie moléculaire et analytique, chimie de synthèse, chimie quantique, etc. D'autres travaux ont montré comment la robotisation augmentée d'algorithmes d'intelligence artificielle peut automatiser tout le processus de découverte de nouveaux procédés. La figure 1 reprise de [4] montre l'évolution de la recherche en chimie computationnelle et ses liens avec l'apprentissage automatique.

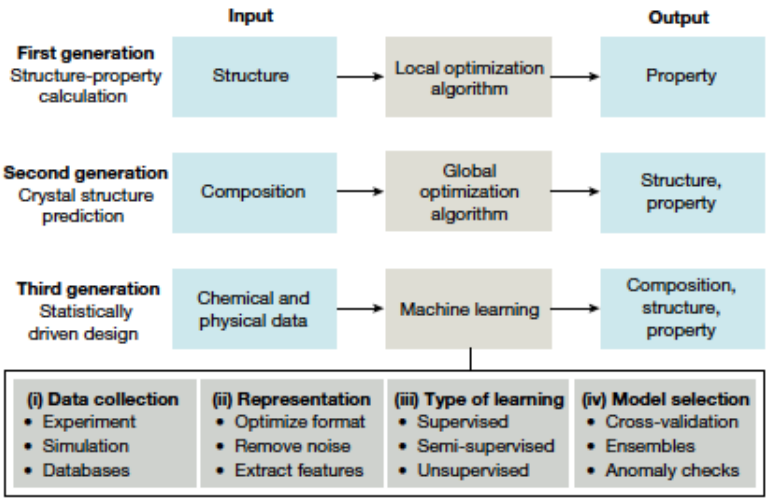

FIG. 1 – *Evolution du workflow scientifique en chimie computationnelle. Source [4]*

Grandes thématiques utilisant l'IA L'ensemble des thématiques relevant de la chimie computationnelle et des sciences des matériaux. On liste ci-après les applications les plus emblématiques.

Découverte de nouvelles molécules Cela inclut la découverte de nouvelles réactions, de nouvelles propriétés physico-chimiques

Prédiction d'activités et découverte automatique de médicaments Inférence de la relation structure-activité/propriété

Synthèse de molécules L'analyse retrosynthétique automatique n'est pas nouvelle en soi, mais les approches par apprentissage par renforcement ouvrent de nouvelles perspectives.

Techniques et outils utilisés L'ensemble des techniques de machine learning et des approches à base de règles sont d'intérêt pour les applications en chimie. La table 2 tirée de [4] résume les outils d'intelligence artificielle dédiés à la recherche en chimie.

Ressources mobilisées actuellement: cloud, mésocentres, etc. La chimie est très impliquée dans des TGIR relevant de la physique (ESRF, ILL, SOLEIL...). D'autres infrastructures nationales existent telles que le a TGIR résonance magnétique nucléaire à très hauts champs (RMN-THC), l'IR RENARD (REseau NAtional de Résonance paramagnétique interDisciplinaire) et le très grand équipement FT-ICR à haut champ. La communauté française ne dispose pas à ce jour de ressources de calcul dédiés à l'utilisation de l'IA en chimie.

3.6 Enseignement

L'un des enjeux du plan national en IA est le doublement des étudiants formés dans ce domaine stratégique. Or l'enseignement des techniques d'apprentissage automatique, et en particulier de l'apprentissage profond, ne peut se faire sans mise en pratique. Par ailleurs, ce domaine, fortement empirique et expérimental, est adapté à un enseignement par projet, handson, hackathon

