Hecker, Dirk; Voss, Angi; Paaß, Gerhard; Wirtz, Tim

Big Data 2.0 – mit synthetischen Daten KI-Systeme stärken

Wirtschaftsinformatik & Management

Provided in Cooperation with:
Springer Nature

Suggested Citation: Hecker, Dirk; Voss, Angi; Paaß, Gerhard; Wirtz, Tim (2022) : Big Data 2.0 – mit synthetischen Daten KI-Systeme stärken, Wirtschaftsinformatik & Management, ISSN 1867-5913, Springer Fachmedien Wiesbaden GmbH, Wiesbaden, Vol. 15, Iss. 2, pp. 161-167, https://doi.org/10.1365/s35764-022-00437-z

This Version is available at:
https://hdl.handle.net/10419/312223

Big Data 2.0 – mit synthetischen Daten KI-Systeme stärken

Bei der Anwendung von Künstlicher Intelligenz (KI) sind fehlende Daten immer noch eine Kernherausforderung und die Kosten zur Beschaffung ein kritischer Faktor für die Wirtschaftlichkeit vieler Geschäftsmodelle. Synthetische, also künstlich generierte Daten bilden einen Ausweg. Ein vielversprechender Lösungsansatz besteht darin, für die Datensynthese selbst ein KI-Modell einzusetzen.

Dirk Hecker, Angi Voss, Gerhard Paaß und Tim Wirtz

Wirtschaftsinformatik & Management 2023 • 15 (2): 161–167
https://doi.org/10.1365/s35764-022-00437-z
Angenommen: 27. September 2022
Online publiziert: 12. Dezember 2022

Bei datengetriebenen Ansätzen hat sich gezeigt, dass komplexere Problemstellungen große KI-Modelle erfordern und diese wiederum umfangreiche Datenmengen mit ausreichender Varianz (Big Data) benötigen. In der Praxis ist häufig der unzureichende und kostspielige Zugang zu Trainingsdaten eine große Hürde. Hier versprechen algorithmisch erzeugte synthetische Daten Abhilfe. Spätestens 2030 werden 60 % aller KI-Trainingsdaten synthetische Daten sein, prognostizierte Gartner im vorigen Jahr [1]. „Big Data 2.0" wird einen neuen großen Schub an Daten erzeugen.

Dieser Trend lässt sich festmachen an steigenden Publikationszahlen, der Identifikation von Use Cases in vielen Branchen [2] und einer wachsenden Start-up-Szene, vor allem in den USA. Die in [3] veröffentlichte Karte unterscheidet drei Gruppen von Unternehmen, je nachdem, ob sie unstrukturierte, strukturierte oder anonymisierte strukturierte Daten erzeugen (**Abb. 1**).

Bei den unstrukturierten Daten bildet das Bildverstehen den Schwerpunkt. Laut einer Umfrage unter 300 Fachleuten für Computer-Sehen (Computer Vision) haben 96 % schon synthetische Daten für das Modelltraining verwendet [4]. Einen zweiten großen Einsatzbereich für synthetische Daten bilden Fälle, in denen man aufgrund der Datenschutzgesetze oder fehlender vertraglicher Grundlagen die eigentlichen Daten nicht direkt nutzen oder weitergeben darf. Hier handelt es sich meist um strukturierte Daten.

Die Anforderungen und die bisherige Herangehensweise in den beiden Haupteinsatzgebieten sind so unterschiedlich, dass wir sie auch in diesem Beitrag getrennt betrachten wollen. Beiden ist gemein, dass sich für sie mit den neuartigen generativen KI-Modellen ein neuer Ansatz bietet: KI-Modelle sind lange Zeit für analytische Aufgaben wie Klassifikation, Vorhersage und Mustererkennung entwickelt worden. Doch seit einigen Jahren entsteht eine synthetische KI, deren Modelle Text, Bilder, Musik oder multimediale Inhalte generieren. Mit diesen Modellen wurden Kunst und Deepfakes geschaffen. Wie man sie einsetzen könnte, um Trainingsdaten für andere KI-Modelle zu erzeugen, wollen wir im Folgenden herausarbeiten.

