



Commission économique pour l'Europe**Conférence des statisticiens européens****Soixante-neuvième réunion plénière**

Genève, 23-25 juin 2021

Point 5 de l'ordre du jour

Travaux du Groupe de haut niveau**sur la modernisation de la statistique officielle****Projet relatif à l'apprentissage automatique, mené sous les auspices du Groupe de haut niveau sur la modernisation de la statistique officielle****Document établi par l'équipe du projet relatif à l'apprentissage automatique, qui relève du Groupe de haut niveau sur la modernisation de la statistique officielle***Résumé*

Le présent document est le rapport principal sur le projet relatif à l'apprentissage automatique, qui relève du Groupe de haut niveau sur la modernisation de la statistique (Groupe de haut niveau) et dont la réalisation était prioritaire pour 2019 et 2020. Ce projet visait à déterminer dans quels domaines l'apprentissage automatique pouvait donner de la valeur ajoutée, à améliorer la capacité des organismes de statistique d'y recourir et à repérer les obstacles communs à l'utilisation accrue d'une telle méthode dans la statistique.

On trouvera dans le présent rapport un résumé des travaux menés ainsi que des recommandations relatives aux moyens de faire progresser l'apprentissage automatique dans les organismes de statistique. Ce rapport se fonde sur les enseignements et retours d'expérience découlant de trois modules de travail : études pilotes (module 1), qualité (module 2) et intégration (module 3). Les rapports distincts sur les trois modules de travail et leurs supports d'accompagnement (codes, données) sont disponibles sur [la plateforme wiki de la CEE consacrée aux statistiques \(UNECE Statistics Wiki\)](#).

Le présent rapport est soumis à la Conférence des statisticiens européens pour information.



I. Introduction

1. Le projet relatif à l'apprentissage automatique a été lancé par le [Groupe de haut niveau de la CEE](#) sur la modernisation de la statistique officielle en mars 2019 et s'est achevé en décembre 2020. Il a réuni plus de 120 participants de 23 pays, de 33 organismes nationaux et de quatre organisations internationales, qui ont mené des travaux et collaboré en vue de promouvoir le recours à l'apprentissage automatique pour la production de statistiques officielles. À cette fin, les participants ont démontré l'intérêt de produire des statistiques officielles au moyen de l'apprentissage automatique, établi un cadre de qualité pour orienter le développement futur de cette méthode, et répertorié les problèmes posés par l'intégration des solutions d'apprentissage automatique dans les processus de production et tenté d'y apporter des solutions. Le 13 novembre 2020, des rapports, d'autres documents, du code, des données et de nombreuses références ont été placés sur la plateforme wiki publique de la CEE consacrée aux statistiques à l'intention des acteurs de la statistique officielle. Un webinaire a ensuite été organisé les 16 et 17 novembre 2020.

2. Les connaissances, l'expérience et la compréhension acquises au fil du projet permettent d'affirmer que l'apprentissage automatique n'est pas juste un concept à la mode. Selon les études menées, cette méthode peut être intégrée aux opérations de codage et de classification et contribuer ainsi à l'obtention de résultats de meilleure qualité à un coût égal ou moindre. Elle donne de bons résultats en ce qui concerne la préparation et l'imputation, dans certains cas, mais il convient de procéder à de nouvelles études dans ces domaines. Il faut absolument exploiter les mégadonnées, par exemple analyser les images satellite ou aériennes. Le succès de l'apprentissage automatique dépend en grande partie de la faculté de mobiliser les spécialistes de diverses disciplines et de tirer parti de leurs connaissances, notamment aux fins de la production durable de données de grande qualité, en quantités suffisantes, qui permettront d'affiner les algorithmes et de suivre les résultats des opérations bénéficiant de l'apprentissage automatique de façon efficace.

3. Même si les études pilotes et autres faits nouveaux font clairement ressortir la valeur ajoutée de l'apprentissage automatique, l'intégration de celui-ci dans les processus de production continue de poser problème. Le projet prévoit un cadre de qualité pour les algorithmes statistiques, et des solutions à d'autres problèmes d'intégration y sont proposées afin que l'apprentissage automatique soit diffusé et accepté plus facilement dans les organisations.

4. Le présent rapport revient sur les circonstances dans lesquelles le projet a été lancé et en décrit le déroulement. Il énumère les principaux résultats obtenus, puis les enseignements importants tirés en vue de faire accepter et de faire progresser l'apprentissage automatique, met en avant les résultats de projets connexes et contient des suggestions relatives aux travaux futurs.

5. Les rapports distincts sur les trois modules de travail et leurs supports d'accompagnement (codes, données), tels que détaillés ci-après, sont disponibles sur la [plateforme wiki de la CEE \(UNECE Statistics Wiki\)](#) :

- a) Module 1 – Études pilotes :
 - [Résumé analytique des études pilotes](#) ;
 - [Rapports et autres documents relatifs aux 19 études pilotes](#) ayant servi à évaluer la valeur ajoutée de l'apprentissage automatique pour le codage et la classification, la préparation et l'imputation ainsi que l'exploitation de données d'imagerie ;
 - Rapports thématiques analysant les méthodes utilisées et résultats obtenus dans le cadre des études pilotes : Module 1, thème 1 : [Rapport sur le codage et la classification](#) ; Module 1, thème 2 : [Rapport sur la préparation et l'imputation](#) ; Module 1, thème 3 : [Rapport sur l'analyse des données d'imagerie](#) ;
- b) Module 2 – Qualité : Cadre de qualité pour les algorithmes statistiques ;

- c) Module 3 – Intégration : [Rapport](#) qui répertorie et analyse les obstacles communs à l'intégration de l'apprentissage automatique dans les processus de production, établi à partir d'une enquête menée par l'équipe du projet ;
- d) Autres supports visant à aider les utilisateurs à déployer l'apprentissage automatique :
 - Code utilisé dans certaines des études pilotes, disponible sous la rubrique [Studies and Codes](#) ;
 - Deux ensembles de données permettant de se former et de mener des essais en matière d'apprentissage automatique, disponibles sous la rubrique [Learning and Training](#) ;
 - Liens vers des supports d'apprentissage et de formation, disponibles sous la rubrique [Learning and Training](#).

II. Modernisation des organismes de statistique

6. Les organismes nationaux de statistique (ONS) sont mis au défi de mieux répondre au besoin croissant d'informations statistiques plus pertinentes, récentes, détaillées et accessibles et de services de données fiables qui peuvent être utilisés aux fins de la prise de décisions claires et efficaces fondées sur ces données. Qui plus est, on les pousse à faire preuve de la plus grande efficacité et à satisfaire à ces attentes sans dépasser le budget existant. À cela s'ajoute la quantité toujours plus importante de données disponibles, dont la source, le format et le niveau de qualité sont très variés.

