Synthetische Daten für Datenschutz – Testlauf gebührenfrei möglich

Originalartikel

Backup

<html> <header class=„article-header“><h1 class=„articleheading“>Synthetische Daten für Datenschutz – Testlauf gebührenfrei möglich</h1> Daniel AJ Sokolov </header><figure class=„aufmacherbild“><figcaption class=„akwa-caption“>Im Gesundheitsbereich können synthetische Daten einerseits für Datenschutz, andererseits zur Erzeugung hilfreicher Bildvarianten eingesetzt werden.(Bild: lenetstan/Shutterstock.com)</figcaption></figure> Hilfe für Datenschutz bei Big Data, Schuhlöffel für autonome Autos, Forschung für Kinderwunsch: Synthetische Daten und einschlägige Start-ups sind in. Synthetische Daten sind groß im Kommen, genauer gesagt von Künstlicher Intelligenz (KI) erzeugte, strukturierte synthetische Daten. Von der Klimaforschung zur Suizidverhütung, von selbstfahrenden Autos bis zur Big-Data-Datenschutzlösung – synthetische Daten ziehen in immer mehr Bereiche ein. Für Datenschutzprojekte sind Testläufe mit strukturierten synthetischen Datensätzen dank des österreichischen Start-ups Mostly.ai gebührenfrei möglich. Einsatz finden synthetische Daten insbesondere dort, wo es nicht genügend (variantenreiche) Daten gibt, oder wo vorhandene Originaldaten aus Gründen des Datenschutzes nicht direkt eingesetzt werden können. Beispielsweise im Bereich der Entwicklung selbstfahrender Autos kommt synthetische Datenerzeugung zum Einsatz, um mehr Varianten ähnlicher Situationen virtuell ausprobieren zu müssen. So können einige bereits vorhandene Aufnahmen plötzlich auf die Fahrbahn laufender Kinder um tausende synthetische Aufnahmen erweitert werden, in denen unterschiedlich aussehende Kinder aus unterschiedlichsten Richtungen zu unterschiedlichen Tageszeiten und Witterungsverhältnissen auf Fahrbahnen unterschiedlicher Ausgestaltung laufen, ohne dass dafür echte Kinder aufgeopfert werden müssten. Software für autonome Fahrzeuge kann dann mit allerlei synthetisch erzeugten Situationen virtuell konfrontiert werden und sich beweisen. Das US-Unternehmen Parallel Domain wurde 2017 dazu gegründet, virtuelle Welten aus echten Straßenkarten zu kreieren. Inzwischen füllt es diese Welten mit vielerlei Licht- und Wetterverhältnissen sowie synthetischen Fahrzeugen und Menschen, die sich mitunter überraschend verhalten. Zu den Kunden zählen beispielsweise Continental, Google, sowie Woven Planet (ehemals Toyota Research Institute, das Lyfts Selbstfahrtochter übernommen hat [1]). Bekanntere Beispiele synthetischer Daten für mehr Vielfalt sind KI-generierte Bilder von

Menschengesichtern oder Katzen. <h3 class=„subheading“ id=„nav_synthetische0“>Synthetische Daten für Datenschutz</h3> Für Datenschutzbelange eingesetzt, ist Ziel von Datensynthetisierung, vorhandene Daten vollständig und unumkehrbar zu anonymisieren, ohne die Nützlichkeit und Nutzbarkeit der in den echten Daten enthaltenen statistischen Informationen zu verlieren. Die Anonymisierung durch synthetische Daten funktioniert allerdings nur korrekt, wenn wesentliche Schutzmaßnahmen gesetzt werden. Offensichtlichstes Beispiel: Die anhand der echten Daten trainierte KI darf die echten Daten (und Metadaten) natürlich nicht zu exakt nachbilden, sonst könnte man ja einfach die Datenbank kopieren. Auch müssen Originaldaten (samt Metadaten) meist leicht reduziert werden: Spezielle Ausreißer sind zu entfernen, bevor die KI daran trainiert wird (im Fachenglisch rare category protection genannt). Es gibt einfach nicht so viele Deutsche, die mehrfache Formel-1-Weltmeister sind und schwere Ski-Unfälle hatten. Das Risiko, Michael S. in synthetischen Daten neu anzulegen, wäre zu groß. Weitere Schutzmaßnahmen müssen bei der Erstellung von Datenbanken aus synthetischen Daten greifen – die Fortschritte, die KI-Experten bei De-Anonymisierung gemacht haben, sind beachtlich. Details dazu sprengen allerdings den Rahmen dieses Artikels. <h3 class=„subheading“ id=„nav_branchenstandard1“>Branchenstandards sind beabsichtigt</h3><figure class=„a-u-inline-right a-inline-image a-u-inline“> <figcaption class=„a-caption“>Alexandra Ebert, Vorsitzende der IEEE Synthetic Data Industry Connections(Bild: mostly.ai)</figcaption></figure> Solche synthetische Daten, die auf echten personenbezogenen Daten beruhen, sind rechtlich gesehen anonymisierte Daten, technisch gesehen aber vielleicht gar keine personenbezogenen Daten. Standards gibt es für synthetische Daten und deren Einsatz für Datenschutzzwecke noch keine. Bei der IT-Branchenorganisation IEEE Standards Association gibt es eine Arbeitsgruppe, die Vorarbeiten für Standardisierung leistet. Sie heißt Synthetic Data Industry Connections [2] und wird von der Österreicherin Alexandra Ebert organisiert, die im Brotberuf Chief Trust Officer beim Unternehmen Mostly.ai mit Sitz in Wien und New York City ist. Die Firma erzeugt unter besonderer Berücksichtigung des Datenschutzes synthetische Daten für Unternehmen wie Nvidia, Telefonica, Versicherungen, Banken oder die Stadt Wien. Im März war Ebert im c't-Datenschutz-Podcast Auslegungssache 58 [3] zum Thema synthetische Daten zu Gast. „Synthetische Daten funktionieren so, dass Du im Gegensatz zu traditioneller Anonymisierung nicht am original Datensatz herumschraubst, versuchst etwas wegzulöschen, zu ändern oder zu verfälschen, sondern Du nutzt den Originaldatensatz nur dazu, Künstliche Intelligenz zu trainieren. Diese KI hat dann vereinfacht gesagt die Aufgabe, herauszufinden, wie sich die (Erzeuger) der Daten verhalten. Was sind die Statistiken, die Muster, die zeitlichen Abhängigkeiten“, erklärte sie in der Auslegungssache 58. <h2 class=„opt-