Name	Description	URL
General-purpose machine-learning frameworks		
Caret	Package for machine learning in R	https://topepo.github.io/caret
Deeplearning4j	Distributed deep learning for Java	https://deeplearning4j.org
H2O.ai	Machine-learning platform written in Java that can be imported as a Python or R library	https://h2o.ai
Keras	High-level neural-network API written in Python	https://keras.io
Mlpack	Scalable machine-learning library written in C++	https://mlpack.org
Scikit-learn	Machine-learning and data-mining member of the scikit family of toolboxes built around the SciPy Python library	http://scikit-learn.org
Weka	Collection of machine-learning algorithms and tasks written in Java	https://cs.waikato.ac.nz/ml/weka
Machine-learning tools for molecules and materials		
Amp	Package to facilitate machine learning for atomistic calculations	https://bitbucket.org/andrewpeterson/amp
ANI	Neural-network potentials for organic molecules with Python interface	https://github.com/isayev/ASE_ANI
COMBO	Python library with emphasis on scalability and efficiency	https://github.com/tsudalab/combo
DeepChem	Python library for deep learning of chemical systems	https://deepchem.io
GAP	Gaussian approximation potentials	http://libatoms.org/Home/Software
MatMiner	Python library for assisting machine learning in materials science	https://hackingmaterials.github.io/matminer
NOMAD	Collection of tools to explore correlations in materials datasets	https://analytics-toolkit.nomad-coe.eu
PROPhet	Code to integrate machine-learning techniques with quantum-chemistry approaches	https://github.com/biklooost/PROPhet
TensorMol	Neural-network chemistry package	https://github.com/jparkhill/TensorMol

FIG. 2 – *Publicly accessible learning resources and tools related to machine learning. Source [4]*

et autres innovations pédagogiques. A ce jour, les étudiants inscrits en Master en France ou dans les grandes écoles en particulier d'ingénieur, pour des raisons de sécurité, ne peuvent bénéficier des infrastructures de calcul nationales. Les enseignants se tournent vers des solutions gratuites (ou payantes selon les accès et les espaces mémoires utilisés), attractives il faut le dire, telles que Google Colab, Azure notebooks, AWS, Hardis et IBM, dans le cadre d'offres éducation générales ou de partenariats ou conventions de mécenat spécifiques. Des sociétés de revente d'heures de calcul pour l'éducation fleurissent par ailleurs. Via ces clouds, les étudiants de ces formations hors numérique peuvent appliquer les technologies de l'IA sous réserve de bénéficier de services adaptés à une utilisation non experte.

Notons enfin que les écoles d'ingénieurs et UFR spécialisées en informatique forment aussi sur leurs plates-formes de travaux pratiques internes (machines "standard" avec processeurs multicœurs et périphériques spécifiques, accélérateurs GPGPUs, FPGAs ou micro-contrôleurs par exemple); souples d'utilisation en mode interactif, peu coûteuses et évolutives, ces plates-formes internes sont essentielles au développement des compétences techniques et scientifiques dans le numérique. L'accès à une ou des plates-formes mutualisées, via le cloud, qu'il soit privé ou public, apparaît comme un complément essentiel pour l'expérimentation et le calibrage sur des grandes tailles de données.

Malgré tout, ces offres trouvent leurs limites quand des étudiants travaillent sur des projets de plusieurs semaines voire mois, comme c'est désormais l'usage avec des challenges data sciences. On ne peut que regretter qu'une politique nationale à destination des étudiants ne puisse exister. Au delà de la mise à disposition de puissances de calcul, une politique de mutualisation de ressources pédagogiques devrait être encouragée.

4 Positionnement de l'offre Cloud IA par rapport aux infrastructures nationales GENCI et Grid'5000/SILECS

Le tableau suivant compare l'offre actuelle GENCI pour l'IA, l'infrastructure de recherche Grid'5000/SILECS et ce qui pourrait être offert par un Cloud dédié à l'IA.

