Mai 2018 · Dr. Jörg Drechsler und Dr. Nicola Jentzsch

Synthetische Daten

Innovationspotential und gesellschaftliche Herausforderungen

Stiftung
Neue
Verantwortung

Think Tank für die Gesellschaft im technologischen Wandel

Dr. Jörg Drechsler und Dr. Nicola Jentzsch
Mai 2018
Synthetische Daten:
Innovationspotential und gesellschaftliche Herausforderungen

Executive Summary

Die Wettbewerbsfähigkeit von Unternehmen hängt zunehmend davon ab, aus Daten neue Produkte, Dienstleistungen oder Geschäftsmodelle zu entwickeln. Datenbasierte Innovation entscheidet auch in öffentlichen Verwaltungen über eine effizientere Ressourcenallokation und damit einhergehenden Kosteneinsparungen. Viele Fälle der Datenverarbeitung und -weitergabe werden von Bürger:innen nur akzeptiert, wenn dafür anonymisierte Daten verwendet werden. Künftig müssen deshalb verstärkt Verfahren gefunden werden, mit welchen personenbeziehbare Daten anonymisiert werden können, aber der Nutzen der Daten für Analysen erhalten bleibt.Dieser Zielkonflikt wird für Staat und Unternehmen aufgrund der exponentiellen Zunahme von Daten (Big Data) eine immer wichtigere Herausforderung.

Diese Herausforderung betrifft aktuell jene Unternehmen, für die Forschungskooperationen, unternehmensinterne und -externe Datenpools oder grenzüberschreitende Datentransfers eine immer größere Rolle spielen.Bisher wurden personenbeziehbare Daten hierfür beispielsweise durch Verrauschung (Zufügung von Zufallswerten) oder Vergröberung verändert. Beides reduziert die Datenqualität.Bei strikter Anwendung der Verfahren sind die entstehenden Daten für viele Anwendungen nicht mehr zu gebrauchen. Andere Methoden, wie die Synthetisierung von Daten, wurden bislang als zu arbeitsaufwändig angesehen. Aufgrund des zunehmenden Einsatzes des maschinellen Lernens sowie der steigenden Rechnerkapazitäten ändert sich dies nun.

Bei der Daten-Synthetisierung handelt es sich um eine Methode, mit der eine „künstliche“ Repräsentation eines Originaldatensatzes erstellt werden kann. Hierzu wird ein Modell entwickelt, das die Originaldaten so gut wie möglich erklärt. Aus diesem Modell werden neue Daten generiert, die wichtige statistische Eigenschaften des Originaldatensatzes erhalten. Der synthetische Datensatz besteht nicht aus Daten natürlicher Personen, sondern aus Daten synthetischer Einheiten. Je nach Anwendung kann die Daten-Synthetisierung mit mathematischen Garantien der Privatheit kombiniert werden.

Diese Methode ist bereits bei Behörden und Instituten mehrerer Länder im Einsatz und wird dazu benutzt, Mikrodatensätze, also Datensätze mit Daten, die auf Individualebene beobachtet werden, zu anonymisieren. Zu den Anwendern gehört unter anderem das U.S. Census Bureau und das deutsche Institut für Arbeitsmarktforschung. So hat das U.S. Census Bureau Mobilitätsströme von Berufspendlern synthetisiert, um diese an Forscher:innen

weitergeben zu können. Synthetische Datensätze werden beispielsweise auch in der Privatwirtschaft von Finanzdienstleistern benutzt, um mit Forscher:innen in der Betrugserkennung zusammenzuarbeiten. In der Forschung zu synthetischen Daten sind mittlerweile rasche Fortschritte zu beobachten.

Überraschenderweise lassen sich quasi alle Datenarten synthetisieren, darunter auch Bild- und Textdateien oder soziale Graphen. Die Daten können in jeglichen Volumina produziert werden. Ihre Qualität ist messbar und kann mit der des Originaldatensatzes verglichen werden. Zusammenhänge, Cluster oder andere Muster aus dem Originaldatensatz bleiben erhalten.

Synthetische Daten gelten dann als anonym, wenn keine Personenbeziehbarkeit besteht. In diesem Fall finden die Regelungen der Datenschutz-Grundverordnung (DS-GVO) keine Anwendung. Zusammenführung von Daten oder zweckunbestimmtes Lernen von Zusammenhängen wären dann möglich. Dies wirft rechtliche und ethische Fragen auf: Wichtige Datenschutz-Prinzipien könnten durch Nutzung synthetischer Daten unterlaufen werden. Unternehmen müssten in Datenstrategien festlegen, welche Auswertungen synthetischer Daten sie für ethisch vertretbar halten und welche nicht.

Neben diesen Aspekten stellen sich auch Fragen der Informationssicherheit. So könnten synthetische Daten, die kaum von echten zu unterscheiden sind, dazu eingesetzt werden, Produkte oder Prozesse anzugreifen. Bereits heute können beispielsweise synthetische Bilder auf Basis realer Bilder von Personen im Internet erzeugt werden. Diese könnten in Zukunft benutzt werden, um etwa Authentifizierungsverfahren wie die Gesichtserkennung auszuhebeln.

Grundsätzlich gilt, dass jede Form der Datennutzung ein inhärentes Risiko trägt. Dateninnovation wird nur möglich sein, wenn wir robuste Verfahren finden, dieses Risiko zu reduzieren. Daten-Synthetisierung könnte ein solches Verfahren sein, das dann zur vollen Entfaltung kommt, wenn die wichtigsten rechtlichen und ethischen Fragen geklärt sind.

Inhalt

Executive Summary 2

II Grundlagen der Produktion synthetischer Daten 7
2.1 Ursprünge der Daten-Synthetisierung 7
2.2 Skalierung von Synthetisierung in Big-Data-Umgebungen 11

III Einsatz und praktische Anwendungsfälle 14
3.1 SIPP Synthetic Beta File: Einkommens- und Transferdaten 14
3.2 OnTheMap: Synthetisierung von Mobilitätsströmen 15
3.3 Synthetisierung von Betriebsdaten 16

IV Grenzen der Synthetisierungsmethode 17

V Gesellschaftliche Herausforderungen synthetischer Daten 18
5.1 Rechtliche Herausforderungen synthetischer Daten 18
5.2 Ethische Herausforderungen synthetischer Daten 20
5.3 Weitere potentielle Risiken synthetischer Daten 20

VI Schluss 21

Impressum 27

1. Einleitung

Datenbasierte Innovation entscheidet in der Datenökonomie über die Wettbewerbsfähigkeit von Unternehmen. Die Erschließung von persönlichen Daten birgt enormes technologisches und soziales Innovationspotential. Daher müssen Unternehmen praktikable Wege finden, Datenschutz bei gleichzeitiger Maximierung des Datennutzens zu implementieren. Aufgrund der exponentiellen Zunahme von Massendaten *(Big Data)* ist dieser Zielkonflikt eine zentrale Herausforderung für Unternehmen und Staat.

Auch öffentliche Verwaltungen stehen zunehmend vor dem Problem, dass sie in Open-Data-Initiativen Daten publizieren sollen. Gleichzeitig sollen sie Datensubjekten ein hohes Schutzniveau garantieren.

Dieser Zielkonflikt treibt seit Jahren Forschungsaktivitäten in Informatik und Statistik an. Beide Disziplinen suchen seit geraumer Zeit nach Methoden, die das Risiko der Reidentifizierung einzelner Personen im Datensatz senken. Die Verrauschung, Vergröberung oder Vertauschung von Daten reduziert aber deren Qualität. Manche Privatheitsgarantien (zum Beispiel strikte Anwendung von *Differential Privacy*) sind deshalb in der Praxis nur von eingeschränkter Nützlichkeit.

Andere, vielversprechende Ansätze von Statistikbehörden wurden bislang außerhalb dieser Behörden als zu rechenintensiv oder zu kompliziert angesehen. Doch die Skalierung von Rechnerkapazität und die Parallelisierung von *Machine-Learning*-Verfahren in Big-Data-Architekturen machen jetzt den Weg für diese Verfahren frei.