7. Enfin, les ONS sont en concurrence avec un nombre croissant d'organismes publics et privés, de taille variable, qui produisent et diffusent les statistiques plus rapidement et les rendent plus faciles d'accès, attirant ainsi l'attention des décideurs et de nombreux autres utilisateurs malgré certains problèmes de pertinence ou de qualité. Les concurrents peuvent produire ces statistiques pour plusieurs raisons, notamment car ils conçoivent des méthodes différentes pour le traitement des données (par exemple, des algorithmes d'apprentissage automatique) ou bien y ont accès rapidement, et les intègrent dans leur processus de production, car ils ont de meilleures capacités informatiques et car les normes qu'ils appliquent en matière de qualité, de transparence, d'éthique et de confidentialité ne sont pas aussi strictes.

8. Toutefois, il s'agit là des domaines dans lesquels les ONS disposent d'un avantage concurrentiel. En publiant des renseignements relatifs aux sources des données, aux méthodes et aux divers indicateurs, les ONS font preuve d'une grande transparence. Ils ont à eux tous une longue expérience de l'intégration efficace de différentes sources de données et sont tenus par la loi de respecter la confidentialité des données et d'en empêcher la divulgation. En outre, ils ont la possibilité de renforcer encore leurs capacités grâce aux réseaux internationaux de professionnels qui unissent leurs efforts autour de besoins et de priorités communs.

9. En plus de compter sur leur savoir-faire individuel et collectif, les ONS doivent, pour conserver leur raison d'être, avoir une culture axée sur l'adaptation et répondre rapidement, de façon toujours fiable, aux besoins des parties prenantes en matière de données, sous la forme à la fois d'informations et de services statistiques. Pour mener à bien leur mission, ils s'appuient notamment sur l'élaboration de méthodes et de processus solides intégrés à la production de statistiques officielles. Comme on l'a vu plus haut, ils doivent dans ce cadre faire face à une multiplication des demandes de données, des sources et des producteurs, ceux-ci parvenant de plus en plus facilement à trouver les sources de données correspondant aux demandes.

III. Proposition de projet relatif à l'apprentissage automatique

10. Les obstacles susmentionnés sont la priorité du [Groupe de haut niveau sur la modernisation de la statistique officielle \(Groupe de haut niveau\)](#), constitué de statisticiens en chef déterminés à diriger activement la modernisation des ONS. Ces statisticiens ont pour mission de s'employer ensemble à répertorier les tendances, les menaces et les possibilités en matière de modernisation des ONS. Le Groupe de haut niveau offre aux experts une plateforme commune qui leur permet d'élaborer des solutions de manière souple et adaptable.

11. Le Groupe de haut niveau bénéficie du soutien du Conseil d'administration, chargé de la gestion stratégique des activités en cours, à savoir des activités de quatre groupes de travail permanents et de deux projets limités dans le temps. L'un de ces groupes de travail, le Blue-Sky Thinking Network (BSTN), est tourné vers l'avenir et sert de « fabrique à idées » pour les acteurs de la modernisation de la statistique. Il a bien compris que beaucoup de nouveaux producteurs de données exploitaient un large éventail de sources de données au moyen d'approches et de méthodes différentes de celles qu'employaient généralement les organismes de statistique. De nombreuses sources de données supposent l'adoption d'approches et de méthodes innovantes, telles que l'apprentissage automatique et l'intelligence artificielle. Lors de son atelier de novembre 2018, le Groupe de haut niveau a une nouvelle fois admis qu'il importait d'intégrer ces technologies à la production de statistiques officielles et a, dans cette optique, soutenu la proposition du BSTN de lancer un projet relatif à l'apprentissage automatique. La note d'information du BSTN justifiait le projet en ces termes :

« L'intérêt suscité par le recours à l'apprentissage automatique pour la statistique officielle grandit rapidement. Pour traiter certaines sources de données secondaires (y compris les sources administratives, les mégadonnées et l'Internet des objets), il semble essentiel d'étudier les possibilités qu'offrent les techniques modernes d'apprentissage automatique, sachant que celles-ci sont également susceptibles de présenter un intérêt en ce qui concerne les données primaires, comme le montre la note d'information. Même si l'apprentissage automatique semble prometteur, les statisticiens de la CEE n'ont qu'une expérience limitée de ses applications concrètes, et certains problèmes relatifs à la qualité et à la transparence des résultats obtenus par ces méthodes doivent encore être réglés. ».

IV. Informations concernant le projet relatif à l'apprentissage automatique

12. La CEE a recruté un responsable de projet fin février 2019 et a lancé le projet, qui réunissait alors 11 participants de six organisations, en mars. Des réunions en ligne ont permis de délimiter et de planifier le projet. La dernière main a été mise au projet au cours d'une première réunion en présentiel organisée par l'Office national de statistique du Royaume-Uni en mai 2019. À cette époque, le projet comptait déjà 27 participants issus de 14 organisations. Les participants sont convenus des principaux objectifs décrits ci-après.

13. Le projet, qui reposait sur un intérêt commun et s'inscrivait dans le prolongement des progrès accomplis à l'échelle nationale, avait pour objectif de faire avancer la recherche, la mise au point et l'application des techniques d'apprentissage automatique en vue de donner de la valeur ajoutée à la production de statistiques officielles. Pour atteindre cet objectif, l'équipe du projet relatif à l'apprentissage automatique devait chercher :

- À étudier et à démontrer la valeur ajoutée de l'apprentissage automatique pour la production de statistiques officielles, correspondant à l'augmentation de la pertinence, à l'amélioration de la qualité globale ou à la réduction des coûts ;
- À accroître la contribution de l'apprentissage automatique à la production de statistiques officielles ;
- À renforcer la capacité des organismes nationaux de statistique d'utiliser l'apprentissage automatique pour produire des statistiques officielles ;

- À resserrer la collaboration entre les organismes de statistique aux fins de la mise au point et de l'application de l'apprentissage automatique.

14. Compte tenu des travaux préliminaires menés par le BSTN, des attentes du Groupe de haut niveau et des intérêts exprimés par les membres de l'équipe, le projet a été scindé en trois modules de travail :

- Module 1 – Mener des études pilotes sur les thèmes suivants :
 - Codage et classification ;
 - Préparation et imputation ;
 - Utilisation de données d'imagerie ;
- Module 2 – Élaborer un cadre de qualité dans lequel s'inscrira le recours à l'apprentissage automatique ;
- Module 3 – Repérer les obstacles à l'intégration et y remédier.

