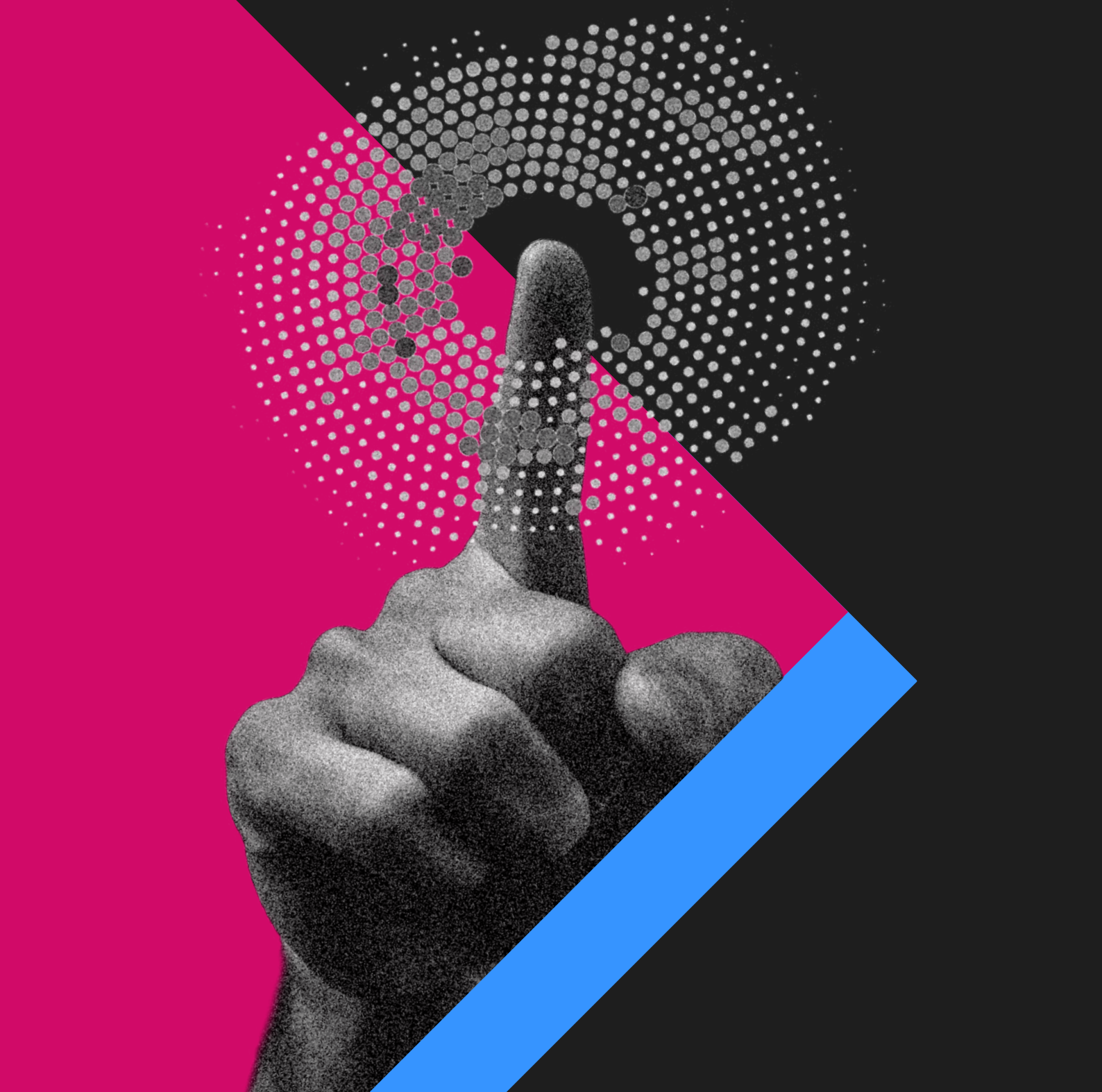
esomar

Esomar 5 sujets de discussion pour aider les acheteurs de données synthétiques augmentées

pour les études de marché et insights



esomar

© 2025. Publié en septembre 2025

Ce document a été rédigé en anglais et le texte anglais (disponible à www.esomar.org) fait foi. Le texte peut être copié, distribué et transmis à condition que la source soit correctement mentionnée et que l'avis suivant soit inclus.

Traduction française par Philippe Guilbert (Esomar Professional Standards Committee).

ISBN : 978-90-8347094-8

Avertissement : ce document ne devrait pas être considéré comme un avis juridique. Il est recommandé de faire appel à un conseiller juridique compétent pour votre situation dans la juridiction appropriée. Malgré une préparation et une révision minutieuses, cette publication peut contenir des erreurs et des imperfections. Esomar et ses contributeurs déclinent toute responsabilité quant aux conséquences qui pourraient en découler.

Les réponses que les entreprises peuvent apporter à ces questions relèvent de leur responsabilité et ne reflètent pas les opinions d'Esomar ou de ses membres. Ces entreprises ont toujours le contrôle total des déclarations qu'elles fournissent et, par conséquent, Esomar ne peut garantir ni être tenu responsable de leur véracité et de leur conformité, ni des conséquences qui pourraient en découler.

Publié par Esomar, Amsterdam, Pays-Bas.
Pour plus d'informations, veuillez envoyer un e-mail à professional.standards@Esomar.org

Esomar 5 sujets de discussion pour aider les acheteurs de données synthétiques augmentées

pour les études de marché et insights

À propos d'

esomar

Depuis 1947, Esomar est le centre mondial de la recherche, de l'insight et de l'analytique. Avec plus de 50 000 membres, plus de 750 entreprises et plus de 130 pays, nous sommes une organisation mondiale qui permet aux professionnels de l'insight et aux entreprises de libérer leur potentiel à l'échelle mondiale et locale, en favorisant les connexions, la collaboration, la croissance et les connaissances.