In diesem Impuls soll die Synthetisierung von Daten einer breiten Öffentlichkeit vorgestellt werden. Synthetische Daten werden auch als „Surrogatdaten“ oder „artifizielle Daten“ bezeichnet. Die Synthetisierung ist ein Verfahren, mit welchem Originaldaten (zum Beispiel sensible personenbezogene Daten) in eine synthetische Repräsentation überführt werden. Je nachdem, ob die Synthetisierung Teilmengen oder alle Daten des Originaldatensatzes umfasst, kann das Risiko der Reidentifizierung von Datensubjekten erheblich gesenkt werden. Ein voll synthetisierter Datensatz besteht aus Daten *synthetischer Subjekte* und nicht *realer Personen.* Diese Daten wurden nicht durch direkte Messung erhoben, sondern sind das Resultat eines algorith-

mischen Mechanismus. Dieser bezieht sich allerdings auf Daten, die mit direkter Messung erhoben wurden.

Diese größtenteils nur in Fachkreisen bekannte Methode wurde schon vor mehr als 25 Jahren erfunden. Sie basiert ursprünglich auf der multiplen Imputation, einem statistischen Verfahren zur Ersetzung fehlender Werte in Datensätzen, bei welchem diese durch mehrere plausible Werte ersetzt werden. Die Erzeugung mehrerer plausibler Werte erlaubt es, die Unsicherheit bei der Schätzung der fehlenden Werte korrekt zu berücksichtigen.

Heute ist die Synthetisierung bei verschiedenen Statistikbehörden und Institutionen im Praxiseinsatz, einige Beispiele werden hier vorgestellt. In den USA veröffentlicht das U.S. Census Bureau synthetisierte Ströme von Berufspendlern (OnTheMap) oder Einkommensdaten und Transferleistungen (SIPP Synthetic Beta). In Schottland erhalten Forscher:innen, die mit den Daten der *Scotish Longitudinal Study* arbeiten wollen, synthetische Daten für ihre Analysen. In Deutschland ist das Institut für Arbeitsmarktforschung (IAB) der Vorreiter in der Synthetisierung. Dort wurde das IAB-Betriebspanel für Forschungszwecke synthetisiert. Mittlerweile arbeiten verschiedene Statistikbehörden in Kanada, England, Australien und Neuseeland an synthetischen Datensätzen.

Überraschenderweise lassen sich quasi alle Datenarten synthetisieren, darunter auch Bild- oder Text-Dateien oder soziale Graphen. Die Datensätze können in sehr großem Umfang zum jeweils gewünschten Präzisionsgrad hergestellt werden. Die Zusammenhänge oder Cluster im Originaldatensatz bleiben bei diesem Verfahren erhalten – ein großer Vorteil des Verfahrens.

Die Qualität der synthetischen Daten ist mess- und damit vergleichbar. Qualität und Umfang solcher Daten können auf den Abnehmer angepasst werden. Kollaborationspartner können die Codes zur Synthetisierung austauschen. Dies ermöglicht eine Einschätzung, ob aussagekräftige Ergebnisse auf Basis der synthetischen Daten zu erwarten sind. Gütekriterien von Modellen, die parallel auf Originaldaten und synthetischen Daten laufen, können ebenfalls verglichen werden. Daten-Synthetisierung kann außerdem mit mathematischen Garantien der Privatheit kombiniert werden.

Unternehmen könnten Synthetisierung in Testumgebungen oder für den internen und externen Datenaustausch einsetzen. Letzteres betrifft unternehmensübergreifende Forschungskollaborationen und Datenpools. Für die öffentliche Verwaltung sind diese Verfahren als Grundlage der Bereitstellung von Mikrodatensätzen interessant. Grundsätzlich könnte die Daten-Synthe-

tisierung auch länderübergreifende Kooperationen und Datenweitergabe erlauben, trotz unterschiedlicher Datenschutzregime.

Obwohl weltweit zunehmend an synthetischen Verfahren geforscht wird, gibt es eine Reihe von Herausforderungen, die wir kurz erläutern wollen. Eine Herausforderung ist die Effizienz im praktischen Einsatz. Wir beschreiben, welche Lösungsansätze hierzu aktuell entwickelt werden und wo derzeit Erkenntnisgrenzen liegen.

Der Einsatz synthetischer Daten wirft eine Reihe rechtlicher, ethischer und technischer Fragen auf. Wir sehen die hier genannten Punkte lediglich als erste Diskussionsanstöße. Mit der zunehmenden Verbreitung dieser Methode, insbesondere in der Privatwirtschaft, ist eine breitere gesellschaftliche Diskussion notwendig.

II Grundlagen der Produktion synthetischer Daten

Genaugenommen gibt es die Daten-Synthetisierung in mehreren Varianten. Diese wollen wir im Folgenden vorstellen. Hierzu gehen wir zunächst auf parametrische Verfahren ein[1] und daran anschließend auf neuere, nicht-parametrische Verfahren, die nicht von einer bestimmten Verteilung der Variablen ausgehen.[2]

2.1 Ursprünge der Daten-Synthetisierung

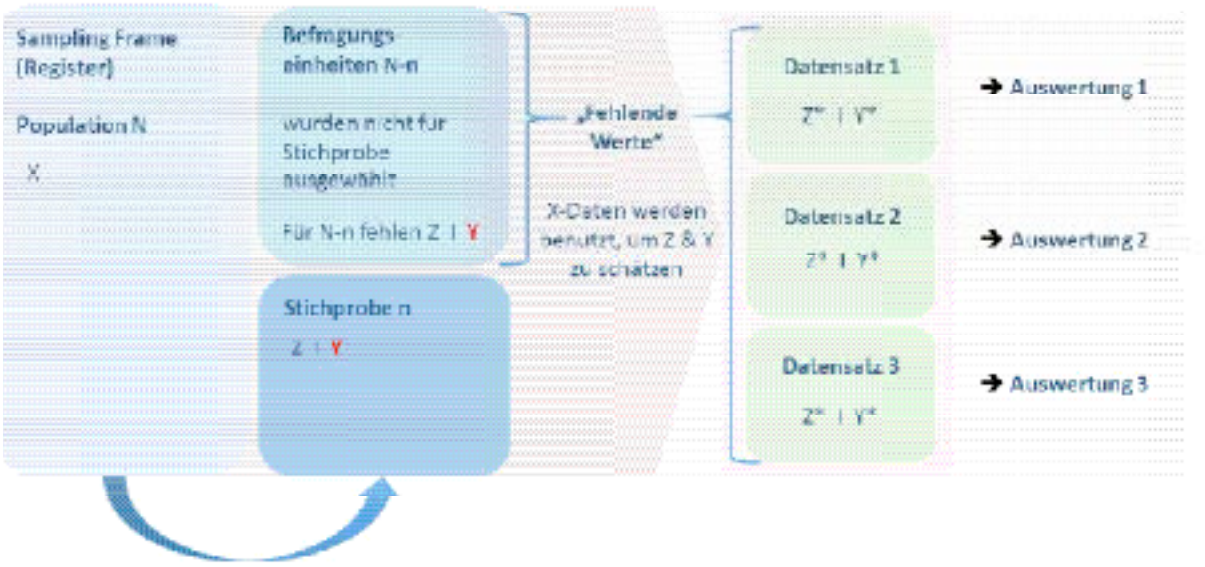
Die Datenschutzproblematik bei der Weitergabe von Befragungsdaten war schon 1993 zentraler Ausgangspunkt für die Arbeiten von Rubin.[3] Auf ihn geht die Idee der Erstellung synthetischer Daten zurück. Die Originaldaten werden bei diesem Verfahren durch mehrere künstliche Werte ersetzt, die möglichst ähnliche Verteilungseigenschaften wie die Originaldaten aufweisen.

1 Hier muss die Verteilung von Variablen bekannt sein, um bestimmte statistische Verfahren anwenden zu können.

2 Eine ausführliche Einführung in die Methodik findet sich außerdem bei Drechsler, J. (2011). Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Lecture Notes in Statistics 201, Springer: New York.

3 Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. Journal of Official Statistics 9, 462–468; Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys, John Wiley & Sons: New Jersey.

