Réflexions sur la création de données synthétiques sûres et conformes à la réglementation

Helena Canever [1]

Le marché des données

Ces dernières années, les données sont devenues un actif de plus en plus précieux pour les entreprises de divers secteurs. La capacité à collecter, stocker et analyser de grandes quantités de données a permis aux organisations d'obtenir des informations, d'améliorer la prise de décision, de stimuler l'innovation et de rester pertinentes sur des marchés dynamiques et concurrentiels. En conséquence, l'économie des données a connu une croissance rapide et sa valeur est estimée à 271,83 milliards de dollars [2], la croissance annuelle moyenne du volume de création de données étant estimée à 40 % [3]. En fait, certains ont même qualifié les données de « nouveau pétrole » en raison de leur importance et de leur valeur, croissantes dans l'économie actuelle.

Par exemple, des entreprises comme Google et Meta collectent des données sur les requêtes de recherche et les habitudes de navigation des utilisateurs pour améliorer leur ciblage publicitaire, mais aussi pour créer de nouveaux produits. Les appareils IoT tels que les appareils ménagers intelligents, les trackers de fitness et autres appareils connectés collectent également des données sur les activités et les routines quotidiennes des utilisateurs, qui peuvent être utilisées pour améliorer les performances

1. Chercheur dans le Centre de recherche et innovation de Talan, http://www.talan.com.

2. https://www.fortunebusinessinsights.com/big-data-analytics-market-106179.

3. https://fr.statista.com/infographie/17800/big-data-evolution-volume-donnees-numeriques-genere-dans-le-monde.

DOI:10.48556/SIF.1024.22.45 **1024** – Bulletin de la SIF, numéro 22, novembre 2023, p. 45–53

des appareils ou pour créer de nouveaux produits et services. En outre, ces données peuvent également être vendues à d'autres entreprises à des fins diverses telles que le marketing ciblé, la gestion des risques ou même la recherche et le développement, ce qui en fait une marchandise précieuse et échangeable.

Cette situation crée une tension entre la croissance des données et la protection de la vie privée. D'une part, les entreprises doivent collecter et analyser les données afin de rester compétitives dans l'économie actuelle. D'autre part, elles doivent également se conformer aux réglementations sur la confidentialité des données et protéger les informations personnelles des individus.

Ce dilemme est d'autant plus évident à la lumière des montants impressionnants des amendes infligées aux entreprises en cas de non-respect du règlement général sur la protection des données (RGPD).

Entré en vigueur en mai 2018, le RGPD [4] établit des règles strictes pour la collecte, l'utilisation et le stockage des données personnelles, donnant aux individus de l'Union européenne un plus grand contrôle sur leurs données personnelles et obligeant les entreprises à obtenir le consentement explicite des individus avant de collecter, utiliser ou diffuser leurs données. Depuis sa conception, cette autorité européenne a émis près de 3 milliards d'euros d'amendes. Dans le cadre du traitement sûr et approprié des informations personnelles, elle a infligé 272 amendes pour un montant total de 375 millions d'euros pour « *insuffisance des mesures techniques et organisationnelles visant à assurer la sécurité des informations* », 476 amendes pour un montant total de 457 millions d'euros pour « *insuffisance de la base juridique du traitement des données* », et 363 amendes pour un montant total de près de 1,7 milliard d'euros pour « *non-respect des principes généraux de traitement des données* » [5].

En outre, si le RGPD est actuellement l'autorité de réglementation des données la plus stricte et la plus importante avec près d'un demi-milliard de citoyens de l'UE, ce n'est pas la seule réglementation de ce type, et d'autres pays et régions ont également introduit une législation similaire pour protéger les données personnelles des individus. Par exemple, en 2018, l'état de Californie a introduit la loi californienne sur la protection de la vie privée des consommateurs (CCPA).

En plus de la pression exercée par les organismes de réglementation, les entreprises et les organisations sont plus que jamais exposées aux cyberattaques, qui entraînent non seulement des violations de données personnelles sensibles, mais qui, dans leur ensemble, érodent la confiance du consommateur. Depuis 2004, plus de 15 milliards de comptes ont été violés, dont près d'un demi-milliard en France, et les dommages causés par les *ransomwares*, qui ont tendance à cibler des institutions

4. https://gdpr-info.eu.
5. https://www.enforcementtracker.com/?insights.

particulièrement sensibles comme les hôpitaux, devraient dépasser 30 milliards de dollars dans le monde en 2023 [6].