15. En plus de démontrer la valeur ajoutée de l'apprentissage automatique, les études pilotes avaient pour objet de permettre aux participants de se former, d'échanger et de collaborer. Au début du projet, le fait qu'il n'existait aucun cadre de qualité applicable à l'apprentissage automatique était considéré comme le principal obstacle à l'acceptation, par les ONS, de solutions éprouvées en la matière. Des obstacles supplémentaires, liés à la mise en pratique et à l'extension future de l'apprentissage automatique, ont également été mis au jour lors de la phase initiale, et d'autres encore ont été répertoriés au fil des études pilotes.

16. Pour résumer en une phrase la relation entre les objectifs du projet et les modules de travail, on peut dire que les activités ont amorcé « l'intégration, dans les processus de production (module 3), des solutions éprouvées d'apprentissage automatique (module 1) de façon rationnelle et efficace (module 2) ». Des participants d'Allemagne, du Canada, du Mexique et du Royaume-Uni se sont proposés pour prendre la tête de 5 sous-groupes (3 pour le module 1, 1 pour le module 2 et 1 pour le module 3). Les sous-groupes ont tenu des réunions en ligne afin de se concerter et de faire part des avancées et résultats obtenus. Des réunions mensuelles ont été organisées pour faire le point sur le projet.

17. Au fil du temps, le projet a attiré de nouveaux participants. D'autres personnes, qui travaillaient dans le cadre des études pilotes (collaborateurs) ou s'intéressaient à l'avancement du projet (abonnés), s'y sont également associées. Au moment où le présent rapport a été établi, le projet comptait 124 membres de 23 pays, de 33 organismes nationaux et de quatre organisations internationales. Les réunions mensuelles étaient donc régulièrement suivies par 40 à 60 personnes. Même si le projet était axé sur l'apprentissage automatique, le grand nombre de participants a été à son avantage, certains étant membres d'autres groupes internationaux travaillant dans des domaines qui recoupaient les questions ou intérêts relatifs à l'apprentissage automatique, par exemple le Groupe mondial de l'ONU sur les mégadonnées et le Groupe des capacités et de la communication, qui dépend du Groupe de haut niveau.

18. L'organisation de quatre réunions relatives au projet a permis aux participants d'échanger et de se concerter plus facilement :

- En mai 2019, l'Office national de statistique du Royaume-Uni a accueilli une réunion destinée à arrêter les objectifs, les modalités d'organisation et les premiers résultats attendus ;
- En septembre 2019, l'Office de statistique de la République de Serbie a organisé une réunion qui avait pour but de faire part des progrès accomplis, de renforcer la collaboration et de convenir des résultats attendus et des délais ;
- En avril 2020, une réunion, qui devait être organisée par l'Office polonais de statistique (Statistics Poland) mais s'est finalement tenue en ligne en raison de la pandémie, a permis de mettre en commun et d'examiner les résultats des études pilotes ;

- En octobre 2020 s'est tenue une autre réunion en ligne, qui visait à débattre des rapports sur les études pilotes, des rapports thématiques, des résultats du projet et des orientations à suivre. À cette occasion, des membres de l'équipe du projet ont été invités à présenter les autres fins auxquelles leurs organisations respectives utilisaient l'apprentissage automatique.

19. Les membres de l'équipe du projet ont publié tous les documents pertinents sur la plateforme wiki de la CEE. On y trouvera notamment les documents de travail, l'état de situation mensuel et la liste des actions à mener, ainsi que de nombreuses autres références communiquées par les membres. Les exposés sur le projet et ses résultats que des membres ont faits à l'occasion de certaines manifestations extérieures, telles que la série de conférences BigSurv2020 ou le sommet ministériel Geo Week 2019, ainsi que les rapports distribués à l'atelier annuel du Groupe de haut niveau sont également disponibles sur la plateforme wiki. Enfin, les membres ont utilisé différents moyens pour échanger le code utile au projet (essentiellement la plateforme GitHub) ainsi que certaines données leur permettant de se former et de mener des essais plus facilement et rapidement.

20. Bon nombre des documents et autres éléments susmentionnés ont été publiés le 13 novembre 2020 sur la plateforme wiki publique de la CEE. Un webinaire s'est ensuite tenu les 16 et 17 novembre 2020. Il portait sur les résultats du projet et s'est conclu par une discussion ouverte concernant les orientations à suivre. Il a réuni 203 personnes de 33 pays et de 60 organismes nationaux et organisations internationales.

21. Le projet a permis la création d'un vaste groupe de spécialistes de diverses disciplines et organisations, déterminés à faire progresser l'apprentissage automatique et à collaborer à cette fin. Le campus universitaire de la science des données de l'Office national de statistique du Royaume-Uni a proposé de diriger un groupe sur l'apprentissage automatique (ci-après, groupe sur l'apprentissage automatique 2021), avec l'appui du secrétariat de la CEE et du Conseil d'administration du Groupe de haut niveau.

V. Résultats du projet, enseignements à tirer et autres questions

22. Les trois modules de travail ont débouché sur 21 études pilotes¹, un cadre de qualité et un rapport qui répertorie et analyse les obstacles communs à l'intégration de l'apprentissage automatique dans les processus de production. Les nombreux enseignements tirés de ces activités ont été publiés sur la plateforme wiki de la CEE (lien : [HLG-MOS Machine Learning Project](#)). Les résultats de chaque étude pilote et les enseignements en découlant (lien : [Pilot study reports](#)) sont examinés et résumés dans les rapports généraux par module (liens : [Coding and Classification Report](#) ; [Edit and Imputation Report](#) ; [Imagery Analysis Report](#)). Un résumé analytique (lien : [WPI executive summary](#)) inspiré de ces rapports dégage les principaux résultats obtenus et enseignements tirés de l'application de l'apprentissage automatique aux processus statistiques. Les modules de travail sur la qualité et l'intégration ont permis d'établir les principales conditions qu'il convenait de réunir pour faire progresser et faire accepter l'apprentissage automatique (liens : [Quality Framework for Statistical Algorithms](#) ; [Integration report](#)).

23. Le présent rapport consigne les enseignements et autres réflexions découlant du projet global et répète le moins possible les informations déjà contenues dans les autres rapports. Le groupe est convenu :

- Que le projet avait atteint ses objectifs ;
- Que la plupart des participants avaient acquis des connaissances et une expérience précieuses en matière d'apprentissage automatique ;

¹ Parmi les 21 études figurent un document décrivant une approche générale pour l'application de l'apprentissage automatique à l'analyse des données d'imagerie et un document contenant des pistes et idées relatives à l'utilisation de l'apprentissage automatique aux fins de la préparation des données. Ces documents ne sont pas liés à une application de l'apprentissage automatique en particulier.