Schaubild 1. Synthetische Daten nach dem Rubin-Verfahren



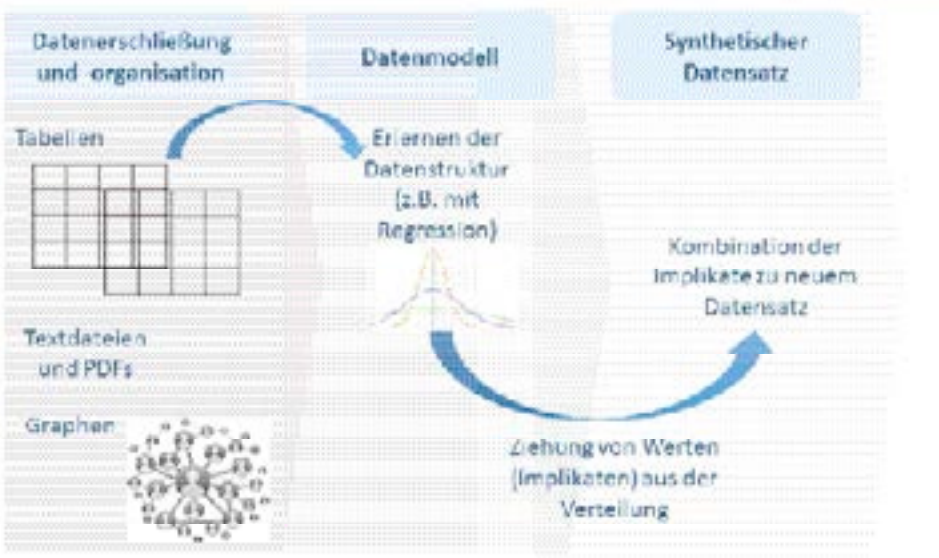
Wie in Schaubild 1 dargestellt, nimmt Rubin an, dass es eine Population gibt, aus der eine Stichprobe gezogen werden kann (sog. Sampling Frame). Für die Population liegen Daten in Form einer Matrix X vor, dies können Adresse, Alter oder ähnliches sein. Nachdem aus dem Frame eine Stichprobe gezogen wurde, werden für diese die Daten Z und Y, zum Beispiel durch Befragung, erhoben. Z könnte das Geschlecht der Befragten sein und Y ihr Einkommen. Letzteres ist ein sensibles Datum (hier in Rot und fett gesetzt). Rubin schlug vor, die Daten der Befragungseinheiten, die *nicht* für die Befragung ausgewählt wurden (N-n), einfach als fehlende Werte zu betrachten. Für sie wurden die Zusatzinformationen Z und Y nicht erhoben.

Nun können auf Basis des X Datensatzes die Variablen Z und Y gleich mehrfach geschätzt werden. Wie dies gemacht wird, hängt von den Daten ab, beispielsweise kann ein Regressionsmodell verwendet werden, um das Einkommen für die ‚fehlenden' Befragten zu schätzen. So können zum Beispiel drei Datensätze erzeugt werden. Ausgehend von den einzelnen erzeugten synthetischen Datensätzen erhält der Anwender des Verfahrens die fina-

len Ergebnisse, indem er zunächst jeden Datensatz getrennt auswertet und dann die Ergebnisse unter Verwendung sehr einfacher Formeln kombiniert.[4]

Dies ist das ursprünglich von Rubin vorgeschlagene Vorgehen, mit welchem sich statistisch valide Schätzergebnisse erzielen lassen. Ein Vorteil des Verfahrens liegt darin, dass quasi keine Originalwerte in den bereitgestellten Daten enthalten sind.[5] Insbesondere bei der vollständigen Synthetisierung kann von einem hohen Datenschutzniveau ausgegangen werden. Da es verschiedenen Verfahren gibt, u.a. die teilweise Synthetisierung (siehe unten), haben wir den Prozess nochmals in Schaubild 2 in einer generischen Version dargestellt.

Schaubild 2. Synthetisierung von Daten: Allgemeine Darstellung



Bei der Synthetisierung lassen sich zwei Ansätze unterscheiden:

- Vollständige Synthetisierung: Alle Daten werden durch synthetische Werte (sogenannte Implikate) ersetzt
- Teilweise Synthetisierung: Nur eine Teilmenge der Daten wird durch synthetische Werte ersetzt

Zeitgleich mit dem Erscheinen des Rubin-Aufsatzes, wurde auch der Ansatz

4 Zum Vorgehen, siehe auch Raghunathan, T. E., J.P. Reiter und D.B. Rubin (2003). Multiple imputation for statistical disclosure limitation, Journal of Official Statistics 19, 1–16; Reiter, J. P. (2005). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata, Journal of Statistical Planning and Inference 131, 365–377.

5 Dies wäre allerdings extra zu prüfen.

zur teilweisen Synthetisierung von Daten publiziert.[6] Bei diesem Ansatz werden nur sensible Merkmale und/oder Merkmale, die dazu verwendet werden können, einzelne Befragungseinheiten zu reidentifizieren, ersetzt. Der Datenanbieter muss hier selbst entscheiden, welche Daten synthetisiert werden müssen, um einen ausreichenden Datenschutz zu gewährleisten. So kann es in manchen Fällen ausreichend sein, lediglich *einzelne Beobachtungen einer einzelnen Variable* durch synthetische Werte zu ersetzen. Ein Beispiel wären Angaben zum Einkommen für Personen mit besonders hohem Einkommen. In anderen Fällen kann es notwendig sein, sämtliche Beobachtungen mehrerer Variablen zu synthetisieren. Es können auch – analog zum Vorgehen bei vollständig synthetisierten Daten – sämtliche Beobachtungen im Originaldatensatz durch synthetische Werte ersetzt werden.

Da bei diesem Ansatz ein Teil der Daten nicht verändert wird und zudem synthetische Daten nur für die Individuen erzeugt werden, die bereits in den ursprünglichen Daten enthalten waren, ist das Reidentifikationsrisiko bei teilweise synthetischen Daten größer als bei vollständig synthetischen Datensätzen. Dafür ist die Datenqualität in der Regel höher.[7]

Allerdings kann davon ausgegangen werden, dass das Risiko der Reidentifikation trotzdem deutlich unter dem Risiko liegt, das bei Einsatz traditioneller Anonymisierungsverfahren (zum Beispiel Aggregation) verbleibt.[8] Es gibt verschiedenen Möglichkeiten zur Bewertung des verbleibenden Risikos teilweise synthetischer Verfahren. Diese werden ausführlich in der Literatur diskutiert.[9]

Bei der Wahl der Modelle, die zur Erzeugung synthetischer Daten eingesetzt werden, sind den Daten-Analysten kaum Grenzen gesetzt. Letztlich muss das gewählte Modell zwei Eigenschaften erfüllen: es muss die Zusammenhänge zwischen den Variablen (oder andere Aspekte von Interesse) mög-

6 Little, R.J.A. (1993). Statistical analysis of masked data. Journal of Official Statistics 9, 407–426.

7 Drechsler, J., S. Bender und S. Rässler (2008). Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel, Transactions on Data Privacy 1, 105 – 130.

8 Durch Kombination von Synthetisierung mit anderen Verfahren (zum Beispiel Differential Privacy) kann außerdem eine formale und damit prüfbare Garantie von Privatheit gegeben werden. Die Autoren bedanken sich bei Omar Ali Fdal und Mikhail Dyakov für diese Anmerkung.

9 Drechsler, J. und J.P. Reiter (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In J. Domingo-Ferrer and Y. Saygin, eds., Privacy in Statistical Databases, 227–238. Springer, New York; Reiter, J. P. und R. Mitra (2009). Estimating risks of identification disclosure in partially synthetic data. Journal of Privacy and Confidentiality 1, 99–110.

lichst gut abbilden und es muss möglich sein, auf Basis des Modells neue Werte zu erzeugen.

Welche Modelle die besten Ergebnisse liefern hängt davon ab, für welche Auswertungen die synthetischen Daten später genutzt werden sollen. So kamen bei den synthetischen Datensätzen von statistischen Ämtern in der Vergangenheit überwiegend parametrische Modelle zum Einsatz, da derartige Modelle auch von den Nutzern der Daten verwendet werden. Hier sind statistisch valide Ergebnisse von besonderem Interesse.