Dans ce contexte, les entreprises cherchent à concilier cette tension en mettant en œuvre des technologies capables de protéger la confidentialité des données tout en leur permettant de les utiliser et de les analyser à des fins diverses. Cet article vise à explorer les données synthétiques en tant que solution technologique à cette question, et à élaborer le pipeline de gestion des données approprié pour minimiser les risques et maximiser la conformité au RGPD.

Données synthétiques

Selon une définition large, les données synthétiques sont toutes les données structurées (par exemple, tableaux, bases de données relationnelles) ou non structurées (par exemple, images, vidéos, texte) qui sont générées artificiellement et qui ne représentent pas un objet ou un événement réel. Les données synthétiques sont généralement créées par l'intelligence artificielle soit en ajoutant du bruit soit d'autres transformations aux données originales. Dans cet article, nous ferons strictement référence aux données synthétiques comme étant des données générées par l'IA.

L'adoption des données synthétiques a connu une augmentation au cours des dernières années et de nouvelles applications pour celles-ci sont continuellement développées. Cette tendance peut être attribuée à l'innovation dans le domaine de l'apprentissage automatique, en particulier les réseaux neuronaux, et à l'augmentation de la puissance de traitement disponible pour entraîner ces algorithmes, et ce pour deux raisons.

6. https://surfshark.com/research/data-breach-monitoring, https://dl.acronis.com/u/rc/White-Paper-Acronis-Cyber-Protect-Cloud-Cyberthreats-Report-Mid-year-2022-EN-US-220811.pdf.

Premièrement, les développements dans le domaine de l'apprentissage automatique, et notamment de l'apprentissage profond, nous ont permis de créer et d'entraîner efficacement des réseaux neuronaux pour créer des données synthétiques très proches des données originales. Un excellent exemple est l'utilisation du *Generative Adversarial Network* pour créer des photos d'individus qui n'existent pas avec un niveau de réalisme qui était auparavant inatteignable [7].

La deuxième raison de l'adoption des données synthétiques est que la demande croissante de solutions d'IA nécessitant l'entraînement de modèles d'IA s'accompagne de l'exigence d'une grande quantité de données de haute qualité, étiquetées et pertinentes. Pour la plupart des entreprises et des startups non GAFAM (Google, Apple, Facebook, Amazon, Microsoft), répondre à cette exigence reste un processus coûteux en temps et en argent. En ce sens, les données synthétiques peuvent produire des données de haute qualité sur demande et à une fraction du coût. Les données synthétiques sont également utiles pour augmenter les cas limites dans les ensembles de données déséquilibrés. Par exemple, dans le domaine de la santé, les individus malades sont généralement minoritaires par rapport aux individus sains du groupe témoin, mais ils constituent souvent la catégorie la plus importante. Dans le secteur bancaire, les cas de fraude sont rares, mais les fournisseurs de cartes de crédit ont un intérêt particulier à les identifier. Dans ce cas, les données synthétiques peuvent augmenter artificiellement le nombre de cas limites dans un ensemble de données, ou reproduire des environnements ou des scénarios rares tels qu'une radiographie de poumon malade ou une fraude à la carte de crédit [3]. Enfin, les données synthétiques ont été adoptées comme outil d'anonymisation des informations sensibles ou personnelles.

Le compromis confidentialité-utilité

Les données synthétiques sont souvent considérées comme une solution aux réglementations relatives à la vie privée et à la protection des données, car elles sont conçues pour ne pas représenter des personnes ou des événements réels. Toutefois, il est important de noter que les bases de données synthétiques de haute dimension et riches en informations peuvent toujours être exposées à des attaques contre les données personnelles. Ces attaques, appelées attaques par inférence, permettent de discerner l'identité d'une personne ou d'un enregistrement particulier en utilisant les informations circonstancielles relatives à l'enregistrement qui, autrement, ne seraient pas identifiables, et sont particulièrement efficaces contre les données aberrantes. La conservation des informations de l'ensemble de données originales augmente la vulnérabilité des données synthétiques aux attaques par inférence. Cela a été démontré par Stadler et al. [5] qui ont constaté que les données synthétiques générées à partir de modèles sans protection explicite de la vie privée laissent les valeurs aberrantes

