Personenbezogene Datenanonymisierung: Theorie und Praxis

Bharat Ahuja
IT-P Information Technology-Partner GmbH
Hannover

Schlüsselworte

DSGVO, Anonymisierung, Pseudonymisierung, Datenschutz.

Einleitung

Die durch die EU eingeführten Regelungen in der DSGVO definieren den zulässigen Umgang mit personenbezogenen Daten. In dieser Verordnung wird genau definiert, in welchen Fällen Daten als personenbezogen zu betrachten sind, und wie die Daten durch die Datenverantwortlichen (Data Controller) verarbeitet werden können.

Zusätzlich sind Pflichten an den Datenverantwortlichen durch diese Verordnung entstanden, wie zum Beispiel das Löschen der Daten nach bestimmten Aufbewahrungsfristen und die Auskunftserteilungs-pflicht.

Mit dem Begriff „Personenbezogene Daten“ werden hier Datenmengen gemeint, die Individuen potenziell eindeutig identifizieren können. In solchen Datenmengen kommen die Datensubjekte, diejenigen Personen die in den Datensätzen beschrieben werden, mit eindeutig identifizierenden Attributen wie Vor- und Nachname, Steueridentifikationsnummer und Adresse vor.

Durch Data Analytics Ansätze wird die Gefahr immer höher, dass Daten zur eigenen Person kritische Entscheidungen beeinflussen können. Einerseits sind die Ergebnisse dieser Techniken unheimlich präzise und können für die Gesellschaft wichtige Ergebnisse liefern (z.B. medizinische Forschung), andererseits können die Ergebnisse gewisse Benachteiligungen durch genau diese Präzision bzw. durch Fehler in der Datenverarbeitung zur Folge haben. Genau hier versucht die DSGVO den Datensubjekten eine Art Kontrolle über ihre Daten zu schaffen.

Gleichzeitig wird versucht, den Schaden bei einem möglichen Datenschutzvorfall zu minimieren. Diebstahl von personenbezogenen Daten stellt für die Gesellschaft ein nicht zu vernachlässigendes Risiko dar. Daher versucht diese Regelung die Anzahl der „Points-Of-Failure“ und den potenziellen Schaden zu kontrollieren.

Allerdings möchten Datenverantwortliche die Daten aus der realen Welt verwenden, um ihre Systeme zu verbessern. Zudem können personenbezogene Daten aus der Vergangenheit potenziell nützlich sein, um Analysen über einen längeren Zeitraum zu ermöglichen.

Die DSGVO sieht einige mögliche Lösungen für dieses Problem vor. Personenbezogene Daten können so verarbeitet werden, um eine Identifizierung der Individuen aus resultierenden Datenmenge auszuschließen. Datenmengen, die keine Individuen eindeutig beschreiben können, werden durch die DSGVO gesondert betrachtet.

Werden die identifizierenden Attribute verarbeitet, um die Originalwerte durch Aliase (Pseudonyme) zu ersetzen, so spricht man von einer Pseudonymisierung. Die Trennung der resultierenden Datenmenge und der Umkehrfunktion zur Pseudonymisierungsfunktion sorgt dafür, dass die Individuen nach der Verarbeitung nicht identifiziert werden können.

Üblicherweise werden hierfür Aliase und Hashverfahren verwendet. Beim Einsatz von Aliasen müssen die Zuordnungen getrennt von der Datenmenge gesichert werden. Bei Hashverfahren geht man davon aus, dass es keine einfach zu berechnende Umkehrfunktion gibt.

Während einerseits die Sicherung der Identität durch die Sicherung der Aliase gewährleistet werden kann, besteht immer noch ein Risiko einer Identifikation, wenn der Zuordnungsschlüssel mit der pseudonymisierten Datenmenge kombiniert wird, oder wenn verschiedene Datenmengen verlinkt werden.

Andererseits können andere Verarbeitungsprozesse dafür sorgen, dass keine eindeutige Zuordnungsfunktion zwischen der resultierenden Datenmenge und Individuen überhaupt hergestellt werden kann. Hier spricht man von der Anonymisierung.