Neuerdings werden Machine-Learning-Verfahren mit guten Ergebnissen eingesetzt. Sie reduzieren den Arbeitsaufwand erheblich.[10] Diese können auch dann angewandt werden, wenn dem Datenanbieter bzw. -nutzer nicht klar ist, welche Aspekte der Daten von Interesse sind. Bei diesen Methoden handelt es sich um die automatisierte Erkennung von Mustern in Datensätzen. Dazu gehören Klassifikations- und Regressionsmodelle, wie *Classification and Regression Trees* (CART) oder *Support Vector Machines*. Diese sind bezüglich ihrer Eignung für die Daten-Synthetisierung bereits untersucht worden.[11]

Der Vorteil dieser Verfahren liegt darin, dass keine expliziten Verteilungsannahmen zu Beginn der Synthetisierung getroffen werden müssen (sogenannte nicht-parametrische Verfahren). Dies reduziert den Modellierungsaufwand erheblich.

2.2 Skalierung von Synthetisierung in Big-Data-Umgebungen

In Anwendungsfällen, bei denen es weniger darum geht, statistisch valide Rückschlüsse auf eine den Daten zugrunde liegende Gesamtpopulation zu erhalten, sondern vielmehr interessante Zusammenhänge in Daten zu finden oder zu prüfen (*Data Mining*), kann auf eine Vielzahl weiterer, sehr flexibler Verfahren zurückgegriffen werden. Hierbei reicht es oftmals aus, abgewandelte Verfahren der Synthetisierung anzuwenden, die beispielsweise nur

10 Drechsler, J. und J.P. Reiter (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. Computational Statistics & Data Analysis, Vol. 55, 3232–3243. Drechsler, J. (2010). Using Support Vector Machines for Generating Synthetic Datasets. In: J. Domingo-Ferrer & E. Magkos (Ed.), Privacy in Statistical Databases (Lecture Notes in Computer Science 6344), Springer, Berlin, 148–161.

11 Siehe unter anderem Drechsler, J. (2010). Using Support Vector Machines for Generating Synthetic Datasets. In: J. Domingo-Ferrer & E. Magkos (Ed.), Privacy in Statistical Databases (Lecture Notes in Computer Science 6344), Springer: Berlin, 148–161.

einen synthetischen Datensatz erzeugen (nicht mehrere) oder synthetische Werte einfach mitteln.

Synthetisierung kann beispielsweise für Textdateien angewendet werden. Die ‚Stimmung' des Textes (sogenanntes Sentiment) bleibt dabei erhalten.[12] Andere wenden CART an, um Daten aus Biodatenbanken zu synthetisieren.[13] Dritte nutzen mathematische Optimierungsverfahren.[14] Es kann gezeigt werden, dass auch in diesem Fall die Unterschiede zwischen synthetischen Daten und Rohdaten gering sind.[15]

In den meisten dieser Fälle wird die Synthetisierung spezialisiert angewandt (zum Beispiel auf Text). Schon vor Jahren haben Forscher:innen der Universität Konstanz gezeigt, dass auch generelle Synthetisierer entwickelt werden können, die unabhängig von einem Spezialfall verschiedene Arten der synthetischen Daten generieren können.[16] Adä und Berthold erklären außerdem, dass realistische Regeln, die auf einem anderen Datensatz gelernt wurden, der Datengenerierung hinzugefügt werden können.

In Big-Data-Umgebungen reicht ein einzelnes Synthetisierungsmodell aufgrund der Vielfalt der Daten oftmals nicht aus. Es müssen mehrere entwickelt werden, die jeweils auf die unterschiedlichen Datentypen angepasst werden, da so bessere Ergebnisse erzielt werden können.[17] Gleichzeitig kann Daten-Synthetisierung mit einem Verfahren kombiniert werden, das über-

12 Maqsud, U. (2015). Synthetic Text Generation for Sentiment Analysis, Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015): 156–161, Lisboa, Portugal, 17 September, 2015.

13 Kuiper, J., E.R. van den Heuvel und M.A. Swertz (2015). The Hybrid Synthetic Microdata Platform: A Method for Statistical Disclosure Control, Biopreservation and Biobanking 13 (3): 178 – 182.

14 Bogle, B. M. und S. Mehrotra (2016). A Moment Matching Approach for Generating Synthetic Data, Big Data, Vol. 4, No. 3, 160 – 178.

15 Visuell ist dies aufbereitet in Erickson, J. (2017). Creating Synthetic Data with SAS/OR, SAS Blog (17. Mai 2017), https://blogs.sas.com/content/operations/2017/05/17/creating-synthetic-data-sasor/.

16 Adä, I. und M.R. Berthold (2010). The New Iris Data: Modular Data Generators, KDD'10, 25.-28. Juli 2010, Washington, DC, USA.

17 Ming, Z., C. Luo, W. Gao, R. Han, Q. Yang, L. Wang und J. Zhan (2014). BDGS: A Scalable Big Data Generator Suite in Big Data Benchmarking, http://prof.ict.ac.cn/BigDataBench/wp-content/uploads/2013/10/BDGS_BigDataBench.pdf.

prüfbare Garantien der Privatheit liefert. Manche dieser Verfahren werden derzeit rege in der Literatur diskutiert (zum Beispiel *Differential Privacy*).[18]

Ming et al. zeigen anhand verschiedener Rohdatensätze (Wikipedia Eintragungen, Amazon Filmbeurteilungen, Google Web Graph, Facebook Graphen, etc.), wie verschiedene Daten-Volumina mit unterschiedlichen Geschwindigkeiten produziert werden können. Die Geschwindigkeit hängt von der Anzahl der parallel implementierten Datengeneratoren ab. Die Autoren zeigen, dass beispielsweise ein Terabyte Wiki Daten in 4.7 Stunden produziert werden kann. Mit dem Anstieg des Datenvolumens können die Autoren große Datenmengen in linearer Bruttozeit erzeugen, wobei die Exekutionszeit pro Dateneinheit sinkt.[19]

Anderson und Kollegen:innen zeigen, dass realistische synthetische Daten auch für das Internet of Things (IoT) produziert werden könnten.[20] Ihr Synthetisierer kann Terabytes an Daten produzieren, welche die hochkomplexen und verschachtelten Strukturen der Originaldaten widerspiegeln.

Auch Methoden des *Deep Learning* werden zunehmend zur Produktion synthetischer Daten eingesetzt.[21] So können Daten mit sogenannten *Generative Models*, einer Unterklasse der Neuronalen Netzen, produziert werden. Ein Beispiel wird von der OpenAI Initiative vorgestellt. Auf Basis eines Samples von Fotos aus dem Internet generiert ihr Modell neue, synthetische Fotos. Das Netz lernt zunächst die offensichtlichsten Merkmale der Bilder, zum Beispiel dass nebeneinander liegende Pixel oft dieselbe Farbe haben. Um zu realistischen Fotos zu gelangen, tritt das Modell, das Daten generiert, gegen ein Diskriminierungsmodell an. Dieses soll die synthetischen Bilder von re-

18 Das Verfahren stellt sicher, dass die Ergebnisse von Datenanalysen quasi gleich sind, unabhängig davon, ob ein Individuum sich in einer Datenbank befindet oder nicht. Kritik an der Methode wird geäußert in Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, L. Vilhuber (2008). Privacy: Theory meets Practice on the Map, Working Paper, http://www.cse.psu.edu/~duk17/papers/PrivacyOnTheMap.pdf und in Bambauer, J., K. Muralidhar, R. Sarathy (2014). Fool's Gold: An Illustrated Critique of Differential Privacy, Vanderbilt Journal of Entertainment & Technology Law, Vol. 16, No. 4, http://www.jetlaw.org/wp-content/uploads/2014/06/Bambauer_Final.pdf

19 Dies ruft Skaleneffekte hervor: Die Konfigurationszeit für Synthetisierung fällt nur einmal zu Beginn des Prozesses an. Die jeweilige Generierungszeit pro Dateneinheit bleibt konstant, aber die durchschnittliche Konfigurationszeit pro Einheit sinkt. Damit sinkt die durchschnittliche gesamte Exekutionszeit pro Einheit.