Animés par nos valeurs fondamentales d'inclusion, d'attention, d'innovation et de confiance, nous guidons ce secteur en constante évolution depuis plus de 75 ans. Nous poursuivons notre engagement à renforcer les standards professionnels et éthiques, à faciliter la formation, à défendre nos intérêts auprès des législateurs, à partager les meilleures pratiques et à promouvoir des solutions basées sur des faits pour les décideurs.

www.esomar.org

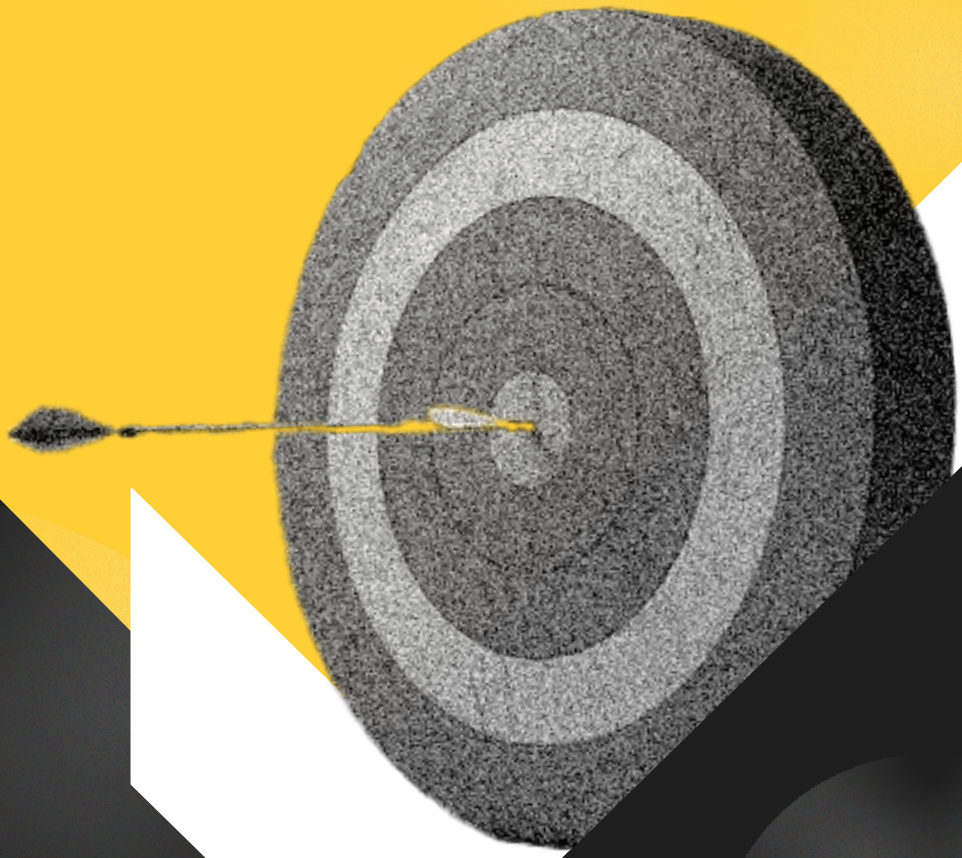
Contenu

Objectif et champ d'application 5

1. Rôle des données synthétiques augmentées 7
2. Avec quels types de données votre système peut-il fonctionner ? 8
3. Quelle méthode votre système utilise-t-il pour synthétiser les données ? 9
4. Comment votre système peut-il être évalué ? 10
5. Utilisation d'informations supplémentaires 13

Considérations de conformité lors de l'utilisation de données synthétiques 14

Remerciements 18



Objectif et champ d’application

Objectif et champ d'application

L'objectif de cette liste de contrôle est d'aider les chercheurs en études de marché, études d'opinion et sociales, ainsi que les analystes de données, à s'y retrouver dans le domaine des données synthétiques. Elle vise à aider les acheteurs et les fournisseurs à appliquer le niveau de contrôle approprié lorsqu'ils commandent un service de données synthétiques, afin de favoriser la transparence, la confiance et la sécurité dans son utilisation et dans le domaine de la recherche en général. Ce cadre en cinq sujets vise à faciliter la discussion sur les avantages et les inconvénients potentiels de l'approche adoptée. Il permettra aux acheteurs de mieux comprendre comment une solution synthétique proposée par un fournisseur peut être utilisée dans le contexte de l'entreprise et/ou de la prise de décision.

Cette liste de contrôle accompagne les *20 questions d'Esomar destinées à aider les acheteurs de services basés sur l'IA pour les études de marché et insights.*

Ces cinq sujets sont importants pour évaluer les avantages et les risques liés à l'utilisation d'approches synthétiques. Esomar recommande d'examiner attentivement les sections pertinentes des 20 questions destinées à aider les acheteurs de services basés sur l'IA dans le cadre de tout exercice visant à évaluer l'utilisation et la valeur des données synthétiques. Il devrait également consulter la mise à jour 2025 *du Code international ICC/Esomar des études de marché, études d'opinion et sociales, et de l'analytique des données.* Esomar recommande vivement aux fournisseurs de souscrire à ce Code, qui contient des exigences spécifiques relatives à l'utilisation des données synthétiques en termes de transparence, de divulgation, de fourniture d'informations techniques et de validation.

L'équipe du projet reconnaît la nature hautement technique du sujet. En réfléchissant à la meilleure façon de rédiger ce cadre, nous nous sommes efforcés de trouver un équilibre entre les exigences des chercheurs techniciens/scientifiques des données et celles des directeurs Etudes & Insights/clients finaux qui cherchent à distinguer les différentes approches, à traiter les problématiques d'entreprise et à éclairer la prise de décision.