- Que le projet avait favorisé le recours accru à l'apprentissage automatique pour la production de statistiques officielles dans la plupart des organisations participantes ;
- Qu'il existait toutefois encore peu d'éléments permettant d'affirmer que des solutions d'apprentissage automatique avaient effectivement été intégrées aux processus de production ;
- Qu'il devenait encore plus utile de transformer une solution théoriquement bonne en une véritable solution de production, et qu'un nombre croissant de parties prenantes étaient intéressées.

24. Les connaissances et l'expérience acquises dans le cadre du projet permettent de résumer en deux mots les concepts nécessaires au développement de l'apprentissage automatique aux fins de la production de statistiques officielles : acceptation et facilitation. Il incombe essentiellement aux personnes qui conçoivent et appliquent les méthodes de faire accepter l'apprentissage automatique au sein d'un organisme de statistique. En outre, il incombe à l'organisme de faciliter la mise au point et la mise en pratique de l'apprentissage automatique. Mais surtout, l'acceptation comme la facilitation nécessitent le soutien de tous les employés.

25. Les prochaines sections détaillent les principales conditions favorables à l'acceptation et à la facilitation des solutions d'apprentissage automatique. Chaque condition est brièvement analysée et mise en parallèle avec les progrès obtenus dans le cadre du projet, ce qui donne lieu à des suggestions concernant les travaux que mènera le groupe sur l'apprentissage automatique 2021. Le tableau 1 de l'annexe répertorie par ordre chronologique ces conditions principales, de la conception de l'idée au maintien de la production au moyen d'éléments fondamentaux et d'un appui organisationnel. Il met en évidence les aspects couverts par le projet relatif à l'apprentissage automatique et certains domaines que le groupe sur l'apprentissage automatique 2021 pourra éventuellement étudier.

VI. Principales conditions favorables à l'acceptation des solutions d'apprentissage automatique

A. Répondre aux besoins des acteurs

26. Fondamentalement, les solutions d'apprentissage automatique doivent être acceptées par les personnes chargées de produire les données (en général, les analystes du domaine) et surtout par les utilisateurs de ces données. Comme toute autre approche ou technologie, l'apprentissage automatique constitue l'un des moyens de parvenir à une fin donnée. Il ne devrait pas être envisagé ou adopté uniquement pour ce qu'il est, mais pour ce qu'il peut faire pour mieux répondre aux besoins (amélioration de la pertinence, du degré de détail, de la rapidité, de l'exactitude, du rendement, etc.). Les études pilotes sont généralement axées sur l'amélioration de la rapidité et de la précision pour trois processus statistiques (lien : [Pilot study reports](#)). Il existe de nombreuses applications de l'apprentissage automatique visant à répondre à d'autres besoins dans le cadre d'autres processus (des exemples sont disponibles sous le lien [Other applications of Machine Learning](#)).

27. L'approche axée sur la pratique a été l'un des éléments ayant permis la réalisation du projet. Peu après le début du projet, on a réfléchi à l'idée d'élaborer un « livre de recettes » expliquant comment utiliser l'apprentissage automatique. L'équipe du projet s'est employée à concocter certains des éléments de base d'une bonne recette (cadres et bonnes pratiques) alors que de nombreux participants apprenaient la recette en la réalisant à partir d'un mélange d'ingrédients (études pilotes) pour répondre aux besoins nécessaires au développement de leurs organisations respectives. Il importe que les activités du groupe continuent de répondre aux besoins des organisations participantes (applications) et contribuent aux aspects plus fondamentaux du projet (voir la ligne « Les fondements » dans le tableau 1 de l'annexe).

B. Disposer d'un cadre de qualité

28. Les solutions d'apprentissage automatique doivent permettre l'obtention de résultats de bonne qualité, voire de qualité supérieure en vue de répondre aux besoins des acteurs. Pour ce faire, il convient de préciser ce qu'on entend par « qualité ». La qualité est définie dans de nombreux cadres de référence largement reconnus, établis par des organismes nationaux et organisations internationales de statistique. Le cadre de qualité pour les algorithmes statistiques (QF4SA) vient compléter ces cadres et porte sur les conditions déterminantes pour l'acceptation des solutions d'apprentissage automatique (lien : [Quality Framework for Statistical Algorithms](#)). Il donne des indications sur le choix des algorithmes (y compris des algorithmes classiques) à utiliser dans le processus de production. L'emploi du terme « algorithme statistique » est délibéré et rend compte du fait que le document traite à la fois des méthodes classiques et des méthodes modernes généralement employées par les statisticiens des organismes publics de façon à renforcer la compréhension mutuelle entre les partisans de chaque type de méthode. Il n'existe aucun critère préétabli permettant d'affirmer que les solutions d'apprentissage automatique donnent des résultats suffisants ou meilleurs que d'autres solutions. Comme la plupart des cadres de référence sur la qualité, le cadre QF4SA définit cinq dimensions qui doivent être envisagées ensemble. Il est possible de mettre l'accent sur une dimension ou deux, mais aucune ne doit être négligée.

29. Le cadre QF4SA a été conçu pendant que les études pilotes étaient menées. Si ces études ont été très utiles à l'élaboration du cadre, elles n'ont pas, faute de temps, permis d'expérimenter véritablement certaines des pratiques qui y étaient recommandées. Certains participants se sont notamment félicités que le projet ait suivi une approche axée sur la pratique. À l'avenir, il est recommandé de mettre en pratique le cadre à l'occasion des études pilotes afin de recueillir des observations utiles à son amélioration et à son utilisation à plus grande échelle.

