Bild: Adobe Stock Sarah Holmlund

Synthetische Daten:

Die Rettung aus der Anonymisierungskrise

Regelmäßig beschwören in einem ersten Schritt Unternehmen und Behörden, dass die durch sie genutzten Daten anonym seien, sie können auf keinen Fall konkreten Personen zugeordnet werden, zudem interessiere man sich für konkrete Personen ohnehin nicht. In einem zweiten Schritt berichten dann Medien, meist unter Hinzuziehung technisch Versierter, dass irgendwie dann doch ein Personenbezug hergestellt werden kann. Gibt es also etwas wie Anonymisierung überhaupt und könnten synthetische Daten eine Alternative sein?

Von Hans-Christian Schellhase, ANMATHO AG

Bei Lichte betrachtet erschwert die Datenschutz-Grundverordnung (DS-GVO) Unternehmen und Behörden die Verarbeitung personenbezogener Daten erheblich, um personenbezogene Daten für sekundäre Zwecke zu nutzen, etwa für Projekte zur künstlichen Intelligenz oder zum Maschinenlernen, müssen gewisse Hürden überwunden werden. Üblicherweise ist das Interesse bei Unternehmen und Behörden an einer solchen Nutzung jedoch recht groß. Insoweit ist durchaus fraglich, ob die DS-GVO Innovationen verhindert, denn häufig fehlt für eine entsprechende Verarbeitung zu einem anderen Zweck die Rechtsgrundlage und die in Rede stehenden Daten dürfen eben nicht über ihren eigentlichen Erhebungszweck hinaus verwandt werden.

Ist Anonymisierung eine Lösung?

Um nun die DS-GVO als Sparringspartner zu meiden, setzen viele Verantwortliche auf die Anonymisierung solcher Daten; anonyme Daten sind gerade keine personenbezogenen Daten, der Anwendungsbereich der DS-GVO wäre erst gar nicht eröffnet. Erwägungsgrund 26 der DS-GVO führt dazu aus:

„Die Grundsätze des Datenschutzes sollten daher nicht für anonyme Informationen gelten, d.h. für Informationen, die sich nicht auf eine identifizierte oder identifizierbare natürliche Person beziehen, oder personenbezogene Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann. Diese Verordnung betrifft somit nicht die Verarbeitung solcher anonymer Daten, auch für statistische oder für Forschungszwecke."

Im Ergebnis sind anonymisierte Daten eben solche Daten, die so anonymisiert sind, dass die betroffene Person nicht oder nicht mehr identifizierbar ist.

Wichtigste Voraussetzung für eine Anonymisierung ist, dass die jeweiligen Daten keine Identifizierung (mehr) zulassen, wobei hier nicht nur auf das einzelne Datum selbst abgestellt wird, auch die Nutzung verschiedener anderer Informationen darf nicht dazu führen, dass die Möglichkeit der Identifizierung einer Person besteht. Gelingt die Anonymisierung, können Unternehmen und Behörden die Daten regelmäßig ohne größere Einschränkungen frei nutzen, sogar verkaufen, der Schutz der vormals „betroffenen Person" wird nicht mehr gewährt, ist nicht mehr erforderlich.

Die echte Anonymisierung von Daten gestaltet sich jedoch als sehr schwierig und ist daneben mit einem signifikanten Einsatz von Zeit und Ressourcen verbunden. Das ist nicht zuletzt auch dem Umstand geschuldet, weil das Scheitern einer Anonymisierung die jeweiligen Prozesse und Verfahren oftmals grundlegend in Frage stellt. Ferner ist eine

Anonymisierung – wie sie die DS-GVO verlangt – stets auch eine Verfolgte des technischen Fortschritts, der immer wieder dazu führt, dass ein hinreichender Zusammenhang zwischen einem Datum und einer konkreten Person herstellbar wird. Möglicherweise erscheint vor diesem Hintergrund die vorsichtige Annahme, so etwas wie Anonymisierung gäbe es gar nicht (wirklich), zumindest ab und an vertretbar.