Plate-forme	**GENCI actuel (peut évoluer)**	**Grid'5000/SILECS**	**CloudIA**
Type d'accès	DARI et Accès dynamique*	Accès dynamique	À la demande
Temps d'accès < 10kh (Limite pouvant être revue si besoin)	Qques jours	Qques heures	Qques heures
Temps d'accès > 10kh (idem)	1-6 mois (via des campagnes et sélection)	Qques heures (régulé par charte + vérification à posteriori)	?
Recherche	Ouverte avec publication	Obligatoire	Pas d'obligation
Chercheurs	Académiques et industriels	Académiques et industriels	Académiques et industriels
Usage	Développement & Recherche	Développement & Recherche	Recherche & Production
Objectifs	Recherche en IA et utilisation de l'IA	Recherche en IA	Recherche en IA et utilisation de l'IA
Financement de l'accès	Gratuit	Gratuit	Gratuit
Investissement	Machines / 6 ans	Renouvellement régulier	Usage / ponctuel
Respect de la PPST (ZRR)	Oui (imposé HFDS)	Non	Non
Respect de la PSSIE	Oui (imposé organisme)	Oui (imposé organisme)	Non
Hébergement souverain	Oui (par construction)	Oui (par construction)	Si possible
Évaluation	Oui / Non si < 10kh	Non	Non
Architecture	V100 actuellement	Multiple	Multiple
Caractéristiques	Nœuds physiques dédiés et support containers	Nœuds physiques	Nœuds virtuels adaptables
Limite	1044 (sauf si financement additionnel)	88 GPUs actuellement (achats en cours)	"Infinie"
Parallélisme	De 1 GPU à totalité de la machine Réseau d'interconnexion très performant	de 1 GPU à la totalité de l'infrastructure. Réseau d'interconnexion performant Réseau d'interconnexion très performant	De 1 GPU à 4 ou 8 GPU Réseau peu performant
Hébergement	Centres nationaux	Organismes partenaires	Opérateur privé / centres nationaux ? / infra répartie (Grid'5000/SILECS)

5 Conclusions et recommandations

Les sections précédentes justifient la mise en place d'un cloud pour la recherche en IA, destiné en priorité à des communautés d'usagers de cette technologie et des expérimentateurs de nouvelles architectures pour l'IA – en opposition à la communauté des développeurs des algorithmes de l'IA qui devraient se tourner en priorité vers les centres nationaux.

Ce cloud devrait à la fois satisfaire les contraintes de protection du patrimoine scientifique national, et offrir des conditions d'accès plus souples – en termes de sécurité - que les centres nationaux (ex. Jean Zay) pour entre autres accueillir des étudiants de Master. Cela plaide pour un "cloud dédié" (2ème cercle selon la doctrine du cloud computing de l'état) opéré par un fournisseur non assujetti à une législation extra-européenne.

Le dimensionnement de cette offre cloud devrait être pensé pour à la fois satisfaire les besoins des utilisateurs potentiels, mais aussi en prenant en compte les offres disponibles par les centres de calcul nationaux, dans un souci de cohérence globale. Nous recommandons que la solution proposée offre le tiers en capacité de calcul GPU du supercalculateur Jean Zay, à savoir une puissance autour de 3 PetaFlop/s, ou l'équivalent de 300 GPU V100 32Go. Un élément important, et différenciant, de l'offre cloud est la capacité de mettre à disposition des architectures matérielles hétérogènes et évolutives (facilitant ainsi des expérimentations type Hardware as a Service). Elle peut aussi permettre d'offrir un accès à une gamme d'architectures d'intérêt plus large pour les applications relatives à l'IA en incluant des alternatives d'ores et déjà annoncées par les constructeurs à base de processeurs offrant des jeux d'instructions et des opérateurs spécifiques dans le matériel ou des accélérateurs type FPGA ou nouvelles gammes de GPU (e.g. nouvelle gamme AMD).

Une offre cloud dédié doit être accompagnée d'une politique d'accès et d'usage adéquate. Nous préconisons une politique d'accès inspirée du modèle Grid'5000/SILECS. Cependant, mettre à disposition des moyens de calcul ne suffira pas pour accompagner l'appropriation des algorithmes et technologies de l'IA par des communautés scientifiques non spécialisées ni en intelligence artificielle ni en informatique.