7. https://thispersondoesnotexist.com.

vulnérables aux attaques par inférence. D'autre part, la mise en œuvre d'une protection différentielle de la vie privée dans les ensembles de données synthétiques se fait au prix d'une diminution de l'utilité à des fins statistiques. Cet équilibre entre la similarité entre les données synthétiques et les données réelles et la préservation de la vie privée est souvent appelé le compromis utilité-vie privée.

Par exemple, la création d'une version synthétique de la base de données RH d'une entreprise qui conserve la même distribution hiérarchique des employés que l'ensemble de données originales permettrait de déduire, avec une marge d'erreur, le salaire et la prime du PDG, car il serait le seul employé de sa catégorie et son salaire serait probablement nettement supérieur à la médiane. L'augmentation du niveau de confidentialité différentielle dans cet ensemble de données, par exemple en modifiant la distribution, entraînerait le sacrifice de certaines propriétés statistiques des données.

Malgré la question du compromis vie privée-utilité, dans les secteurs de la santé et de la finance, les données synthétiques sont souvent saluées comme une solution totalement anonyme qui peut permettre l'utilisation de données sensibles tout en restant dans les limites des réglementations sur la protection des données comme le RGPD. Les partisans de l'utilisation des données synthétiques font valoir que, puisque les données générées n'identifient plus aucun individu réel spécifique, elles sont totalement anonymes. En réalité, des violations de données personnelles peuvent toujours se produire, même à partir de données synthétiques.

Le texte actuel du RGPD définit les données personnelles comme « *toute information concernant une personne physique identifiée ou identifiable* ».

Les données synthétiques générées à partir de données personnelles concerneront toujours une personne physique puisque les données personnelles sont le substrat sur lequel les données synthétiques sont générées. La clé pour déclencher la réglementation européenne actuelle est l'identifiabilité des individus. Un individu est identifié si les données contiennent des informations personnelles telles que le nom, l'adresse, l'âge, et il est identifiable si d'autres informations sont présentes qui ont le potentiel de se rapporter à un individu en particulier. En ce sens, un ensemble de données synthétiques qui reproduit parfaitement toutes les propriétés statistiques de l'ensemble de données original peut néanmoins exposer l'identité ou les informations personnelles des valeurs aberrantes en raison de leur caractère unique. Le plus souvent, pour que les données synthétiques soient sûres, il faut ajouter un certain niveau de bruit ou de distorsion par rapport à l'original. En d'autres termes, l'utilité de l'ensemble de données, qui correspond à sa similarité avec l'original, est diminuée [2].

Du point de vue de l'entreprise, le problème est double, car la confidentialité et l'utilité peuvent être difficiles à évaluer a priori. En outre, l'identifiabilité des données synthétiques dépend fortement du contexte et peut changer au fil du temps avec l'évolution de la technologie, ce qui accroît l'incertitude. Dans ce contexte, ils doivent soigneusement équilibrer ces exigences contradictoires pour s'assurer qu'ils

sont en mesure de maximiser la valeur de leurs données tout en les protégeant contre un accès et une utilisation non autorisés. Cette tâche peut s'avérer complexe et difficile, en particulier pour les entreprises qui traitent des ensembles de données volumineux et complexes.

Gestion des données

L'utilisation des données personnelles est soumise à une réglementation stricte par le RGPD, même dans les limites d'une entreprise. Plus précisément, selon l'article 5 du RGPD, les données personnelles « *sont collectées pour des finalités déterminées, explicites et légitimes et ne sont pas traitées ultérieurement de manière incompatible avec ces finalités* » et « *ne sont pas conservées sous une forme permettant l'identification des personnes concernées pendant une durée n'excédant pas celle nécessaire à la réalisation des finalités pour lesquelles elles sont traitées* ».