Synthetische Daten als Mittelweg

Sie sind weder „Fisch noch Fleisch", synthetische Daten könnten jedoch der Mittelweg sein. Die Synthetisierung ist ein Verfahren mit dem Originaldaten – etwa hoch sensible besondere Kategorien personenbezogener Daten – in synthetische Daten überführt werden. Vereinfacht ausgedrückt werden Daten künstlich erzeugt, wobei die Identifizierbarkeit von betroffenen Personen – insbesondere vor dem Hintergrund des technischen Fortschrittes – erheblich davon abhängt, ob die Synthetisierung lediglich Teilmengen oder alle Daten des Originaldatensatzes umfasst. Im Ergebnis entsteht damit ein voll synthetisierter Datensatz aus Daten synthetischer Subjekte und nicht realer Personen. Das ist dadurch bedingt, dass diese Daten gerade nicht durch direkte Messungen erhoben wurden, sondern das Resultat von Algorithmen sind, wodurch diese allerdings Daten, die mit direkten Messungen erhoben wurden, verarbeiten.

Die Synthetisierung als zunehmend genutzte Methode wurde schon vor mehr als 25 Jahren erfunden, sie beruht ursprünglich auf der multiplen Imputation, einem statistischen Verfahren zur Ersetzung fehlender Werte in Datensätzen, bei dem diese durch mehrere plausible Werte ersetzt werden. Heute ist die Synthetisierung bei verschiedenen Statistikbehörden und in der Finanzwelt sehr verbreitet, innerhalb Deutschlands ist das Institut für Arbeitsmarktforschung (IAB) Vorreiter in der Synthetisierung.

Ein weiterer großer Vorteil der Synthetisierung ist, dass sich alle Datenarten synthetisieren lassen, darunter auch Bilder und Texte. Die Datensätze können darüber hinaus in sehr großem Umfang zum jeweils gewünschten Präzisionsgrad hergestellt werden, die Zusammenhänge oder Cluster des Originaldatensatzes bleiben dabei erhalten. Die Qualität der synthetischen Daten ist mess- und damit auch vergleichbar, selbst die Qualität und der Umfang solcher Daten ist anpassbar. Eine Synthetisierung von Daten kann zudem mit mathematischen Garantien für die Privatheit kombiniert werden.

Im Übrigen ermöglicht die hier dargestellte Synthetisierung natürlich auch länderübergreifende Kooperationen und Datenweitergaben, trotz unterschiedlicher Datenschutzregime.

Durch die rasanten Entwicklungen im Bereich der künstlichen Intelligenz (KI) könnte es für immer mehr Unternehmen und Behörden attraktiv werden auf synthetische Daten zurückzugreifen. Künstliche Intelligenzen, wie etwa das sogenannte Deep Generative Model, generieren hier synthetische Daten, wobei maschinelle Lernalgorithmen zum Einsatz kommen, die auf einen Datensatz trainiert werden und die statistischen Informationen sowie Strukturen dieser Originaldaten erlernen. Aus diesem trainierten Verständnis des Datensatzes können dann wiederum neue, synthetische Datensätze geschaffen werden. Interessant ist hier vor allem auch, dass die genutzten Modelle zudem noch unentdeckte Zusammenhänge erkennen können.

Grenzen der Synthetisierung

Eine Synthetisierung weist aber auch Grenzen auf. Hinzuweisen ist insbesondere auf den Umstand, dass die Qualität und Plausibilität synthetisierter Daten sehr stark mit der Qualität der Synthetisierung selbst verknüpft ist. Weiter entbindet die Arbeit mit synthetischen Daten die Verantwortlichen nicht per se von der Pflicht, die erzeugten Daten mit Blick auf eine Identifizierung betroffener Personen zu prüfen, selbst bei vollsynthetisierten Datensätzen ist diese nicht vollständig ausgeschlossen. Möglicherweise könnten in den in Rede stehenden Daten statistische Ausreißer abgebildet sein, die eine Identifizierung möglich machen. Darüber hinaus besteht immer auch eine gewisse Skepsis gegenüber synthetischen Daten dergestalt, dass Zweifel bestehen, ob sich Erkenntnisse, die sich auf Grundlage der synthetischen Daten ergaben, auch auf der Grundlage der Originaldaten ergeben hätten. Abschließend muss immer auch beachtet werden, dass die Qualität synthetischer Daten von der Qualität der „Rohdaten" abhängt, so werden Fehler und Auslassungen in den Ursprungsdaten grundsätzlich reproduziert.

Ist die Zukunft synthetisch, also künstlich?

Die Nutzung synthetischer Datensätze wird für Unternehmen und Behörden gerade durch den Einsatz von KI zunehmend attraktiver, wobei man bei einer Verarbeitung synthetischer Daten regelmäßig nur sehr begrenzt den Zwängen der DS-GVO unterworfen sein wird. ■