Il sera nécessaire de mettre en place des actions incitatives (formation, challenges, etc.), un portail facile d'accès, une strate logicielle MLaaS (*Machine Learning as a Service*), AIaaS (*Artificial Intelligence as a Service*) et HaaS (*Hardware as a Service*). Mais sans accompagnement humain, les membres du GT pensent que cela sera largement insuffisant.

Pour faire aboutir l'effort de la puissance publique pour réussir le plan national en IA, nous recommandons fortement la mise en place d'un centre de compétences en Intelligence Artificielle, indépendant des futurs centres 3IA afin de servir la communauté française dans son ensemble. Ce centre de compétences, composé d'une dizaine de personnes, et commun au supercalculateur Jean-Zay et l'offre cloud, aura pour vocation de former et accompagner les communautés d'utilisateurs, de porter des codes, etc.

5.1 Souveraineté et sécurité

5.1.1 Contrainte de souveraineté pour la protection du patrimoine scientifique

L'intelligence artificielle est au centre d'enjeux stratégiques, industriels mais aussi géostratégiques. La protection du patrimoine scientifique, qu'il relève de découvertes majeures ou de production de données expérimentales mérite une attention particulière. Le groupe de travail

recommande d'avoir recours à un fournisseur français ou européen non soumis à des législations extra-européennes, et garantissant une protection des données et algorithmes hébergés sur le territoire européen.

5.1.2 Contraintes d'accès

Le Cloud pour l'IA doit permettre un accès aisé pour tous et en particulier pour les étudiants des différentes formations intéressées par l'intelligence artificielle. La sécurité doit être néanmoins assurée avec un suivi adéquat de l'utilisation des ressources et des moyens mis à disposition.

En ce qui concerne l'accès au Cloud par les startups et PMEs, un accès spécifique pourra être mis en place pour leur permettre de faire des expérimentations de faisabilité technologiques avec un modèle tarifaire analogue à celui pratiqué dans le programme SIMSEO.

5.2 Dimensionnement

Même s'il est encore difficile de chiffrer précisément ce que sera l'utilisation des différentes plates-formes dédiées à l'IA en France, compte tenu de la taille de la plate-forme Jean Zay et des besoins actuels exprimés par les communautés, il nous semble important d'avoir accès à minima à une puissance de traitement équivalente à un tiers de Jean Zay, soit environ 3 Pétaflops. Cette taille sera sans aucun doute appelée à évoluer avec l'adhésion et la formation des différentes communautés.

Concernant la taille du volume de données stockées, il est actuellement difficile de l'évaluer. De plus, il serait intéressant que ces données soient partagées entre les diverses plates-formes mise à disposition afin d'éviter une réplication inutile et des durées de transferts prohibitifs.

Un autre aspect important d'un cloud pour l'IA est la capacité à être évolutif et réactif à l'arrivée de nouveaux processeurs spécifiques (processeurs spécialisés, FPGA, GPU de nouvelles générations, ...). L'évolution des processeurs actuels, et en particulier des GPU, est très dynamique et il est fort peu probable que les grands centres de calcul puissent acquérir des processeurs pour effectuer des tests et des mises au point d'algorithmes. Par exemple, plusieurs sociétés proposent maintenant des processeurs dédiés à l'IA (Graphcore ou Cerebras par exemple) et Intel sortira bientôt un processeur appelé Spring Crest (après avoir acquis la société Nervana) aux performances prometteuses. Un cloud pour l'IA pourrait offrir aux chercheurs et aux entreprises de telles plates-formes de tests en quantité suffisante, comme ce qui est fait par exemple dans le cadre de la plate-forme Grid'5000/SILECS ou GENCI, en attendant des évolutions notables dans les mésocentres ou chez GENCI.

5.3 Politique d'accès et d'usage

5.3.1 Souplesse d'accès

Les demandes de comptes pourront être au fil de l'eau avec des demandes simplifiées, permettant l'accès au plus grand nombre et en particulier aux étudiants. La création des comptes devrait être dématérialisée et conforme aux usages en informatique moderne.