intitle“>Empfohlener redaktioneller Inhalt</h2> Mit Ihrer Zustimmmung wird hier ein externer Podcast (Podigee GmbH) geladen. <label class=„opt-incta-persistence“><input class=„opt-inpersistence-checkbox“ data-should-persist=„“ type=„checkbox“ /> Podcasts immer laden</label> <button class=„opt-incta“ data-opt-in=„“>Podcast jetzt laden</button> Ich bin damit einverstanden, dass mir externe Inhalte angezeigt werden. Damit können personenbezogene Daten an Drittplattformen (Podigee GmbH) übermittelt werden. Mehr dazu in unserer Datenschutzerklärung [4]. <noscript> </noscript> Traditionelle Anonymisierung nutzt destruktive Verfahren, die auf originalen Datensätzen beruhen und Teile wegstreichen. Oft bleibt nicht viel über. Das schränkt dann den Nutzen der Daten ein. „So etwas wie KI (auf traditionell anonymisierte Daten) zu trainieren, ist nicht mehr sinnvoll möglich“, stellte Ebert fest. Gleichzeitig bliebe das Risiko der Re-Identifizierung bestehen: Denn bei Verhaltensdaten aus Big Data, beispielsweise Banktransaktionen oder Gesundheitsdaten, funktioniere traditionelle Anonymisierung nicht mehr. KI sei einfach zu gut bei Re-Identifizierung. <h3 class=„subheading“ id=„nav_beispiel_projekt2“>Beispiel-Projekte</h3> Im Bereich Gesundheitsdaten sind leicht Beispiele zu finden. Zur Förderung künstlicher Befruchtungen könnte es helfen, die Qualität von Embryos im Frühstadium (Blastoyzsten) besser zu bewerten [5]. An entsprechender KI forschen das Kinderwunschzentrum am Kepler Universitätsklinikum im oberösterreichischen Linz gemeinsam mit dem Software Competence Center Hagenberg,. Weil nicht so viele Bilder von Blastozysten zur Verfügung, hat eine KI (konkret Generative Adversarial Networks) weitere Varianten erzeugt. Nicht unähnlich hat BMW eine KI für Qualitätssicherung in der Produktion – trainiert wurde sie anhand hunderttausender auf Knopfdruck erzeugter, synthetischer Bilder [6]. Das US-Veteranenministerium hat mit dem Wettbewerb „Mission Daybreak [7]“ 20 Millionen Dollar ausgelobt, um Mittel und Wege zu finden, die Suizidraten unter Ex-Soldaten zu senken. In der ersten Runde des Wettbewerbs wurden 20 Projekte ausgesucht, die nun Zugriff auf synthetische Daten über Veteranen und deren Gesundheit erhalten. Die Echtdaten können aus Datenschutzgründen nicht preisgegeben werden. Die Preisträger des Wettbewerbs sollen dieser Tage bekanntgegeben werden. Dann wird sich zeigen, ob und wie sie die synthetischen Daten nutzen. <div class=„a-u-inline ho-text c3“><header class=„mb-4“><h3 class=„inline-flex pb-2 pr-8 text-xl font-bold leading-none border-b-4 border-gray-800 dark:border-white“>Lesen Sie auch</h3></header><section data-component=„TeaserList“ class=„grid gap-6 md:gap-y-8“ data-sneak-peek-elements-container=„true“><article data-component=„TeaserContainer“ data-cid=„6338468“ data-content-id=„3263590“ class=„flex ho-text“ data-teaser-name=„MinimalHorizontalTeaser“ data-upscore-object-id=„6338468“><a data-component=„TeaserLinkContainer“ href=„https://www.heise.de/news/Googles-Algorithmen-stufen-Ziffern-als-Copyright-Verletzung-ein-6338468.html“ class=„group flex“ data-google-interstitial=„true“ data-upscore-url=„true“><figure data-component=„Image“ class=„w-24 mr-2 md:w-40 shrink-0 md:mr-4“> <img data-ff-replacement=„1“ width=„1280“ height=„720“ src=„https://www.heise.de/imgs/18/3/2/6/3/5/9/0/Copyright_Infringement_500-bffcb6624ba4d227.png“ alt=„Email: Your file violates Google Drive's Terms of Service Your file ‘500-nonewline.txt’ contains content that violates Google Drive's Copyright Infringement policy and hence, some features related to this file may have been restricted. Thanks for