20 Anderson, J., K.E. Kennedy, L.B. Ngo, A. Luckow und A.W. Apon (2015). Synthetic data generation for the internet of things, Working Paper, 2014 IEEE International Conference on Big Data, DOI: 10.1109/BigData.2014.7004228

21 Bei diesen Methoden handelt es sich unter anderem um Muster- oder Objekterkennung durch den Einsatz Neuronaler Netze.

alen unterscheiden: „Am Ende wirft das Generator-Netzwerk Bilder aus, die für das Diskriminierungsmodell von realen Bildern ununterscheidbar sind."[22]

Auch Apple setzt Synthetisierung ein und zeigt, wie diese genutzt werden kann, um sehr schnell mit Labels versehene Datensätze zu produzieren. In der Vergangenheit mussten Fotos beispielsweise arbeitsintensiv von Menschen als ‚Auge' oder ‚kein Auge' klassifiziert werden. Mit Synthetisierung können realistische Augen produziert werden, die gleich als solche gelabelt werden. Dies löst Skaleneffekte beim Training von Neuronalen Netzen aus.[23] Synthetisierung erfordert zusätzliche Rechenoperationen. Es ist offen, inwiefern synthetische Daten in Umgebungen mit sehr hohem Datendurchsatz ‚*on the fly*' produziert werden können.

III Einsatz und praktische Anwendungsfälle

Die Synthetisierung von Daten ist bereits bei verschiedenen Statistikbehörden und Forschungszentren sowie in der Privatwirtschaft im Einsatz. Im Folgenden wollen wir auf drei Anwendungsfälle eingehen. Die drei Beispiele sind aufgrund übergeordneter Gesichtspunkte interessant. Im ersten Fall (*SIPP Synthetic Beta File*) wird gezeigt, wie Behörden Datensätze zusammenführen und synthetisieren, um detaillierte Informationen an Forscher:innen weitergeben zu können. Im zweiten Fall (OnTheMap) wird gezeigt, wie dünn besetzte Datensätze synthetisiert werden können.[24] Im dritten Fall (SynLBD) geht es darum, besonders sensitive Betriebsdaten einer breiten Fachöffentlichkeit zur Verfügung zu stellen. Dies wird verbunden mit dem Fall der GsynLBD für die grenzüberschreitende Vergleichbarkeit von Datensätzen.

3.1 SIPP Synthetic Beta File: Einkommens- und Transferdaten

In der *Survey of Income and Program Participation (SIPP)* sammelt das U.S. Census Bureau Einkommens-, Steuer- und Transferleistungsdaten einer repräsentativen Stichprobe von U.S.-Amerikaner:innen. Es handelt sich um eine Serie von nationalen Panel-Datensätzen, also Befragungen.[25] Im Rahmen eines Forschungsprojekts wurden diese Daten mit Daten der Sozialversicherungsbehörde und der Steuerbehörde der USA verknüpft. Da die verknüpften

22 S. OpenAI Blog https://blog.openai.com/generative-models/

23 Apple (2017). Improving the Realism of Synthetic Images, Apple Machine Learning Journal 1 (1) https://machinelearning.apple.com/2017/07/07/GAN.html

24 Nimmt man alle existierenden Webseiten der Welt als Gesamtpopulation an, kann ein Internetnutzer nur eine verschwindend geringe Anzahl der Webseiten besuchen. Für alle anderen Webseiten muss der Besuch auf den Wert Null gesetzt werden. Daraus ergibt sich ein dünn besetzter Datensatz.

25 Weitere Erklärungen: https://www.census.gov/programs-surveys/sipp/about.html

Daten eine Vielzahl sensibler Informationen enthalten und außerdem Auszüge aus dem *SIPP* allgemein verfügbar und somit zur Reidentifikation nutzbar sind, konnte ein ausreichender Datenschutz auf Basis herkömmlicher Anonymisierungsverfahren nicht gewährleistet werden. Daher wurde eine synthetische Version (*SIPP Synthetic Beta*) erstellt, die über einen Server der Cornell University zugänglich ist.[26]

Das Census Bureau offeriert allen Datennutzern zusätzlich die Möglichkeit, ihre Datenanalyse-Protokolle einzureichen. Die Behörde kann die Ergebnisse dann auf den vertraulichen Originaldaten (Echtdaten) validieren.

3.2 OnTheMap: Synthetisierung von Mobilitätsströmen

In den USA sammelt das U.S. Census Bureau innerhalb des *Longitudinal Employer-Household Dynamics Program* Daten, die das Berufspendlerverhalten der Bevölkerung zeigen. Für die Analyse teilt das U.S. Census Bureau die USA in circa acht Millionen geographische Zensus-Blöcke ein. Die Befragungsdaten zeigen, von welchem Block eine Person startet und welcher Block ihr Ziel ist. Dieser Datensatz ist dünn besetzt, da bei acht Millionen Blöcken auf der US-Karte nur wenige hundert Pendler existieren, die in die einzelnen Richtungen fahren.[27] Mobilitätsdaten gelten als sensibel und sollten daher nicht im Original veröffentlicht werden.

Aus diesem Grund teilsynthetisiert das U.S. Census Bureau die Daten. Seit 2006 ist die zugehörige web-basierte Anwendung online, die jährlich aktualisiert wird. Der Datensatz zeigt detaillierte Wohn-Arbeitsstätte-Kombinationen auf Zensus-Block-Ebene, obwohl die Vertraulichkeit strikt gesichert ist.“[28]

Für die Synthetisierung nutzt das U.S. Census Bureau unter anderem die Originaldaten, um Wahrscheinlichkeitsverteilungen der Anzahl von Arbeitern, die in den einzelnen Blöcken wohnen, basierend auf Attributen wie Alter, Einkommen, Branche und Wohnort zu schätzen.[29] Aus diesen Verteilungen werden neue Werte gezogen, die regelbasiert zu synthetischen Endwerten

26 Weitere Erklärungen: https://www2.vrdc.cornell.edu/news/data/sipp-synthetic-beta-file/

27 Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, L. Vilhuber (2008). Privacy: Theory meets Practice on the Map, Working Paper, http://www.cse.psu.edu/~duk17/papers/PrivacyOnTheMap.pdf.

28 U.S. Census Bureau (2009). OnTheMap: Synthetic Data Protection OnTheMap, https://lehd.ces.census.gov/doc/help/OTMSyntheticData%2005262009-jma.pdf.

29 Für vergrößerten Vertraulichkeitsschutz werden außerdem die Parameter der Wahrscheinlichkeitsverteilungen durch Zugabe eines Zufallsfaktors (Noise) verändert.

kombiniert werden. Die synthetischen Daten entsprechen nicht immer exakt den Originaldaten,[30] so wird beispielsweise die Anzahl langer Pendelstrecken überschätzt. Aber dieser Schätzfehler reduziert sich mit Anzahl der Arbeiter pro Zielblock.

3.3 SynLBD und GsynLBD: Synthetisierung von Betriebsdaten

Zwei weitere Datenprodukte sollen hier ebenfalls kurz dargestellt werden. So bietet das U.S. Census Bureau die *Longitudinal Business Database (LBD)* in einer synthetischen Variante an. Bei der LBD handelt es sich um administrative Betriebsdaten, die sämtliche Betriebe nahezu aller Wirtschaftszweige seit 1976 umfassen. Die Daten enthalten unter anderem Informationen zu Betriebsgründung und -schließung, Wirtschaftszweigen, Umsätzen, und Beschäftigtenzahlen.[31]

Ein teilweise synthetischer Datensatz *(synLBD)* wurde erstellt, indem zunächst die Lebensdauer jedes Betriebs basierend auf der Wirtschaftszweiginformation synthetisiert wurde. Danach wurde für die so erzeugten Betriebe die jeweilige Betriebsgröße für die Jahre, in denen der Betrieb existiert, modelliert. Derzeit untersucht das Institut für Arbeitsmarkt- und Berufsforschung (IAB), ob sich der Merkmalskanon der LBD mit den Daten des IAB Betriebshistorikpanels nachbilden lässt. Ziel des Projekts ist es, unter Verwendung der Synthetisierungsmodelle, die für die *LBD*-Daten entwickelt wurden, eine synthetische Version der deutschen Daten zu erstellen (GsynLBD).[32] Auf diese Weise wäre es erstmals möglich, länderübergreifende Vergleichsstudien auf Basis von Betriebsdaten durchzuführen. Diese Daten könnten unter Einhaltung des Datenschutzes mit der synthetischen Version der U.S. Daten gemeinsam auf einem Server angeboten werden. Länderübergreifende Vergleiche sind bisher aufgrund der Sensitivität und des hohen Reidentifikationsrisikos kaum durchführbar.