5.3.2 Actions incitatives

Le groupe recommande la mise en place d'une structure de gouvernance très légère, pour éviter l'accumulation des strates administratives. Celle-ci doit être en constante concertation avec GENCI. Des actions de communication sont primordiales pour attirer les différentes communautés.

5.3.3 Collecte de résultats

Le groupe recommande la collecte de résultats sous forme de publications soutenues par l'outil. Les utilisateurs du Cloud IA exploitant les ressources mises à leur disposition doivent s'engager à produire leurs résultats. Il serait intéressant d'avoir également la trace des travaux effectués sur le cloud (durée, matériels utilisés) et ceci relié au projet l'utilisant.

5.3.4 Politique de stockage et de partage des données

Les données sont centrales pour les algorithmes et les applications de l'IA. Leur ouverture au monde de la recherche est cruciale et recommandée par les principaux organismes et par la commission européenne (recommandation des principes FAIR notamment). Celà n'est pas spécifique au cloud IA mais doit être abordé de manière globale et concertée pour toutes les plates-formes mises à disposition des chercheurs et industriels. Par ailleurs il est important que ces données puissent être partagées entre les plates-formes de manière simple et efficace.

5.3.5 Régulation de demande

Le cloud IA est avant tout dédié à des expérimentations d'une durée réduite dans un premier temps. De la même manière que des heures sont demandées sur les plates-formes des mésocentres nationaux, il conviendra de mettre en place une politique de régulation pour les demandes d'une durée supérieure à 10kh d'utilisation. Par ailleurs, afin de pouvoir chiffrer l'utilisation (et éventuellement la facturer), le cloud IA devra fournir un détail d'utilisation par utilisateur et par projet.

5.4 Strates logicielles

5.4.1 Portail

Pour de nombreuses communautés interrogées dans le groupe de travail, l'accès aux infrastructures de traitement et de stockage doit pouvoir se faire de la manière la plus simple possible, comme par exemple via un portail web donnant accès à différentes applications sans programmation ni scripts.

Une offre globale permettant un accès simple aux différentes solutions matérielles et logicielles pour les utilisateurs/développeurs en IA, académiques ou industriels, pourrait être proposée sous forme d'un portail d'accès unique, et en fonction des besoins (depuis l'expérimentation, l'évangélisation jusqu'aux activités lucratives). Ce portail servirait à la fois à orchestrer de façon transparente les besoins sur l'infrastructure la plus appropriée (en termes de niveaux de service, confidentialité/accréditation, disponibilité) mais aussi à proposer des offres intégrées de type EaaS (*Expertise as a Service*), IAaaS/MLaaS (*IA et ML as a service*) mais aussi HaaS (*Hardware as*

a Service). Ce continuum permettrait aussi d'avoir potentiellement une interface mobile autorisant par exemple le débordement d'activités de recherche ouverte vers le Cloud (*burst mode*). Enfin ce portail permettrait d'avoir un recensement des utilisateurs et de mettre en place des actions de formation et d'échanges techniques et scientifiques.

5.4.2 MLaaS, AIaaS

Les fournisseurs de Clouds privés (Amazon, Google, OVH) ont des offres d'*IA as a Service* ou de *Machine Learning as a Service* pour les utilisateurs de ce type de technologies logicielles et matérielles.

La liste des outils logiciels disponibles de manière simplifiée via le cloud devra être déterminée avec les communautés et en fonction de l'évolution du domaine. De plus le Cloud permettra de fournir des ressources de traitement et de stockage de manière élastique, en fonction des besoins croissants ou décroissants, et ceci de manière transparente pour les utilisateurs.

5.4.3 HaaS

Pour certains informaticiens expérimentateurs, il est important de pouvoir avoir accès aux couches les plus basses, et ceci sans virtualisation. Une plate-forme offrant donc un accès directement au matériel peut être requis, avant de passer dans un second temps sur des plates-formes comme Jean Zay.