Die zugehörigen Datenmengen heißen anonymisiert und werden von der Verordnung explizit ausgeschlossen und als nicht personenbezogen betrachtet. Allerdings hat es sich als schwer erwiesen, Anonymisierungsgrade zu gewährleisten, da alle Identifizierungsmöglichkeiten berücksichtigt werden müssen.

Üblicherweise werden Randomisierung und Generalisierung hierfür verwendet.

Um aus der personenbezogenen Datenmenge eine neue pseudonymisierte bzw. anonymisierte Datenmenge zu erzeugen, ist die DSGVO weiterhin wirksam, da ein Zweck bzw. eine Grundlage für diese weitere Verarbeitung der Daten bestehen muss. Insbesondere trifft die DSGVO auf die Erhebung und Speicherung der Daten weiterhin zu.

Dieser Vortrag beschäftigt sich mit den Techniken personenbezogene Datenmengen zu verarbeiten, so dass eine Zuordnung zwischen der resultierenden Datenmenge und den Datensubjekten nicht mehr möglich ist. Oft wird irrtümlich die Pseudonymisierung für ein Anonymisierungsverfahren gehalten. Daher werden diese Techniken verglichen und danach die Konsequenzen für die Datenverantwortlichen erläutert. Es werden im Anschluss wichtige Erkenntnisse gezogen, die mit realistischen Kontexten und Zwecken einer solchen Verarbeitung verbunden werden.

Identifizierung von Individuen in Datensätzen

Erwägungsgrund 26 der DSGVO spezifiziert, dass Daten nicht mehr als personenbezogen zu betrachten sind, wenn es nicht möglich ist, diese dem zugehörigen Individuum zuzuweisen. Weiterhin besagt die Verordnung, dass man alle „relativ wahrscheinlichen“ Mittel berücksichtigen muss, die zur Identifizierung verwendet werden müssten.

Da sich diese Mittel kaum vorhersehen lassen, weil die Forschung in diesem Bereich immer weiter fortschreitet, das Know-How sich verbreitet und die Kosten für Rechenleistung niedriger werden, fokussiert sich die Verordnung auf das Ergebnis eines Identifizierungsversuchs, nämlich die Zuordnung zwischen Individuen und Datenmengen. Somit muss man die Verordnung nicht mit jedem einzelnen Fortschritt aktualisieren, sondern die Aktualisierung ist Teil der Verordnung selbst.

Daraus folgt aber, dass eine Datenmenge, die heute als „anonymisiert“ gilt, in einigen Jahren nicht zwingend eine Identifizierung ausschließen muss. Die Trennung dieser Zuordnung ist also ein kontinuierlicher Prozess. Darüber hinaus gilt, dass jeder Anonymisierungsprozess spezifisch zum

Kontext und mit einem bestimmten Zweck entworfen werden muss. Mit Systemkontext ist gemeint, wie die personenbezogenen Daten durch das System und über die Systemgrenzen hinaus fließen. Unter Zweck der Anonymisierung versteht man die Prozesse, die auf diesen anonymisierten Datenmengen basieren sollten.

Genau dasselbe wird in einer Veröffentlichung der Artikel-29-Datenschutzgruppe [Opinion on Anonymisation Techniques (2014); Seite 3] ermittelt. Es gibt damit kein Anonymisierungsprodukt als universelle Lösung, sondern dieser Prozess muss individuell entworfen werden.

Ziel ist das Hindern einer Identifizierung, in der versucht wird, genau diese Eigenschaften des Systems auszunutzen. Daher muss jede Lösung dem Systemkontext und Zweck angepasst werden.

Zum Schluss kann man aus dieser Definition erkennen, dass die Anonymisierung ein irreversibler Prozess ist. Die Identifizierung darf für keinen (inkl. Datenverantwortlichen) möglich sein.

Die Datenschutzgruppe beschreibt 3 Kriterien, die bei der Berücksichtigung eines Verfahrens in Erwägung gezogen werden sollten.

- Kann ein Individuum eindeutig erkannt werden?
- Können Datensätze über ein Individuum verlinkt werden?
- Kann Information über ein Individuum gefolgert werden?