Auch die Statistikbehörden in Kanada und Brasilien haben großes Interesse an diesem Forschungsprojekt geäußert und überlegen ebenfalls, ihre Betriebsdaten auf diesem Weg zugänglich zu machen.

30 Es gibt verschiedene Kriterien, mit welchen der Qualitätsunterschied zwischen Originaldaten und synthetischen Daten gemessen werden kann. Eines davon ist beispielsweise die Überlappung der Konfidenzintervalle.

31 Miranda, J. und R.S. Jarmin (2002). The longitudinal business database. Discussion Paper CES-WP-02-17, U.S. Census Bureau, Center for Economic Studies.

32 Drechsler, J. und L. Vilhuber (2014). A first step towards a German SynLBD: Constructing a German Longitudinal Business Database. Statistical Journal of the IAOS, Vol. 30, 137–142.

IV Grenzen der Synthetisierungsmethode

Wir wollen nachfolgend die Grenzen der Methode aufzeigen. Grundsätzlich steht und fällt die Qualität und Plausibilität der synthetisierten Daten mit dem Arbeitsaufwand, der in die Synthetisierung gesteckt wird. Eine größere Investition in die Erstellung eines guten Datenmodells steigert die Qualität synthetisierter Daten. Umgekehrt gilt, dass nur die Zusammenhänge zwischen den Variablen im Originaldatensatz, die im Modell zur Synthetisierung berücksichtigt wurden, auch in den synthetischen Daten gefunden werden. Bei Synthetisierungsmethoden, die auf Verteilungsannahmen basieren, resultiert eine präzisere Repräsentation der Verteilung in qualitativ höherwertigen Daten. Dies steigert aber den Arbeitsaufwand. Das sogenannte *overfitting*,[33] eine Überanpassung des Modells an Trainingsdaten, ist kein Problem, solange das Reidentifizierungsrisiko nicht steigt. Allerdings sollten Datenanbieter wissen, welche Zusammenhänge oder Fragestellungen für die späteren Nutzer:innen der synthetischen Daten interessant sind.

Die Arbeit mit synthetischen Daten entbindet Anwender:innen auch nicht von der Aufgabe, das Reidentifizierungsrisiko zu analysieren. Bei vollsynthetisierten Datensätzen ist dieses Reidentifizierungsrisiko zwar sehr gering, aber nicht gleich Null. So können statistische Ausreißer abgebildet sein, die eine Reidentifizierung im Einzelfall doch ermöglichen. Für einen erhöhten Schutz können die Verfahren mit prüfbaren Garantien kombiniert werden, wie bereits erwähnt.[34]

Eine besondere Herausforderung stellt die Synthetisierung hochkomplexer und hochdimensionaler Daten dar. In der Praxis ist es ein schwieriges Unterfangen, sämtliche Zusammenhänge oder Cluster adäquat abzubilden und dabei noch komplexe Restriktionen zwischen den Variablen korrekt zu replizieren.

Eine weitere Hürde stellt die Akzeptanz durch die anvisierten Nutzer:innen der Daten dar. Viele potenzielle Nutzer:innen sind skeptisch, wenn sie mit synthetischen Daten arbeiten sollen. Wer garantiert ihnen, dass die Erkenntnisse, die sie auf Grundlage der synthetischen Daten erzielen, zu ähnlichen

33 Das Modell kann die Trainingsdaten sehr gut erklären, funktioniert aber nicht so gut auf neuen Daten.

34 Dies wird gezeigt in Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke und L. Vilhuber (2008). Privacy: Theory meets Practice on the Map, Work. Paper, http://www.cse.psu.edu/~duk17/papers/PrivacyOnTheMap.pdf.

Schlussfolgerungen führen, wie sie auf Basis der Originaldaten gezogen worden wären?

Eine Möglichkeit, dieser Kritik zu begegnen, sind Verifikationsserver.[35] Diese Server vergleichen die Ergebnisse auf Basis der synthetischen und der originalen Datensätze. Sie geben Nutzer:innen der synthetischen Daten lediglich die Information zurück, wie nahe ihre Ergebnisse (aus den synthetischen Daten) an denen der Originaldaten liegen. Allerdings ist auch hier auf eventuell neu entstehende Datenschutzrisiken zu achten, da der Verifikationsserver zusätzliche Informationen bezüglich der Originaldaten preisgibt.

Streng genommen ist dies keine Schwäche der Methode, aber es soll hier nicht unerwähnt bleiben, dass die Qualität synthetischer Daten von der Qualität der Rohdaten abhängt. Fehler und Auslassungen in den Rohdaten werden im synthetischen Datensatz reproduziert.

V Gesellschaftliche Herausforderungen synthetischer Daten

Wir haben die Synthetisierung zunächst im Licht der datengetriebenen Innovation und ihrer Potentiale für Kollaborationen und Open Data diskutiert. Mit der Methode entstehen aber auch eine Reihe rechtlicher, ethischer und technischer Herausforderungen, die wir hier nur anschneiden können. Mit einer weiteren Verbreitung der Methode, müsste diese Debatte auf einer breiten gesellschaftlichen Basis geführt werden.

5.1 Rechtliche Herausforderungen synthetischer Daten

Institutionen, die personenbezogene Daten akkumulieren und verwerten, unterstehen mit diesen Tätigkeiten der Datenschutz-Grundverordnung (DS-GVO), die ab 25. Mai 2018 Geltung erlangt. Mit dieser Verordnung treten eine Reihe von Rechten und Pflichten für Unternehmen in Kraft, welche die Transparenz von Datenverarbeitungsvorgängen erhöhen sollen. Datenschutzverstöße können außerdem mit empfindlichen Strafen geahndet werden.

Ein Prinzip der Verordnung ist die Zweckbindung (Artikel 5 Abs. 1 b) DS-GVO). Sinngemäß dürfen personenbezogene Daten nur für eindeutig festgelegte und zulässige Zwecke erhoben werden. Auch die Weiterverarbeitung darf nur

35 McClure, D.R. und J.P Reiter (2012). Towards providing automated feedback on the quality of inferences from synthetic datasets. Journal of Privacy and Confidentiality, 4(1), Article 8.

an bestimmte Zwecke gebunden sein. Bei einer den Zweck verfremdenden bzw. zweckändernden Weiterverarbeitung (ohne vorherige Einholung einer Einwilligung) muss der Verantwortliche einen Test der Vereinbarkeit von alten und neuen Zwecken durchführen und darf die neue Verarbeitung nur vornehmen, wenn diese Vereinbarkeit besteht. Ansonsten handelt es sich um einen Datenschutzverstoß.

Der Status von synthetischen Daten steht und fällt mit der Personenbeziehbarkeit. Um Personenbezug festzustellen, werden von Gerichten hypothetische Ketten (der Datenweitergabe) gebildet, bei welchen gefragt wird, wie groß die Wahrscheinlichkeit ist, dass irgendwo in dieser Kette eine Reidentifizierung stattfinden kann bzw. vorgesehen ist.[36] Eine andere juristische Einschätzung geht davon aus, dass die Analyse personenbezogener Daten einer Rechtsgrundlage bedarf, nicht aber die Erzeugung synthetischer Daten. Die Personenbeziehbarkeit synthetischer Daten wäre also zu prüfen und eventuell rechtlich zu klären, ab wann der Status des Personenbezugs gegeben ist oder eventuell entfällt.

