

# Appel à manifestation d'intérêt : Ensemble de données d'apprentissage machine pour de meilleurs résultats en matière de santé

## Lacuna Fund : Our Voice on Data

15 June 2021

### Table des matières

1 – INTRODUCTION .....	2
OBJET ET OBJECTIFS DU FONDS .....	2
PRINCIPES DU FONDS .....	2
PHILOSOPHIE EN MATIÈRE D'OCTROI DE SUBVENTIONS.....	3
2- VUE D'ENSEMBLE.....	3
CRITÈRES D'ÉLIGIBILITÉ APPLICABLES AUX ORGANISATIONS .....	3
PROCESSUS DE SÉLECTION ET CRITÈRES D'ÉVALUATION.....	4
TIMELINE.....	5
3 – OBJET ET BESOINS .....	6
OBJET.....	6
BESOINS.....	7
4 – INFORMATIONS DE LA MANIFESTATION D'INTÉRÊT .....	8



This document is licensed under a [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) (CC-BY 4.0) license.

# 1 – Introduction

## Objet et objectifs du Fonds

Le Lacuna Fund soutient la création, le développement et la tenue à jour d'ensembles de données d'évaluation et de formations équitables permettant une application fiable des outils d'apprentissage machine présentant une valeur sociale élevée.

Le Fonds poursuit plusieurs objectifs :

- verser des fonds aux institutions en vue de créer, de développer et/ou de tenir à jour des ensembles de données qui comblent les lacunes et réduisent les biais dans les données étiquetées utilisées pour l'apprentissage machine ;
- permettre aux populations mal desservies de profiter des avancées offertes par l'IA ;
- favoriser une meilleure compréhension par la communauté de l'apprentissage machine et les organisations philanthropiques de la manière de financer le plus efficacement et économiquement possible le développement et la tenue à jour d'ensembles de données étiquetées dans le respect des principes d'équité.

## Principes du Fonds

Les principes suivants guideront la gouvernance et les opérations du Lacuna Fund.

- **Accessibilité** – le Lacuna Fund s'engage à garantir que les ensembles de données créés grâce à son financement sont accessibles aux communautés mal desservies et leur bénéficient au service des objectifs énoncés précédemment. Les ensembles de données et la propriété intellectuelle dont ils font l'objet seront couverts par des licences de données ouvertes appropriées pour favoriser une utilisation responsable en aval. (Voir la Politique en matière de propriété intellectuelle pour plus de détails.)
- **Équité** – le Lacuna Fund vise à rendre l'IA plus équitable en favorisant des ensembles de données qui sont créés par des personnes aux profils sous-représentés dans le monde et qui répondent à leurs besoins. Ces ensembles de données ne doivent pas créer ou renforcer des biais ou préjugés (par exemple, ils doivent être inclusifs en termes de genre et représentatifs des personnes de couleur à l'échelle mondiale) ni soutenir des systèmes ou des technologies susceptibles de nuire.
- **Éthique** – le Lacuna Fund financera la collecte de données d'une manière conforme aux normes éthiques du travail et exigera des bénéficiaires de subventions qu'ils précisent les mesures qu'ils prendront pour protéger la vie privée et éviter tout préjudice dans la collecte, l'octroi de licences et l'utilisation des ensembles de données créés avec les fonds des subventions.
- **Approche participative** – le Lacuna Fund s'efforce de répondre aux besoins des parties prenantes concernées en encourageant le leadership ou un engagement fort des experts locaux, des bénéficiaires et des utilisateurs finaux dans la gouvernance du Fonds et dans les

projets soutenus. Le Fonds examinera la participation au regard de ses principes d'équité et d'éthique.

- **Qualité** – les données produites grâce aux efforts financés par Lacuna doivent être d'une qualité élevée permettant des applications bénéfiques pour la recherche, les communautés et l'industrie.
- **Impact transformationnel** – en finançant des ensembles de données qui comblent des lacunes fondamentales de l'IA, le Lacuna Fund entend libérer les avancées dont celle-ci est porteuse au bénéfice des communautés pauvres et mal desservies.

## Philosophie en matière d'octroi de subventions

Le Lacuna Fund privilégie une approche collaborative et pilotée localement pour la création, le développement et la tenue à jour d'ensembles de données. Nous considérons que l'utilité et l'actualisation pérennes des données ouvertes imposent de s'appuyer sur une communauté investie et concernée par ces données.