En d'autres termes, une entreprise ne peut traiter des données à caractère personnel que si elle dispose d'une base légale pour le faire, pour une finalité spécifique et limitée et pour une durée limitée. Il arrive souvent que l'utilisation interne de données à caractère personnel pour développer des solutions d'IA ne soit pas justifiée, car un modèle d'apprentissage automatique ne peut pas être formé à partir de données à caractère personnel si une entreprise ne dispose pas de motifs légitimes, comme le consentement, l'exécution du contrat ou un intérêt légitime pour le faire. L'article 5 du RGPD limite également la conservation à long terme de la plupart des données personnelles. Par exemple, une entreprise ne peut pas conserver les données de ses anciens employés une fois leur contrat terminé ou un fournisseur de services ne peut pas conserver les données personnelles des utilisateurs finals une fois leur abonnement au service terminé au-delà de ce qui est strictement nécessaire.

Par conséquent, la formation des générateurs de données synthétiques et l'utilisation conséquente des données générées, en particulier si les données de formation ne peuvent être conservées, peuvent nécessiter le consentement explicite des personnes et une mise à jour des conditions générales.

Un problème de cybersécurité

Les modèles d'apprentissage automatique peuvent être vulnérables aux cyberattaques qui provoquent des violations de données et certaines attaques constituent un risque particulier pour les données personnelles. Dans les attaques par inversion de modèle, la connaissance du modèle peut conduire à la connaissance des données d'entraînement avec un certain degré de précision. En revanche, dans les attaques par inférence d'appartenance, les connaissances sur les données d'entraînement ne sont pas récupérées, mais il est possible de déduire si un individu particulier figurait

ou non dans l'ensemble d'entraînement. Les deux types d'attaques peuvent être réalisés uniquement sur la base de l'accès aux requêtes, par exemple par le biais d'une API (attaques de type boîte noire), ou avec des connaissances sur l'architecture du modèle (attaques de type boîte blanche) [6].

En raison de ces vulnérabilités, non seulement les données synthétiques, mais aussi leur générateur lui-même, doivent être considérés avec le même niveau de sécurité que celui qui serait accordé à l'ensemble de données originales.

Un pipeline pour la manipulation de données synthétiques

Dans cet article, j'ai mis en évidence les problèmes potentiels concernant la confidentialité, la sécurité et la fidélité des données synthétiques, l'objectif global étant la création de générateurs de données synthétiques qui soient sûrs à utiliser sans diminuer la valeur des données originales.

Pour y parvenir, je propose un pipeline représentant les étapes de la conception, de la formation et de l'utilisation réussie des données synthétiques au sein d'une entreprise :

(1) avant d'adopter une IA générative et d'utiliser des données sensibles comme données d'entraînement, il est important que les professionnels des données, mais aussi les dirigeants et les propriétaires de produits, connaissent les possibilités et les pièges des données synthétiques. L'organisation de courtes formations ou d'ateliers sur le sujet pourrait éviter les idées fausses et une utilisation plus éclairée de cette technologie ;

(2) les responsables des données au sein de l'entreprise doivent évaluer les pipelines existants et les pratiques de gouvernance des données pour s'assurer que la structure actuelle des données est conforme à la réglementation et sûre. Si ce n'est pas le cas, l'entreprise doit réviser son architecture et sa réglementation pour s'adapter au traitement des données synthétiques et des générateurs de données synthétiques. Cette étape est cruciale pour minimiser le risque de sanctions réglementaires ;

(3) comme je l'ai mentionné dans le paragraphe précédent sur le compromis vie privée-utilité, les garanties de confidentialité explicites réduisent la vulnérabilité d'un ensemble de données synthétiques aux attaques par inférence. Les garanties formelles de confidentialité peuvent être mises en œuvre comme un paramètre de confidentialité par rapport auquel l'enregistrement de sortie d'un modèle génératif est testé. L'enregistrement n'est libéré que si un certain nombre d'enregistrements d'entrée sont indiscernables [1] ;

(4) une fois les données synthétiques générées, leur utilité doit être évaluée pour s'assurer que la qualité des données est préservée et que les données sont toujours utiles pour les applications en aval. Une première approche consiste