Sollten persönliche Daten in synthetische Daten überführt werden, bei denen die Wahrscheinlichkeit der Reidentifzierung gleich Null (zum Beispiel technisch ausgeschlossen) ist, dann wären diese als anonym zu klassifizieren. Die Regelungen der DS-GVO würden in diesem Fall keine Anwendung finden: Zweckbindung, Dokumentations- oder Auskunftspflichten würden entfallen.

Eine Zusammenführung verschiedener Datensätze mit synthetisierten Daten wäre in diesem Fall ebenfalls möglich. Das Matching könnte über ‚statistische Zwillinge' möglich sein, also synthetische Fälle, die große Ähnlichkeiten aufweisen. Dieses Vorgehen ermöglicht ein Erlernen neuer statistischer Zusammenhänge auf vormals getrennten Datensätzen.[37] Die so gewonnenen Erkenntnisse könnten auf Subjekte in dem Originaldatensatz transferiert werden.

Obwohl Daten ohne Personenbezug anonym sind, ergeben sich indirekt Konsequenzen für die Privatsphäre der Datensubjekte im *Originaldatensatz*. Solche Konsequenzen werden in der Literatur als ‚Inferenzrisiko' bezeich-

36 Die Autoren bedanken sich bei Carlo Piltz für diese Ausführungen.

37 Das ist ein erheblicher Wettbewerbsvorsprung für jene Unternehmen, die als erste in den Besitz von neuen Arten von Originaldaten gelangen.

net.[38] So ist es möglich, den Wert eines sensitiven Attributes (zum Beispiel Krankheit einer Person) aus einem Set anderer Attribute (Bewegungsprofil, Ernährungsprofil, etc.) zu prognostizieren. Solche Prognosen können auf Basis sehr kleiner Gruppen gemacht werden (sog. *‚private inference'*).

Sobald aber ein:e Datenanalyst:in diese Zusammenhänge auf Basis des synthetischen Datensatzes gelernt hat, kann er:sie diese Erkenntnisse auf die Datensubjekte im Originaldatensatz übertragen.

5.2 Ethische Herausforderungen synthetischer Daten

Sollte der rechtliche Rahmen im Hinblick auf synthetische Daten bestimmte Datenanalysen nicht eindeutig legitimieren, stellen sich für Unternehmen und öffentliche Verwaltungen Fragen der Datenethik. Unter ‚Datenethik' sind Abwägungen zu verstehen, wofür Daten verwendet werden dürfen und wofür nicht. Solche Fragen müssen Unternehmen im Rahmen ihrer Datenstrategie insbesondere im Hinblick auf potentielle Reputationsrisiken beantworten. Und in vielen Unternehmen hat dieser Diskurs bereits begonnen.

Auch im Bereich der Bereitstellung öffentlicher Daten muss in der Politik diskutiert werden, inwiefern der Zugang zu solchen Daten nur auf Basis vertraglicher Vereinbarungen und Protokolle, nebst der Überprüfung derselben (Due Diligence) geschehen darf.

5.3 Weitere potentielle Risiken synthetischer Daten

Auf der technischen Ebene stellen sich Fragen der Informationssicherheit. Die Originalität der Daten – im Sinne von Echtheit – spielt in vielen digitalen Anwendungen eine kritische Rolle. So können synthetische Daten nicht nur produziert werden, um Produkte oder Prozesse zu verbessern, sondern auch, um diese anzugreifen. Beispielsweise könnten synthetische Gesichter eingesetzt werden, um biometrische Authentifizierungsverfahren im Online-Banking auszuhebeln.[39] Und vor etwa zwei Jahren zeigte eine Forschergruppe an der University of North Carolina at Chapel Hill, wie aus öffentlich zugäng-

38 Castelluccia, C., G. Acs und D. Le Metayer (2016). Testing the Robustness of Anonymization Techniques, Presentation at the Brussels Privacy Symposium Identifiability: Policy and Practical Solutions for Anonymization and Pseudonymization, https://fpf.org/brussels-privacy-symposium-agenda/.

39 Dieses Szenario wurde von Herrn Professor Scheuermann (Lehrstuhl für Technische Informatik, Humboldt-Universität zu Berlin) bei einem BMBF-Fachgespräch diskutiert.

lichen Bildern realer Personen synthetische Gesichter derselben generiert werden.[40] Es handelt sich dabei um synthetische Bilder echter Personen.

Für die Integrität von Online-Diensten ist deshalb eine Diskussion über das Gefährdungspotential solcher Verfahren und wie dieses reduziert werden kann, notwendig. Letzten Endes müssten hier Verfahren entwickelt werden, die es ermöglichen, die Vertrauenswürdigkeit eines Datensatzes robust zu verifizieren.

VI Schluss

In der Datenökonomie müssen State-of-the-art Methoden entwickelt werden, die den Datennutzen maximieren *und* Compliance mit Datenschutzvorschriften gewährleisten. Der bekannte Weg ist, Daten unter Anwendung verschiedener Verfahren zu anonymisieren. Je nach Anwendung kann dies aber die Nützlichkeit der Daten reduzieren. Aufgrund der gestiegenen Rechnerkapazitäten ist es heute möglich, sehr große Datenmengen (Big Data) zu synthetisieren. Dies geschieht mit verschiedenen Ansätzen, durch die eine qualitativ hochwertige, ‚artifizielle' Repräsentation der Originaldaten geschaffen werden kann.

In der Vergangenheit wurde Daten-Synthetisierung als zu arbeitsaufwändig angesehen. Dies ändert sich mit dem vermehrten Einsatz von Machine-Learning-Verfahren. Mittlerweile werden synthetische Datensätze für das autonome Fahren erstellt, in der Betrugserkennung, für medizinische Daten oder für das IoT. Das Anwendungsgebiet synthetischer Daten hat sich in den vergangenen Jahren stark erweitert. Der Einsatz wird sich künftig aller Voraussicht nach noch steigern.

Insgesamt wären deshalb klare Regeln notwendig, unter welchen Bedingungen und anhand welcher Kriterien synthetische Daten als anonyme Daten gelten. Dies könnte mit der Entwicklung von automatisierten Methoden, mit welchen Personenbezug festgestellt werden kann, einhergehen.

Deutschland sollte wesentlich mehr als bisher in die Forschung an synthetischen Daten investieren. Diese Förderung sollte automatisierte Verfahren der Synthetisierung ins Zentrum stellen, welche die Effizienz im praktischen Einsatz steigern. Die wichtigsten Forschungsimpulse in dem Bereich kom-

40 Xu, Y., T. Price, J.-M. Frahm und F. Monrose (2016). Virtual U: Defeating Face Liveness Detection by Building Virtual Models from Your Public Photos, Proceedings of the 25th USENIX Security Symposium August 10–12, 2016, https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/xu.

men auch hier immer noch aus den USA. Deutschland kann nur vereinzelt auf Forscher:innen und Projekte verweisen.

Daten-Synthetisierung könnte Forschungskooperationen befördern. Synthetische Daten könnten künftig beispielsweise in Datenpools zusammengeführt werden. Dies sollte von Privatheits-Garantien für die Datensubjekte begleitet werden, insbesondere wenn es sich um eine Teilsynthetisierung der Echtdaten handelt. Nutzer:innen der Daten sollten außerdem über eine Data Governance verfügen und forschungsethische Verpflichtungen unterschreiben, bevor sie auf gepoolte synthetische Daten zugreifen dürfen. Diese Verpflichtungen sollten permanent überprüft werden. Zuwiderhandlungen sollten außerdem empfindlich geahndet werden.

Wir haben nur einige der rechtlichen und ethischen Fragen, die sich in Bezug auf synthetische Daten stellen, aufgeworfen. Wir sehen diese Diskussion erst ganz am Anfang, denn auch andere Fragen, beispielsweise bezüglich des Wettbewerbs und der Marktdominanz, stellen sich hier.

Zusätzlich sollten Fragen der Informationssicherheit und des Einsatzes dieser Verfahren jenseits der Dateninnovation eingehender untersucht werden. Insgesamt stellt sich die Frage, ob der Datenschutz in seiner gängigen Fassung noch ausreicht oder durch weitere Verpflichtungen flankiert werden muss.