C. Démontrer la valeur ajoutée de l'apprentissage automatique

30. La plupart des études pilotes se concentraient sur cette condition. En ce qui concerne le codage et la classification, les études ont démontré que l'apprentissage automatique pouvait donner de meilleurs résultats que les opérations exclusivement manuelles. Elles se sont toutes heurtées à l'absence de référence statistiquement valable pour la comparaison des résultats obtenus au moyen de l'apprentissage automatique. En conséquence, l'objectif premier de nombreuses études a été de répliquer une opération existante, par exemple obtenir les mêmes classes de produits qu'une opération de classification manuelle, et de faire reposer la valeur ajoutée sur la rapidité et, indirectement, les coûts. Cet objectif présente trois inconvénients sérieux. Premièrement, la précision de l'opération existante (ou mise en concurrence) est souvent soit inconnue soit non attestée par une méthode d'évaluation valable. Deuxièmement, l'apprentissage automatique ne peut jamais répliquer parfaitement une opération. Les études menées dans le cadre du projet révèlent que les solutions d'apprentissage automatique peuvent répliquer, avec un degré de fiabilité élevé, entre 40 % et 85 % des résultats d'une opération (manuelle ou autrement automatisée) existante. Troisièmement, et il s'agit là de l'aspect le plus important, l'objectif de l'apprentissage automatique ne devrait, en tout état de cause, pas être uniquement de reproduire une opération, à moins qu'il soit possible de réaliser cette opération bien plus rapidement et à un coût nettement inférieur. Au lieu de cela, on devrait chercher à améliorer l'opération en combinant les avantages des différentes méthodes. Dans le cas d'une opération de classification, on pourrait ainsi s'appuyer sur les prévisions de l'apprentissage automatique pour attribuer automatiquement une classe (lorsqu'on sait que ces prévisions sont très précises, par exemple à plus de 98 %), utiliser les prévisions précises mais pas assez fiables pour aider les programmeurs et, si les prévisions ne sont pas assez précises (en général pour des classes plus rares), compter exclusivement sur les programmeurs pour l'attribution des classes restantes. Différentes versions de cette stratégie sont utilisées pour la production (liens : [Workplace injury and illness](#) ; [Industry and occupation](#) ; [Standard Industrial Classification](#)). La stratégie a également fait l'objet d'une expérience reposant sur le partage de codes et de données destinés à l'apprentissage automatique (liens : [A user's experiences with the ML code and data shared](#) ; [Shared code](#) ; [Product description dataset](#)).

31. En ce qui concerne la préparation et l'imputation, les études ont montré que l'apprentissage automatique pouvait n'apporter aucune valeur ajoutée (une méthode d'imputation simple donnait de meilleurs résultats que toutes les autres options) et, au mieux, être prometteur. Rien n'indique que les méthodes d'apprentissage automatique ne peuvent pas s'appliquer. Elles peuvent nécessiter moins de programmation et être plus rapides à appliquer que les méthodes actuelles. En revanche, il est difficile de créer et d'avoir en permanence de bonnes données de formation pour les algorithmes correspondants, et même si ces méthodes sont plus rapides ou plus précises, il peut être très compliqué d'expliquer ce qu'elles produisent et comment, et donc de convaincre les parties prenantes de les adopter. Il faut procéder à d'autres études et développer certains aspects fondamentaux (voir, par exemple, le lien [Hints and ideas on data cleaning](#)) pour donner des indications sur la façon d'utiliser l'apprentissage automatique dans ce domaine et déterminer les conditions propices à son application.

32. Dès le début du projet, on a estimé que l'apprentissage automatique était essentiel à l'exploitation efficace de grands volumes de données. Cette thèse a été confirmée par les études pilotes sur l'analyse des données d'imagerie (photographie satellite et aérienne). L'un des problèmes que posent ces données, auxquelles on a de plus en plus accès en grandes quantités, consiste à fournir à leurs utilisateurs des renseignements sur les processus complexes qu'il convient d'exécuter pour les exploiter correctement et efficacement, notamment lorsqu'on doit faire appel à l'apprentissage automatique. L'équipe du projet a conçu une approche générale de la production de statistiques officielles au moyen de données satellite et de l'apprentissage automatique (lien : [Generic pipeline](#)) afin de communiquer une partie de ces renseignements. Cette approche a été appliquée à la description de deux études sur l'imagerie satellite et aérienne (liens : [Aerial Image Address Use Classification](#) ; [Integrating EO with Official Statistics using Machine Learning](#)).

33. Les organisations sont invitées à poursuivre le développement de leurs solutions d'apprentissage automatique en vue de les mettre en pratique, et de continuer à collaborer et à échanger à cet égard. Cela pourrait s'appliquer à d'autres domaines intéressant les organisations (des exemples sont disponibles sous le lien [Other applications of Machine Learning](#)), en particulier pour les besoins qui sont associés à une forte intensité de main d'œuvre, qui sont stables dans le temps et pour lesquels on dispose de grands volumes de données permettant d'affiner les algorithmes. On pourrait ainsi envisager d'expérimenter certaines des pratiques décrites dans le cadre QF4SA et les documents relatifs à l'intégration et obtenir des informations utiles.

D. Assurer la pérennité et la fiabilité des résultats

34. L'objectif premier des études pilotes était d'évaluer la valeur ajoutée des différents algorithmes d'apprentissage automatique et de déterminer quel était, à partir des données disponibles, le meilleur modèle (algorithme et paramètres). Comme il a été indiqué plus haut, de nombreux obstacles empêchent encore l'intégration de solutions d'apprentissage automatique éprouvées dans les processus de production. De plus, au fil du temps, ces solutions doivent non seulement continuer à assurer leur fonction, mais aussi s'améliorer encore à mesure qu'elles « apprennent » davantage et s'adaptent à l'évolution des données saisies. Tandis qu'ils avançaient dans leurs recherches relatives aux solutions d'apprentissage automatique dans le cadre du projet, les participants ont demandé quand ils devaient mettre à jour ou rafraîchir les algorithmes d'apprentissage automatique et les paramètres correspondants, à quelle fréquence et suivant quelle méthode. D'autres applications de l'apprentissage automatique soulèvent des questions semblables.

35. Seule une application faisant partie du projet est intégrée au processus de production depuis suffisamment longtemps pour permettre de disposer d'une vaste expérience en ce qui concerne la mise à jour des algorithmes d'apprentissage automatique (lien : [Workplace injury and illness](#)). La mise en place et le maintien d'une solution d'apprentissage automatique efficace reposent sur les données utilisées pour la formation, non seulement au départ, lorsque l'on définit l'algorithme initial et ses paramètres, mais aussi longtemps que cette solution est appliquée. Les données utilisées pour l'évaluation, parfois désignées sous le terme « étalon-or », sont également importantes. On en a besoin pour évaluer non seulement

l'efficacité de l'algorithme d'apprentissage automatique, mais également celle de l'opération toute entière, celle-ci comprenant généralement des tâches de bureau. Ces données doivent être indépendantes des données destinées à la formation. Elles sont essentielles mais représentent souvent un coût important et sont soumises à certains critères ; il faut par exemple que des données de terrain soient recueillies et que les textes soient classifiés par des spécialistes du domaine en question. L'équipe du projet mesure désormais l'importance de disposer de bonnes données de formation et d'évaluation, et a pu caractériser ces données.

36. À l'avenir, il est recommandé de documenter et de mettre en commun ces connaissances.