Synthetische Daten-Proxys ohne Personenbezug

Die europäische Datenschutzgrundverordnung (DSGVO) schränkt die Nutzung personenbeziehbarer Daten bei fehlenden vertraglichen Grundlagen stark ein. Das betrifft das Gesundheitswesen und die Pharmabranche, die Finanz- und

Dr. Dirk Hecker (✉)
ist stellvertretender Institutsleiter des Fraunhofer IAIS. Er studierte und promovierte im Bereich Geoinformatik an den Universitäten Köln und Bonn. Seit vielen Jahren beschäftigt sich Dirk Hecker mit Aspekten der Digitalisierung und deren Einfluss auf Wirtschaft und Gesellschaft.
dirk.hecker@iais.fraunhofer.de

Dr. Angi Voss
beschäftigt sich im Fraunhofer IAIS mit Big Data, künstlicher Intelligenz und Datenwissenschaft. Sie hat das Schulungsprogramm „Data Science" der Fraunhofer Allianz „Big Data und KI" mitgestaltet, die Kompetenzplattform KI.NRW mit ins Leben gerufen und arbeitete mit an den Grundzügen des Prüfkatalogs für vertrauenswürdige KI. Mit Wissensrepräsentation und Knowledge Engineering stieg sie 1986 in die KI ein.

Dr. Gerhard Paaß
ist Diplom-Mathematiker mit Promotion in Makroökonomie an der Universität Bonn. Seine Forschungsschwerpunkte liegen im Bereich der neuronalen Netze, des Textmining und der Statistik. Als Wissenschaftler in der Gesellschaft für Mathematik und Datenverarbeitung und Gruppenleiter im Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS war er an zahlreichen Forschungs- und Industrieprojekten zum maschinellen Lernen beteiligt.

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Schloss Birlinghoven,
Sankt Augustin, Deutschland

Dr. Tim Wirtz
ist Physiker mit Promotion an der Universität Duisburg-Essen. Schon während seiner Promotion hat er sich mit komplexen Zusammenhängen in Datensätzen und deren Auswirkungen auf die Modellierung auseinandergesetzt. Seit 2015 arbeitet er am Fraunhofer IAIS als Senior Data Scientist, seit 2022 auch als Abteilungsleiter mit seinen Teammitgliedern an den Themen des Machine und Deep Learning mit Anwendungen in diversen Branchen.

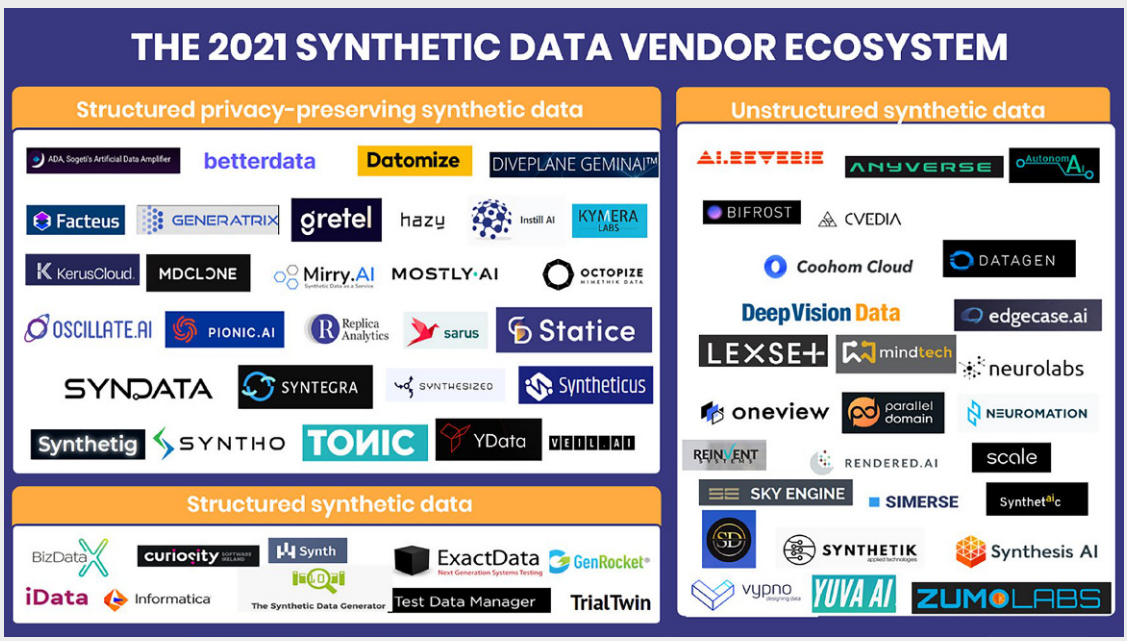
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Schloss Birlinghoven,
Sankt Augustin, Deutschland