Le Lacuna Fund espère financer des ensembles de données qui contribuent aux multiples applications à forte valeur sociale, que ce soit au travers de la recherche, de l'innovation commerciale ou de l'amélioration des services du secteur public. **Bien que la section 3 « Objet et besoins » expose les besoins définis par le Groupe consultatif technique (GCT), le Lacuna Fund accueille toutes les idées novatrices dans ce domaine mettant en évidence un avantage clairement articulé aligné sur les critères de sélection présentés ci-après.**

Le Lacuna Fund est soutenu par la Fondation Rockefeller, Google.org et le Centre de recherche pour le développement international du Canada. Par ailleurs, cet appel relatif à des ensembles de données étiquetées dans le respect des principes d'équité pour de meilleurs résultats en matière de santé est soutenu par les fondations Wellcome Trust et The Gordon and Betty Moore Foundation.

## 2- Vue d'ensemble

### Critères d'éligibilité applicables aux organisations

**Le Lacuna Fund entend rendre son financement accessible à un maximum d'organisations dans l'univers de l'IA pour le bien social et cultiver les capacités et les organisations émergentes dans ce domaine.**

Pour pouvoir bénéficier d'un financement, les organisations doivent remplir les critères suivants :

- être une entité à but non lucratif, un institut de recherche, une entreprise sociale à but lucratif ou une équipe composée de ce type d'organisations. Pour présenter leur projet, les particuliers doivent recourir à un promoteur institutionnel. Les partenariats sont fortement encouragés, mais seul le candidat principal recevra des fonds ;
- avoir une mission de soutien du bien sociétal, au sens large ;

- être une organisation dont le siège se situe dans le pays ou la région dans lequel les données seront collectées ou avoir un partenariat important avec celui-ci/celle-ci ;
- disposer de toutes les autorisations requises, nationales ou autres, pour poursuivre les recherches proposées, ainsi que des accords d'utilisation des données ou des projets pour les obtenir. Le cas échéant, le processus d'autorisation pourra être mené en parallèle de la demande de subvention. Les éventuels frais d'autorisation sont à la charge du candidat ;
- disposer de la capacité technique – ou de la faculté de développer cette capacité grâce au partenariat décrit dans la manifestation d'intérêt – pour procéder à l'étiquetage, à la création, à la consolidation, au développement et/ou à la tenue à jour des ensembles de données, y compris la capacité à mettre en œuvre les bonnes pratiques et les normes établies dans le domaine visé (par ex. les résultats en matière de soins de santé) afin de permettre à plusieurs entités de réaliser des travaux analytiques de haute qualité pour l'IA/l'apprentissage machine.

## Processus de sélection et critères d'évaluation

Le Lacuna Fund est à la recherche d'organisations désireuses de répondre à un appel à propositions visant à débloquer, créer, consolider et/ou améliorer des ensembles de données étiquetées susceptibles de favoriser des résultats plus équitables en matière de soins de santé. Le Lacuna Fund et ses partenaires procèdent à une sélection initiale des manifestations d'intérêt pour vérifier l'éligibilité organisationnelle et la faisabilité du concept original présenté dans la manifestation d'intérêt. À l'issue de cette sélection initiale, un Groupe consultatif technique composé d'experts, d'utilisateurs de données et de parties prenantes pour le domaine visé évalue les manifestations d'intérêt sur la base des critères de sélection énoncés ci-après. Le Lacuna Fund invite alors quelques organisations à fournir une proposition complète fondée sur les éléments présentés dans la manifestation d'intérêt. Les membres du Groupe consultatif technique ne peuvent pas soumettre une manifestation d'intérêt ou une proposition pour répondre à un appel à propositions pour lequel ils sont évaluateurs (voir la [Politique du Lacuna Fund en matière de conflit d'intérêts](#)).

**Le Groupe consultatif technique pour cet appel à propositions évaluera les manifestations d'intérêt pour sélectionner les quelques organisations qui seront invitées à présenter des propositions complètes en vue d'un financement. Les candidats invités seront sélectionnés sur la base du niveau de satisfaction des critères suivants :**

- **Qualité** – l'organisation ou l'équipe proposant le projet démontre une collaboration interdisciplinaire entre des experts qualifiés dans le domaine de la santé, la recherche, l'apprentissage machine et la gestion de données. Les partenariats entre des systèmes de soins de santé bien dotés en ressources et disposant d'une forte capacité de recherche et des systèmes desservant des populations marginalisées sont encouragés.
- **Impact transformationnel** – les ensembles de données doivent améliorer l'apprentissage machine et procurer un avantage social tangible aux populations mal desservies. Les exemples incluent, sans s'y limiter : a) l'étiquetage, le nettoyage ou la collecte de données de validation ou la mise en commun des ensembles de données existants pour dégager une valeur supplémentaire ou assurer une plus grande précision de l'ensemble de données existant ; b) la création d'un nouvel ensemble de données étiquetées de grande valeur à destination d'une population mal desservie ou concernant un problème lié aux résultats en

matière de soins de santé ; c) l'amélioration de la représentativité d'un ensemble de données existant et de la prise en compte de la race, du genre, de l'origine ethnique, des capacités, etc. ; ou d) l'association des ensembles de données cliniques existants aux données sur les déterminants sociaux de la santé afin de créer des ensembles de données d'information plus solides.