Les acheteurs et les fournisseurs sont encouragés à utiliser ce cadre comme un outil dynamique. À mesure que les pratiques en matière de données synthétiques évoluent, ce cadre devra être mis à jour, en tenant compte des outils supplémentaires et des sessions de discussion fournis dans le cadre de la contribution d'Esomar à l'utilisation responsable des données synthétiques dans le secteur de la recherche.

[1]Les termes « acheteur » et « fournisseur » sont utilisés dans différents contextes. Par exemple, un « acheteur » peut être un client qui fournit des données internes à un fournisseur afin de générer des données synthétiques ; un deuxième scénario peut être celui où un acheteur sous-traite des services à une entreprise qui dispose de données internes pour créer des données synthétiques ; un autre scénario peut être celui où un fournisseur acquiert des données auprès d'un tiers afin de générer des données synthétiques pour un acheteur. Aux fins de la présente liste de contrôle, le terme « utilisateurs » est considéré comme synonyme de « acheteurs ». Il est recommandé

Aux fins du présent cadre, les données synthétiques constituent un sous-domaine de l'IA. Elles sont définies comme suit :

- *L'intelligence artificielle (IA)* est un ensemble de technologies conçues pour simuler l'intelligence humaine et les capacités de résolution de problèmes.
- *Les données synthétiques* désignent les informations générées pour reproduire les caractéristiques des données du monde réel.

En ce qui concerne le champ d'application actuel des données synthétiques dans le domaine de la recherche, nous distinguons trois domaines à prendre en considération.

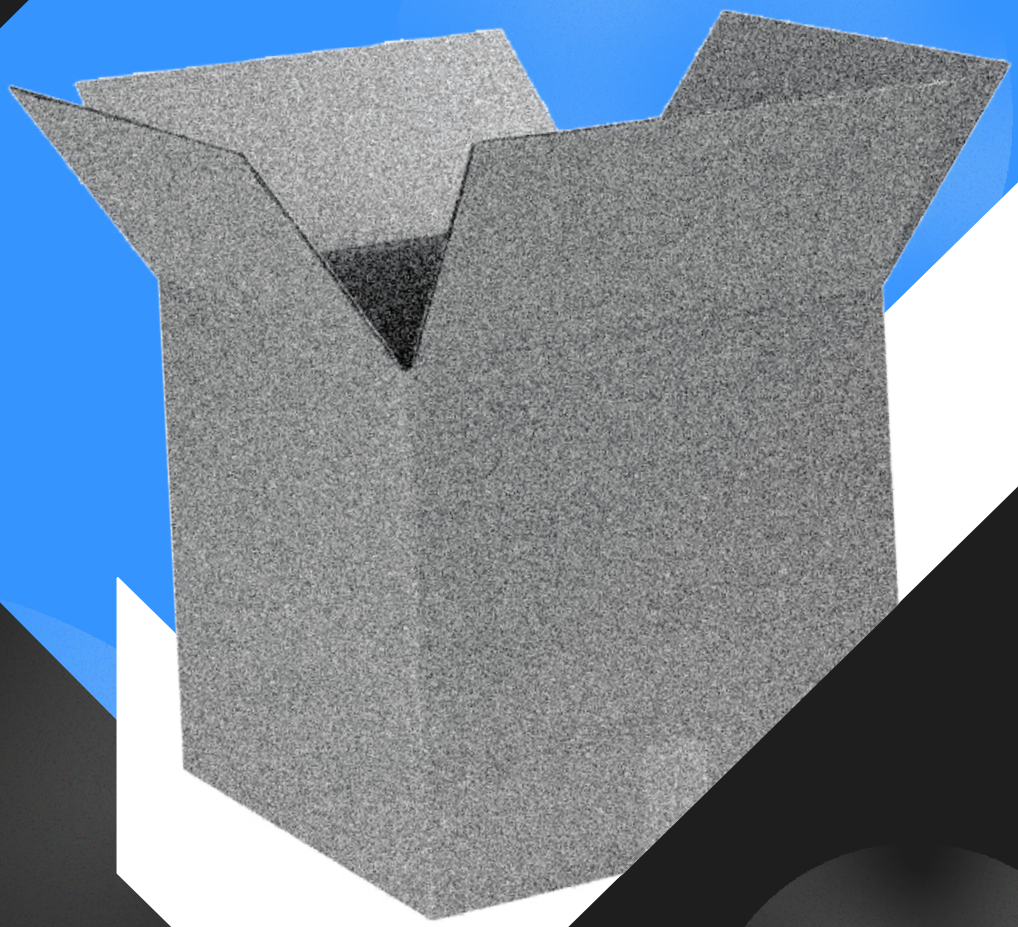
1. *Boost et imputation des données synthétiques (également appelées données synthétiques augmentées)*
   Les données synthétiques sont générées pour être utilisées en complément des données primaires collectées auprès de personnes réelles. Cela inclut la création de cas supplémentaires et/ou de valeurs supplémentaires pour des personnes existantes. Cela s'applique aux augmentations d'échantillon qui combinent les données provenant de participants humains avec les données produites par un modèle.

2. *Données entièrement synthétiques ou données synthétiques complètes*
   Un ensemble de données entièrement nouveau est généré et n'inclut aucune donnée primaire. Ce terme peut également s'appliquer aux jumeaux numériques, des données créées pour reproduire les caractéristiques de personnes réelles, par exemple en s'appuyant sur plusieurs sources et études.

3. *Personas synthétiques*
   Représentations numériques d'une personne ou d'un groupe de personnes créées pour imiter les comportements, les préférences et les caractéristiques de personnes ou de groupes réels.

Le champ d'application de ce cadre couvre le premier domaine (boost et imputation des données synthétiques) et exclut les deuxième et troisième domaines.

Ce document est conçu pour servir de cadre de discussion, au sein ou en dehors d'une organisation, pour les conversations avec les fournisseurs. Il présente les raisons pour lesquelles ce sujet a été inclus, ainsi que des informations supplémentaires.