Versicherungsbranche, die öffentliche Verwaltung, aber auch alle Funktionsbereiche in Unternehmen, die Kundendaten verarbeiten und mit KI effizienter oder kundenfreundlicher werden oder ihr Personal besser unterstützten möchten.

Unternehmen erkennen oft erst mit der Zeit, welche Anwendungsmöglichkeiten in „ihren" Daten stecken. Nachträglich das Einverständnis der Betroffenen einzuholen kann schwierig sein. Einen Ausweg bietet die Anonymisierung der Daten. Das Problem mit Anonymisierungsmethoden wie „differential privacy" oder „k-anonymity" besteht aber darin, dass sie so viele Daten durch Zufallswerte ersetzen (Randomisierung) oder sie so stark generalisieren, dass die geplanten Analysen oder KI-Modelle zu ungenau werden und keine nützlichen Erkenntnisse oder Ergebnisse liefern.

Maschinelle Lernverfahren wie Klassifikation und Regression können in vorher pseudonymisierten Datenbeständen die Verteilung eines einzelnen Merkmals lernen und dessen Werte passend neu generieren. Aber könnte man nicht darüber hinaus die Gesamtverteilung lernen und eine statistisch ähnlich verteilte neue Datenmenge generieren? Solche anonymisierten Daten-Proxys hätten keine Personenbezüge und wären dem Original sehr viel ähnlicher als bei einer klassischen Anonymisierung. Man könnte daraus die gleichen sta-

Abb. 1 Anbieter von synthetischen Daten. (Mit Genehmigung aus [3])



Kernthesen

- Die Mehrzahl der Trainingsdaten für KI-Projekte wird demnächst synthetisch generiert.
- Die Hauptanwendungsbereiche sind der Ersatz personenbeziehbarer Daten und die Generierung von Trainingsdaten für das Computer-Sehen.
- Beim Computer-Sehen können sich generative KI-Modelle und 3-D-Simulationsumgebungen gut ergänzen.

tistischen Schlüsse ziehen wie aus dem Original und sie für Business Intelligence, Data Mining, maschinelles Lernen oder What-if-Simulationen heranziehen. Datenschutzbehörden wie die europäische EDPS [5] oder die norwegische NDPA fördern und empfehlen den Ansatz [6].

Anonyme synthetische Daten-Proxys haben viele Einsatzmöglichkeiten: Man kann sie schlichtweg verkaufen oder an Partner weitergeben, die damit Software, KI-Anwendungen, digitale Zwillinge des Unternehmens oder neue digitale Geschäftsmodelle entwickeln. Man kann sie als Benchmark für den Vergleich verschiedener Softwareangebote herausgeben. Man kann sie für die Schulung von angehendem Personal nutzen. Man kann sie vorsorglich für zukünftige Ideen speichern, bevor die Speicherfrist der Originaldaten abläuft. Oder man kann sie für die Forschung freigeben, wie dies im Fall des Krebsdatenregisters des britischen Gesundheitsdienstes und des niederländischen Krebszentrums geschehen ist [7].