Grundsätzlich gilt, dass jede Form der Datennutzung ein inhärentes Risiko trägt. Informationsgewinn kann nur zustande kommen, indem der Einzelne einen Teil seiner Privatsphäre preisgibt. Letztlich ist es eine gesamtgesellschaftliche Frage, wie viel Privatsphäre wir aufzugeben bereit sind, um den Erkenntnisgewinn für die Allgemeinheit zu maximieren. Die Nutzung von synthetischen Daten könnte ein weiterer wichtiger Schritt sein, einen Kompromiss zwischen beidem zu finden.

Referenzen

Adä, I. und M.R. Berthold (2010). The New Iris Data: Modular Data Generators, KDD'10, July 25–28, 2010, Washington, DC, USA.

Anderson, J., K.E. Kennedy, L.B. Ngo, A. Luckow und A.W. Apon (2015). Sythetic data generation for the internet of things, Working Paper, 2014 IEEE International Conference on Big Data, DOI: 10.1109/BigData.2014.7004228.

Apple (2017). Improving the Realism of Synthetic Images, Apple Machine Learning Journal 1 (1) https://machinelearning.apple.com/2017/07/07/GAN.html

Bambauer, J., K. Muralidhar und R. Sarathy (2014). Fool's Gold: An Illustrated Critique of Differential Privacy, Vanderbilt Journal of Entertainment & Technology Law, Vol. 16 (Summer 2014), No. 4, http://www.jetlaw.org/wp-content/uploads/2014/06/Bambauer_Final.pdf.

Benedetto, G. und M. Stinson (2015). Disclosure Review Board Memo: Second Request of Release of SIPP Synthetic Beta Version 6.0, January 15, 2015, https://www2.vrdc.cornell.edu/news/wp-content/papercite-data/pdf/drbmemo2015.pdf.

Bogle, B.M. und S. Mehrotra (2016). A Moment Matching Approach for Generating Synthetic Data, Big Data, Vol. 4, No. 3, 160 – 178.

Castelluccia, C., G. Acs und D. Le Metayer (2016). Testing the Robustness of Anonymization Techniques, Presentation at the Brussels Privacy Symposium Identifiability: Policy and Practical Solutions for Anonymization and Pseudonymization, https://fpf.org/brussels-privacy-symposium-agenda/.

Drechsler, J. (2010). Using Support Vector Machines for Generating Synthetic Datasets. In: J. Domingo-Ferrer & E. Magkos (Ed.), *Privacy in Statistical Databases* (Lecture Notes in Computer Science 6344), Springer: Berlin, 148–191.

Drechsler, J. (2011). Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Lecture Notes in Statistics 201, Springer: New York.

Drechsler, J., S. Bender und S. Rässler (2008). Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Es-

tablishment Panel, Transactions on Data Privacy 1: 105 – 130.

Drechsler, J. und J.P. Reiter (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In J. Domingo-Ferrer and Y. Saygin, eds., Privacy in Statistical Databases, 227–238. Springer: New York.

Drechsler, J., und J.P. Reiter (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, Vol. 55, 3232–3243.

Drechsler, J. und L. Vilhuber (2014). A first step towards a German SynLBD: Constructing a German Longitudinal Business Database. *Statistical Journal of the IAOS*, Vol. 30, 137–142.

Erickson, J. (2017). Creating Synthetic Data with SAS/OR, SAS Blog (17. Mai 2017). https://blogs.sas.com/content/operations/2017/05/17/creating-synthetic-data-sasor/.

Kuiper, J., E.R. van den Heuvel und M.A. Swertz (2015). The Hybrid Synthetic Microdata Platform: A Method for Statistical Disclosure Control, Biopreservation and Biobanking 13 (3), 178 – 182.

Little, R. J. A. (1993). Statistical analysis of masked data. Journal of Official Statistics 9, 407–426.

Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke und L. Vilhuber (2008). Privacy Theory meets Practice on the Map, Working Paper, http://www.cse.psu.edu/~duk17/paper/PrivacyOnTheMap.pdf.

Maqsud, U. (2015). Synthetic Text Generation for Sentiment Analysis, Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015): 156–161, Lisboa, Portugal, 17 September, 2015.

McClure, D.R. und J.P Reiter (2012). Towards providing automated feedback on the quality of inferences from synthetic datasets. *Journal of Privacy and Confidentiality*, 4(1), Article 8.

Ming, Z., C. Luo, W. Gao, R. Han, Q. Yang, L. Wang und J. Zhan (2014). BDGS: A Scalable Big Data Generator Suite in Big Data Benchmarking, http://prof.ict.ac.cn/BigDataBench/wp-content/uploads/2013/10/BDGS_BigDataBench.

pdf.

Miranda, J. und R.S. Jarmin. The longitudinal business database. Discussion Paper CES-WP-02-17, U.S. Census Bureau, Center for Economic Studies, 2002.

Raghunathan, T. E., J.P. Reiter und D.B. Rubin (2003). Multiple imputation for statistical disclosure limitation, Journal of Official Statistics 19, 1–16.

Reiter, J. P. und R. Mitra (2009). Estimating risks of identification disclosure in partially synthetic data. Journal of Privacy and Confidentiality 1, 99–110.

Reiter, J. P. (2005). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata, Journal of Statistical Planning and Inference 131, 365–377.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. Journal of Official Statistics 9, 462-468.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys, John Wiley & Sons: New Jersey (USA).

U.S. Census Bureau (2009). OnTheMap: Synthetic Data Protection OnTheMap, https://lehd.ces.census.gov/doc/help/OTMSyntheticData%2005262009-jma.pdf.

Xu, Y., T. Price, J.-M. Frahm und F. Monrose (2016). Virtual U: Defeating Face Liveness Detection by Building Virtual Models from Your Public Photos, Proceedings of the 25th USENIX Security Symposium August 10–12, 2016, https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/xu.

Die Autoren

Dr. Jörg Drechsler ist Wissenschaftler am Institut für Arbeitsmarkt- und Berufsforschung (IAB) in Nürnberg und Adjunct Assistent Professor im Joint Program in Survey Methodology an der University of Maryland. Er habilitierte sich 2015 an der Ludwig-Maximilians-Universität in München im Fach Statistik und wurde 2009 an der Otto-Friedrich-Universität in Bamberg promoviert. Seine Forschungsschwerpunkte sind Datenschutz und fehlende Werte bei Befragungen. Seine Forschung zu synthetischen Daten wurde mehrfach ausgezeichnet und vielfach in internationalen Fach-Journalen publiziert. Eine Monographie zu diesem Thema ist im Springer-Verlag erschienen.

Dr. Nicola Jentzsch leitet das Arbeitsgebiet Datenökonomie an der Stiftung Neue Verantwortung (SNV). Sie ist eine Ökonomin, die Ansätze der Wettbewerbs- und Verhaltensökonomie verbindet. Sie wurde 2005 in Wirtschaftswissenschaft an der Freien Universität Berlin promoviert. Vor der SNV arbeitete sie mehrere Jahre am DIW Berlin und war unter anderem Research Fellow an der Yale University und der Georgetown University. Ihre Forschungsschwerpunkte sind Wettbewerb und Datenschutz und Privacy-Garantien (Privacy by Design). Sie ist Gewinnerin eines Google Research Awards. Ihre Forschung zu Privatsphäre wurde in internationalen ökonomischen Fachzeitschriften publiziert.

Korrespondenz bezüglich des vorliegenden Impulses ist an Dr. Nicola Jentzsch zu richten.

Die Autoren bedanken sich bei Iris Adä, Mikhail Dyakov, Omar Ali Fdal, Stefan Heumann, Carlo Piltz und Alexander Sängerlaub für Kommentare und Anmerkungen.

Impressum

Stiftung Neue Verantwortung e. V.

Beisheim Center
Berliner Freiheit 2
10785 Berlin

T: +49 (0) 30 81 45 03 78 80
F: +49 (0) 30 81 45 03 78 97

www.stiftung-nv.de
info@stiftung-nv.de

Design:
Make Studio
www.make-studio.net

Layout:
Johanna Famulok

Free Download:

www.stiftung-nv.de