Les cinq sujets à aborder avec les fournisseurs

1. Rôle des données synthétiques

Cette section permettra de clarifier le rôle et le potentiel des données synthétiques.

a) Quel est le rôle prévu des données synthétiques ? Par exemple, s'agit-il d'augmenter la taille de l'échantillon, d'améliorer la définition et la robustesse de l'échantillon, de compenser les lacunes de l'échantillon ou d'aider à modéliser et à prédire les réponses ? Qu'est-ce que vous pouvez désormais réaliser qui n'était pas possible avant l'inclusion des données synthétiques et pourquoi ?

Contexte

Les réponses à cette question aideront l'utilisateur à comprendre comment le fournisseur envisage l'utilisation des données synthétiques, les raisons qui le poussent à les utiliser, les domaines dans lesquels il estime qu'elles peuvent le mieux compléter ou remplacer les pratiques de recherche habituelles et/ou résoudre des problèmes ou limites connus tels que le biais d'échantillon, les faibles taux de réponse ou la rareté des données, et pourquoi. Elles donneront une indication sur l'ampleur des expérimentations menées, ce qui a fonctionné et ce qui n'a pas fonctionné, ainsi que les conclusions tirées.

(b) Vos données d'origine peuvent-elles être synthétisées ? La taille et/ou la nature des données d'origine et leurs propriétés statistiques sont-elles telles qu'elles peuvent être modélisées pour créer des données synthétiques utiles ?

(c) Quelles parties des données se prêtent à la synthétisation ?

Contexte

Certains sujets sont beaucoup plus faciles à modéliser que d'autres, en fonction de la nature de ce qui est mesuré et des interrelations entre les données. Les données attitudinales, par exemple, peuvent être plus faciles à synthétiser que les informations factuelles (comme la couleur des cheveux) ; les données sensibles au facteur temps (comme la notoriété publicitaire) peuvent être plus difficiles à modéliser que les données plus stables (comme la notoriété spontanée d'une marque).

(d) Quel est l'impact des biais sur vos données ? Quels sont les biais présents dans les données et quel est leur impact sur les résultats ? La partie synthétique de l'ensemble de données reflète-t-elle les biais présents dans la partie de données réelles et amplifie-t-elle les biais des données réelles ?

Contexte

Les résultats des données synthétiques peuvent parfois reproduire et amplifier les biais des ensembles de données de base, ce qui peut avoir une incidence sur les résultats. Les

[2]**Le biais temporel** consiste à prendre en compte des questions telles que la période de l'année/la saisonnalité et tout impact connexe.
[3]**La validité conceptuelle** est la capacité d'un ensemble de mesures à refléter quelque chose qui n'est pas

données synthétiques ne pourront pas combler les lacunes des données sous-déclarées dans l'ensemble de données d'origine, même dans les cas où elles peuvent compenser le sous-échantillonnage. La transparence concernant les biais et les méthodes permettant de les surmonter est essentielle pour clarifier la nature de l'offre synthétique et son utilisation. Les biais peuvent inclure le biais d'échantillonnage, le biais temporel et la validité conceptuelle.[3]

(e) Pourriez-vous fournir des exemples de cas où votre système a produit des résultats médiocres ou trompeurs ? Comment ceux-ci ont-ils été identifiés et quelles mesures ont été prises ?

Contexte

Il est important de comprendre où et quand le système présente des défaillances, comme celles liées au seuil d'échantillonnage et au type de questions qui posent problème. La discussion à ce sujet permettra également d'évaluer le degré de transparence du fournisseur en tant que partenaire professionnel.

2. Avec quels types de données votre système peut-il fonctionner ?

Cette section aidera l'acheteur à évaluer si le système proposé correspond à ses objectifs d'entreprise et est susceptible d'apporter un bénéfice clair.

(a) Quels types de données votre système peut-il traiter en termes d'entrées et de sorties ? Quelle est la taille minimale requise pour l'échantillon et quels éléments de données sont utilisés dans le processus de modélisation ?

Contexte

Cette question vous aidera à comprendre les fondements de l'approche et sa validité, et vous donnera une indication du type et de la quantité de données requises. Elle clarifiera les types de données spécifiques et les tailles minimales et maximales requises. Par exemple, le système est-il capable de traiter des variables nominales, ordinales , d'intervalles, de ratio et des données textuelles ? Existe-t-il des exigences minimales ou maximales en matière de taille d'échantillon ? (Par exemple, est-il efficace avec 100 cas ? Peut-il atteindre 100 000 cas ?) Existe-t-il des limites à l'expansion des données (par exemple, les données synthétiques peuvent-elles augmenter l'ensemble de données uniquement d'un certain pourcentage ?) Les réponses à ces questions vous aideront également à déterminer si les résultats correspondent à vos besoins et vous indiqueront l'éventail complet des résultats possibles proposés par le fournisseur.

Voici d'autres questions à prendre en considération :

(b) Quels éléments des données sont utilisés dans la modélisation ?

Contexte

Même si le système est capable de lire toutes vos données, il est important de comprendre quelles parties spécifiques sont utilisées dans le processus de modélisation. Certaines méthodes de modélisation se concentrent sur l'isolement de certains paramètres à modéliser, ce qui peut introduire des biais involontaires, tandis que d'autres adoptent une approche plus holistique qui peut introduire davantage de bruit dans le modèle. Pour les techniques telles que la modélisation par similarité, où des répondants similaires sont comparés, combien de répondants similaires sont utilisés dans la comparaison ?

(c) La nature de l'augmentation repose-t-elle sur des données chronologiques/séries temporelles, et votre modèle est-il capable de les traiter ?

Contexte

La modélisation des données chronologiques/séries temporelles nécessite des aspects et des considérations de modélisation spécifiques. Cela vous aidera à comprendre à quelle vitesse le processus deviendra obsolète et si les données peuvent être améliorées en tenant compte de facteurs tels que la saisonnalité et les changements au fil du temps.