E. Tenir compte des questions éthiques et juridiques

37. « Ces dix dernières années, l'apprentissage automatique a gagné du terrain et donné lieu à l'essor de nouvelles applications. Certaines de ces applications relèvent du domaine social, de sorte que les modèles fondés sur des profils de données peuvent avoir des conséquences importantes pour la vie des personnes concernées. Pour éviter que ces modèles entraînent une discrimination indésirable, différentes méthodes ont été proposées en matière d'équité algorithmique. » Ce texte est un extrait du résumé d'un document de travail établi par le Centre pour la statistique fondée sur les mégadonnées de l'Office néerlandais de statistique (lien : [Fair algorithms in context](#)). L'importance des questions éthiques y est soulignée. Ces questions ont été soulevées à l'occasion de plusieurs discussions dans le cadre du projet mais n'ont pas été traitées à titre individuel.

38. À l'avenir, ces questions pourraient être traitées soit par le groupe sur l'apprentissage automatique 2021 soit en collaboration avec d'autres groupes de travail qui les envisagent dans un contexte plus général. Dans le cadre des travaux qui seront menés, il importera d'opérer une distinction entre les questions relatives aux sources de données et les méthodes d'exploitation. Il importera également de se concentrer sur les questions propres aux statistiques officielles, qui sont principalement des agrégats, et non sur les conséquences directes pour les personnes concernées, par exemple l'obtention d'un prêt ou d'un diagnostic médical, qui reviennent souvent dans les débats.

F. Concevoir des solutions reposant sur des fondements scientifiques solides issus de nombreuses disciplines

39. Les informations émanant des organismes nationaux et organisations internationales de statistique ont toujours été pertinentes et fiables car elles sont produites moyennant des méthodes et processus solides. En s'inspirant des mêmes principes de base pour concevoir et appliquer les méthodes d'apprentissage automatique, on peut faciliter considérablement l'acceptation de ces méthodes ainsi que la réalisation des conditions susmentionnées. Les fondements scientifiques de ces processus comprennent des connaissances et des compétences relevant de nombreuses disciplines, telles que le domaine concerné, les statistiques, l'informatique, la méthodologie, la science des données et les activités opérationnelles. Par rapport aux méthodes classiques, les méthodes d'apprentissage automatique exigent une collaboration transversale encore plus étroite, des premières étapes (développement d'une idée et mise en correspondance avec un besoin) aux dernières (mise en pratique). C'est surtout vrai en ce qui concerne les connaissances du domaine concerné, car l'apprentissage automatique n'est pas qu'une solution comme une autre qui doit fonctionner pour répondre aux besoins de l'activité exercée dans ce domaine, c'est aussi une solution qui a cruellement besoin des connaissances du domaine pour fonctionner. Si l'idée d'utiliser l'apprentissage automatique peut venir d'une seule personne (comme dans le cas de certaines études pilotes), on doit rapidement faire intervenir des acteurs d'autres disciplines, notamment du domaine concerné, pour la développer correctement et efficacement. De même que pour la qualité, on peut se concentrer sur une discipline ou deux, mais aucune ne doit être négligée. Cela ressort clairement d'une expérience au cours de laquelle l'équipe du projet a partagé du code et des données destinés à l'apprentissage automatique. Cette expérience a été réalisée par une personne qui a une connaissance modeste de l'apprentissage automatique et a, en cours de route, tiré de nombreux enseignements et commis beaucoup d'erreurs (lien : [A user's experiences with the ML code and data shared](#)).

40. Le projet relatif à l'apprentissage automatique a bénéficié de la participation de spécialistes de nombreuses disciplines. Ceux-ci ont pu apprendre, débattre et examiner les aspects à prendre en considération dans le cadre de la conception, de l'évaluation et du développement des solutions d'apprentissage automatique. À l'avenir, le groupe sur l'apprentissage automatique devrait chercher à attirer plus de participants spécialisés dans la science des données, les domaines concernés et l'informatique, en fonction des plans établis et objectifs poursuivis. Les organismes de statistique continueront de se heurter à des difficultés pour acquérir, développer et gérer les différentes compétences spécialisées qu'il leur faut pour utiliser l'apprentissage automatique de façon efficace et rationnelle en vue de répondre à leurs besoins. L'acquisition et le développement de ces connaissances (par exemple, par la formation) ont été désignés comme le besoin le plus urgent dans le cadre d'un sondage réalisé pendant le webinaire (liens : [ML Webinar](#) ; [Poll results](#)). Cette condition est examinée plus en détail ci-après dans le cadre des moyens de faciliter la mise en œuvre des solutions d'apprentissage automatique.

VII. Principales conditions favorables à la mise en œuvre des solutions d'apprentissage automatique

A. Se doter de compétences dans de multiples disciplines

41. Depuis toujours, la production des statistiques officielles associe des connaissances et compétences issues de nombreuses disciplines. C'est d'autant plus vrai que les sources de données (grands, moyens ou petits volumes), les utilisateurs désireux d'exploiter ces données et les technologies qui permettent de le faire sont de plus en plus nombreux. Même si la science des données, qui est une discipline assez récente, englobe la plupart de ces compétences, la variété et la complexité des compétences requises dans chaque discipline sont telles qu'il n'est pas possible de les trouver chez une seule personne, ni même quelques personnes. Réunir les compétences requises est l'un des principaux problèmes auxquels sont confrontés les organismes de statistique. Quatre étapes peuvent être distinguées : la définition, l'acquisition, le développement et la gestion des compétences. Certaines étapes ont été traitées dans le cadre du projet (lien : [Integration](#)). On a ainsi pu mettre le doigt sur de nombreuses mesures concrètes que les ONS pouvaient adopter en vue de promouvoir l'apprentissage automatique et d'y recourir davantage. Beaucoup de ces mesures sont très récentes (lien : [Initiatives to accelerate the integration of machine learning solutions](#)).

42. Les initiatives visant à accélérer l'intégration des solutions d'apprentissage automatique consistent notamment à créer des organismes indépendants consacrés à la science des données, des laboratoires et des forums internes ou externes d'échange sur la science des données et l'apprentissage automatique. Les responsables de telles entités collaborent vraisemblablement à titre informel. À l'avenir, il leur est recommandé de créer un réseau formel leur permettant de faire part des difficultés rencontrées, des pratiques appliquées, des expériences menées et des résultats obtenus. Un tel réseau devrait se concentrer sur les aspects liés à la gestion, par exemple sur les stratégies d'entreprise, l'adéquation avec les besoins, le changement culturel ou la communication. Il pourrait aussi collaborer étroitement avec le groupe sur l'apprentissage automatique 2021, par exemple pour suggérer des projets et établir les priorités. Le réseau ou le groupe devrait également se rapprocher d'autres groupes qui se penchent sur des aspects analogues, par exemple du Groupe des capacités et de la communication, qui dépend du Groupe de haut niveau et s'occupe notamment de la gestion du changement, des cadres organisationnels pour la collaboration et du renforcement des compétences.