Synthetische Daten für das Computer-Sehen

Zu viele KI-Projekte scheitern an fehlenden Daten. In der bereits erwähnten Umfrage [4] unter Fachleuten für Computer-Sehen haben das schon 99 % der Befragten erlebt. Derzeit ist die Nachfrage nach besseren Daten besonders ausgeprägt in der Robotik für Manipulation und Orientierung, bei autonomen Fahrzeugen für Orientierung und Fortbewegung, in der Überwachung von Arealen zwecks Vorbeugung oder Rekonstruktion von Geschehnissen, in der bildgebenden Medizin zur Diagnose, für die Qualitätsbeurteilung in der Fertigung, für Schadensprognose an Bauwerken und Schadensbewertung in Versicherungsfällen. In Zukunft wird man Bilddaten auch benötigen, um künstliche Agenten im Metaverse zu trainieren.

In vielen Einsatzgebieten ist es schlichtweg unmöglich, genug Daten durch Sammeln oder Messen zu gewinnen. Für die Bilderkennung im autonomen Fahrzeug müsste man verschiedenste Umgebungen unter diversen Licht- und Wetterverhältnissen und mit immens vielen Kombinationen von Verkehrsteilnehmern und Hindernissen und diversen Kameratypen filmen.

Nicht nur, dass man Daten mit einer größeren Variationsbreite benötigt, an eine vertrauenswürdige KI-Anwendung werden weitere Anforderungen gestellt. Ethikkommissionen in vielen Staaten und großen Unternehmen haben dafür Richtlinien aufgestellt. Je nach Einsatzzweck und Risiken entstehen verschiedene Anforderungen, sei es an Fairness (unterrepräsentierte Personengruppen?), Korrektheit auch in seltenen, aber kritischen Fällen (Unfälle, Fertigungsschäden, seltene Krankheiten) oder Robustheit (gegenüber Angriffen, Rauschen, im Randbereich des Einsatzgebietes). Fast immer sind ergänzende Daten Teil der Lösung, oder sie werden zur Validierung der Anforderungen benötigt [8].

Beseitigt man Defizite von KI-Modellen durch gezielte Erweiterung der Trainingsdaten, dann ändert man damit auch die statistischen Verteilungen in Bezug auf die Ausgangsdaten. Es entsteht bewusst kein statistischer Zwilling – denn der hätte ja immer noch dieselben Defizite.

Möchte man Roboter oder Fahrzeuge interaktiv durch Reinforcement-Lernen trainieren, kommt man schon heute kaum an einfach kontrollierbaren, effizienten Generatoren mit Feedbackkanal vorbei [9]. Aber auch für überwachte und unüberwachte Lernverfahren oder andere Aufgaben des Computer-Sehens kann man mit 3D-Simulationsumgebungen und Werkzeugen der 3D-Computeranimation Trainingsdaten erzeugen. Gut geeignet sind „Game Engines“ wie Unreal und Unity 3D, mit denen man beindruckende VR-Welten, mobile Spiele und Animationsfilme schaffen kann.

In derartigen 3D-Simulatoren können Objekte extern gesteuert werden, während sich das Verhalten der anderen Objekte programmieren lässt. Alle wichtigen Parameter zum Berechnen der Grafik (Rendering) wie Lichtverhältnisse, Texturen, Kameraposition lassen sich genau einstellen. Der Simulator liefert zusätzlich genaue und objektive Annotationen der Objekte und Entfernungen.

Leider ist es auch aufwendig, sehr realistische Simulationen zu erzeugen. Ein KI-Modell, das auf weniger realistisch simulierten Daten trainiert wird, könnte sich durch diese „Realitätslücke“ irritieren lassen [9]. Gibt es also für die Generierung synthetischer Bild- und Videodaten eine Alternative zu Simulatoren oder eine Möglichkeit, die Realitätslücke in nicht fotorealistischen Simulationen zu schließen?

Handlungsempfehlungen

- Europa braucht eine eigene Start-up-Szene mit spannenden Geschäftsmodellen für synthetische Datum, um diesen Trend mitzugestalten.
- Data Scientists sollten sich in Zukunft intensiv mit generativen Modellen auseinandersetzen.

Synthetische KI zur Datensynthese?