[4]**Données ordinales** – données nominales ordonnées

3. Quelle méthode votre système utilise-t-il pour synthétiser les données ?

Cette section vous aidera à clarifier la méthode qui sous-tend le système et le ou les types de situations dans lesquels il est le plus ou le moins adapté.

(a) Quelle méthode votre système utilise-t-il, par exemple, est-il basé sur des méthodes de modélisation par similarité telles que les techniques du plus proche voisin, ou sur du bootstrapping, des algorithmes statistiques (par exemple, les réseaux bayésiens) ou utilise-t-il un modèle génératif profond (tel que GANS, Diffusion, Autorégression) et/ou d'autres méthodes d'apprentissage automatique ?

Contexte

Cette question permettra de déterminer s'il s'agit d'une technique statistique ou de modélisation qui cherche à utiliser l'IA pour générer de nouvelles données. Chaque méthode présente un compromis entre la quantité de données d'apprentissage nécessaires et la sophistication de la modélisation.

Que devriez-vous rechercher dans une réponse ?

Les réponses à cette question vous aideront à comprendre si l'approche proposée fonctionne dans le cas des données de base et pour le problème que vous cherchez à résoudre. Vous devriez rechercher spécifiquement la mention de toutes les approches et techniques déployées au sein du système et la manière dont elles interagissent pour obtenir le résultat. Les paramètres utilisés par un fournisseur peuvent être considérés comme propriétaires, mais les acheteurs ont besoin d'une description des algorithmes et des techniques appliqués pour pouvoir évaluer quand la technique est plus susceptible de fonctionner ou d'échouer. Par exemple :

- Les techniques du plus proche voisin peuvent fonctionner avec des échantillons de taille assez réduite, car elles constituent une forme de pondération à micro-échelle. En contrepartie, cette méthode n'apporte aucune nouvelle information et amplifie les biais ou les anomalies de l'échantillon que vous synthétisez.

- Les approches de modélisation statistique tiennent compte des tendances et des relations statistiques plus larges dans les données et nécessitent un échantillon plus important pour cartographier de manière fiable ces relations. Elles peuvent également avoir du mal à différencier les facteurs causaux et corrélatifs, ce qui nécessite des compétences et une expertise pour identifier les relations.

- Les techniques de modèle génératif profond les plus sophistiquées peuvent identifier des schémas récurrents et isoler les facteurs causaux. Cependant, ces méthodes nécessitent généralement beaucoup plus de données d'entraînement et de ressources informatiques (potentiellement un ordre de grandeur bien supérieur à celui d'autres approches) pour fonctionner de manière fiable. Elles exigent également un niveau d'expertise plus élevé pour être mises en œuvre correctement."

[5]La modélisation par similarité est un processus qui consiste à créer des répliques d'individus existant dans votre ensemble de données d'origine. Il peut s'agir, par exemple, d'une « personne moyenne » créée à partir de deux ou plusieurs individus figurant dans l'ensemble de données de base.

D'autres questions doivent être prises en compte ici :

(b) Pour quels types de données ou de situations votre système fonctionne-t-il bien, et dans quelles situations est-il moins efficace ?

(c) Comment votre approche gère-t-elle les changements dynamiques continus, comment les données modélisées peuvent-elles être mises à jour et à quelle fréquence ?

(d) Votre système peut-il s'adapter à plusieurs zones géographiques présentant des caractéristiques, des catégories de produits et des types d'études différents, tout en conservant sa validité et sa comparabilité ?

(e) Votre système peut-il prendre en charge la réalisation simultanée ou fréquente de nombreuses études, par exemple des études de suivi multi-pays, des études continues sur la valeur de marque ou des projets de segmentation à la demande ? Quelles sont les limites techniques ou procédurales ?

Contexte

Il s'agit d'une considération pratique si vous envisagez d'utiliser le boost et l'imputation de données synthétiques à grande échelle. Elle permet de déterminer si le fournisseur est en mesure de gérer ce volume de travail.

4. Comment votre système peut-il être évalué ?

Cette section aidera l'acheteur à comprendre comment le système proposé peut être évalué et validé.

(a) Comment avez-vous évalué votre système ? Quelle théorie sous-jacente utilisez-vous ? Comment vos méthodes globales sont-elles validées et comment validez-vous chaque ensemble de données ? Comment l'évaluation de votre système est-elle liée aux données et aux besoins de ce projet spécifique ? Les validations sont-elles effectuées pour chaque projet ou ensemble de données spécifique ? À quel niveau des données les tests sont-ils effectués ?

Contexte

Les utilisateurs ont besoin d'un moyen d'évaluer si les données synthétiques sont adaptées à l'usage prévu. Cette question vise à clarifier l'approche adoptée pour l'évaluation, à déterminer si les processus sont génériques et s'ils sont appliqués à chaque projet. Certains ensembles de données sont plus faciles à synthétiser que d'autres en fonction de la force des relations dans les données d'apprentissage et de la diversité sous-jacente des réponses à toute question. Il ne suffirait donc pas qu'un fournisseur affirme que ses méthodes fonctionnent et ont été validées. Il est important de déterminer comment un utilisateur peut comprendre dans quelle mesure elles peuvent fonctionner ou ont fonctionné pour son projet spécifique.

Pour évaluer la fiabilité de l'offre, les tests de base tels que les tests d'exactitude et de précision ne sont pas suffisants. Il est nécessaire d'examiner la reproductibilité des schémas de données et la fidélité des données, pour lesquelles il existe une gamme spécialisée de tests statistiques.

Que devraient rechercher les acheteurs dans une réponse/discussion ?

Les acheteurs auront besoin de réponses qui se rapportent à des éléments qu'ils peuvent vérifier, plutôt que de références à des performances passées. Ils devraient s'efforcer de faire la distinction entre une réponse qui met l'accent sur l'idée que la solution a fonctionné, par exemple dans X cas sur Y, et qu'elle devrait donc fonctionner X/Y % du temps, et une réponse théorique.