- **Équité** – la proposition est assortie d'un argumentaire convaincant démontrant que les ensembles de données seront mis en œuvre afin d'aider les communautés vulnérables et mal desservies.
- **Approche participative** – les ensembles de données doivent être centrés sur les besoins des communautés affectées et il convient de travailler avec les partenaires pour déterminer les avantages pour la communauté. Plus précisément, les projets doivent associer les patients, les cliniciens et les membres de la communauté aux décisions relatives à la gouvernance des données (par exemple, quelles données sont conservées et comment elles sont utilisées). Lorsque l'ensemble de données a une portée géographique, l'équipe est majoritairement implantée dans la région concernée et/ou entretient des liens étroits avec les acteurs locaux pour garantir une tenue à jour et un usage pérennes de l'ensemble de données par la communauté locale.
- **Éthique** – le projet peut passer avec succès un examen sur les questions d'éthique (par exemple par un comité d'examen institutionnel) portant sur les aspects suivants : a) les questions liées au respect de la vie privée ; b) le risque d'utilisation abusive de l'ensemble de données en aval ; c) les vecteurs de discrimination potentiels (par exemple la discrimination fondée sur le genre) ; et d) les conditions de travail justes et équitables si des étiqueteurs rémunérés participent au projet.
- **Efficacité** – le porteur de projet a examiné les ensembles de données existants et propose l'utilisation d'outils et techniques efficaces de collecte et d'étiquetage de données permettant d'accélérer la collecte, le nettoyage et le partage des données.
- **Faisabilité** – le projet est réalisable en fonction du budget et de l'étendue des travaux proposés. Bien que la manifestation d'intérêt ne doive pas inclure un budget détaillé, nous nous attendons à ce que les manifestations d'intérêt incluent une estimation générale du budget global prévu pour mener à bien le projet.
- **Accessibilité** – l'ensemble de données sera largement mis à disposition dans le cadre d'une licence ouverte ou, en cas d'impossibilité, un argumentaire convaincant expose les raisons d'un système d'octroi de licence plus strict dans le but de protéger la vie privée ou de prévenir tout préjudice, ainsi qu'un mécanisme pour fournir l'accès en vertu de la licence proposée.
- **Durabilité** – le projet comporte un plan visant à garantir la durabilité et la tenue à jour future de l'ensemble de données, par exemple par une communauté spécifique ou un groupe de parties intéressées (organisations à but lucratif ou non lucratif), et un modèle de gouvernance robuste pour l'ensemble de données ouvert.

## Timeline

Appel à manifestation d'intérêt diffusé publiquement sur le site Web de Lacuna Fund	15 June 2021
---	--------------

<b>Date limite pour les questions/réponses</b> Veuillez adresser vos questions à <a href="mailto:secretariat@lacunafund.org">secretariat@lacunafund.org</a>	<b>25 June 2021</b>
<b>Publication des réponses</b>	<b>2 July 2021</b>
<b>Date limite pour le dépôt des candidatures</b>	<b>15 July 2021</b>

**Période de questions/réponses** : toutes les questions concernant cet appel à manifestation d'intérêt doivent être adressées par courrier électronique à l'adresse [secretariat@lacunafund.org](mailto:secretariat@lacunafund.org) en précisant « Question Équité & Santé 2021 » dans l'objet. Les questions adressées d'ici le 25 juin seront anonymisées et leur réponse sera publiée le 2 juillet dans un document posté sur la page « [Soumettre un projet](#) » du site Web de Lacuna Fund.