Die spektakulären Fortschritte der modernen KI wurden durch tiefe künstliche neuronale Netze erzielt (Deep Learning), die besonders gut mit unstrukturierten Daten – Sprache und Text, Bild und Video – umgehen können. Standen zunächst analytische Aufgaben für das Sprach-, Text- und Bildverstehen im Vordergrund, rücken inzwischen kreative Aufgaben ins Rampenlicht [10]. Im Jahr 2018 wurde das erste KI-generierte Bild für 380.000 € versteigert und 2020 veröffentlichte der Guardian einen von einer KI verfassten Artikel. Deepfakes und stilistisch transformierte Fotos kennen mittlerweile fast alle, die ein Smartphone haben.

Für die Bilderzeugung – und Transformation – werden oft generative adversarielle Netze (GANs) und „Variational Autoencoder" (VAEs) eingesetzt. Sie können lernen, Bilder stilistisch zu transformieren, die Perspektive zu ändern, Objekte zu animieren und Lücken auszufüllen [11]. In der Medizin sind Bilddaten aus GANs von Fachleuten kaum noch von echten zu unterscheiden – sie bestehen den „visuellen Turing-Test" [12]. Die GANs werden nicht nur für Bilder, sondern auch zur Generierung von synthetischen Tabellendaten genutzt [13]. So lassen sich also die oben gesuchten anonymisierten Daten-Proxys erzeugen.

Den Sprachtechnologien gelang der Durchbruch mit gewaltigen Sprachmodellen wie BERT von Google, USA, und GPT-3 von OpenAI, ebenfalls USA, [14]. Sie und ihre Nachfolger wurden auf riesigen Textmengen aus dem Internet und weiteren Textkorpora trainiert. Als Trainingsansatz hat es sich etabliert, in jedem vorgelegten Textstück das nächste Wort oder irgendein ausgelassenes Wort prognostizieren zu lassen. Da ja die fehlenden Worte in dem vollständigen Text stehen, benötigt man keine zusätzlichen Annotationen. Die KI-Modelle lernen „selbstüberwacht". Für spezielle Anwendungen können sie nachtrainiert werden, oder noch einfacher, man gibt ihnen wenige Beispiele und sie ergänzen das nächste Beispiel analog (Few-Shot Learning) [10].

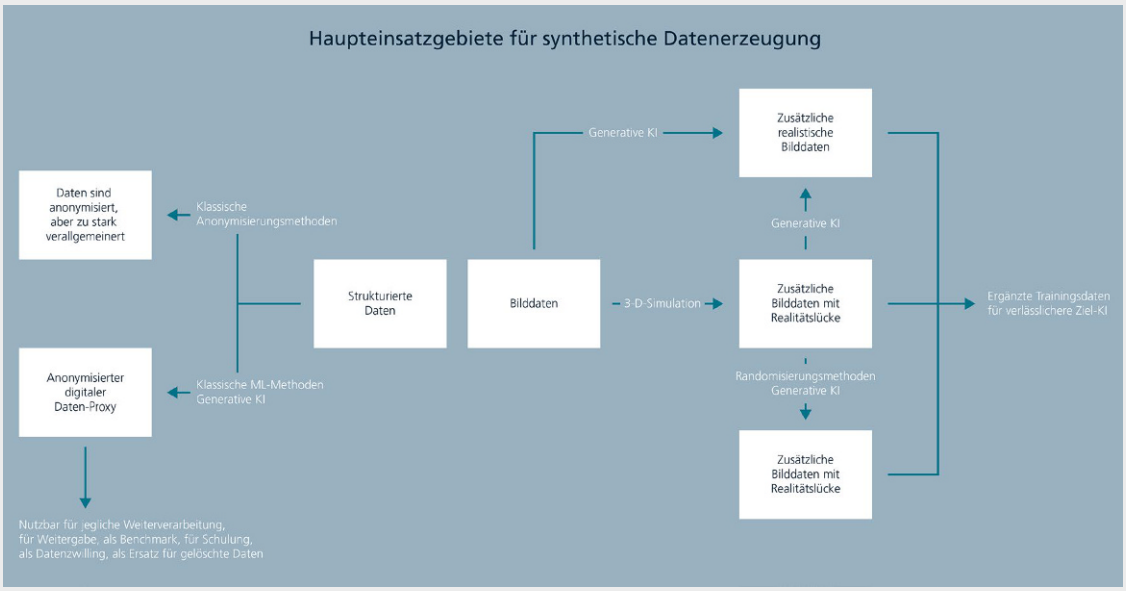
Man könnte ein Modell konstruieren, welches Restaurantkritiken als gut, mittel, oder schlecht klassifizieren soll. Dann kann man einem großen Sprachmodell wenige Beispiele von Restaurantkritiken mit der zugehörigen Klasse und der Aufgabenstellung „Restaurantkritik erzeugen" geben (few shots). Als Antwort kann das Modell Tausende unterschiedlicher Restaurantkritiken zu diesen Klassen produzieren. Nach diesem Schema lassen sich Trainingsdaten für vielfältige Textanalysen generieren.