De nombreux systèmes prétendent montrer que les données synthétiques sont similaires aux données d'origine, mais il est important de comprendre ce que signifie « similaire » et comment cela est mesuré. Une similarité trop grande peut indiquer que la méthode de synthétisation copie des données plutôt que de créer de nouveaux points de données. Un autre domaine d'intérêt est la génération de réponses ouvertes. Le système génère-t-il des réponses ouvertes ? Si oui, s'agit-il de doublons de réponses ouvertes existantes ou l'IA est-elle utilisée pour les générer et, si oui, comment est-elle utilisée ?

Les acheteurs devraient également demander des conseils sur la manière d'évaluer les informations fournies par les nouvelles données (d'origine et synthétiques). Par exemple, quels types de différences pourraient être significatifs, sachant que les données synthétiques ont tendance à ne pas avoir les mêmes propriétés statistiques que les données primaires, ce qui a un impact sur des statistiques aussi fondamentales que l'écart-type. Il peut être utile de déterminer si chaque modèle est un nouveau cas d'utilisation ou s'il est « actualisé » à chaque fois. Il peut également être utile de comprendre si le fournisseur utilise des tests de validation pour comparer la qualité des valeurs synthétiques aux approches modélisées.

D'autres questions à prendre en compte ici sont les suivantes :

(b) Dans quelle mesure les données synthétiques correspondent-elles aux données qui n'ont pas été utilisées pour les produire ?

Contexte

Par exemple, quelle est la distribution des valeurs dans l'ensemble de données augmentées et comment se compare-t-elle à un ensemble de données de taille comparable constitué entièrement de données primaires collectées auprès de personnes réelles qui n'ont pas été utilisées dans le processus de modélisation/synthétisation des données ? Les corrélations, les tendances et la saisonnalité sont-elles préservées ? Comment des tests tels que ceux de significativité statistique peuvent-ils être effectués ? Les données capturent-elles les valeurs aberrantes ou les événements inhabituels/ponctuels ? Il peut être utile de déterminer comment la qualité/fiabilité des données d'origine a été évaluée par le fournisseur et son système.

(c) Comment les acheteurs peuvent-ils évaluer le système ?

Contexte

Lorsqu'un utilisateur reçoit les données augmentées, comment peut-il les évaluer ? Par exemple, un benchmarking a-t-il été effectué par rapport à des ensembles de données réelles ? Cette question permettra de clarifier le degré de transparence et d'ouverture des approches de validation et la manière dont elles peuvent être testées.

(d) Les insights sont-ils stables sur plusieurs cycles du processus de génération de données ? Existe-t-il des sources connues de variabilité et comment sont-elles communiquées ?

(e) Les données synthétiques sont-elles marquées, étiquetées ou filigranées dans les résultats ? La provenance de chaque enregistrement ou point de données est-elle traçable jusqu'à sa source d'origine (données primaires et données dérivées synthétiquement) ?

Contexte

Les chercheurs et les clients devraient être en mesure de vérifier l'ensemble de données et de comprendre sa composition[6].

(f) Quelle documentation est fournie avec l'ensemble de données synthétiques ?

Contexte

Les acheteurs peuvent trouver utile de savoir si le fournisseur dispose de métadonnées expliquant comment les données ont été générées et modélisées, ainsi que d'informations relatives aux journaux d'audit, rapports de validation ou limites connues, qui peuvent aider les utilisateurs à interpréter et à diffuser les résultats de manière responsable au sein de leur propre organisation.own organisations.

[6]Cf l'article 7 du Code international ICC/Esomar 2025 des études de marché, études d'opinion et sociales, et de l'analytique des données, qui définit des exigences spécifiques en matière de transparence pour l'utilisation de données synthétiques dans la recherche.

5. Utilisation d'informations supplémentaires

Cette section aidera l'acheteur à comprendre si des données ou des sources d'informations supplémentaires ont été ajoutées et, le cas échéant, pourquoi et comment cela aura un impact sur les données synthétiques augmentées.

Votre méthode fonctionne-t-elle uniquement avec les données d'échantillon collectées ou intègre-t-elle des données ou informations supplémentaires ?

Contexte

Cette question permettra de déterminer si le fournisseur a pris en compte et a accès à l'ensemble complet des informations pertinentes, ainsi qu'à anticiper une dépendance future en matière de données. Certaines méthodes de modélisation peuvent s'appuyer sur des données d'apprentissage provenant de tiers, ce qui peut soulever des questions quant à la propriété des données synthétisées.

Que devriez-vous rechercher dans une réponse ?

Quelles sont les informations supplémentaires ? S'agit-il, par exemple, de données d'enquête, de données transactionnelles, de données secondaires ou de paramètres de modèles d'IA ? D'où proviennent les données supplémentaires, par exemple du fournisseur ou d'un tiers ? Quels biais sont susceptibles d'être contenus dans les données supplémentaires, notamment les biais d'échantillonnage, les biais temporels et la validité conceptuelle ? Tout biais contenu dans les données supplémentaires, par exemple les données provenant d'études antérieures, est susceptible d'avoir un impact sur les données synthétiques augmentées.

Outre les problèmes tels que la représentativité de l'échantillon et le biais des questions dans les données supplémentaires, si celles-ci sont anciennes, elles pourraient masquer les changements présents dans les nouvelles données après la génération des données synthétiques. Il existe un risque d'amplification des biais des données de base, par exemple, le manque de diversité peut amplifier les déséquilibres latents des données d'origine.

Les données supplémentaires devraient être répertoriées et les biais potentiels de chaque

[7]**Le biais temporel** consiste à prendre en compte la période de l'année/la saisonnalité et tout impact connexe.
[8]**La validité conceptuelle** est la capacité d'un ensemble de mesures à refléter quelque chose qui n'est pas directement mesurable.