à mesurer la corrélation entre les caractéristiques des données originales et celles de l'ensemble de données synthétiques. On peut également comparer des mesures statistiques comme la moyenne, la médiane, l'écart-type, les valeurs manquantes, etc. La corrélation entre les caractéristiques doit également être comparée si le type d'information est disponible dans l'ensemble de données originales. Enfin, si des mesures de prédiction comme le score F1 [8] ou la précision d'un modèle entraîné sur l'ensemble de données originales sont disponibles, le modèle doit être ré-entraîné sur les données synthétiques et évalué par rapport à un ensemble de données réelles. Une comparaison des performances du second modèle devrait nous permettre de savoir si les caractéristiques des données originales ont été capturées par le générateur de données synthétiques [4] ;

(5) une fois que votre générateur a été créé et que ses données synthétiques sont utilisées, que ce soit pour le test d'applications, l'analyse ou l'apprentissage automatique, une supervision constante est importante. Les données doivent toujours être testées par rapport à des données réelles si ces dernières sont mises à jour. Les générateurs peuvent également être ré-entraînés pour capturer les changements dans le monde réel. De plus, il est important de recevoir un retour constant de la part des ingénieurs de données pour garantir la fidélité des données synthétiques à l'original, non seulement dans leur nature mais aussi dans leur forme. Les formats et les structures sont-ils respectés de manière à permettre à ces données de s'intégrer de manière transparente à l'architecture MLOps ?

(6) enfin, bien que le texte actuel du RGPD ne mentionne pas explicitement les données synthétiques et les générateurs de données synthétiques, les responsables des données doivent toujours se tenir au courant des changements possibles, car ceux-ci ont tendance à ne pas suivre le même rythme que le développement technologique. Si des changements surviennent, l'ensemble du pipeline doit être réévalué.

Conclusions

Les données synthétiques ont fait l'objet d'un certain engouement ces derniers temps, qu'il s'agisse d'aider à la formation de voitures à conduite autonome [9] ou d'améliorer la détection des fraudes sur les cartes de crédit [10]. Ce battage est basé sur la tendance actuelle qui fait des données un actif très précieux avec une grande

8. en.wikipedia.org/wiki/F-score.

9. *Waymo, formerly Google self-driving car project*, https://waymo.com.

10. https://www.forbes.com/sites/johnkoetsier/2020/09/21/50-less-fraud-how-amex-uses-ai-to-automate-8-billion-risk-decisions/?sh=7301903f1a97.

valeur potentielle mais qui est encore parfois difficile, laborieux et coûteux à traiter. Dans cet article, je me suis concentrée sur les données synthétiques qui peuvent être générées à partir d'ensembles de données tabulaires et relationnelles et qui, à mon avis, sont plus susceptibles de contenir des informations personnelles sensibles. Dans ce contexte, j'ai souligné le compromis confidentialité-utilité et la façon dont l'équilibre entre les deux peut être un processus délicat dicté par les besoins de sécurité et les réglementations officielles en matière de protection des données, et qui est motivé par la nécessité de préserver et d'amplifier la valeur intrinsèque des données que l'on veut synthétiser.

À la lumière de ce que je pense être les obstacles les plus courants au déploiement des données synthétiques, j'ai présenté à la fin de l'article les principales étapes générales à suivre pour garantir une exploitation sûre et efficace de la technologie des données synthétiques. Pour chaque ensemble de données et chaque cas d'utilisation, les détails de ces étapes seraient probablement très différents afin de s'adapter aux risques et aux besoins spécifiques de l'application.

Références

[1] Vincent Bindschaedler, R. Shokri, and Carl A. Gunter. Plausible deniability for privacy-preserving data synthesis. *ArXiv*, abs/1708.07975, 2017.

[2] Cesar Augusto Fontanillo López. On the legal nature of synthetic data. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.

[3] Sergey I. Nikolenko. *Synthetic Data for Deep Learning*, volume 174. Springer, 2021.

[4] Joshua Snoke, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and Specific Utility Measures for Synthetic Data. *Journal of the Royal Statistical Society Series A : Statistics in Society*, 181(3) :663–688, 03 2018.

[5] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data – anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1451–1468, Boston, MA, August 2022. USENIX Association.

[6] Michael Veale, Reuben Binns, and Lilian Edwards. Algorithms that remember : model inversion attacks and data protection law. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 376, 2018.