B. Choisir l'infrastructure informatique qui convient

43. Dès le départ, l'équipe du projet relatif à l'apprentissage automatique a décidé de mettre l'accent sur la démonstration de la valeur ajoutée, la qualité et l'intégration. La question de l'infrastructure informatique a été abordée à plusieurs occasions mais n'a pas été approfondie. À l'avenir, le groupe sur l'apprentissage automatique 2021 devrait y réfléchir

mais devrait d'abord effectuer des recherches pour déterminer s'il existe d'autres groupes de travail étudiant la question et s'en rapprocher, ou si des travaux ont déjà été menés en la matière.

C. Utiliser la recherche-développement

44. Ci-avant, on a vu que la première condition à remplir pour faire accepter des solutions d'apprentissage automatique était de les adapter aux besoins des acteurs. Les débats tenus au sein du sous-groupe sur l'intégration n'ont pas permis de dégager un consensus parfait en la matière. Certains membres ont souligné qu'il importait de partir d'un besoin pour ensuite passer à la recherche-développement, concevoir un prototype et faire intervenir les autres domaines, tels que l'informatique. D'autres ont insisté sur l'intérêt d'acquérir d'abord une expérience en matière d'apprentissage automatique, grâce à la recherche-développement, pour ensuite pouvoir définir les problèmes susceptibles d'être résolus par l'apprentissage automatique. Quelle qu'elle soit, l'option retenue pour développer l'utilisation de l'apprentissage automatique devrait être motivée, sinon par un besoin propre à l'activité (par exemple, émanant d'un seul programme statistique), au moins par une stratégie claire, adoptée à l'échelle de l'entreprise, qui lui permettra de gagner en pertinence grâce à des informations de meilleure qualité, disponibles plus rapidement et, éventuellement, à un coût moindre.

D. Échanger et collaborer

45. Dans le cadre du projet, les membres ont fait circuler des documents de travail, des références méthodologiques et techniques ainsi que des liens vers des supports de formation et ont fait des exposés à l'occasion de réunions et de sessions « Sprint ». L'équipe a souvent pu compter sur la structure d'appui du Groupe de haut niveau pour consulter des spécialistes ou trouver un lieu de réunion. La mise à disposition du code et des données destinés à l'apprentissage automatique a considérablement facilité et accéléré la formation et la réalisation d'expériences. Bon nombre des documents et autres éléments susmentionnés ont été compressés et publiés sur la plateforme wiki de la CEE le 13 novembre, afin de permettre à tous les acteurs de la statistique officielle de tirer parti des connaissances et des supports réunis par l'équipe du projet pendant deux ans (liens : [Pilot study reports](#) ; [Shared code](#) ; [Learning, references and shared data](#)).

46. Dorénavant, l'échange et la collaboration contribueront non seulement à faire avancer l'apprentissage automatique mais également à décider rapidement de ne pas l'appliquer à des domaines ou à des contextes pour lesquels il n'a pas été établi que cela présentait un intérêt. Même si, au vu des circonstances, les sessions « Sprint » et les autres réunions en ligne ont été couronnées de succès, il importera de continuer à plaider en faveur de la tenue occasionnelle de réunions en présentiel sous la forme de séminaires ou d'ateliers d'experts, propices à une collaboration plus étroite qui peut ensuite être renforcée par des échanges à distance. La structure d'appui du Groupe de haut niveau devrait continuer à promouvoir le développement de l'apprentissage automatique en permettant la circulation d'informations au sein d'un vaste réseau de parties intéressées.

E. Obtenir l'appui des hauts responsables

47. Le projet relatif à l'apprentissage automatique n'aurait pas été possible et n'aurait pas été couronné de succès sans la participation de nombreuses personnes de multiples organisations ni l'appui que les statisticiens en chef ont fourni par l'intermédiaire du Groupe de haut niveau. Pour les prochaines étapes, il sera essentiel de pouvoir continuer à compter sur le soutien de ces acteurs afin de poursuivre la recherche-développement dans leurs organisations respectives et avec d'autres partenaires dans le cadre d'initiatives de collaboration. En contrepartie, les auteurs des initiatives seront tenus de rendre des comptes concernant l'exécution des travaux répondant aux priorités des organismes de statistique. À cet égard, il est recommandé au groupe sur l'apprentissage automatique 2021 de continuer de faire rapport au Groupe de haut niveau (via le Conseil d'administration).

F. Mobiliser tous les employés

48. Les nouvelles technologies, telles que l'apprentissage automatique ou l'intelligence artificielle, ont des effets importants sur la culture d'une organisation. L'apprentissage automatique influe sur la marge d'action de l'organisation et de chaque employé à titre individuel ainsi que sur les méthodes de travail. Toutes les études menées dans le cadre du projet faisaient intervenir des méthodes d'apprentissage supervisé qui nécessitaient des contributions essentielles de la part des spécialistes et des employés de bureau, susceptibles d'être les plus touchés par le changement. Selon ces études, les méthodes d'apprentissage automatique ne peuvent complètement se substituer aux tâches effectuées par le personnel et ne devraient donc pas être vues comme telles mais plutôt comme un moyen d'introduire des processus automatisés ou de renforcer ces processus lorsqu'ils existent, et d'obtenir ainsi de meilleurs résultats, à un coût égal ou moindre, tout en permettant au personnel d'avoir du temps à consacrer à des tâches à plus forte valeur ajoutée. Comme il est établi que les solutions d'apprentissage automatique apportent une valeur ajoutée à un plus grand nombre d'opérations, les employés, quelles que soient leur fonction et leur position hiérarchique, doivent être encouragés à les envisager pour répondre à leurs besoins. Ils devraient également pouvoir consulter des spécialistes ou avoir accès à un centre d'expertise sur l'apprentissage automatique afin de savoir si cette solution devrait être examinée plus avant.

VIII. Conclusion : l'apprentissage automatique va-t-il s'essouffler, s'imposer ou échouer ?

49. Dans la note d'information qu'il a établie en décembre 2018, où il proposait de créer le projet relatif à l'apprentissage automatique, le BSTN a écrit : « Bien que l'apprentissage automatique semble prometteur, les statisticiens de la CSE ne disposent que d'une expérience limitée de ses applications concrètes, et certains problèmes notamment liés à la qualité et à la transparence des résultats obtenus par cette méthode doivent encore être résolus. ». À l'époque, on aurait pu résumer cette problématique à la question suivante, dans le cadre de la production de statistiques officielles : « L'apprentissage automatique va-t-il s'essouffler, s'imposer ou échouer ? »

50. Deux ans plus tard, les travaux menés dans le cadre du projet nous donnent la réponse : l'apprentissage automatique n'est pas un simple phénomène de mode ; il est indispensable lorsqu'il peut donner de la valeur ajoutée, sans quoi on ne devrait pas y avoir recours (pour éviter qu'il échoue) ; toutefois, il n'est pas encore assez accepté et facilité par certains.