Considérations de conformité lors de l'utilisation de données synthétiques

Considérations de conformité lors de l'utilisation de données synthétiques

Objectif et champ d'application

Cette section est conçue pour aider les acheteurs et les fournisseurs à prendre en compte les aspects juridiques pouvant s'appliquer à l'utilisation des données synthétiques. Basée sur une section similaire tirée du document ***« 20 questions pour aider les acheteurs de services basés sur l'IA » publié par Esomar,*** elle peut être utilisée en complément du document « 5 sujets de discussion pour aider les acheteurs de données synthétiques augmentées pour les études de marché et insights » publié par Esomar, ou comme liste de contrôle autonome pour les données synthétiques. Avec le développement des sources de données synthétiques, l'acheteur devrait anticiper les répercussions potentielles de la législation, et cette liste de contrôle peut aider à identifier ces problèmes.

Objet de cette section

Cette liste de contrôle aidera les acheteurs et les fournisseurs à déterminer s'ils connaissent suffisamment les cadres juridiques qui régissent les activités basées sur l'IA. Les fournisseurs d'IA et leurs clients sont soumis aux exigences en matière de protection des données et de sécurité de l'information imposées par les lois et réglementations sur la protection des données. Ces lois et réglementations varient selon les juridictions, avec des lois et réglementations différentes s'appliquant dans différents pays ou États au sein d'un même pays, et sont généralement interprétées en fonction du lieu où les données ont été collectées ou de l'emplacement du fournisseur .

En outre, les fournisseurs et les acheteurs peuvent être soumis à des lois et réglementations relatives à la législation sur la confidentialité, l'intelligence artificielle, la propriété intellectuelle, les droits d'auteur et d'autres domaines pertinents.

Les réponses aux questions de cette section peuvent aider les acheteurs à comprendre les politiques, procédures et pratiques mises en œuvre par un fournisseur en matière de protection des données, de sécurité de l'information et de conformité.

1. Provenance des données : documentez-vous l'origine et le traitement des données d'entrée et ces sources sont-elles mises à disposition ?

Contexte

Les données devraient être obtenues et traitées de manière responsable, légale et éthique. Il est donc important de comprendre d'où proviennent les données d'entrée, comment et par qui elles

[9]Les lois et réglementations applicables en matière de protection des données comprennent, sans s'y limiter : la loi sur la protection des informations personnelles ou APPI (Japon) ; la loi australienne sur la protection de la vie privée (Australie) ; la loi californienne sur la protection des consommateurs ou CCPA (État de Californie aux États-Unis) ; la loi sur la protection de la vie privée des enfants en ligne ou COPPA (États-Unis) ; la loi sur la protection des données (Royaume-Uni) ; les modifications relatives aux exigences de localisation des données apportées à la loi sur la protection des données (Fédération de Russie) ; la loi générale sur la protection des données (Brésil) ; le règlement général sur la protection des données ou RGPD (UE/EEE) ; la loi sur la portabilité et la responsabilité en matière d'assurance maladie ou HIPAA (États-Unis) ; la loi Graham-Leach Bliley ou GLBA (États-Unis) ; et la LPRPDE (Canada).

2. Avis de confidentialité : avez-vous vérifié ou demandé l'avis de confidentialité (parfois appelé « politique de confidentialité ») du fournisseur de la source d'origine de toutes les données personnelles utilisées pour créer le modèle synthétique, ainsi qu'un résumé des concepts clés qu'il aborde ?

Contexte

Les lois et réglementations en matière de protection des données ainsi que les codes de conduite du secteur des études de marché exigent qu'un avis de confidentialité accompagne la collecte de données. Celui-ci devrait renseigner sur toutes les données personnelles collectées et traitées, la finalité pour laquelle elles sont utilisées et la manière dont elles sont utilisées, divulguées, gérées et traitées. L'examen de l'avis de confidentialité d'un fournisseur peut aider les acheteurs à comprendre ses procédures et pratiques en matière de données personnelles et à déterminer dans quelle mesure il se conforme aux lois, réglementations et codes de conduite applicables.

Dans le cadre de la production de données synthétiques, des informations claires et significatives devraient être fournies aux personnes concernées sur l'intention de collecter leurs données personnelles en vue de la génération ultérieure de données synthétiques. Lorsque les données personnelles sont anonymisées avant d'être introduites dans des systèmes de données synthétiques, une bonne pratique consiste à informer les personnes concernées que leurs données personnelles seront anonymisées puis utilisées à ces fins. Dans tous les cas, l'explication des données synthétiques doit être claire, rédigée dans un langage simple et éviter le jargon technique.

3. Quelles mesures ont été prises pour se conformer aux lois sur la protection des données et mettre en œuvre des mesures visant à protéger la vie privée des participants à la recherche ? Avez-vous évalué les risques pour les personnes concernées, comme l'exige la législation sur la protection de la vie privée ? Avez-vous obtenu le consentement pour le traitement des données lorsque cela était nécessaire ? Quelles mesures ont été prises pour supprimer les identifiants des personnes d'origine lors de la production de données synthétiques ?

Contexte

Il est essentiel de comprendre la position du fournisseur en matière de conformité afin de garantir le respect des lois et de contribuer à la protection des données personnelles des individus.

Les lois mondiales sur la confidentialité exigent des chercheurs qu'ils évaluent les risques que le traitement des données personnelles peut présenter pour les personnes et, dans de nombreux cas, qu'ils obtiennent leur consentement pour traiter leurs données à des fins particulières. Cela se fait généralement par le biais d'une analyse d'impact sur la protection des données (AIPD), et des conseils sont généralement disponibles auprès de l'autorité nationale de protection des données.