Diese Faktoren werden immer abgewogen, um den Prozess gemäß dem Systemkontext und dem Anonymisierungszweck zu gestalten.

Pseudonymisierung

Ein Pseudonym bedeutet letztlich ein Alias bzw. eine Umbenennung eines Individuums. Hier werden die Attribute, die einen Datensatz genau klassifizieren könnten, mit Aliasen ersetzt. Diese sogenannten Identifier, wie Steuernummer zum Beispiel, würden sonst eine Person sofort identifizieren.

Die Sicherung der Identitäten wird durch die Trennung der Aliaszuordnungen und der resultierenden Datenmenge gewährleistet.

Leider wurde festgestellt, dass Attribute, die Individuen nicht eindeutig identifizieren, trotzdem in Kombination verwendet werden können, um die Identität doch sehr präzise einzugrenzen. [Sweeney 1997] hat festgestellt, dass sich 87% der US-Bevölkerung eindeutig über die Kombination von Geschlecht, Postleitzahl und Geburtstag eindeutig identifizieren ließ. Da sich ein solches Verhalten der Bevölkerungsdaten in erfassten personenbezogenen Daten wiederspiegeln würde, ist jede personenbezogene Datenmenge anfällig dafür, Datensätze anhand solcher Quasi-Identifikatoren Individuen zuzuordnen.

Es ist sofort ersichtlich, warum Pseudonymisierung die Definition von Anonymisierung laut der DSGVO nicht erfüllt, da dieser Prozess reversibel ist (anhand der Zuordnungsschlüssel), und es zudem nicht unwahrscheinlich ist, dass Individuen in pseudonymisierten Datenmengen über Quasi-Identifier wieder identifiziert werden können.

Anonymisierung

Die Anonymisierung hingegen versucht Datenmengen zu produzieren, die versuchen, die drei Kriterien der Datenschutzgruppe zu erfüllen. Es ist nie möglich alle Kriterien zu 100% zu erfüllen,

und trotzdem nützliche Daten zu erzielen. Die Kunst liegt darin, das Risiko zu erkennen, zu quantisieren und zu klassifizieren.

Auch für die Anonymisierung müssen die Identifier entfernt werden. Aber man versucht noch dazu eine Identifizierung anhand Kombinationen verschiedener Quasi-Identifier zu verhindern.

Es gibt i.A. zwei Arten von Verfahren: Randomisierung und Generalisierung.

Bei der Randomisierung versucht man anhand von kontrolliertem Rauschen, Datensätze zu modifizieren bzw. hinzuzufügen, so dass das Gesamtverhalten der Datenmenge relativ unverändert bleibt bzw. sich der Bias in der Aggregatsfunktion korrigieren lässt, während einzelne Datensätze über die Quasi-Identifier nicht mehr Individuen zugeordnet werden können.

Zu den bekannten Techniken zählen Noise Addition, Tauschen von Werten und Differential Privacy.

Bei der Generalisierung versucht man hingegen, Quasi-Identifier soweit zusammenzufassen, so dass man Information verliert. Damit bildet man Äquivalenzklassen der Quasi-Identifier, die ein präziseres Identifizieren hindern. Haben zum Beispiel Person A und B jeweils die Postleitzahl 30159 und 30173, so haben diese nach der Generalisierung dieselbe Postleitzahl 301** und lassen sich nicht innerhalb dieser Klasse anhand der Postleitzahl unterscheiden.

Hier wird sehr oft k-Anonymity, l-Diversity und t-Closeness verwendet.

Im Vortrag werden die Kriterien auf die Verfahren angewandt, anhand von Beispielen wird erläutert, wann man sich für die Anonymisierung bzw. für die Pseudonymisierung entscheiden sollte und wie das Risiko betrachtet werden sollte.

Kontaktadresse:

Bharat Ahuja

IT-P Information Technology-Partner GmbH

Seligmannallee 6

D-30173 Hannover

Telefon: **+49 (0) 511 6168040 320**

E-Mail **bharat.ahuja@it-p.de**

Internet: **www.purplestack.de**