51. Le projet ne cesse de s'agrandir et compte désormais 124 membres de 23 pays, de 33 organismes nationaux et de quatre organisations internationales, ce qui atteste de l'intérêt porté à l'apprentissage automatique et des progrès accomplis dans son utilisation. La forte mobilisation de nombreux participants et l'aide considérable reçue ont permis à l'équipe du projet d'établir et de faire circuler beaucoup de rapports, d'autres documents, de codes, de données, de supports de formation et de références utiles aux acteurs de la statistique officielle. Au-delà du projet, les débats, les sessions et les manifestations entières consacrés à l'apprentissage automatique et à la science des données se multiplient rapidement (voir la liste des sessions et débats organisés à l'occasion de la série de conférences BigSurv20, sur le thème « Mégadonnées et sondages », sous le lien [Other applications of Machine Learning](#) ; le prochain colloque organisé par Statistique Canada portera sur la science des données²).

52. L'apprentissage automatique est incontournable dans les domaines pour lesquels on a établi qu'il contribuait à produire des données plus pertinentes, de meilleure qualité, plus rapidement ou de façon plus rentable, sans qu'aucune de ces dimensions ne soit sensiblement réduite. On a plus de chances d'obtenir de meilleurs résultats en l'utilisant pour des opérations à forte intensité de main-d'œuvre, répétitives et stables, telles que le codage et la classification. D'aucuns pourraient dire que cela vaut pour toute automatisation introduite dans de telles opérations. En réalité, dans le cas de l'apprentissage automatique, l'automatisation pourrait être réalisée plus rapidement. L'apprentissage automatique est

² Un lien vers l'annonce du colloque sera communiqué ultérieurement.

souvent essentiel au maniement de grands volumes de données. Il est plus compliqué de l'appliquer à des processus plus subjectifs, tels que la préparation et l'imputation des données.

53. Toute technologie nouvelle et évolutive se heurte à la résistance, plus ou moins grande, de différentes parties. Certaines vont lui opposer des arguments scientifiques qui ne feront que la rendre plus robuste. D'autres rejettent tout bonnement les changements en général. Les unes seront convaincues dès lors que le développement des solutions d'apprentissage automatique reposera sur des fondements scientifiques solides issus des différentes disciplines concernées, un cadre de qualité et des considérations éthiques. En ce qui concerne les autres, une gouvernance claire et forte pourra lever les obstacles. L'échange et la collaboration entre les organismes de statistique et au sein de ces organismes sont également essentiels en vue d'accroître le recours à l'apprentissage automatique, sur la base des enseignements tirés concernant les domaines où celui-ci donne de la valeur ajoutée, se montre prometteur ou ne devrait pas être envisagé.

54. En l'espace d'un an et demi, le projet relatif à l'apprentissage automatique a suscité un formidable élan favorisant un emploi responsable de cette méthode dans la production de statistiques officielles. Il a permis aux organismes participants d'avoir davantage recours à l'apprentissage automatique, et de nombreux enseignements en ont été tirés et ont été communiqués aux acteurs de la statistique officielle. Fait tout aussi important, au terme du projet, un groupe énergique d'experts de différents domaines s'est dit déterminé à poursuivre les travaux, une organisation s'est proposée pour gérer ce groupe et d'autres personnes ont indiqué qu'elles souhaitaient faire partie du groupe ou le soutenir dans ses efforts visant à mettre en pratique les concepts envisagés pour l'apprentissage automatique.

Tableau

Modules de travail du projet relatif à l'apprentissage automatique en 2020 ([liens, anglais seulement](#)) et thèmes pouvant être étudiés par le groupe sur l'apprentissage automatique 2021 (**en rouge**)

	Passer d'un concept à une solution valable (démonstration)		Mettre la solution en production (mise en pratique)		Veiller à la stabilité de la production (maintien)		
Les étapes	Toutes les études pilotes du module 1		Certaines études pilotes du module 1		Un tout petit nombre d'études pilotes du module 1		
	Les autres applications de l'apprentissage automatique		Certaines autres applications de l'apprentissage automatique		Un tout petit nombre d'autres applications de l'apprentissage automatique		
	Module 3 : Intégration (trimestres 5 et 6)		Module 3 : Intégration (trimestre 5)				
<i>Volet 1 : Appuyer les études en cours axées sur la production ; intégrer de nouvelles études portant sur d'autres processus (par exemple, couplage de fichiers) et/ou sources de données (par exemple, données satellite)</i>							
	Qualité (exactitude, actualité, efficacité, explicabilité et reproductibilité)	Données utiles à la formation	Connaissances/ Compétences	Infrastructure informatique	Interopérabilité/ Processus interne	Aspects éthiques et juridiques	Sécurité
Les fondements	Module 2 : Qualité		Module 3 : Intégration (trimestres 3 et 4)				
	<i>Volet 2 : Mettre à l'essai des pratiques et méthodes relatives à certaines dimensions du cadre QF4SA (Module 2) ;</i>	<i>Volet 4 : rechercher des moyens d'obtenir des données utiles à la formation et de les tenir à jour, déterminer quand il faut se former à nouveau à un modèle, définir ce qu'on entend par « utile » et comment mesurer l'utilité</i>	<i>Volet 5 : Quelles compétences, comment les acquérir, où les trouver ?</i>	<i>À définir</i>	<i>À définir</i>	<i>Volet 6 : Manuel d'éthique, règlements, etc.</i>	<i>À définir</i>
	<i>Volet 3 : Revoir et améliorer le cadre</i>						

	Organisation	Échange et collaboration
Les outils	<u>Module 3 : Intégration (trimestres 1 et 2)</u>	<u>Projet relatif à l'apprentissage automatique du Groupe de haut niveau</u>
	<u>Initiatives visant à accélérer l'intégration des solutions d'apprentissage automatique</u>	<u>Études et codes</u>
	<i>Volet 7 : Créer et maintenir un réseau de chefs d'unités chargées de la science des données</i>	<u>Outils de formation</u>
	<i>Volet 8 : Après 2021, comment mieux se préparer pour les deux à cinq prochaines années, à quelles technologies et sources de données s'attendre, quelles sont les compétences nécessaires ?</i>	<u>Webinaire tenu dans le cadre du projet</u>