COMMUNIQUÉ DE PRESSE

Paris, le 27 avril 2023

Données synthétiques : Octopize et le Health Data Hub publient un guide à la création et l'évaluation de données de synthèse

Les données synthétiques sont l'une des solutions pour la recherche en données de santé car elles peuvent suppléer des données réelles. Encore faut-il s'assurer de leur qualité. Afin de permettre à l'écosystème de se saisir de ce sujet, la start-up Octopize, spécialisée dans la génération de données de synthétiques anonymes , et le Health Data Hub, plateforme nationale de données de santé, ont collaboré pour réaliser un notebook pédagogique comparant différentes méthodes de génération de données synthétiques. Ce notebook met à disposition des outils d'évaluation de l'anonymat et de la qualité des données synthétiques générées. La méthode de la start-up Octopize permet à la fois de prouver l'anonymat et d'assurer la reproductibilité des analyses. De plus, elle s'applique à tous les cas d'usage avec une faible difficulté dans l'entraînement des données.

Les données de synthèse cherchent à reproduire les caractéristiques structurelles et/ou statistiques de données réelles: même granularité même nombre de lignes et de colonnes si il s'agit d'une table), même valeur statistiques (elles aboutissent à des résultats similaires), elles sont même indiscernables des données d'origines générées à l'aide d'un modèle, elles concernent tous types de données et peuvent être de nature et de complexité diverses.

L'intérêt scientifique des données de synthèse peut se manifester au travers de nombreux cas d'usage. En effet, les données synthétiques peuvent permettre d'évaluer l'intérêt d'une base de données, de créer du contenu pédagogique, d'augmenter la taille d'une cohorte d'étude, voire de conduire une étude scientifique dans son intégralité sans avoir accès aux données réelles, ou uniquement de façon ponctuelle.

Les données synthétiques, outils précieux à la recherche en données de santé

Afin de permettre aux utilisateurs de comprendre comment et pourquoi les données synthétiques peuvent être utilisées dans des projets de recherche médicale, la start-up nantaise Octopize et le Health Data Hub ont collaboré sur la rédaction d'un notebook pédagogique sur ce sujet.

Ce dernier fournit un guide étape par étape pour créer des données synthétiques à l'aide de différents outils, offrant chacun leurs avantages et leurs inconvénients. En s'appuyant sur l'expertise combinée d'Octopize et du Health Data Hub, ce guide met à disposition des outils d'évaluation des données synthétiques générées, du point de vue de l'utilité des données et du niveau de confidentialité assuré. Il propose de plus une illustration pratique de ces étapes, via leur mise en application sur un jeu de données open data.

Un travail permis par la mise en commun des compétences d'Octopize et du Health Data Hub

Ce notebook compare trois méthodes de génération de données de synthèse. La "génération structurelle", développée par le HDH, construit les données de synthèse à partir du schéma structurel de la base de données réelles, de façon aléatoire. La méthode Avatar mise au point par Octopize, recrée les caractéristiques des individus étudiés, et les mélange pour éviter la réidentification des données. Elle a été publiée récemment dans NPJ Digital Medicine (https://www.nature.com/articles/s41746-023-00771-5). La troisième méthode mobilise des techniques de *deep learning* (CT-GAN), où l'algorithme est entraîné sur les données pour en comprendre la structure et en générer de nouvelles sur cette base.

En se basant sur un jeu de données réelles accessible en open source, les équipes d'Octopize et du HDH ont ainsi comparé les données de synthèses créées selon les différentes méthodes de génération, aux données réelles de base.

Ces travaux permettent de démontrer les avantages et inconvénients de chacune de ces méthodes, et de déterminer dans quels cas d'usages chacune est la plus pertinente. Ainsi, si la génération structurelle ne requiert ni l'accès aux données ni aucun entraînement, les données ne peuvent être utilisées qu'à des fins pédagogiques ou de test. À l'inverse, le *deep learning* nécessite une fort volume de données difficiles à entraîner. Quant à l'avatarisation, elle permet à la fois de prouver l'anonymat et d'assurer la reproductibilité des analyses De plus, elle s'applique à tous les cas d'usage avec une faible difficulté dans l'entraînement des données.

La base de la comparaison entre les méthodes de génération de données de synthèse repose sur l'utilisation de métriques d'évaluation de la qualité et de la confidentialité. Ces métriques sont librement accessibles avec le notebook et donnent des clés aux utilisateurs sur les points de vigilance à garder à l'esprit lors de la production de données de synthèse.

Publié sur Gitlab, ce notebook se veut une ressource utile pour les chercheurs, les professionnels de la santé et les étudiants désireux de comprendre comment les données synthétiques peuvent être utilisées pour conduire des projets de recherche médicale tout en protégeant la vie privée des patients.

Lien vers le notebook

Destiné à l'écosystème de la recherche en données de santé et entre autres aux start-ups, le notebook est un exemple des actions d'accompagnement développées par le Health Data Hub à leur égard. Ces dernières feront par ailleurs l'objet d'un événement le 29 juin prochain à PariSanté Campus.

À PROPOS DU HEALTH DATA HUB

Le Health Data Hub est un groupement d'intérêt public créé par la Loi du 24 juillet 2019 relative à l'organisation et la transformation du système de santé. Il associe 56 parties prenantes, en grande majorité issues de la puissance publique (CNAM, CNRS, France Assos Santé…) et met en œuvre les grandes orientations stratégiques relatives au Système National des Données de Santé fixées par l'Etat et notamment le ministère des Solidarités et de la Santé. C'est un service à destination de l'écosystème de santé, des acteurs à l'origine de la collecte de données, des porteurs de projets d'intérêt général et de la société civile. En ce sens, il promeut l'innovation en santé et l'accessibilité des données et des connaissances par le biais, entre autres, d'événements fédérateurs comme l'organisation de data challenge et d'appels à projets.

Contact presse :
presse@health-data-hub.fr

Abonnez-vous à l'infolettre sur le site internet Foire Aux Questions du HDH

06 95 66 26 52

Retrouvez nos engagements vis-à-vis des citoyens
Suivez le Health Data Hub sur LinkedIn et Twitter

À PROPOS D'OCTOPIZE

Proven privacy, Unlocked data.

L'ambition d'Octopize, startup deeptech, est de devenir le leader européen de l'anonymisation des données personnelles grâce à sa méthode brevetée : Avatar. Commercialisée depuis 2019 sous forme de logiciel et de service, la méthode Avatar a été expertisée avec succès par la CNIL en 2020, encouragée par une levée de fonds de 1,5 million € en 2021 et récompensée par le Premier ministre lors du concours i-Nov 2022 avec un nouveau financement de 0,5 million €. Après le droit, les investisseurs et l'État, la communauté scientifique valide à son tour notre méthode Avatar en 2023, dans la revue Nature Digital Medicine. Reconnue depuis plusieurs années dans le secteur sensible de la santé (CHU, instituts de recherche, Big Pharma, Medtech), Octopize accélère aujourd'hui sa croissance dans les télécoms, les assurances, l'automobile et la banque.

Pour en savoir plus : octopize.io
Contact : Olivier BREILLACQ, fondateur & directeur – linkedin.com/in/olivier-breillacq
Contact presse : contact@octopize.io