Große Sprachmodelle lassen sich auch auf andere Medientypen verallgemeinern. Worte werden zerlegt in Tokens (Wortteile) und für jedes Token werden Embeddingvektoren berechnet. Ein Bild wird zerlegt in Image Patches (z. B. 16×16 Pixel), für die auch Embeddingvektoren berechnet werden. Dies geschieht mit dem Algorithmus von BERT, d. h. die Embeddings sind kontextsensitiv und berücksichtigen die Inhalte der anderen Tokens und Image Patches. Damit wird gelernt, welche Image Patches in Bildern zusammen auftauchen und wie sie mit Text-Tokens assoziiert sind. Bei der Anwendung wird jeweils das nächste Token bzw. Image Patch zu den bisher gegebenen generiert.

Man kann damit zu bestehenden Teilbildern neue Image Patches erzeugen, oder zu einem Text das zugehörige Bild bzw. zu einem Bild die Textbeschreibung. DALL E-2 von OpenAI [15] und Imagen von Google [16] sind große Sprachmodelle, die auf diese Weise Bilder zu Texten erzeugen. Zusätzlich verwenden sie Diffusionsmodelle, um Bilder kleiner Auflösung in nahezu fotorealistische Bilder größerer Auflösung zu transformieren. Durch den Eingabetext kann man Objekte in beliebige Umgebungen platzieren, zum Beispiel „A photo of an astronaut riding a horse" und hiermit Trainingsdaten für die Erkennung von „Astronaut" produzieren. Mitte 2022 waren die zugehörigen Modelle aber noch nicht öffentlich verfügbar.

3D-Simulatoren und KI-Modelle für die Bilddatensynthese können sich ergänzen. Mit Diffusionsmodellen kann man gezielt fehlende Beispiele erzeugen, um die KI-Anwendung fairer, genauer oder robuster, also insgesamt vertrauenswürdiger zu machen. Bei Simulationen könnte man zur Vereinfachung die Texturen weglassen und die entstehende Realitätslücke von einem GAN-Modell schließen lassen. Man spricht hier auch von „Domänenadaptierung" durch Transferlernen. Bessere Lernergebnisse lassen sich erzielen, wenn man synthetische und Originaldaten zum Training des Zielmodells mischt (Datenaugmentierung) [9], ein Ansatz, der zur Bilderkennung im großen Stil verwendet wird.

Abb. 2 Haupteinsatzgebiete und Ansätze zur synthetischen Datenerzeugung



„Domänenrandomisierung" ist eine andere Möglichkeit, die Realitätslücke zu schließen. Hier verändert man zufällig in einer Szene möglichst viele nicht wesentliche Aspekte. Man tauscht Texturen oder ganze Objekte aus oder fügt neue Objekte in die Szene ein. Auch dazu kann man GANs und Diffusionsmodelle einsetzen. Die Ziel-KI behandelt die Zufälle in der Realität dann hoffentlich wie die anderen zufälligen Variationen, mit denen sie trainiert wurde, nämlich als unwesentlich [9].

Die besprochenen Möglichkeiten zur Datensynthese und die identifizierten Einsatzmöglichkeiten synthetischer KI fasst die **Abb. 2** zusammen.