## 3 – Objet et besoins

### Objet

Le but de cet appel à manifestation d'intérêt est de recenser les projets dont les porteurs seront invités à soumettre des propositions complètes pour le développement d'ensembles de données de formation et d'évaluation ouverts et accessibles pour les applications d'apprentissage machine, qui abordent les inégalités dans les résultats en matière de soins de santé. Une meilleure mise en œuvre des modèles d'apprentissage machine pourrait permettre de fournir de meilleures informations aux médecins au moment de la prise en charge du patient. En fournissant davantage de données et d'informations aux cliniciens, ceux-ci peuvent prendre de meilleures décisions concernant les diagnostics et les options de traitement des patients, tout en comprenant les résultats possibles et le coût de chacun d'eux. La valeur de l'apprentissage machine dans le domaine des soins de santé réside dans sa capacité à traiter d'énormes ensembles de données qui dépassent les capacités humaines, puis à convertir de manière fiable l'analyse de ces données en informations cliniques qui aident les médecins à planifier et à prodiguer des soins, ce qui permet d'obtenir de meilleurs résultats, de réduire les coûts des soins et d'accroître la satisfaction des patients<sup>1</sup>.

Les biais présents dans les ensembles de données qui informent les algorithmes sur lesquels ces systèmes sont basés sont une préoccupation majeure de la mise en œuvre de l'apprentissage machine. Bien souvent, les ensembles de données sur les soins de santé ne sont pas suffisamment représentatifs des groupes démographiques et socio-économiques. Lorsque cette représentation est prise en compte, les données peuvent également se révéler partiellement inexactes, ce qui entraîne une mauvaise compréhension des différences entre ces groupes. Enfin, les données actuelles peuvent omettre

---

<sup>1</sup> Corbett, Ed. « Real-World Benefits of Machine Learning in Healthcare », *Health Catalyst*, 19 novembre 2020, [www.healthcatalyst.com/clinical-applications-of-machine-learning-in-healthcare](http://www.healthcatalyst.com/clinical-applications-of-machine-learning-in-healthcare)

d'importantes données socio-économiques, environnementales et autres qui peuvent déterminer les résultats en matière de santé. L'amélioration de ces ensembles de données selon ces principes pourrait donc améliorer les modèles d'apprentissage machine et les résultats en matière de santé<sup>2</sup>.

## Besoins

Le Lacuna Fund lance un appel à manifestation d'intérêt de la part d'organisations qualifiées désireuses de développer des ensembles de données de formation et d'évaluation ouverts et accessibles pour des applications d'apprentissage machine qui abordent les inégalités dans les résultats des soins de santé. La portée de l'effort est la suivante :

**Portée géographique** : aux États-Unis et dans les pays à revenu intermédiaire – tranche inférieure (PRITI), avec 1 à 2 propositions et entre 30 et 70 % du montant à financer dans chacune de ces zones géographiques, selon le nombre d'ensembles de données financés.

**Portée matérielle** : les projets doivent être axés sur les ensembles de données pouvant être utilisés pour aborder les disparités des soins de santé dans l'un ou plusieurs des domaines suivants : cancer, maladies infectieuses ou maladies chroniques. Par ailleurs, les candidats peuvent également présenter des arguments convaincants concernant un autre domaine dans lequel un ensemble de données pourrait considérablement réduire les inégalités. En tout état de cause, les candidats sont encouragés à envisager si l'ensemble de données qu'ils proposent pourrait être utilisé pour résoudre d'autres problèmes, en plus du cas d'utilisation prévu.

**Portée en termes de diversité** : notre objectif est de soutenir la création, l'augmentation ou la consolidation d'ensembles de données qui sont représentatifs des populations touchées et qui sont donc moins biaisés et plus susceptibles d'aboutir à des résultats équitables en matière de santé. Au vu des inégalités historiques liées à la race aux États-Unis, nous sommes intéressés par les ensembles de données qui pourraient contribuer à réduire les disparités raciales dans les résultats des soins de santé aux États-Unis, et les ensembles de données qui peuvent atténuer les inégalités dans les résultats des soins de santé liés à l'identité dans les PRITI (par exemple l'origine ethnique, l'affiliation tribale, le sexe, etc.). En outre, nous savons qu'un manque de diversité en termes de genre, d'âge, de capacité, d'emplacement géographique et même de type d'établissement de soins (par exemple un système de santé communautaire par rapport à un grand système médical universitaire, des patients hospitalisés par rapport à des patients ambulatoires) peut rendre un ensemble de données moins représentatif et conduire à des résultats inéquitables pour des sous-ensembles de la population. Nous encourageons les candidats à envisager d'adopter une approche intersectionnelle en incluant des données pour plusieurs groupes mal desservis, ou à expliquer pourquoi un certain type de diversité est important pour garantir des résultats équitables pour un cas d'utilisation spécifique.