helping Google keep the web safe. *Restricted file* 500-nonewline.txt A review cannot be requested for this restriction. Google LLC 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA. You have received this email because one of your files violates Google Drive's Terms of Service."“ class=„c1“ referrerpolicy=„no-referrer“ /> </figure> <header data-component=„TeaserHeader“><h3 class=„flex flex-col“>Googles Algorithmen stufen Ziffern als Copyright-Verletzung ein</h3></header> </article></section> Für den Finanzbereich schildert Ebert im c‘t-Podacst das Beispiel von Transaktionsdaten einer Bank. Daraus geht hervor, wie oft Pensionisten zum Bankomaten gehen oder wie häufig Studenten bei Amazon einkaufen. „All das wird auf sehr granularer Ebene (von einer KI) erlernt; und dann, in einem komplett separaten Schritt, wird der Algorithmus genutzt, um neue synthetische Daten zu erzeugen“, sagte Ebert, „Ich habe dann synthetische Konsumenten und deren synthetische Finanztransaktionen. Da gibt es keinen 1:1-Bezug zwischen einem echten (Menschen) und irgendeinem synthetischen Individuum.“ Aber trotzdem seien im Datensatz die gleichen Statistiken zu finden wie in den Originaldaten. Die für die Bank wertvollen Muster bleiben erhalten, jedoch ohne datenschutzrelevanten Personenbezug. <h3 class=„subheading“ id=„nav_kein_simpler3“>Kein simpler Remix</h3> Anders ausgedrückt: Die Geschichten, die die synthetischen Daten erzählen, ähneln den Geschichten der Originaldaten sehr, aber die handelnden Charaktere sind andere. Allerdings soll es sich, richtig synthetisiert, nicht um einen simplen Remix echter Daten handeln, sondern um neu erstellte Datensätze. Mit Synthetisierung sollen über 90% der in einem Datenkonvolut enthaltenen Information erhalten werden, verspricht die Branche. Mit traditioneller Anonymisierung, korrekt umgesetzt, wäre es oft nur ein einstelliger Prozentwert. Die synthetisierten Daten können mit Dritten geteilt oder als Open Data veröffentlicht werden. Und natürlich kann das eigene Unternehmen die synthetischen Daten dort verwenden, wo es die Originaldaten nicht auswerten darf, weil diese zu anderen Zwecken erhoben wurden (juristisches Stichwort: Zweckbindung). Um Unternehmen und Forschern den Einstieg in die Arbeit mit synthetischen Daten für Datenschutzbelange zu erleichtern, stellt Mostly.ai einen gebührenfreien Generator [8] für Versuche zur Verfügung. Mit dem Testdatengenerator kann jeder Nutzer eigene Daten einsetzen und daraus pro Tag bis zu 100.000 Zeilen synthetischer Daten generieren lassen. () URL dieses Artikels:<small><code>https://www.heise.de/-5045353</code></small> Links in diesem Artikel:<small><code>[1] https://www.heise.de/news/Lyft-gibt-selbstfahrende-Autos-auf-6028972.html</code></small><small><code>[2] https://standards.ieee.org/industry-connections/synthetic-data/</code></small><small><code>[3] https://www.heise.de/hintergrund/Auslegungssache-58-EU-Datenstrategie-synthetische-Daten-Bias-und-Datenschutz-6336597.html</code></small><small><code>[4] https://www.heise.de/Datenschutzerklaerung-der-Heise-Medien-GmbH-Co-KG-4860.html</code></small><small><code>[5] https://www.softwarepark-hagenberg.com/partner-news/detail/news/schwanger-dank-kuenstlicher-intelligenz</code></small><small><code>[6] https://www.heise.de/select/ct/2022/20/2221009185134215638</code></small><small><code>[7] https://www.missiondaybreak.net/</code></small><small><code>[8] https://mostly.ai/synthetic-data-platform/generate-synthetic-data/</code></small><small><code>[9] mailto:ds@heise.de</code></small> Copyright © 2022 Heise

Medien</p> </html>

From:
https://schnipsl.qgelm.de/ - **Qgelm**

Permanent link:
https://schnipsl.qgelm.de/doku.php?id=wallabag:wb2synthetische-daten-fr-datenschutz--testlauf-gebhrenfrei-mglich

Last update: **2025/06/27 11:17**