Fazit

Wir haben in den vergangenen Jahren enorme Fortschritte im maschinellen Lernen gesehen. Viele KI-Anwendungen haben es inzwischen in die operative Anwendung, in unseren täglichen Alltag geschafft. Es gibt aber auch zahlreiche Anwendungsideen, die aufgrund unzureichender Daten, wegen einer zu teuren Datenerfassung oder aus Datenschutzgründen scheitern. Synthetisch erzeugte Daten können hier Abhilfe schaffen und der KI-Entwicklung weiteren Schub geben. Dazu haben wir in diesem Artikel zwei Einsatzbereiche beleuchtet, die auch im Fokus einer regen Start-up-Szene stehen.

Die zuletzt besprochenen Bildgeneratoren sind nur der Anfang weiterer, noch mächtigerer multimodaler Modelle wie im Mai 2022 Flamingo von DeepMind, die auch Videos verstehen oder produzieren oder sogar Geräte und digitale Agenten steuern können. Diese neuen, vielseitigen Modelle sind umso spannender, als man sie bereits durch Angabe weniger Beispiele benutzen kann. Jetzt ist also ein guter Zeitpunkt für deutsche und europäische Forschungseinrichtungen und Unternehmen, sich intensiv mit den Möglichkeiten großer KI-Modelle zur Datensynthese zu befassen.

Funding. Open Access funding enabled and organized by Projekt DEAL.

Literatur

[1] https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/. Zugegriffen: 12. Juni 2022

[2] El Emam, K. (2020). *Accelerating AI with synthetic data*. O'Reilly Media, Inc. https://www.oreilly.com/library/view/accelerating-ai-with/9781492045991/

[3] Devaux, E. (2021). List of synthetic data startups and companies—2021. https://elise-deux.medium.com/the-list-of-synthetic-data-companies-2021-5aa246265b42. Zugegriffen: 12. Juni 2022

[4] Datagen (2022). Synthetic data: key to production-ready AI in 2022. https://datagen.tech/ai/synthetic-data-key-to-production-ready-ai-in-2022/. Zugegriffen: 12. Juni 2022

[5] Zerdick, T. (2021). Is the future of privacy synthetic? https://edps.europa.eu/press-publications/press-news/blog/future-privacy-synthetic_en. Zugegriffen: 12. Juni 2022

[6] Hann, T. (2021). The Executive's Guide to Accelerating Artificial Intelligence and Data Innovation with Synthetic Data. https://hbr.org/sponsored/2021/09/the-executives-guide-to-accelerating-artificial-intelligence-and-data-innovation-with-synthetic-data. Zugegriffen: 12. Juni 2022

[7] James, S., Harbron, C., Branson, J., & Sundler, M. (2021). *Synthetic data use: exploring use cases to optimise data utility*. Springer. https://link.springer.com/article/10.1007/s44163-021-00016-y

[8] Poretschkin, M., et al. (2022). Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html. Zugegriffen: 12. Juni 2022

[9] Nikolenko, S. (2021). *Synthetic Data for Deep Learning*. Springer. https://link.springer.com/book/10.1007/978-3-030-75178-4

[10] Paaß, G., & Hecker, D. (2020). *Künstliche Intelligenz – Was steckt hinter der Technologie der Zukunft?* Springer. https://link.springer.com/book/10.1007/978-3-658-30211-5

[11] Jabbar, A., Li, X., & Omar, B. (2021). A survey on generative Adversarial networks: variants, applications, and training. *ACM Computing Surveys*. https://doi.org/10.1145/3463475.

[12] Chen, R., et al. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature*. https://doi.org/10.1038/s41551-021-00751-8.

[13] Zhao et al. 2021: CTAB-GAN: Effective Table Data Synthesizing

[14] Themath, C. (2021). Moderne Sprachtechnologien – Konzepte, Anwendungen, Chancen. https://www.ki.nrw/studie-moderne-sprachtechnologien/#download-studie. Zugegriffen: 12. Juni 2022

[15] Ramesh, A., et al. (2022). Hierarchical text-conditional image generation with CLIP latents, openAI. https://deepai.org/publication/hierarchical-text-conditional-image-generation-with-clip-latents. Zugegriffen: 12. Juni 2022

[16] Saharia, C., et al. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. https://imagen.research.google/paper.pdf. Zugegriffen: 12. Juni 2022