Parmi les autres considérations, citons les suivantes :

**Ensembles de données existants par rapport aux nouveaux** : nous pensons qu'il est très utile de débloquer, d'augmenter ou de consolider les ensembles de données existants et nous sommes

---

<sup>2</sup> Wawira Gichoya J, McCoy LG, Celi LA, *et al.*, « Equity in essence: a call for operationalising fairness in machine learning for healthcare », *BMJ Health & Care Informatics*, 2021 ;**28**:e100289. doi:10.1136/bmjhci-2020-100289

également ouverts aux propositions visant à créer de nouveaux ensembles de données tels que des ensembles de données de référence pour tester la généralisation des algorithmes et les différences dans la qualité de la prédiction pour les populations mal desservies. Plus important encore, nous voulons nous assurer que l'ensemble de données sous-jacent n'est pas fondamentalement vicié afin d'éviter de perpétuer les biais. Toutes les approches décrites ci-après sont intéressantes :

- mettre en commun des données existantes provenant des systèmes de soins de santé, des caisses d'assurance maladie ou des intermédiaires de données sur la santé afin de les rendre accessibles aux chercheurs pour qu'ils puissent déceler et corriger les biais algorithmiques (par exemple MIMIC, PhysioNet) ;
- combler les lacunes/rendre les ensembles de données existants plus représentatifs (par exemple des images du cancer de la peau, des radiographies des poumons pour détecter la COVID, etc.) ;
- associer les ensembles de données cliniques existants aux données sur les déterminants sociaux de la santé afin de créer des ensembles de données plus solides et plus informatifs ;
- nettoyer les ensembles de données existants pour garantir l'exactitude des données sur la race, l'origine ethnique, le genre, le handicap, etc.

**Type d'ensemble de données** : les candidats retenus proposeront la création, l'expansion ou la consolidation d'un ou de plusieurs ensembles de données pouvant être utilisés pour répondre à plusieurs questions de recherche et corriger un biais ou une disparité.

## 4 – Informations de la manifestation d'intérêt

La manifestation d'intérêt doit être soumise exclusivement via le portail spécifique disponible à l'adresse [www.lacunafund.org/fr/soumettre-un-projet/](http://www.lacunafund.org/fr/soumettre-un-projet/). Une description des questions relatives aux candidatures est proposée ci-après à titre d'information uniquement. *Veillez limiter votre manifestation d'intérêt à 4 pages, références non incluses, avec des marges de 2,5 cm et une police de caractères de 11 points minimum. Les annexes ou supports de la manifestation d'intérêt au-delà des 4 pages ne seront pas examinés.*

**Les informations suivantes doivent y figurer :**

- **Informations sur le candidat**
  - **Nom des organisations participantes**
  - **Descriptions et qualifications des organisations participantes** – fournir un bref aperçu de la mission des organisations participantes, des services fournis et des groupes d'intérêt desservis, de la manière dont elles satisfont aux critères d'éligibilité énoncés ci-dessus, et des qualifications particulières du candidat pour entreprendre le travail proposé, y compris son expérience dans le développement et le partage d'ensembles de données sur la santé ou les soins de santé.
- **Résumé de l'objectif de la manifestation d'intérêt**
  - **Problème et solution proposée** – décrire les disparités ou les inégalités dans le domaine de la santé, qui est le plus affecté par ce problème et la solution proposée (étiquetage, consolidation, création, augmentation ou tenue à jour de l'ensemble de données).

- **Méthodologie générale** – donner un bref aperçu des étapes proposées (et des hypothèses clés) pour élaborer et mettre en œuvre l'ensemble de solutions. Expliquer les autorisations en place ou les mesures que vous prendrez pour obtenir des accords d'utilisation des données.
- **Résultats et avantages escomptés après la mise en œuvre du projet** – expliquer comment le projet proposé contribuera à produire les effets souhaités. Le cas échéant, décrire comment les produits peuvent susciter des pistes de recherche ou applications commerciales multiples et durables.
- **Calendrier général et budget global pour la mise en œuvre du projet :**
  - **Calendrier** – proposer un calendrier général pour l'achèvement des étapes incluses dans la méthodologie générale ci-dessus, y compris le nombre total de mois nécessaires à l'achèvement.
  - **Budget** – fournir une vue d'ensemble du budget prévu pour la réalisation des étapes incluses dans la méthodologie générale ci-dessus. Les budgets doivent être présentés en dollars américains (USD). La dotation totale disponible est d'environ 800 000 USD. Nous prévoyons des budgets dans une fourchette de 10 000 à 100 000 USD pour les projets de petite taille à taille moyenne, et des budgets pouvant atteindre 500 000 USD pour les projets de grande ampleur et complexes. Nous pensons pouvoir financer un ou deux grands projets ou un plus grand nombre de petits projets.