5.5 Cohérence de la politique nationale

Le groupe recommande de mettre à disposition du monde scientifique une **offre globale** et permettre à tout utilisateur/développeur en IA, académique ou industriel, de pouvoir accéder le plus simplement en **souplesse** via un portail d'**accès unique**, et en fonction de son besoin (depuis l'expérimentation bas niveau, l'évangélisation et jusqu'aux activités lucratives) à des ressources de calcul variées et compétitives. Ce portail servirait à la fois à orchestrer de façon transparente les besoins sur l'infrastructure la plus appropriée (en termes de niveaux de service, confidentialité/accréditation, disponibilité) mais aussi à proposer des offres intégrées de type EaaS (*Expertise as a Service*), IAaaS/MLaaS (*IA et ML as a Service*). Elle permettrait aussi de coller aux évolutions rapides des logiciels et du matériels. Ce continuum permettrait aussi d'avoir une interface mobile autorisant par exemple le débordement d'activités de recherche ouverte vers le Cloud (burst mode). Enfin elle attirerait les communautés de recherche déjà utilisatrices d'infrastructures comme GRID'5000/SILECS ou France GRILLES (dont le Cloud fédéré, certes de petite taille, est assez chargé) pour des usages qui relèvent de l'IA et sont déjà présents. Cette politique nationale ainsi définie et consolidée constituerait un apport effectif, dans le cadre de l'offre des plates-formes technologiques, à la politique européenne portée en particulier par le consortium AI4EU (`https://www.ai4eu.eu/`).

5.6 Accompagnement humain

Avec la convergence en cours des usages entre HPC, traitement de données instrumentales et computationnelles et l'intelligence artificielle, le défi humain majeur à relever est celui de la

formation initiale et continue et de l'accompagnement des communautés scientifiques dans l'architecture et l'écriture de leurs programmes de modélisation par simulation numérique.

Il sera donc important de créer un Centre de compétences en IA et ceci en coordination entre GENCI et le Cloud IA.

- Faciliter l'accès au supercalculateur avec une aide au portage des codes,
- Se maintenir à l'état de l'art pour évaluer et comparer les méthodes existantes pour l'aide à l'innovation en IA,
- La formation d'utilisateurs non-experts de technologies de l'IA,
- Mettre à jour les logiciels utiles en IA,
- Gestion et mise à disposition des données pour l'IA,
- Suivre l'évolution des architectures dédiées à l'IA,
- Soutien à des hackathons et des écoles d'été en IA,
- Développer des cas d'usage et des démonstrateurs des dernières techniques en IA (en association avec un réseau de living labs ?),
- Réaliser des études de marché, des livres blancs et des études de faisabilité en IA.

Références

[1] Allistène. Infrastructure de recherche pour l'intelligence artificielle, January 2018. https://www.allistene.fr/pour-la-creation-dun-infrastructure-de-recherche-dediee-a-lintelligence-artificielle/.

[2] ALTEXSOFT. Comparaison des offres de Cloud pour l'IA en termes de fonctionnalités. https://www.altexsoft.com/blog/datascience/comparing-machine-learning-as-a-service-amazon-microsoft-azure-google-cloud-ai-ibm-watson/.

[3] BRIDGE-AI. Bridges-AI: A Platform for Deep Learning, other Kinds of Machine Learning, Graph Analytics, and Data Science. https://psc.edu/bridges-ai-early-users.

[4] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547, 2018.

[5] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *arXiv preprint arXiv:1903.10563*, 2019.

[6] Daydé et al. Livre blanc sur les données au CNRS :etat des lieux et pratiques, 2018. http://www.cocin.cnrs.fr/spip.php?article8.

[7] LLBL. Singularity. https://singularity.lbl.gov/.

[8] C. Villani. Donner un sens à l'intelligence artificielle (IA), 2018. http://www.enseignementsup-recherche.gouv.fr/cid128577/rapport-de-cedric-villani-donner-un-sens-a-l-intelligence-artificielle-ia.html.