Lorsque des données personnelles sont traitées, les lois sur la protection de la vie privée exigent généralement des chercheurs qu'ils soient transparents avec les personnes concernées quant à la manière dont leurs données seront traitées, les personnes avec lesquelles elles seront partagées, les fins auxquelles elles seront utilisées et la manière dont elles seront sécurisées. Les acheteurs et les fournisseurs sont soumis aux exigences

en matière de protection des données et de sécurité de l'information imposées par les lois et réglementations relatives à la protection des données.

Dans le cadre de la production de données synthétiques, les réponses à cette question vous aideront à comprendre si le fournisseur a anticipé et atténué les challenges de confidentialité associés aux méthodes qu'il utilise.

4. Quelles mesures prenez-vous pour garantir que les systèmes IA-synthétiques sont résistants aux attaques, au bruit et autres perturbations potentielles ? Quels cadres et standards de sécurité de l'information utilisez-vous ?

Contexte

La réponse devrait inclure un processus permettant d'identifier et de traiter les vulnérabilités et les menaces potentielles pour le système synthétique, sa conception ou son fonctionnement, telles que la manipulation des données d'entrée ou l'utilisation abusive ou inappropriée du système. La réponse devrait également couvrir les mesures visant à détecter et à lutter contre les cyberattaques, les systèmes de cybersécurité et la conformité du modèle à des standards de sécurité spécifiques. Existe-t-il des plans de secours permettant de mettre en place un plan de sauvegarde en cas de problème ? Il peut être utile de comprendre où se trouvent les données et si elles sont stockées dans un cloud partagé ou dédié, ou sur un appareil.

5. Propriété des données : le fournisseur des données d'origine définit-il et communique-t-il clairement la propriété des données (entrées, sorties et livrables), y compris les droits de propriété intellectuelle et les autorisations d'utilisation ?

Contexte

La propriété des données saisies dans le modèle est-elle clairement établie ? Il est utile de préciser dès le départ ce qui peut ou ne peut pas être fait avec les données, et de veiller à ce que cela soit correctement documenté.

Les propriétaires des données peuvent souhaiter que leurs informations restent confidentielles ou secrètes et exiger qu'elles ne soient pas publiées dans le domaine public ou dans des outils basés sur le cloud. Il est important de préciser clairement quelles données sont considérées comme confidentielles et pour lesquelles le téléchargement dans des technologies accessibles au public devrait être évité.

Il peut être utile de déterminer s'il existe des politiques et des avis concernant l'approche à adopter en cas de retrait potentiel de données personnelles, lorsque celles-ci ont ensuite été utilisées pour créer des données synthétiques.

[10]Les exemples incluent, sans s'y limiter, COBIT, HITRUST, ISO 27001, ISO 42000, le cadre de cybersécurité du NIST et SOC 2.

6. Existe-t-il des restrictions quant à l'utilisation des données d'entrée ?

Contexte

Les propriétaires de données peuvent ne pas être à l'aise avec le fait que leurs données soient traitées pour la production de données synthétiques. Un certain nombre de technologies d'IA sont nouvelles, et le propriétaire des données peut ne pas souhaiter que ses données soient téléchargées dans un système qu'il ne connaît pas, qu'il n'a pas examiné ou auquel il ne fait pas confiance. Les propriétaires de données peuvent ne pas autoriser le mélange de leurs données avec d'autres ensembles de données. Il existe toute une série de lois émergentes sur l'IA et les propriétaires de données ne souhaiteront pas que leurs données soient chargées dans des technologies potentiellement non conformes. En outre, il se peut que les propriétaires de données souhaitent stipuler dans quel pays ou quelle région les données ou les modèles devraient être conservés. Les propriétaires de données peuvent être des clients, des fournisseurs et/ou des prestataires de données tierces. Il est conseillé de s'assurer que les relations contractuelles avec les propriétaires de données sont claires et précisent les utilisations autorisées et interdites des données fournies dans un contexte synthétique.

7. Propriété : Savez-vous clairement à qui appartient le résultat ?

Contexte

Les parties peuvent s'interroger sur l'entité propriétaire des résultats, c'est-à-dire sur le détenteur de la propriété intellectuelle (PI) des informations synthétiques produites par le système. Les clients, les fournisseurs et les prestataires de données tierces peuvent avoir leur propre approche en matière de propriété des résultats, compte tenu (par exemple) de leur investissement dans la recherche, la technologie et/ou les processus d'approvisionnement en données. Il est recommandé d'établir des relations contractuelles claires entre les parties concernées, en précisant la propriété des données synthétiques

Remerciements

Esomar tient à remercier tous ces contributeurs pour leur temps et leurs précieux apports.

L'équipe du projet

Ray Poynter	Co-fondateur, ResearchWiseAI | Ancien président d'Esomart
Jon Puleston	Méthodologue en chef, Ipsos | Membre du comité des standards professionnels Esomar
Judith Passingham	Présidente du comité des standards professionnels Esomar

Contributeurs

Des contributions précieuses pour cet exercice ont été fournies par :
Comité des standards professionnels Esomar
Comité des affaires juridiques Esomar

En collaboration avec des personnes possédant une expertise technique spécifique et/ou un point de vue client/utilisateur :

Alina Serbanica	Ipsos
Andrew Jeavons	Signoi
Andrew Thompson	Kantar
Barry Jennings	G3 Consulting
Chris Davies	Kantar
Corrine Moy	Consultant | Membre du Conseil Esomar
George Terhanian	Electric Insights
Kevin Cowan	BBC
Kiran HN	Kantar
Rachael Popoola	Pepsico
Renee Smith	Toluna Group
Richard Timpone	Protopian Works
Vinay Ahuja	Procter & Gamble Europe
Ziad Elmously	MetrixLab

Animé par

Kathy Joe	Consultante du comité des standards professionnels Esomar
Xabier Palacio	Esomar Responsable Esomar Intelligence, Advocacy & Standards

esomar

www.esomar.org